

AI-Based Classification of OSA Severity in Awake Subjects Using Breathing Sound Analysis

By

Ali Mohammad Alqudah

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Biomedical Engineering Program

University of Manitoba

Winnipeg

Copyright © 2026 by Ali Mohammad Alqudah

ABSTRACT

Obstructive sleep apnea (OSA) is a prevalent, underdiagnosed disorder associated with cardiovascular and cognitive complications. The diagnostic gold standard, polysomnography (PSG), is costly and resource-intensive, motivating the development of accessible screening alternatives. This thesis proposes that tracheal breathing sounds (TBS) recorded during wakefulness, combined with anthropometric data, can predict OSA severity using three AI-based approaches: classification, feature interpretation, and non-invasive clinical screening.

The first approach involves developing a six-pairwise base-model framework using TBS combined with anthropometric data. A multidomain feature set comprising spectral, temporal, time-frequency, and nonlinear features was extracted. Features were selected using statistical filtering, SHAP ranking, and Recursive Feature Elimination before training pairwise binary classifiers. Nonlinear and spectral features were the most discriminative indicators of the severity of upper airway obstruction.

The second approach extends the base-models framework by introducing a stacked meta-model architecture for multi-class OSA severity prediction across both three-class and four-class stratification schemes. A conformal prediction (CP) framework was further incorporated to construct adaptive prediction sets with guaranteed statistical coverage, thereby transparently communicating diagnostic uncertainty in ambiguous cases. Stacking pairwise classifiers achieved clinically meaningful multi-class severity stratification. Conformal prediction provided principled uncertainty quantification for point-of-care triage.

The third approach extends the base-models framework by focusing on the clinical interpretability and physiological relevance of the extracted acoustic features. Features were evaluated using discrimination, stability, anthropometric correlation, and SHAP explainability. A Structure—

Function–Symptom framework linked discriminative acoustic descriptors to airway anatomy, airflow turbulence, and AHI-defined disease burden. The results demonstrated that spectral, bispectral, and fractal-based nonlinear features consistently emerged as the most stable and physiologically meaningful biomarkers, linking TBS signatures of wakefulness to progressive upper airway dysfunction.

Together, these approaches form a wakefulness-based framework for OSA screening and severity assessment using TBS recorded in under 10 minutes. For binary OSA screening, the deep learning model achieved 74.9% accuracy, 76.1% sensitivity, and 73.3% specificity. For pairwise severity discrimination using the base models, the best binary classifier (Non-OSA vs. Severe-OSA) reached 88.2% accuracy, 83.3% sensitivity, and 90.9% specificity. The meta-modeling framework further achieved 76.7% accuracy, 77.7% sensitivity, and 87.1% specificity in the three-class setting, and 76.7% accuracy, 75% sensitivity, and 92% specificity in the four-class setting. Conformal prediction further provides calibrated confidence sets, enhancing clinical trustworthiness and uncertainty-aware point-of-care triage.

ACKNOWLEDGMENTS

I would like to acknowledge the support and guidance of my supervisor, **Prof. Zahra Moussavi**, who has been consistently supportive, cooperative, and helpful throughout my PhD journey, enabling me to complete this work.

I would also like to acknowledge and thank my committee members, **Dr. Brian Lithgow**, **Dr. Mirosław Pawlak**, and **Dr. Carson Leung**, for their constructive feedback, valuable discussions, and continued support throughout the various stages of my thesis.

I wish to acknowledge the following entities: 1) the **Natural Sciences and Engineering Council of Canada (NSERC)**, 2) the **Faculty of Graduate Studies of the University of Manitoba**, and 3) **Mathematics of Information Technology and Complex Systems (MITACS)** for providing the financial support that made this research possible.

I also thank **Emil & Lynette Hain** and **Edward R. Toporeck** for creating graduate scholarships at the University of Manitoba, which further aided my academic pursuits.

My sincere thanks to my colleagues in the Acoustic Team for their collaboration and support, and for creating a positive and stimulating research environment.

Finally, I want to thank **the participants** who took part in the experimental and clinical parts of this research. Without their contributions, this study would not have been feasible.

DEDICATION

To my beloved parents, Mohammad Alqudah and Nawal Alqudah, whose love, sacrifices, and constant encouragement have guided me every step of the way.

My wife, Shoroq Abu Qazan, for her endless support, patience, and encouragement throughout this journey.

My beloved son, Nouh Ali Alqudah, whose love and smiles have filled my life with strength, motivation, and meaning.

My brothers (Rami and Ahmad) and sisters (Weam, Reema, Esraa, and Eman), nephews, and niece, whose constant encouragement and support have been instrumental throughout my journey.

CONTRIBUTIONS OF AUTHORS

This dissertation is structured as a combined (sandwich) thesis and includes eleven individual papers; all developed during the author's doctoral research under the supervision of Professor Zahra Moussavi. Six papers (Chapters Five, Six, Seven, and Eight) have been published in peer-reviewed journals, while one (Chapters Three and Four) appears in peer-reviewed conference proceedings as of the time of writing (March 2026).

Mr. Ali Mohammad Alqudah was the primary contributor and first author of all the manuscripts presented in this thesis. **Mr. Alqudah** 's contribution to this work includes developing the research questions, pre-processing the data, designing the diagnosis techniques, assisting with ethics approval, conducting the analyses, writing all manuscripts, submitting all the manuscripts, and responding to reviewers' comments. He was also responsible for integrating clinical and engineering aspects of the studies and ensuring that all methodological and ethical standards were met. **Professor Zahra Moussavi** contributed to the conception, drafting, submission, and review of the papers, interpretation of results, and manuscript revisions. She also contributed to refining research objectives and obtaining ethics approval. **Prof. Brian Lithgow** assisted with the review of the manuscript in Chapter Seven. **Dr. Walid Ashraf** assisted with the review and interpretation of the results in Chapter Seven.

Peer-reviewed journal publications of this dissertation:

- Alqudah, A. M., & Moussavi, Z. (2026). Awake Insights for Obstructive Sleep Apnea: Severity Detection Using Tracheal Breathing Sounds and Meta-Model Analysis. *Diagnostics*. DOI: [10.3390/diagnostics16030448](https://doi.org/10.3390/diagnostics16030448)
- Alqudah, Ali Mohammad, et al. "Interpretable Acoustic Features from Wakefulness Tracheal Breathing for OSA Severity Assessment." *Journal of Clinical Medicine* 15.3 (2026): 1081. DOI: [10.3390/jcm15031081](https://doi.org/10.3390/jcm15031081)
- Alqudah, A. M., & Moussavi, Z. "CUES: A Multiplicative Composite Metric for Evaluating Clinical Prediction Models Theory, Inference, and Properties." *Mathematics* 14.3 (2026): 398. DOI:[10.3390/math14030398](https://doi.org/10.3390/math14030398)
- Alqudah, A. M., & Moussavi, Z. "Bridging Signal Intelligence and Clinical Insight: A Comprehensive Review of Feature Engineering, Model Interpretability, and Machine Learning in Biomedical Signal Analysis." *Applied Sciences* 15.22 (2025). [10.3390/app152212036](https://doi.org/10.3390/app152212036)
- Alqudah AM, Moussavi Z. Assessing Obstructive Sleep Apnea Severity During Wakefulness via Tracheal Breathing Sound Analysis. *Sensors*. 2025; 25(20):6280. [10.3390/s25206280](https://doi.org/10.3390/s25206280)
- Alqudah, A. M., & Moussavi, Z. "A Review of Deep Learning for Biomedical Signals: Current Applications, Advancements, Future Prospects, Interpretation, and Challenges." *Computers, Materials & Continua* (2025). [10.32604/cmc.2025.063643](https://doi.org/10.32604/cmc.2025.063643)
- Alqudah, A. M., et al. "Obstructive sleep apnea detection during wakefulness: a comprehensive methodological review." *Medical & Biological Engineering & Computing* 62.5 (2024): 1277-1311. DOI: [10.1007/s11517-024-03020-3](https://doi.org/10.1007/s11517-024-03020-3)

List of Conference papers and abstracts of this disseration:

- Alqudah, Ali Mohammad, and Zahra Moussavi. "Deep Learning Model for OSA Detection using Tracheal Breathing Sounds During Wakefulness." *CMBES Proceedings* 45 (2023).
- Alqudah, Ali Mohammad, and Zahra Moussavi. "Comparative Analysis of Power Spectrum and Bispectrum Features for Obstructive Sleep Apnea Severity Classification." *2025 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. IEEE, 2025. DOI: [10.1109/3ict68299.2025.11442042](https://doi.org/10.1109/3ict68299.2025.11442042)
- Alqudah, Ali Mohammad, and Zahra Moussavi. "Towards Transparent Diagnostics: Explainable AI Across Biomedical Signals." *2025 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. IEEE, 2025. DOI: [10.1109/3ict68299.2025.11442144](https://doi.org/10.1109/3ict68299.2025.11442144)

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGMENTS	iii
DEDICATION	iv
CONTRIBUTIONS OF AUTHORS	v
CHAPTER ONE: MOTIVATION AND RESEARCH OBJECTIVES	28
1.1 Motivation	28
1.2 Goals and Objectives	29
1.3 Thesis Organization	33
References	36
CHAPTER TWO: INTRODUCTION AND BACKGROUND	38
2.1 Background	38
2.1.1 Clinical Presentation of OSA	38
2.1.2 Clinical Challenges in OSA Diagnosis	39
2.1.3 Anatomy and Physiology of The Human Upper Airway	40
2.1.4 Pathophysiology of OSA	42
2.2 Awake OSA Techniques	43
2.3 Artificial Intelligence for Biomedical Signals	47
2.4 Explainable AI for Biomedical Signals	50
References	54
CHAPTER THREE: DEEP LEARNING MODEL FOR OSA DETECTION USING TRACHEAL BREATHING SOUNDS DURING WAKEFULNESS	68
Abstract	68
3.1. Introduction	69
3.2. Methodology	70
3.2.1. Dataset	70
3.2.2. Preprocessing and Power Spectrum	71
3.2.3. Deep Learning Model	71
3.3. Results	73

3.4.	Discussion.....	75
3.5.	Conclusion	76
	References	77
CHAPTER FOUR: COMPARATIVE ANALYSIS OF SPECTRAL AND BISPECTRAL ANALYSES FOR OBSTRUCTIVE SLEEP APNEA SEVERITY CLASSIFICATION.....		80
	Abstract.....	80
4.1.	Introduction.....	81
4.2.	Literature Review	83
4.3.	Methodology	84
4.3.1.	Dataset.....	84
4.3.2.	5-Fold Stratified Cross-Validation.....	84
4.3.3.	Power Spectrum Feature Extraction.....	86
4.3.4.	Bispectrum Feature Extraction	87
4.3.5.	Statistical Analysis and Evaluation Metrics	88
4.4.	Results	89
4.4.1.	Summary of Power Spectrum Feature Performance Across All Models and Folds	89
4.4.2.	Summary of Bispectrum Feature Performance Across All Models and Folds..	92
4.5.	Discussion.....	95
4.6.	Conclusion	96
	References	97
CHAPTER FIVE: ASSESSING OBSTRUCTIVE SLEEP APNEA SEVERITY DURING WAKEFULNESS VIA TRACHEAL BREATHING SOUND ANALYSIS		100
	Abstract.....	100
5.1.	Introduction.....	101
5.2.	Literature Review of Tracheal Breathing Sounds Analysis	102
5.3.	Materials and Methods.....	104
5.3.1.	Tracheal Breathing Sounds Dataset	105
5.3.1.1.	Splitting Dataset for Training and Testing	106
5.3.1.2.	Splitting Dataset for K-Fold.....	108
5.3.2.	Tracheal Breathing Pre-Processing	109

5.3.3.	Anthropometric Missing Value Imputation.....	110
5.3.4.	Feature Extractions	110
5.3.4.1.	Finding Gaps in Power Spectrum and Bispectrum Using Confidence Intervals	111
5.3.5.	Automatic Feature Normalization.....	114
5.3.6.	Selecting Best Features	114
5.3.7.	Experiments Models Training.....	115
5.3.7.1.	Classifier Configuration and Hyperparameter Optimization	116
5.3.7.2.	Bootstrap Aggregation with OOB Validation.....	117
5.3.7.3.	Class Imbalance Mitigation	118
5.3.7.4.	Robust Model Selection via Repeated Trials	119
5.3.7.5.	Final Model Selection	119
5.3.8.	Model Evaluation	120
5.3.8.1.	Evaluation Protocol	120
5.3.8.1.1.	Out-of-Bag (OOB) Validation	120
5.3.8.1.2.	Independent Test Evaluation	121
5.3.8.1.3.	Performance Metrics	121
5.4.	Results	121
5.4.1.	Feature Selection Results	122
5.4.2.	Experiments Models Results.....	123
5.4.2.1.	Training and Testing Results.....	123
5.4.2.2.	K-Fold Results.....	126
5.5.	Discussion.....	127
5.6.	Limitations.....	132
5.7.	Conclusions.....	133
	References	134
CHAPTER SIX: AWAKE INSIGHTS FOR OBSTRUCTIVE SLEEP APNEA: SEVERITY DETECTION USING TRACHEAL BREATHING SOUNDS AND META-MODEL ANALYSIS.....		145
	Abstract.....	145
6.1.	Introduction and Literature Review	146

6.2. Materials and Methods	148
6.2.1. Dataset	149
6.2.2. Dataset Split and Cohort Stratification for Training and Testing	150
6.2.3. Dataset Split and Cohort Stratification for K-Fold Cross Validation	151
6.2.4. Preprocessing and Feature Engineering	153
6.2.5. Base Models	154
6.2.6. Meta-Model Construction	155
6.2.7. Conformal Prediction	156
6.2.8. Model Evaluation	159
6.2.8.1. Evaluation Protocol	159
6.2.8.2. Performance Metrics	159
6.2.8.3. Conformal Prediction Metrics	160
6.3. Results	161
6.3.1. Training-Testing Meta-Model Performance Summary	161
6.3.2. K-Fold Meta-Model Results	169
6.3.3. Statistical Significance Analysis	171
6.4. Discussion	173
6.4.1. Meta-Model Performance for Training-Testing	174
6.4.2. Meta-Model Performance for k-Fold	177
6.4.3. Statistical Significance of Meta-Model Improvements	179
6.4.4. Comparison to Prior TBS Studies	180
6.4.5. Clinical Utility and Limitations	181
6.5. Conclusions	182
References	183
CHAPTER SEVEN: INTERPRETABLE ACOUSTIC FEATURES FROM WAKEFULNESS TRACHEAL BREATHING FOR OSA SEVERITY ASSESSMENT	190
Abstract	190
7.1. Introduction	191
7.2. Materials and Methods	192
7.3. Results	199
7.3.1. Clinical Interpretation of Top Features	201

7.3.2.	Top-Ranked Features.....	202
7.3.3.	Feature Stability Across Folds and Models	203
7.3.4.	Correlation with Anthropometric Data	204
7.4.	Discussion.....	205
7.4.1.	Clinical Relevance of Key Features.....	206
7.4.1.1.	Non-OSA vs. Mild-OSA.....	206
7.4.1.2.	Non-OSA vs. Moderate-OSA	207
7.4.1.3.	Non-OSA vs. Severe-OSA.....	208
7.4.1.4.	Mild-OSA vs. Moderate-OSA	209
7.4.1.5.	Mild-OSA vs. Severe-OSA.....	210
7.4.1.6.	Moderate-OSA vs. Severe-OSA	211
7.4.1.7.	Physiological Themes Across Models	212
7.4.2.	Rationale for Multi-Class OSA Severity Stratification	212
7.4.3.	Physiological and Clinical Interpretation of Feature Linkage to Severity	213
7.4.3.1.	Acoustic Signatures of Airway Chaos and Ventilatory Effort (AHI Correlation)	214
7.4.3.2.	Morphological and Spectral Markers of Flow Limitation and Airway Dynamics	215
7.4.3.3.	Validation Through Established Anatomical Risk Factors	216
7.4.4.	Correlation with Anthropometric Data	216
7.4.5.	Alignment with Prior Wakefulness-Based OSA Studies	217
7.4.6.	Comparison with Other Awake Screening Modalities.....	219
7.4.7.	Limitations and Future Work.....	220
7.5.	Conclusions.....	222
	References	222
CHAPTER EIGHT: CUES: A MULTIPLICATIVE COMPOSITE METRIC FOR EVALUATING CLINICAL PREDICTION MODELS THEORY, INFERENCE, AND PROPERTIES.....		232
	Abstract.....	232
8.1.	Introduction.....	233
8.1.1.	The Need for a New Evaluation Metric	234
8.1.2.	What Makes This Score Different from Current Scores	234

8.2. Materials and Methods	236
8.2.1. Overview of the CUES Framework	236
8.2.1.1. Assumptions and conventions	239
8.2.2. Estimation of Individual Components	239
8.2.2.1. Utility (U): Integrated Net Benefit	240
8.2.2.2. Calibration (C): Brier Skill Score	241
8.2.2.3. Equity (E): Worst-Case Disparity with Variance Penalty	242
8.2.2.3.1. Extended Equity (E) for Demographic and Anthropometric Fairness	243
8.2.2.4. Stability (S): Normalized Confidence Interval	244
8.2.3. Practical Estimation Algorithm for CUES	245
8.2.4. CUES Curve and Visualization as Interpretation	245
8.2.4.1. Net Benefit Computation	245
8.2.4.2. Normalization	246
8.2.4.3. Area Under the CUES Curve (AUC-CUES)	246
8.2.4.4. Visualization	246
8.2.4.5. Theoretical Comparison with Existing Metrics (Decoupling from AUROC)	247
8.2.4.6. Interpretation	247
8.2.5. Gradient-Based Optimization of CUES (Optional)	248
8.2.5.1. Convexity	248
8.2.5.2. Practical approach	248
8.2.6. Theoretical Properties and Statistical Inference	249
8.2.6.1. Boundedness	249
8.2.6.2. Zero Sensitivity	250
8.2.6.3. Monotonicity	250
8.2.6.4. Relative Sensitivity	250
8.2.6.5. Continuous Mapping Theorem	250
8.2.6.6. Asymptotic Distribution (Delta Method Sketch)	251
8.2.6.7. Bootstrap-Based Inference	251
8.2.6.8. Statistical Testing and Confidence Statements	251
8.2.6.8.1. Bootstrap confidence intervals for CUES	251

8.2.6.8.2.	Hypothesis test for difference between two models (paired bootstrap)	252
8.2.7.	CUES as Feature Selection as Application	252
8.2.8.	Computational Considerations and Scalability.....	253
8.2.9.	Datasets and Experiment Setup.....	254
8.3.	Validation and Results	257
8.3.1.	Binary Classification Performance	257
8.3.2.	Multi-class Classification Performance	261
8.3.3.	Correlation Analysis	264
8.3.4.	Detailed Analysis of Binary and Multi-Class Findings.....	265
8.3.4.1.	Binary Classification Analysis	265
8.3.4.2.	Multi-class Classification Analysis	266
8.3.4.3.	Cross-Dataset and Correlation Analysis	267
8.3.5.	Interpretation of CUES Components.....	267
8.3.6.	CUES Features Selection.....	270
8.3.6.1.	Binary Classification Results	270
8.3.6.2.	Multi-class Classification Results	271
8.3.6.3.	Results of Retraining Using Top Selected Features	272
8.3.7.	Optimization Results	275
8.3.8.	Results of Demographic and Anthropometric Fairness	278
8.4.	Discussion.....	281
8.5.	Conclusions.....	286
	References	288
	CHAPTER NINE: CONCLUSION, LIMITATIONS, AND FUTURE WORK.....	295
9.1	Summary and Conclusion	295
9.2	Limitations.....	299
9.3	Future Work	301
	APPENDIX A	305
A.1	Anthropometric Missing Value Imputation.....	305
A.2	Automatic Feature Normalization.....	306
A.3	Selecting Best Features	307
A.3.1	Stage 1: Filtering by Univariate t-Test.....	308

A.3.2	Stage 2: SHAP-Based Feature Ranking.....	308
A.3.3	Embedded Recursive Feature Elimination (RFE) with Ensemble Model	309
	References	318
APPENDIX B		321
B.1	Top Selected Features for Each Model.....	321
B.2	Analysis of Clinical Interpretation of Top Features for All Models	323
B.2.1	Appendix B.1. Non-OSA vs. Mild-OSA	323
B.2.2	Non-OSA vs. Moderate-OSA	325
B.2.3	Non-OSA vs. Severe-OSA.....	326
B.2.4	Mild-OSA vs. Moderate-OSA	327
B.2.5	Mild-OSA vs. Severe-OSA.....	328
B.2.6	Moderate-OSA vs. Severe-OSA	329
B.3	Analysis for Non-OSA vs. Mild-OSA	330
B.4	Analysis for Non-OSA vs. Moderate-OSA.....	331
B.5	Analysis for Non-OSA vs. Severe-OSA	332
B.6	Analysis for Mild-OSA vs. Moderate-OSA.....	334
B.7	Analysis for Mild-OSA vs. Severe-OSA	335
B.8	Analysis for Moderate-OSA vs. Severe-OSA.....	336
APPENDIX C		339
C.1	Proof Sketches for Theoretical Properties of the CUES Metric	339
C.1.1	Definition and Notation	339
C.1.2	Boundedness the CUES Metric Satisfies $0 \leq CUES \leq 1$	339
C.1.3	Failure Dominance and Non-Compensability	339
C.1.4	Monotonicity in Each Component.....	340
C.1.5	Relative Sensitivity and Balanced Contribution to First Order, CUES Responds Equally to Relative Changes in Each Component	340
C.1.6	Continuity and Consistency of the CUES Estimator.....	341
C.1.7	Asymptotic Normality via the Delta Method Assume	341
C.1.8	Bootstrap Validity for Finite-Sample Inference	341

LIST OF TABLES

Table 3-1 Dataset Distribution.....	70
Table 3-2 The proposed architecture layer details.....	74
Table 3-3 Performance of the Proposed Methodology.....	75
Table 3-4 A Comparison Between Proposed Work and Previous Works.....	76
Table 4-1 Anthropometric data distribution over folds.....	85
Table 4-2 PowerSpectrum Features Statistics.....	90
Table 4-3 Most Frequent Power Spectrum Features.....	91
Table 4-4 Bispectral Features Statistics.....	92
Table 4-5 Most Frequent Bispectral Features.....	94
Table 5-1 Participants' severity groups and anthropometric information.....	106
Table 5-2 Participants' severity groups and anthropometric information for training set.....	107
Table 5-3 Participants' severity groups and anthropometric information for the testing set.....	107
Table 5-4 Participants' severity groups and anthropometric information for K-folds.....	108
Table 5-5 OOB results for experimental models.....	124
Table 5-6 OOB Test results for experimental models.....	124
Table 5-7 OOB 3-fold cross-validation results of experimental models.....	126
Table 5-8 Test 3-fold cross-validation results of experimental models.....	127
Table 6-1 Severity groups and participants' anthropometric information. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m ²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).....	150
Table 6-2 Severity groups and Participants' anthropometric information of the training set. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m ²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).....	151
Table 6-3 Severity groups and Participants' anthropometric information of the test set. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m ²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).....	151
Table 6-4 Demographic, anthropometric, and clinical characteristics of participants across OSA severity groups and cross-validation folds. Values are reported as mean ± standard deviation. Each row corresponds to one fold within a given severity group, illustrating the balance of subject distribution and key confounding variables across folds. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m ²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).....	152
Table 6-5 Results for the Meta-Model for three classes (mean ± 95% CI).....	163
Table 6-6 Results for the Meta-Model for four classes (mean ± 95% CI).....	164
Table 6-7 Results for the Meta-Model for three classes.....	167
Table 6-8 Results for the Meta-Model for four classes.....	168
Table 6-9 Train 3-fold cross-validation results of experimental models.....	169
Table 6-10 Test 3-fold cross-validation results of experimental models.....	169

Table 6-11 Summary comparison of OOB, independent test, and 3-fold cross-validation mean performance for the proposed meta-model.	171
Table 6-12 Statistical significance analysis comparing the proposed meta-model to the base classifier.	173
Table 6-13 Quantitative comparison of the proposed meta-model framework with representative prior TBS-based OSA studies in terms of dataset size, task formulation, performance, recording condition, and uncertainty modeling.....	180
Table 7-1 Participants' Severity Groups and Anthropometric Information. AHI: apnea–hypopnea index, BMI: body mass index, NC: neck circumference, MPS: Mallampati score, M/F: male/female.....	195
Table 7-2 Participants' Severity Groups and Anthropometric Information for K-folds. AHI: apnea–hypopnea index, BMI: body mass index, NC: neck circumference, MPS: Mallampati score, M/F: male/female.....	196
Table 7-3 Structure–Function–Symptom Interpretation of Dominant Acoustic Features Across OSA Severity Comparisons.	202
Table 7-4 Top 10 Ranked Tracheal Breathing Sound Features for OSA Severity Classification (Lower is better).....	203
Table 7-5 Top 10 Most Stable Tracheal Breathing Sound Features. Lower Absolute Delta AUC indicates higher feature stability across cross-validation folds, while Absolute Delta Correlation reflects the consistency of feature correlation. Features are considered stable/significant if Abs Delta AUC ≤ 0.01 and Abs Delta Correlation ≤ 0.1 , reflecting minimal variation and robust performance across folds. Features with Abs Delta Correlation between 0.1 and 0.15 are considered slightly unstable but still statistically acceptable, as this slight variation does not substantially affect their overall predictive reliability.....	204
Table 7-6 Top 10 Pearson correlations between breathing sound features and anthropometric measures (NC, BMI, Sex, MPS). Each row indicates which feature is correlated with which anthropometric variable, along with the fold and severity comparison.....	205
Table 7-7 Comparison of the proposed framework with representative wakefulness-based OSA screening modalities. Acc: accuracy, Sens: sensitivity, Spec: specificity.	220
Table 8-1 Computational complexity and scalability of CUES components.	254
Table 8-2 Binary classification: multiple classifiers CUES performance.	258
Table 8-3 Binary classification: multiple classifiers performance comparison between CUES and traditional metrics.	259
Table 8-4 Multi-class classification: multiple classifiers CUES performance.	262
Table 8-5 Multi-class classification: multiple classifiers performance comparison between CUES and traditional metrics.....	262
Table 8-6 Correlations between CUES and performance metrics.	265
Table 8-7 Multiple classifiers performance comparison between CUES and traditional metrics using two selected features subsets.....	274
Table 8-8 Values for hyperparameter optimization grid search.....	276

Table A-1 Summary of extracted features for all models.	311
Table A-2 Summary of selected features for each model.	314
Table B-1 Summary of Top Selected Features for Each Model.....	321
Table B-2 Top 10 ranked features for Non-OSA vs. Mild-OSA classification based on AUC Test and correlation performance.	330
Table B-3 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Non-OSA vs. Mild-OSA classification.	330
Table B-4 Top 10 correlations between selected features and anthropometric characteristics in Non-OSA vs. Mild-OSA participants.	330
Table B-5 Top 10 features showing the highest Pearson correlations with AHI in Non-OSA vs. Mild-OSA comparison.	331
Table B-6 Top 10 ranked features for Non-OSA vs. Moderate-OSA classification based on AUC Test and correlation performance.....	331
Table B-7 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Non-OSA vs. Moderate-OSA classification.....	331
Table B-8 Top 10 correlations between selected features and anthropometric characteristics in Non-OSA vs. Moderate-OSA participants.....	332
Table B-9 Top 10 features showing the highest Pearson correlations with AHI in Non-OSA vs. Moderate-OSA comparison.	332
Table B-10 Top 10 ranked features for Non-OSA vs. Severe-OSA classification based on AUC Test and correlation performance.....	332
Table B-11 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Non-OSA vs. Severe-OSA classification.....	333
Table B-12 Top 10 correlations between selected features and anthropometric characteristics in Non-OSA vs. Severe-OSA participants.....	333
Table B-13 Top 10 features showing the highest Pearson correlations with AHI in Non-OSA vs. Severe-OSA comparison.....	333
Table B-14 Top 10 ranked features for Mild-OSA vs. Moderate-OSA classification based on AUC Test and correlation performance.....	334
Table B-15 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Mild-OSA vs. Moderate-OSA classification.....	334
Table B-16 Top 10 correlations between selected features and anthropometric characteristics in Mild-OSA vs. Moderate-OSA participants.....	334
Table B-17 Top 10 features showing the highest Pearson correlations with AHI in Mild-OSA vs. Moderate-OSA comparison.	335
Table B-18 Top 10 ranked features for Mild-OSA vs. Severe-OSA classification based on AUC Test and correlation performance.....	335
Table B-19 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Mild-OSA vs. Severe-OSA classification.....	335

Table B-20 Top 10 correlations between selected features and anthropometric characteristics in Mild-OSA vs. Severe-OSA participants.	336
Table B-21 Top 10 features showing the highest Pearson correlations with AHI in Mild-OSA vs. Severe-OSA comparison.	336
Table B-22 Top 10 ranked features for Moderate-OSA vs. Severe-OSA classification based on AUC Test and correlation performance.	336
Table B-23 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Moderate-OSA vs. Severe-OSA classification.	337
Table B-24 Top 10 correlations between selected features and anthropometric characteristics in Moderate-OSA vs. Severe-OSA participants.	337
Table B-25 Top 10 features showing the highest Pearson correlations with AHI in Moderate-OSA vs. Severe-OSA comparison.	337

LIST OF FIGURES

Fig. 2-1 Anatomy of the upper airway [25] (CC BY 4.0).....	41
Fig. 3-1 The proposed deep learning model architecture.....	72
Fig. 3-2 The output of Phases-based training; A) Confusion matrix, B) ROC.....	74
Fig. 3-3 The output of Subject-based training; A) Confusion matrix, B) ROC.....	75
Fig. 4-1 Average PSD for different OSA severities.	86
Fig. 4-2 Average Bispectrum for different OSA severities.	87
Fig. 5-1 Block diagram of the proposed methodology. The process includes (1) preprocessing of tracheal breathing sounds (segmentation, filtering, normalization), (2) extraction of spectral, temporal, nonlinear, and morphological features, (3) feature selection using t-test, SHAP ranking, and RFE, and (4) classifier training with bootstrap aggregation and OOB validation.	104
Fig. 5-2 Experimental setup for TBS recordings. A Sony ECM77B condenser microphone was positioned at the suprasternal notch using a 2 mm custom plastic chamber to ensure consistent skin coupling and minimize ambient noise. Signals were sampled at 10,240 Hz using the Biopac DA100C, while participants were in a supine position during controlled wakefulness breathing maneuvers.	105
Fig. 5-3 The results of preprocessing techniques on a sample breathing phase.	110
Fig. 5-4 Flow chart of feature extraction from training data.	111
Fig. 5-5 A sample of the detected gaps regions of both PSD and bispectrum where (a) shows the PSD detection gaps, highlighted in yellow, using the proposed method; (b) shows the regions containing bispectrum detection gaps, highlighted as red boxes, using the proposed method....	113
Fig. 5-6 A Expanded classifier performance metrics across six binary experiments. For each experiment, the top six classifiers were selected based on their mean performance. Grouped bar plots display both out-of-bag (OOB) and test set results for accuracy, sensitivity, and specificity. This visualization highlights the differences between training (out-of-bag, OOB) and generalization (test) performance.	125
Fig. 5-7 A Receiver Operating Characteristic (ROC) curves of the top six classifiers for representative binary experiments. The curves illustrate the discrimination ability of each classifier across sensitivity–specificity trade-offs, complementing Tables 4 and 5 by providing a visual comparison of performance beyond single-value metrics.	125
Fig. 6-1 General steps of the proposed model.	149
Fig. 6-2 Meta-model for multi-class OSA severity classification.....	155
Fig. 6-3 Conformal prediction workflow for OSA severity classification. The training set is split into a calibration set (20%) for computing nonconformity scores and a model-fitting set (80%) for ensemble training. OOB predictions are used only for internal diagnostics and are not involved in calibration or testing, ensuring proper separation and reliable uncertainty estimation.	156
Fig. 6-4 Conformal prediction pipeline for uncertainty-calibrated OSA severity classification.	158

Fig. 6-5 Confusion matrix for three classes of OSA severity classification. Blue indicates correctly classified samples, red indicates misclassifications, with color intensity proportional to the number of samples.	162
Fig. 6-6 Confusion matrix for four classes of OSA severity classification. Blue indicates correctly classified samples, red indicates misclassifications, with color intensity proportional to the number of samples.	164
Fig. 6-7 Three-class testing conformal prediction set map.....	166
Fig. 6-8 Four-class testing conformal prediction set map.....	166
Fig. 6-9 ROC curves and AUC for the three-class testing set.	167
Fig. 6-10 ROC curves and AUC for the four-class testing set.....	168
Fig. 6-11 Learning curves for accuracy (top), sensitivity (middle), and specificity (bottom) for the three-class (left) and four-class (right) OSA severity models. Solid lines denote independent test performance and dashed lines denote k-fold cross-validation. All metrics improve monotonically with increasing training data and show no saturation, indicating a data-limited regime and consistent generalization across evaluation strategies.	170
Fig. 7-1 Overview of the machine learning pipeline for wakefulness-based tracheal breathing sound analysis and OSA severity classification.....	193
Fig. 7-2 Schematic illustration of the proposed stratified k-fold cross-validation strategy. The full dataset is partitioned into k folds while preserving the proportional distribution of OSA severity classes (Non, Mild, Moderate, Severe) and key anthropometric risk factors (age, BMI, neck circumference, sex, and Mallampati score). Each fold maintains comparable joint distributions, ensuring balanced subgroup representation during training and validation and minimizing bias due to population heterogeneity.....	196
Fig. 7-3 Framework for evaluating and interpreting tracheal breathing sound features in obstructive sleep apnea (OSA). Extracted acoustic and morphological features are analyzed for discriminative power, consistency with clinical variables, robustness across cross-validation folds, and model explainability. The results are integrated into a multidimensional feature ranking, producing tables, plots, and visual maps that link sound descriptors to airway dynamics, airflow turbulence, and anatomical variability across OSA severity levels.....	198
Fig. 7-4 Representative examples of gap regions identified between two classes based on non-overlapping 95% confidence intervals. (a) Power spectral density (PSD): the green shaded areas indicate frequency bands where the 95% confidence intervals of the average PSD for Moderate-OSA and Severe-OSA do not overlap, defining statistically significant spectral gaps between the two classes. (b) Bispectrum: red boxes indicate time–frequency regions in the bispectral domain where the 95% confidence intervals of the averaged bispectrum for Moderate-OSA and Severe-OSA do not overlap. The bispectrum, a third-order spectral measure, captures quadratic phase coupling and nonlinear interactions between frequency components; gap regions therefore indicate class-specific differences in nonlinear spectral coupling.....	200
Fig. 8-1 CUES component visualization for breast cancer (Logistic Regression). Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility	

curve, U) across decision thresholds; (b) reliability diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data. 260

Fig. 8-2 CUES component visualization for diabetes (Bagging). Each panel summarizes complementary aspects of model performance: (a) Normalized net benefit (utility curve, U) across decision thresholds; (b) reliability diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data..... 260

Fig. 8-3 CUES component visualization for diabetes (KNN). Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, U) across decision thresholds; (b) reliability diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data..... 261

Fig. 8-4 Macro-OVR CUES visualization for heart diseases (logistic model). Each panel summarizes complementary aspects of model performance: (a) macro-averaged normalized net benefit curve (utility, U); (b) macro-OVR reliability diagram (calibration, C); (c) mean CUES component scores (C, U, E, S); (d–e) bootstrap-based macro-utility stability analysis; (f) stability band (± 1 SD) around the Macro-OVR CUES curve; (g) per-class CUES curves illustrating class-wise trade-offs; (h) bar chart of per-class utility (U) values; (i) probability distributions across classes. This figure summarizes how multi-class models balance calibration, utility, equity, and robustness under the Macro-OVR framework. 263

Fig. 8-5 Macro-OVR CUES visualization for heart diseases (logistic model). Each panel summarizes complementary aspects of model performance: (a) macro-averaged normalized net benefit curve (utility, U); (b) macro-OVR reliability diagram (calibration, C); (c) mean CUES component scores (C, U, E, S); (d–e) bootstrap-based macro-utility stability analysis; (f) stability band (± 1 SD) around the Macro-OVR CUES curve; (g) per-class CUES curves illustrating class-wise trade-offs; (h) bar chart of per-class utility (U) values; (i) probability distributions across

classes. This figure summarizes how multi-class models balance calibration, utility, equity, and robustness under the Macro-OVR framework.	263
Fig. 8-6 Macro-OVR CUES visualization for heart diseases (KNN). Each panel summarizes complementary aspects of model performance: (a) macro-averaged normalized net benefit curve (utility, U); (b) Macro-OVR reliability diagram (calibration, C); (c) mean CUES component scores (C, U, E, S); (d–e) bootstrap-based macro-utility stability analysis; (f) stability band (± 1 SD) around the Macro-OVR CUES curve; (g) per-class CUES curves illustrating class-wise trade-offs; (h) bar chart of per-class utility (U) values; (i) probability distributions across classes. This figure summarizes how multi-class models balance calibration, utility, equity, and robustness under the Macro-OVR framework.	264
Fig. 8-7 Stacked CUES component scores for multiple classifiers trained on the breast cancer dataset. The figure demonstrates how overall CUES values reflect critical dimensions of clinical reliability beyond traditional metrics, highlighting differences in model stability and equity. ..	268
Fig. 8-8 Stacked CUES component scores for classifiers on the heart disease binary dataset. The components illustrate the importance of calibration and equity, showing that moderate traditional performance does not guarantee clinical suitability when important reliability factors are lacking.	268
Fig. 8-9 Stacked CUES component scores for classifiers on the dermatology dataset. Here, interpretable and straightforward models, such as Logistic Regression, achieve high composite CUES scores, demonstrating the benefits of transparency and stability for trustworthy clinical predictions.	269
Fig. 8-10 Stacked CUES component scores for classifiers on the heart disease multi-class dataset. The figure shows how subgroup variability and equity deficiencies lower CUES values, prompting a refined model selection to achieve more equitable clinical impact.	269
Fig. 8-11 Stacked CUES component scores for classifiers on the diabetes dataset. The visualization reveals cases in which stability or equity gaps reduce the overall CUES score, reinforcing CUES's role in identifying reliable and fair models for clinical deployment.	269
Fig. 8-12 Overall, CUES features importance analysis for the binary datasets. The bar chart (A) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance for the heart disease dataset using the KNN classifier. The bar chart (B) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance for the diabetic dataset using a logistic classifier.	271
Fig. 8-13 Overall, CUES features importance analysis for the multi-class datasets. The bar chart (A) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance in the dermatology dataset using a logistic classifier. The bar chart (B) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance for the heart disease dataset using a bagging classifier.	272
Fig. 8-14 CUES component visualization for breast cancer (Bagging) using only the top 50% features selected using the CUES Score. Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, U) across decision thresholds; (b) reliability	

diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.	273
Fig. 8-15 CUES component visualization for dermatology (Bagging) using only the top 50% features selected using the CUES Score. Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, U) across decision thresholds; (b) reliability diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.	273
Fig. 8-16 Accuracy vs. CUES (cross-validation scores): shows differences in optimal hyperparameters when tuned for accuracy versus CUES. CUES-optimized models often achieved slightly lower accuracy but higher overall CUES, reflecting improved calibration and stability.	277
Fig. 8-17 Components and metrics comparison: displays bar plots of accuracy, sensitivity, specificity, and individual CUES components (C, U, E, S). This allows inspection of trade-offs between classical metrics and explainable utility metrics.	277
Fig. 8-18 Heatmap: a comprehensive heatmap of all metrics and CUES components highlights each classifier's performance profile. Darker colors indicate higher values, allowing rapid comparison across classifiers and optimization strategies.	277
Fig. 8-19 Comprehensive anthropometric equity analysis across classifiers. The figure includes overall CUES performance, decomposition into utility, calibration, equity, and stability, a heatmap of feature-wise equity, the relationship between equity and CUES, utility disparities, and a summary of feature-wise equity across all models.	280
Fig. A-1 Flow chart of missing anthropometric data filling.	305
Fig. A-2 Flow chart of automatic feature normalization.	307
Fig. A-3 Flow chart of feature selection methodology.	308
Fig. A-4 Shapley values calculated for selected features of base model 6.	309

LIST OF ABBREVIATIONS

AbsDeltaAUC	Absolute Change in AUC
AHI	Apnea–Hypopnea Index
AI	Artificial Intelligence
AUC	Area Under the Curve
AUPRC	Area Under the Precision–Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BBox	Bounding Box
BCa	Bias-Corrected and Accelerated
BMI	Body Mass Index
BS	Bootstrap
BSS	Between-Subjects Sum of Squares
CAM	Class Activation Mapping
CDF	cumulative distribution function
CDSS	Clinical Decision Support Systems
CFD	Computational Fluid Dynamics
CI	Confidence Interval
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CP	Conformal Prediction
CPAP	Continuous Positive Airway Pressure
CQT	Constant-Q Transform
CSV	Comma-Separated Values
CT	Computed Tomography
CUES	Calibration, Utility, Equity, Stability Metric
CV	Cross-Validation
DCA	Decision Curve Analysis
DFA	Detrended Fluctuation Analysis
DIR	Direction histogram
DL	Deep Learning
ECG	Electrocardiogram
ECOC	Error-Correcting Output Codes
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
EU	European Union
FD	Fractal Dimension
FDA	Food and Drug Administration
FFT	Fast Fourier Transform
FM	Frequency Mean
FM2M	Mean Frequency Ratio
FM2MF	frequency modulation–to–modulation frequency
FN	False Negative
FP	False Positives
FPR	False Positive Rate

GBM	Gradient Boosting Machine
GPU	Graphics Processing Unit
GRU	Gated recurrent unit
HP	Harmonic–Percussive
HST	Home Sleep Testing
IG	Integrated gradient
INB	Incremental Net Benefit
IQR	Interquartile Range
k-Fold CV	K-Fold Cross-Validation
KNN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LES	Large Eddy Simulation
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
LSTM	Long Short-Term Memory
MDR	Medical Device Regulation
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multi-layer Perceptron
MPS	Mallampati Score
MRI	Magnetic Resonance Imaging
NB	Naive Bayes
NC	Neck Circumference
NEP	Negative Expiratory Pressure
NHR	Noise-to-Harmonic Ratio
NPV	Negative Predictive Value
ODI	Oxygen Desaturation Index
OOB	Out-of-Bag
OSA	Obstructive Sleep Apnea
OVR	One-Versus-Rest
PBP	Probabilistic Binary Patterns
PCA	Principal Component Analysis
PPG	Photoplethysmography
PPV	Positive Predictive Value
PR	Precision–Recall
PSD	Power Spectral Density
PSG	Polysomnography
PSR	Phase Space Reconstruction
QDA	Quadratic Discriminant Analysis
RANS	Reynolds-Averaged Navier–Stokes
RBF	Radial Basis Function
RF	Random Forest
RFE	Recursive Feature Elimination
RMS	Root Mean Square

RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RQA	Recurrence Quantification Analysis
SaMD	Software as a Medical Device
SD	Standard Deviation
SGDM	Stochastic Gradient Descent with Momentum
SHAP	SHapley Additive exPlanations
SNR	Signal-to-Noise Ratio
STOP-BANG	Snoring, Tiredness, Observed apnea, high blood Pressure, Body mass index, Age, Neck circumference, Gender
SVM	Support Vector Machine
TBS	Tracheal Breathing Sounds
TP	True Positive
TP	True Positives
TQWT	Tunable Q-Factor Wavelet Transform
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UA	Upper Airway
UAR	Unweighted Average Recall
VMD	Variational Mode Decomposition
WHR	Waist-to-Hip Ratio
XAI	Explainable Artificial Intelligence
Δ AUC	Change in AUC

CHAPTER ONE: MOTIVATION AND RESEARCH OBJECTIVES

1.1 Motivation

Obstructive sleep apnea (OSA) is a globally prevalent disorder, affecting up to 1 billion people, with nearly 80% of moderate to severe cases remaining undiagnosed [1,2]. This underdiagnosis leads to profound health consequences, including increased risk for cardiovascular and cerebrovascular disease, impaired cognitive function, depression, and perioperative complications [3]. Despite the high burden, the diagnostic infrastructure for OSA remains limited. Although Polysomnography (PSG) remains the gold standard, its practical and resource-intensive nature limits accessibility. PSG is not available in every city; for example, there are only two sleep centers equipped with full PSG in Winnipeg and Brandon in Manitoba. The alternative, home sleep testing (HST), is less resource-intensive but also less comprehensive and user-dependent, which compromises diagnostic reliability; in addition, its cost is not covered by general healthcare.

Given these limitations, the development of accurate, fast, and cost-effective daytime OSA severity detection tools is imperative. These tools are especially crucial for preoperative assessments, as undiagnosed OSA significantly increases the risk of anesthesia-related complications. Standard screening tools, such as STOP-BANG, remain limited by poor specificity, prompting the need for objective alternatives [4,5].

The integration of non-invasive acoustic techniques and advanced signal processing represents a promising solution. Breathing sounds captured during wakefulness [6-13] may reflect subtle upper airway (UA) anatomical and mechanical differences indicative of OSA severity [14]. Therefore, there is a growing need for comprehensive, real-time systems that leverage these signals alongside anthropometric features to enable reliable OSA severity detection without overnight monitoring.

1.2 Goals and Objectives

Although tracheal breathing sounds (TBS) recorded during wakefulness have been proposed as a promising non-invasive biomarker for obstructive sleep apnea (OSA) [6-13], critical limitations persist in existing classification methodologies. Most studies rely on a narrow spectral feature space [6-12], neglecting the nonlinear, turbulent, and tissue-vibrational phenomena that collectively govern TBS characteristics across different OSA severity. The absence of principled feature selection strategies further compounds this issue, introducing redundant or non-discriminative features that compromise model generalizability and clinical interpretability. Critically, existing approaches are predominantly limited to binary OSA detection [6-13], failing to address the clinically essential task of severity stratification required for individualized treatment planning and perioperative risk assessment. Furthermore, the near-universal absence of predictive uncertainty quantification in current frameworks fundamentally limits their trustworthiness and readiness for deployment in safety-critical clinical environments.

Therefore, the primary goal of this research is to develop a fast, non-invasive, and clinically reliable computational framework for detecting OSA severity using TBS recorded during wakefulness, combined with anthropometric data. To address the identified limitations, the proposed framework moves beyond binary classification by targeting three and four clinically meaningful severity classes, employs a principled multidomain feature extraction and selection strategy, integrates ensemble learning architectures capable of capturing complex inter-class acoustic relationships, and incorporates uncertainty quantification to enhance the trustworthiness and deployability of model predictions in clinical settings. Beyond classification performance, the framework further undertakes a systematic clinical interpretation of the extracted acoustic features, assessing their discriminative power, cross-validation stability, and physiological relevance

through a Structure–Function–Symptom lens that links acoustic biomarkers to underlying upper airway anatomy and disease burden. To achieve these goals, the specific objectives of this research are as follows:

Primary goal: Development of a machine learning framework for OSA severity classification using tracheal breathing sounds (TBS) recorded during wakefulness and anthropometric features.

Objective 1.1: Develop a signal preprocessing and feature extraction framework for TBS.

The recorded TBS undergo preprocessing steps, including noise reduction, segmentation, and band-pass filtering in the 75–3000 Hz frequency range, to isolate relevant respiratory acoustic components. A diverse set of acoustic features is then extracted, including spectral, nonlinear, and time-frequency descriptors, to capture the physiological and aerodynamic characteristics of the upper airway reflected in the breathing sounds.

Objective 1.2: Identify the most informative acoustic and anthropometric features using a multi-stage feature selection strategy.

To improve model performance and interpretability, a multi-step feature selection pipeline is developed. This process integrates statistical significance testing, SHAP-based feature importance analysis, and recursive feature elimination to identify the most discriminative features for OSA severity classification while reducing redundancy within the feature space.

Objective 1.3: Develop optimized base classifiers for pairwise OSA severity comparisons.

A set of binary classifiers is constructed to model pairwise comparisons among the OSA severity categories. Each classifier is trained with hyperparameters optimized via Bayesian

optimization and validated on Out-of-Bag (OOB) samples to ensure robustness and minimize overfitting.

Objective 1.3: Integrate base classifiers using a stacked ensemble meta-model for multi-class severity prediction.

To enhance overall classification performance, the outputs of the base classifiers are aggregated using a stacked ensemble learning approach. A meta-model is trained to learn the relationships among the base classifier outputs, enabling improved multi-class OSA severity prediction by leveraging complementary information across models.

Objective 1.5: Incorporate conformal prediction to quantify uncertainty in model outputs.

A conformal prediction framework is implemented to provide calibrated confidence estimates for model predictions. By generating prediction sets with guaranteed statistical coverage, this approach quantifies prediction uncertainty and enhances the reliability of the proposed framework for potential clinical use.

Objective 1.6: Assess the predictive performance, reliability, and robustness of the proposed framework.

Evaluate the clinical interpretability and physiological relevance of acoustic features within the base-models framework. Features will be assessed for discriminative power (AUC), cross-validation stability (Absolute Delta AUC), anthropometric correlation, and SHAP-based explainability. A Structure–Function–Symptom framework will link the most discriminative descriptors to upper airway anatomy, airflow patterns, and disease burden (AHI) to identify physiologically meaningful biomarkers of airway dysfunction.

Secondary goal: Develop a robust metric for evaluating clinical prediction models

Objective 2.1: Design and implement a novel metric for clinical model evaluation.

A novel metric is developed to evaluate clinical prediction models, improving sensitivity to clinically relevant outcomes, even in imbalanced or heterogeneous datasets. It provides a single, interpretable score reflecting both discriminative performance and clinical usefulness across models and patient cohorts.

Objective 2.2: Benchmark the proposed metric against conventional evaluation measures.

The metric is evaluated across multiple clinical prediction models to assess its performance relative to standard metrics such as AUC, accuracy, sensitivity, and specificity. Comparative analyses focus on scenarios with class imbalance or heterogeneous patient populations, evaluating whether the new metric more effectively captures clinically meaningful differences in model performance.

Objective 2.3: Explore correlations between the metric and clinical variables to assess interpretability and relevance.

The metric's values are analyzed in relation to patient demographics, disease severity, and outcome prevalence to understand its clinical implications. The objective is to verify that higher metric scores correspond to models that are not only statistically robust but also aligned with clinical priorities, enabling reliable model selection and deployment in real-world healthcare settings.

Objective 2.4: Apply the metric for feature selection

The metric is applied to guide feature selection, identifying the most informative predictors that maximize clinical relevance and model performance.

1.3 Thesis Organization

This dissertation is organized into nine chapters and three appendices.

Chapter One describes rationale for this work and outlines the goals, specific research objectives, and the scope of the study.

Chapter Two establishes the foundational context for the research. It opens with an overview of obstructive sleep apnea (OSA) covering its clinical presentation, upper airway anatomy and physiology, and underlying pathophysiology before critically examining the limitations of conventional diagnostic approaches. Awake OSA assessment techniques are then reviewed, followed by a survey of artificial intelligence methods for biomedical signal analysis. The chapter concludes by addressing the emerging role of Explainable AI (XAI) and identifying key research gaps in interpretability and clinical translation that motivate this work.

The following six chapters comprise 6 individual manuscripts: four published in peer-reviewed journals and two in peer-reviewed conference proceedings.

Chapter Three partially addresses **Objective 1.1** and presents a deep learning-based study for OSA detection using tracheal breathing sounds recorded during wakefulness. Tracheal breathing sounds are preprocessed and transformed into power spectral representations, which serve as inputs to a deep learning classifier. This chapter establishes the initial signal processing and modelling pipeline and demonstrates the clinical feasibility of wakefulness-based acoustic screening.

Chapter Four addresses **Objective 1.1** and partially **Objective 1.2** by presenting a comparative analysis of spectral and bispectral feature extraction methods for OSA severity classification. Power spectrum and bispectrum features are systematically extracted and evaluated across multiple classifiers using stratified cross-validation. The chapter identifies the relative

discriminative power of linear versus nonlinear acoustic descriptors, motivating the broader multi-feature framework developed in subsequent chapters.

Chapter Five addresses **Objectives 1.1, 1.2,** and partially **Objective 1.3** by presenting a comprehensive framework for OSA severity assessment from tracheal breathing sounds. The methodology covers signal preprocessing, confidence-interval-based gap analysis of spectral features, multi-stage feature selection, and optimized classifier training using bootstrap aggregation and OOB validation. The chapter provides the core methodological foundation for the ensemble and explainability layers of the framework.

Chapter Six addresses **Objectives 1.3 and 1.5** by extending the base classification framework through a stacked ensemble meta-model architecture augmented with **conformal prediction**. Pairwise binary base classifiers are integrated via a meta-learner to enable robust multi-class OSA severity prediction, while conformal prediction sets provide statistically guaranteed uncertainty quantification. This chapter demonstrates meaningful performance improvements over individual classifiers and introduces a clinically relevant mechanism for quantifying prediction confidence.

Chapter Seven addresses **Objective 1.6** by focusing on the clinical interpretability and physiological relevance of the acoustic features driving model decisions. A **Structure–Function–Symptom** framework is applied to link the most discriminative spectral, nonlinear, and time-frequency descriptors to upper airway anatomy, airflow dynamics, and disease burden as measured by the Apnea–Hypopnea Index (AHI). Features are assessed for discriminative power (AUC), cross-validation stability (Absolute Delta AUC), anthropometric correlation, and SHAP-based explainability across all pairwise severity comparisons.

Chapter Eight addresses the **secondary goal** and all associated objectives (**Objectives 2.1–2.4**) by introducing **CUES**, a novel multiplicative composite metric for evaluating clinical prediction models. The metric integrates four clinically grounded components: **Utility** (integrated net benefit), **Calibration** (Brier Skill Score), **Equity** (worst-case disparity with variance penalty), and **Stability** (normalized confidence interval). CUES is benchmarked against conventional metrics in both binary and multi-class scenarios, validated for its correlation with clinical variables, and applied as a feature selection criterion to identify predictors that maximize both statistical performance and clinical relevance.

Chapter Nine synthesizes the thesis's contributions, discusses overarching limitations, and outlines directions for future research.

Appendix A provides supplementary details on three methodological components shared across multiple manuscripts: anthropometric missing-value imputation, automatic feature normalization, and the multi-stage feature-selection pipeline, including univariate t-test filtering, SHAP-based feature ranking, and recursive feature elimination (RFE).

Appendix B presents supplementary results supporting Chapter Seven, including the top-selected features for each pairwise model and detailed clinical interpretation of acoustic descriptors across all six OSA severity comparisons.

Appendix C provides formal proof sketches for the theoretical properties of the CUES metric, including boundedness, failure dominance, monotonicity, relative sensitivity, continuity, asymptotic normality via the delta method, and bootstrap validity for finite-sample inference.

References

- [1] Young, T.; Finn, L.; Peppard, P.E.; Szklo-Coxe, M.; Austin, D.; Nieto, F.J.; Stubbs, R.; Hla, K.M. Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort. *Sleep* 2008, 31, 1071, doi:10.1016/s8756-3452(08)79181-3.
- [2] American Academy of Sleep Medicine Hidden Health Crisis Costing America Billions: Underdiagnosing and Undertreating Obstructive Sleep Apnea Draining Healthcare System; 1st ed.; Frost & Sullivan : Darien, IL, 2016;
- [3] American Society of Anesthesiologists Practice Guidelines for the Perioperative Management of Patients with Obstructive Sleep Apnea: An Updated Report. *Anesthesiology* 2014, 120, 268–286, doi:10.1097/ALN.0000000000000053.
- [4] El-Sayed, I.H. Comparison of Four Sleep Questionnaires for Screening Obstructive Sleep Apnea. *Egyptian Journal of Chest Diseases and Tuberculosis* 2012, 61, 433–441, doi:10.1016/J.EJCDT.2012.07.003.
- [5] Nagappa, M.; Liao, P.; Wong, J.; Auckley, D.; Ramachandran, S.K.; Memtsoudis, S.; Mokhlesi, B.; Chung, F. Validation of the STOP-Bang Questionnaire as a Screening Tool for Obstructive Sleep Apnea among Different Populations: A Systematic Review and Meta-Analysis. *PLoS One* 2015, 10, e0143697, doi:10.1371/JOURNAL.PONE.0143697.
- [6] Elwali, A.; Moussavi, Z. Obstructive Sleep Apnea Screening and Airway Structure Characterization During Wakefulness Using Tracheal Breathing Sounds. *Ann. Biomed. Eng.* 2016, doi:10.1007/s10439-016-1720-5.
- [7] Moussavi, Z.; Elwali, A.; Soltanzadeh, R.; MacGregor, C.A.; Lithgow, B. Breathing Sounds Characteristics Correlate with Structural Changes of Upper Airway Due to Obstructive

Sleep Apnea. In Proceedings of the Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE; IEEE, 2015; pp. 5956–5959.

[8] Moussavi, Z.; Elwali, A. System and Methods for Screening Obstructive Sleep Apnea During Wakefulness Using Anthropometric Information and Tracheal Breathing Sounds 2021.

[9] Moussavi, Z.; Elwali, A. System and Methods for Screening Obstructive Sleep Apnea During Wakefulness Using Anthropometric Information and Tracheal Breathing Sounds 2021.

[10] Elwali, A.; Moussavi, Z. Determining Breathing Sound Features Representative of Obstructive Sleep Apnea During Wakefulness with Least Sensitivity to Other Risk Factors. *J. Med. Biol. Eng.* 2019, 39, 230–237, doi:10.1007/s40846-018-0410-1.

[11] Hajipour, F.; Jozani, M.J.; Elwali, A.; Moussavi, Z. Regularized Logistic Regression for Obstructive Sleep Apnea Screening during Wakefulness Using Daytime Tracheal Breathing Sounds and Anthropometric Information. *Med. Biol. Eng. Comput.* 2019, 57, 2641–2655, doi:10.1007/s11517-019-02052-4.

[12] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea Using Anthropometric Information and Tracheal Breathing Sounds. *Sci. Rep.* 2019, 9, 11467, doi:10.1038/s41598-019-47998-5.

[13] Soltanzadeh, R.; Moussavi, Z. Breathing Sounds Spectral and Higher Order Statistics Changes from Wakefulness to Sleep in Apneic and Non-Apneic People. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014 2014, 4228–4231, doi:10.1109/EMBC.2014.6944557.

[14] Ryan, C.M.; Bradley, T.D. Pathogenesis of Obstructive Sleep Apnea. *J. Appl. Physiol.* (1985) 2005, 99, 2440–2450, doi:10.1152/JAPPLPHYSIOL.00772.2005.

CHAPTER TWO: INTRODUCTION AND BACKGROUND

2.1 Background

Obstructive sleep apnea (OSA) is one of the most prevalent sleep-related breathing disorders and represents a significant global health concern [1–3]. Sleep apnea is a sleep disorder characterized by recurrent upper airway collapse, manifesting as complete (apnea) or partial (hypopnea) cessation of breathing during sleep [4,5]. It is classified into three types: central sleep apnea, OSA, and mixed sleep apnea. Central sleep apnea arises from a temporary cessation of respiratory effort due to diminished neural output from the brainstem, while OSA, accounting for approximately 84% of all sleep apnea cases, results from recurrent partial or complete collapse of the upper airway during sleep [6]. Mixed sleep apnea features characteristics of both. These repeated breathing interruptions fragment normal sleep architecture and are associated with a wide range of adverse physiological and neurocognitive consequences [7].

Despite its serious health implications, a large proportion of individuals with OSA remain undiagnosed, partly due to the limited accessibility of current diagnostic methods [3]. This has driven growing interest in developing rapid, non-invasive screening alternatives [8–10].

A comprehensive understanding of the disorder requires an examination of its clinical presentation, the challenges of current diagnostic practices, and the anatomical and physiological mechanisms contributing to airway obstruction. This dissertation focuses on OSA because of its higher prevalence and clinical significance.

2.1.1 Clinical Presentation of OSA

OSA is a chronic respiratory disorder characterized by recurrent episodes of upper airway (UA) collapse during sleep, resulting in repeated apnea and hypopnea events [4,5]. These events lead to intermittent hypoxemia, sympathetic nervous system activation, and sleep fragmentation.

OSA presents with a range of nocturnal and daytime symptoms. Common nocturnal symptoms include habitual loud snoring, gasping or choking during sleep, and frequent awakenings, while daytime signs include excessive sleepiness, morning headaches, impaired concentration, irritability, and fatigue [11,12]. In severe cases, OSA is linked to cognitive dysfunction, mood issues, and a higher risk of motor vehicle accidents. Over time, untreated OSA contributes to cardiovascular disease, stroke, type 2 diabetes, and metabolic problems, leading to a lower quality of life and a greater healthcare burden [7,11,12].

Additionally, several demographic and body measurements, such as age, sex, body mass index (BMI), and neck circumference, have also been shown to correlate significantly with the presence and severity of OSA [13]. Clinically, OSA severity is typically assessed using the Apnea–Hypopnea Index (AHI), which quantifies the number of apnea and hypopnea events occurring per hour of sleep. AHI is usually measured during an overnight sleep study called polysomnography (PSG) [6,14]. According to established clinical guidelines, AHI values are categorized as follows: $AHI < 5$ indicates no OSA; 5–15 indicates mild OSA; 15–30 indicates moderate OSA; and $AHI > 30$ indicates severe OSA [5,14]. This classification framework plays a central role in guiding clinical decision-making and treatment planning.

2.1.2 Clinical Challenges in OSA Diagnosis

While the gold standard PSG measures a variety of physiological parameters, including airflow, respiratory efforts, blood oxygen saturation, and sleep stages, it is still a costly and less accessible tool for OSA diagnosis [6,14]. In fact, a comprehensive PSG study requires special equipment, well-trained technologists, and a sleep laboratory, making it impractical for widespread application, especially in underserved communities. As a result, access to PSG is often limited, leading to long waiting times and delayed diagnosis in many healthcare systems. These logistical

constraints significantly hinder the timely identification and management of OSA [6]. Home Sleep Testing (HST) has been introduced as a more convenient alternative; however, it measures fewer physiological signals and may fail to detect complex or atypical sleep-disordered breathing patterns [11,12].

Consequently, OSA remains substantially underdiagnosed worldwide. Epidemiological studies estimate that more than 80% of individuals with moderate-to-severe OSA remain unaware of their condition [1–3]. This underdiagnosis contributes to considerable healthcare and societal costs. In the United States alone, the annual economic burden of untreated OSA, including healthcare expenditures, productivity losses, and accident-related costs, has been estimated at \$65-\$165 billion [9,15,16].

To address these limitations, there is increasing interest in alternative screening methods that are rapid, cost-effective, and suitable for deployment outside specialized sleep laboratories [17–22]. Common clinical screening tools, such as the STOP-BANG questionnaire and the Epworth Sleepiness Scale, are widely used for their simplicity. However, these questionnaires rely on subjective patient responses and demonstrate relatively low specificity, with false-positive rates reported as high as 75% [23,24]. As a result, they often lead to unnecessary PSG referrals and fail to provide objective physiological evidence of airway obstruction.

2.1.3 Anatomy and Physiology of The Human Upper Airway

The upper airway serves as the primary conduit for airflow between the external environment and the lower respiratory tract. It includes several anatomical structures such as the nasal cavity, oral cavity, pharynx, and larynx. The geometry and mechanical properties of these structures play a critical role in determining airflow dynamics during respiration.

Numerous anatomical factors have been associated with increased susceptibility to OSA. These include enlarged tongue and soft palate structures, thickened lateral and posterior pharyngeal walls, elongated airway passages, and increased compliance of soft tissues within the airway [17–22]. Such structural characteristics increase the likelihood of airway narrowing or collapse during sleep.

Morphological differences in airway geometry also contribute to airflow instability. In healthy individuals, the velopharyngeal airway typically exhibits a laterally oriented elliptical cross-sectional shape. In contrast, individuals with OSA often present with a more circular airway configuration, which may predispose the airway to collapse under negative inspiratory pressure [18,19].

These anatomical variations influence airflow behavior within the airway, including turbulence and resonance characteristics during breathing. As a result, the airflow passing through the airway can generate distinctive acoustic signatures that reflect underlying structural and physiological conditions [17–22]. The anatomical distinctions are illustrated in Figure 2-1.

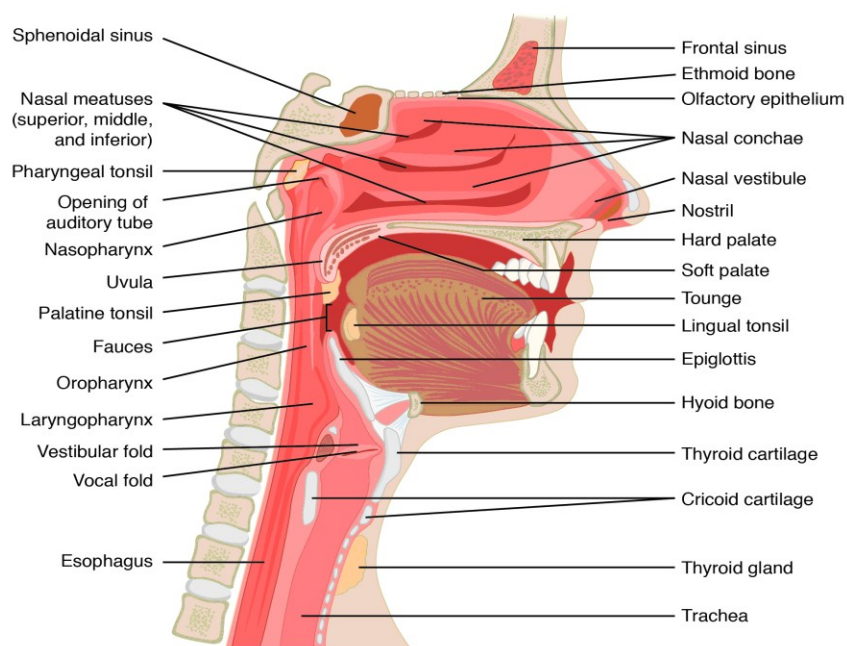


Fig. 2-1 Anatomy of the upper airway [25] (CC BY 4.0).

2.1.4 Pathophysiology of OSA

OSA is a multifactorial disorder in which recurrent upper airway obstruction during sleep arises from the interplay of anatomical vulnerability, neuromuscular dysregulation, and ventilatory control instability. Maintenance of upper airway patency requires a balance between the mechanical compliance of the airway, governed by its shape, size, and soft tissue composition, and the activation of pharyngeal dilator muscles [26]. Airway compliance, defined as the change in volume per unit change in transmural pressure, determines the airway's susceptibility to collapse under the negative intraluminal pressures generated during inspiration [27]. A highly compliant airway deforms more readily under sub-atmospheric pressures, particularly in individuals with structural abnormalities such as enlarged soft tissue structures, craniofacial skeletal restriction, or increased fat deposition in the parapharyngeal region [28,29].

During wakefulness, tonic activation of upper airway dilator muscles, most notably the genioglossus, compensates for these structural vulnerabilities by stabilizing and enlarging the pharyngeal lumen [26]. This neuromuscular compensation effectively counteracts the collapsing forces acting on the pharyngeal walls, which is why individuals predisposed to OSA typically do not exhibit overt airway obstruction while awake. However, during sleep, particularly during rapid eye movement (REM) phases, dilator muscle activity is substantially reduced, diminishing neuromuscular support and rendering the airway increasingly susceptible to collapse [27].

The resulting cyclical obstruction triggers a cascade of pathophysiological responses. Airway collapse produces apnea or hypopnea events, causing intermittent hypoxemia and hypercapnia. These alterations in blood gas composition activate peripheral and central chemoreceptors, eliciting cortical arousals that restore muscle tone and reestablish airflow [30–

33]. While effective in terminating individual events, these repeated arousals fragment sleep architecture and prevent the attainment of deep, restorative sleep stages. Over time, chronic sympathetic nervous system activation resulting from repetitive hypoxic episodes contributes to systemic hypertension, cardiovascular disease, and metabolic dysfunction [34,35].

An additional contributor to OSA pathophysiology is ventilatory control instability, quantified by loop gain, which measures the respiratory control system's tendency to overreact to ventilation perturbations [36]. A high loop gain system responds to an apnea-induced hypoxic event with excessive hyperventilation, driving CO₂ below the apneic threshold and precipitating subsequent hypoventilation or apnea [37]. This oscillatory instability promotes periodic breathing and sustains recurrent apneic episodes, particularly in individuals whose anatomical or neuromuscular vulnerabilities already compromise airway patency [38]. Patients with severe OSA have been shown to exhibit higher loop gain values, reflecting a less stable ventilatory control system [39].

2.2 Awake OSA Techniques

The conventional diagnosis of OSA relies primarily on overnight monitoring techniques such as PSG or HST [11,12]. Although these approaches provide reliable diagnostic information, they are expensive, time-consuming, and resource-intensive. Consequently, long waiting lists and limited accessibility often delay diagnosis and treatment. To address these limitations, researchers have explored methods capable of predicting OSA during wakefulness [18,40–46]. These techniques aim to provide rapid, non-invasive, and cost-effective screening tools that can identify individuals at high risk for OSA without requiring overnight monitoring [40].

The rationale for awake OSA detection is that individuals with OSA exhibit anatomical and physiological differences in the upper airway even while awake [47]. Structural characteristics

such as increased pharyngeal length, an enlarged soft palate, and greater airway collapsibility have been reported in patients with OSA [48,49]. These anatomical variations influence airflow, tissue vibration, and airway mechanics, which can be captured through various physiological or acoustic measurements during wakefulness [18,22].

Although patients with OSA typically do not exhibit overt breathing difficulty while awake due to compensatory activation of upper airway dilator muscles, these underlying structural abnormalities can still affect measurable signals, including speech, breathing sounds, and airway pressure patterns [50]. Recent advances in signal processing and machine learning have enabled the extraction of subtle features from such signals, making wakefulness-based screening increasingly feasible [51].

A comprehensive methodological review identified several major categories of awake OSA detection techniques. These approaches differ in their underlying physiological bases and measurement modalities but share the common goal of providing rapid, non-invasive screening. The primary categories include imaging techniques, negative expiratory pressure testing, facial landmark analysis, acoustic pharyngometry and nasal airway pressure measurement, breathing sound analysis, speech signal analysis, and questionnaire-based screening tools [40].

Imaging techniques were among the earliest approaches used to study OSA during wakefulness. Modalities such as magnetic resonance imaging (MRI) [52] and computed tomography (CT) [53] have been used to visualize structural differences in the upper airway between individuals with OSA and those without. These studies have demonstrated that OSA patients often present with anatomical features such as increased soft tissue mass, altered airway geometry, and increased airway compliance [52,53]. Although imaging methods provide valuable

physiological insights, their high cost and complexity limit their practicality as routine screening tools [52,53].

Negative expiratory pressure (NEP) testing represents another technique used to evaluate airway collapsibility during wakefulness [45,54,55]. In this method, a brief negative pressure is applied to the airway during expiration to assess the upper airway's tendency to collapse [55]. Individuals with OSA generally exhibit greater airflow limitation under negative pressure due to increased airway collapsibility [55]. While NEP testing provides useful physiological information, specialized equipment and controlled testing conditions are often required [45,54,55].

Facial landmark analysis is a more recent approach that utilizes facial imaging and computer vision techniques to estimate craniofacial characteristics associated with OSA risk [44,56]. Structural features such as mandibular position, facial width, neck circumference, and other craniofacial measurements can be extracted automatically from facial images [45,54,55]. Machine learning algorithms can then analyze these features to estimate the likelihood of OSA [44]. This approach offers advantages in terms of convenience and potential integration with mobile devices.

Acoustic pharyngometry and nasal airway pressure measurements are additional techniques that assess the geometry and resistance of the upper airway [57-59]. Acoustic pharyngometry uses reflected sound waves to estimate the cross-sectional area and volume of the airway, providing insight into airway obstruction patterns [57-59]. Nasal airway pressure measurements evaluate airflow resistance during breathing and can help detect abnormalities associated with airway narrowing [57-59].

Breathing sound analysis is another promising technique for awake OSA detection. Airflow passing through the upper airway generates acoustic signals that reflect airway geometry

and tissue vibration characteristics [18,21,22,60–63]. Differences in airway structure between OSA and healthy individuals can lead to detectable changes in breathing sound patterns [18,21,22,60–63]. Signal-processing methods combined with machine-learning classifiers have been used to extract features from these sounds for automated OSA screening [18,21,22,60–63].

Similarly, speech signal analysis has gained considerable attention in recent years. Speech production involves complex interactions between airflow, vocal fold vibration, and upper airway resonance [42,43,64–73]. Structural abnormalities associated with OSA can influence vocal tract acoustics, leading to measurable changes in speech characteristics [42,43,64–73]. Acoustic features such as formants, spectral energy distribution, and Mel-frequency cepstral coefficients (MFCCs) have been widely used to identify patterns associated with OSA [42,43,64–73]. These techniques offer the advantage of requiring only short speech recordings, making them highly suitable for rapid screening applications.

Finally, questionnaire-based screening tools remain widely used in clinical practice due to their simplicity and minimal cost [46,74–78]. Instruments such as the STOP-BANG questionnaire and the Epworth Sleepiness Scale estimate OSA risk based on demographic, clinical, and symptomatic information [42,43,64–73]. However, these questionnaires typically exhibit relatively low specificity and high false-positive rates, limiting their effectiveness as standalone diagnostic tools.

Overall, research into awake OSA detection techniques has expanded significantly in recent years. Many of these approaches show promising results, particularly when combined with modern machine learning methods. However, further validation using larger datasets and standardized protocols is necessary before these techniques can be widely adopted in clinical practice. The development of accurate, portable, and user-friendly wakefulness-based screening

tools could significantly reduce the burden on sleep laboratories and facilitate earlier diagnosis and treatment of OSA.

2.3 Artificial Intelligence for Biomedical Signals

Biomedical signals represent dynamic physiological processes generated by the human body and serve as critical indicators for health monitoring and disease diagnosis. Examples include electrocardiograms (ECGs), electroencephalograms (EEGs), electromyograms (EMGs), respiratory sounds, oxygen saturation signals, and other physiological measurements [79,80]. These signals provide valuable insights into cardiovascular, neurological, and respiratory function and have long been used for clinical assessment and monitoring. However, their inherent complexity, high dimensionality, and susceptibility to noise present significant challenges for traditional signal processing and manual interpretation techniques [79,80].

Recent advances in Artificial Intelligence (AI) have dramatically transformed the analysis of biomedical signals. AI algorithms, particularly machine learning (ML) and deep learning (DL) models, can learn complex patterns directly from raw or minimally processed data, enabling automated detection of physiological abnormalities and disease states [79,80]. Unlike conventional statistical approaches that rely heavily on handcrafted features and predefined thresholds, AI-based systems can capture nonlinear relationships and subtle signal characteristics that may be difficult for human observers to identify [79,80].

The general pipeline for AI-based biomedical signal analysis typically consists of several stages. First, signals are acquired through sensors or medical devices such as electrodes, microphones, or wearable systems [81]. This is followed by preprocessing steps designed to remove noise and artifacts, including filtering, normalization, and segmentation. Feature extraction techniques are then applied to transform raw signals into meaningful representations [79–81].

These features may be derived from time-domain statistics, frequency-domain analysis, time–frequency representations such as spectrograms or wavelets, or more advanced nonlinear measures such as entropy and fractal characteristics [79–81]. Finally, machine learning algorithms are trained to classify, predict, or interpret physiological conditions based on the extracted features [79].

Traditional machine learning models frequently used in biomedical signal analysis include support vector machines (SVM), random forests (RF), k-nearest neighbors (KNN), logistic regression, and gradient boosting methods [79]. These algorithms have demonstrated strong performance across many clinical applications due to their robustness and effectiveness with relatively small datasets [79]. However, recent developments in deep learning have further expanded the capabilities of biomedical signal analysis. Deep neural networks, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based architectures, can automatically learn hierarchical feature representations directly from raw signals or time–frequency images [80].

Deep learning has been particularly successful in applications involving complex temporal patterns and high-dimensional biomedical data [82]. For instance, CNNs have been widely applied to ECG classification, enabling automated detection of arrhythmias with performance comparable to expert cardiologists [83]. Similarly, RNNs and long short-term memory (LSTM) networks have been used to analyze sequential biomedical signals such as EEG recordings for seizure detection and sleep stage classification [84]. In respiratory medicine, machine learning models have been employed to analyze breathing sounds and airflow patterns to detect conditions such as asthma, chronic obstructive pulmonary disease (COPD), and obstructive sleep apnea (OSA) [85].

In the context of sleep medicine, AI has become increasingly important for improving diagnostic efficiency and accessibility [79,80,82]. Traditional diagnostic methods, such as polysomnography (PSG), generate large volumes of multimodal physiological data, including EEG, electrooculography (EOG), electromyography (EMG), airflow, respiratory effort, and oxygen saturation [86]. Manual scoring of these recordings is time-consuming and requires specialized expertise [14]. AI-based systems can automate this process by learning patterns associated with sleep stages, respiratory events, and pathological conditions [86].

Beyond sleep laboratory environments, AI has also enabled the development of portable and wearable diagnostic technologies [79,80,82]. Advances in sensor design and embedded computing have enabled the collection of biomedical signals outside clinical settings, including via smartphones and low-cost wearable devices [79,80,82]. AI algorithms can process these signals in real time to detect physiological abnormalities, enabling remote health monitoring and early disease detection [79,80,82]. This paradigm shift has contributed to the emergence of digital health and telemedicine, where continuous physiological monitoring can support proactive healthcare management [79,80,82].

Despite the promising capabilities of AI in biomedical signal analysis, several challenges remain. Biomedical datasets are often limited in size, heterogeneous across populations, and subject to measurement variability due to differences in sensor placement, sampling rates, and environmental conditions [79,80,82]. These factors can affect model generalizability and robustness when systems are deployed in real-world clinical environments [79,80,82]. Additionally, the increasing complexity of deep learning models has raised concerns about transparency and interpretability, as many models operate as opaque “black boxes” that offer little insight into their internal decision-making processes [79,80,82].

These challenges highlight the need for more transparent and trustworthy AI systems in healthcare [79,80,82]. While AI models may achieve high predictive accuracy, clinicians require clear explanations of how these predictions are generated before they can confidently incorporate them into medical decision-making [79,80,82]. As a result, the integration of interpretability mechanisms has become a critical research direction, leading to the development of Explainable Artificial Intelligence (XAI), which is discussed in the following section.

2.4 Explainable AI for Biomedical Signals

The increasing use of Artificial Intelligence (AI) in healthcare has created a growing demand for transparency and interpretability in machine learning models [87–89]. Although modern AI systems, particularly deep learning models, have achieved remarkable performance in analyzing biomedical signals, their complex architectures often operate as “black boxes.” In these systems, predictions are generated through intricate internal computations that are difficult for humans to interpret [79,80]. This lack of transparency poses significant challenges in clinical environments where diagnostic decisions directly impact patient safety and treatment outcomes [89].

Explainable Artificial Intelligence (XAI) has emerged as a critical field aimed at addressing these challenges [53]. XAI refers to a collection of techniques and methodologies designed to make the decision-making processes of AI models understandable to human users [87,88]. In healthcare, explainability is essential not only for improving trust in AI systems but also for ensuring accountability, regulatory compliance, and safe clinical deployment [87–89].

The importance of XAI is particularly evident in biomedical signal analysis [79,80,82]. Signals such as ECG, EEG, respiratory sounds, and oxygen saturation measurements contain complex physiological information that clinicians rely on to diagnose and monitor disease

[79,80,82]. When AI models are used to interpret these signals, clinicians must be able to verify that the model is basing its predictions on physiologically meaningful features rather than spurious correlations or noise [79,80,82]. Without such validation, the adoption of AI-based diagnostic systems in clinical practice remains limited [79,80,82].

XAI methods can generally be categorized into two broad groups: inherently interpretable models and post-hoc explanation techniques [89]. Inherently interpretable models are designed so that their decision-making process can be directly examined [87–89]. Examples include linear regression models, decision trees, and rule-based systems [87–89]. While these models offer transparency, they may lack the predictive power required for complex biomedical data [87–89]. Consequently, many high-performing deep learning models require post hoc explanation methods to analyze and interpret predictions after training [87–89].

Several XAI techniques have been developed for interpreting AI models applied to biomedical signals [87–89]. One widely used approach is SHAP (Shapley Additive Explanations), which is based on cooperative game theory [90]. SHAP assigns an importance value to each feature in the input data, quantifying its contribution to the model’s prediction [90]. In biomedical signal analysis, SHAP can highlight specific waveform segments or frequency components that strongly influence diagnostic outcomes [87–89]. For example, in ECG analysis, SHAP may identify segments of the QRS complex or ST segment that are most relevant to detecting arrhythmias [91].

Another commonly used method is LIME (Local Interpretable Model-agnostic Explanations), which generates local approximations of complex models [92]. LIME works by perturbing input data and observing how the model’s predictions change [92]. It then constructs a simple surrogate model that explains the prediction for a specific instance [88,89]. In biomedical

signal applications, LIME can highlight time–frequency regions of spectrograms or signal segments that influence the model’s classification decision [93].

Uncertainty quantification techniques such as Conformal Prediction (CP) have also gained attention in medical AI applications [80]. Unlike traditional explanation methods that focus solely on feature importance, conformal prediction provides statistically valid measures of prediction uncertainty [94]. This capability is particularly valuable in healthcare settings, where clinicians must assess the reliability of automated predictions before making treatment decisions [95,96]. By providing prediction intervals or label sets with guaranteed confidence levels, CP allows AI systems to express uncertainty and flag ambiguous cases for further clinical evaluation [80,94–96].

Gradient-based attribution techniques have also been widely used for interpreting deep learning models applied to biomedical signals [80]. Integrated Gradients (IG) calculates feature importance by integrating gradients of the model’s output with respect to the input features [97,98]. This method is particularly useful for time-series data, as it can reveal which portions of a signal contribute most strongly to the model’s prediction [91]. Similarly, Grad-CAM (Gradient-weighted Class Activation Mapping) generates visual explanations by highlighting important regions in time–frequency representations such as spectrograms [80,99].

Counterfactual explanations provide another form of interpretability by identifying minimal changes to input data that would alter the model’s prediction [89]. In biomedical signal analysis, counterfactual explanations can reveal how physiological signals would need to change for a different diagnostic outcome to occur [89]. For instance, a counterfactual explanation might indicate how small modifications in respiratory airflow patterns could shift a classification from

obstructive sleep apnea to normal breathing [89]. Such explanations can provide actionable insights for clinicians and support personalized patient management [89].

The integration of XAI into biomedical signal analysis offers several important clinical benefits. First, it enhances diagnostic confidence by allowing clinicians to verify that AI predictions align with known physiological mechanisms [87,88]. Second, it facilitates improved decision-making by providing interpretable insights that complement traditional clinical knowledge [89]. Third, XAI can assist in medical education by illustrating how complex signals are interpreted by AI models [89]. Finally, explainability supports regulatory compliance by ensuring that AI systems meet transparency requirements established by healthcare authorities and regulatory frameworks [80].

Despite these advantages, the application of XAI in biomedical signal analysis also presents several challenges. Some explanation methods can be computationally expensive, particularly when applied to large datasets or real-time monitoring systems [80]. Additionally, explanations generated by certain techniques may vary depending on data sampling or model sensitivity, potentially leading to inconsistent interpretations [80]. Therefore, careful validation and evaluation of XAI methods are necessary before deploying them in clinical environments [80].

Overall, Explainable Artificial Intelligence plays a crucial role in bridging the gap between high-performance machine learning models and clinical usability. By providing transparent insights into AI-driven predictions, XAI enables healthcare professionals to better understand, trust, and effectively utilize AI technologies in biomedical signal analysis. As AI continues to evolve and integrate into medical practice, the development of robust, reliable, and clinically meaningful explainability techniques will remain a fundamental research priority.

References

- [1] Rizzo, D.; Baltzan, M.; Sirpal, S.; Dosman, J.; Kaminska, M.; Chung, F. Prevalence and Regional Distribution of Obstructive Sleep Apnea in Canada: Analysis from the Canadian Longitudinal Study on Aging. *Canadian Journal of Public Health* 2024, 115, 970–979, doi:10.17269/s41997-024-00911-8.
- [2] Lechat, B.; Naik, G.; Reynolds, A.; Aishah, A.; Scott, H.; Loffler, K.A.; Vakulin, A.; Escourrou, P.; McEvoy, R.D.; Adams, R.J.; et al. Multinight Prevalence, Variability, and Diagnostic Misclassification of Obstructive Sleep Apnea. *Am. J. Respir. Crit. Care Med.* 2022, 205, 563–569, doi:10.1164/rccm.202107-1761OC.
- [3] Singh, M.; Liao, P.; Kobah, S.; Wijeyesundera, D.N.; Shapiro, C.; Chung, F. Proportion of Surgical Patients with Undiagnosed Obstructive Sleep Apnoea. *Br. J. Anaesth.* 2013, 110, 629–636, doi:10.1093/bja/aes465.
- [4] Colten, H.R.; Altevogt, B.M. *Sleep Disorders and Sleep Deprivation*; National Academies Press: Washington, D.C., 2006; ISBN 978-0-309-10111-0.
- [5] American Academy of Sleep Medicine *International Classification of Sleep Disorders: Diagnostic & Coding Manual*; 2nd ed.; American Academy of Sleep Medicine: Westchester, IL, 2005; ISBN 0965722023.
- [6] Kushida, C.A.; Littner, M.R.; Morgenthaler, T.; Alessi, C.A.; Bailey, D.; Coleman, J.; Friedman, L.; Hirshkowitz, M.; Kapen, S.; Kramer, M.; et al. Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005. *Sleep* 2005, 28, 499–523, doi:10.1093/sleep/28.4.499.
- [7] Yoshihisa, A.; Takeishi, Y. Sleep Disordered Breathing and Cardiovascular Diseases. *J. Atheroscler. Thromb.* 2019, 26, 315–327, doi:10.5551/JAT.RV17032.

- [8] Bradley, T.D.; Floras, J.S. Sleep Apnea: Implications in Cardiovascular and Cerebrovascular Disease; Bradley, T.D., Floras, J.S., Eds.; 2nd ed.; CRC Press: Boca Raton, 2013; ISBN 978-0-429-11867-8.
- [9] Espiritu, J.R.D. Health Consequences of Obstructive Sleep Apnea. In Management of Obstructive Sleep Apnea; Springer International Publishing: Cham, 2021; pp. 23–43.
- [10] Butt, M.; Dwivedi, G.; Khair, O.; Lip, G.Y.H. Obstructive Sleep Apnea and Cardiovascular Disease. *Int. J. Cardiol.* 2010, 139, 7–16, doi:10.1016/J.IJCARD.2009.05.021.
- [11] Corral-Peñafiel, J.; Pepin, J.-L.; Barbe, F. Ambulatory Monitoring in the Diagnosis and Management of Obstructive Sleep Apnoea Syndrome. *European Respiratory Review* 2013, 22, 312–324, doi:10.1183/09059180.00004213.
- [12] Walia, R.; Achilefu, A.; Crawford, S.; Jain, V.; Wigley, S.D.; McCarthy, L.H. Are At-Home Sleep Studies Performed Using Portable Monitors (PMs) as Effective at Diagnosing Obstructive Sleep Apnea (OSA) in Adults as Sleep Laboratory-Based Polysomnography (PSG)? *J. Okla. State Med. Assoc.* 2014, 107, 642–644.
- [13] Povitz, M.; Bray Jenkyn, K.; Kendzerska, T.; Allen, B.; Pendharkar, S.R.; Ouedraogo, A.; George, C.F.P.; Shariff, S.Z. Clinical Pathways and Wait Times for OSA Care in Ontario, Canada: A Population Cohort Study. *Canadian Journal of Respiratory, Critical Care, and Sleep Medicine* 2019, 3, 91–99, doi:10.1080/24745332.2018.1512841.
- [14] Berry, R.B.; Brooks, R.; Gamaldo, C.E.; Harding, S.M.; Marcus, C.; Vaughn, B. V; others The AASM Manual for the Scoring of Sleep and Associated Events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine 2012, 176, 2012.

- [15] American Academy of Sleep Medicine Hidden Health Crisis Costing America Billions: Underdiagnosing and Undertreating Obstructive Sleep Apnea Draining Healthcare System; 1st ed.; Frost & Sullivan : Darien, IL, 2016;
- [16] The Harvard Medical School Division of Sleep Medicine The Price of Fatigue: The Surprising Economic Costs of Unmanaged Sleep Apnea; Harvard Medical School Division of Sleep Medicine Boston: Boston, MA, 2010;
- [17] Yadollahi, A.; Moussavi, Z. Acoustic Obstructive Sleep Apnea Detection. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; IEEE, September 2009; pp. 7110–7113.
- [18] Elwali, A.; Moussavi, Z. Obstructive Sleep Apnea Screening and Airway Structure Characterization During Wakefulness Using Tracheal Breathing Sounds. *Ann. Biomed. Eng.* 2016, doi:10.1007/s10439-016-1720-5.
- [19] Montazeri, A.; Giannouli, E.; Moussavi, Z. Assessment of Obstructive Sleep Apnea and Its Severity during Wakefulness. *Ann. Biomed. Eng.* 2012, 40, 916–924, doi:10.1007/s10439-011-0456-5.
- [20] Hajipour, F.; Jozani, M.J.; Moussavi, Z. A Comparison of Regularized Logistic Regression and Random Forest Machine Learning Models for Daytime Diagnosis of Obstructive Sleep Apnea. *Med. Biol. Eng. Comput.* 2020, 58, 2517–2529, doi:10.1007/s11517-020-02206-9.
- [21] Hajipour, F.; Jozani, M.J.; Elwali, A.; Moussavi, Z. Regularized Logistic Regression for Obstructive Sleep Apnea Screening during Wakefulness Using Daytime Tracheal Breathing Sounds and Anthropometric Information. *Med. Biol. Eng. Comput.* 2019, 57, 2641–2655, doi:10.1007/s11517-019-02052-4.

- [22] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea Using Anthropometric Information and Tracheal Breathing Sounds. *Sci. Rep.* 2019, 9, 11467, doi:10.1038/s41598-019-47998-5.
- [23] Chen, L.; Pivetta, B.; Nagappa, M.; Saripella, A.; Islam, S.; Englesakis, M.; Chung, F. Validation of the STOP-Bang Questionnaire for Screening of Obstructive Sleep Apnea in the General Population and Commercial Drivers: A Systematic Review and Meta-Analysis. *Sleep and Breathing* 2021, 25, 1741–1751, doi:10.1007/S11325-021-02299-Y.
- [24] Mazzotti, D.R.; Keenan, B.T.; Thorarinsdottir, E.H.; Gislason, T.; Pack, A.I.; Schwab, R.; Maislin, G.; Jafari, N.; Boland, M.; Pack, F.; et al. Is the Epworth Sleepiness Scale Sufficient to Identify the Excessively Sleepy Subtype of OSA? *Chest* 2022, 161, 557–561, doi:10.1016/J.CHEST.2021.10.027.
- [25] J. Gordon Betts; Peter DeSaix; Eddie Johnson; Jody E. Johnson; Oksana Korol; Dean H. Kruse; Brandon Poe; James A. Wise *Anatomy and Physiology*; 2nd ed.; OpenStax: Houston, TX, 2022;
- [26] Horner, R.L. Motor Control of the Pharyngeal Musculature and Implications for the Pathogenesis of Obstructive Sleep Apnea. *Sleep* 1996, 19, 827–853, doi:10.1093/sleep/19.10.827.
- [27] White, D.P. The Pathogenesis of Obstructive Sleep Apnea. *Am. J. Respir. Cell Mol. Biol.* 2006, 34, 1–6, doi:10.1165/rcmb.2005-0317OE.
- [28] Mezzanotte, W.S.; Tangel, D.J.; White, D.P. Waking Genioglossal Electromyogram in Sleep Apnea Patients versus Normal Controls (a Neuromuscular Compensatory Mechanism). *Journal of Clinical Investigation* 1992, 89, 1571–1579, doi:10.1172/JCI115751.
- [29] Jordan, A.S.; McSharry, D.G.; Malhotra, A. Adult Obstructive Sleep Apnoea. *The Lancet* 2014, 383, 736–747, doi:10.1016/S0140-6736(13)60734-5.

- [30] Dempsey, J.A.; Veasey, S.C.; Morgan, B.J.; O'Donnell, C.P. Pathophysiology of Sleep Apnea. *Physiol. Rev.* 2010, 90, 47–112, doi:10.1152/physrev.00043.2008.
- [31] Guyenet, P.G. Regulation of Breathing and Autonomic Outflows by Chemoreceptors. In *Comprehensive Physiology*; Wiley, 2014; pp. 1511–1562.
- [32] Gleeson, K.; Zwillich, C.W.; White, D.P. The Influence of Increasing Ventilatory Effort on Arousal from Sleep. *American Review of Respiratory Disease* 1990, 142, 295–300, doi:10.1164/ajrccm/142.2.295.
- [33] Sagaut, Pierre.; Deck, Sebastien.; Terracol, Marc. Multiscale and Multiresolution Approaches in Turbulence. 2006, 340.
- [34] Narkiewicz, K.; Montano, N.; Cogliati, C.; van de Borne, P.J.H.; Dyken, M.E.; Somers, V.K. Altered Cardiovascular Variability in Obstructive Sleep Apnea. *Circulation* 1998, 98, 1071–1077, doi:10.1161/01.CIR.98.11.1071.
- [35] Somers, V.K.; Dyken, M.E.; Clary, M.P.; Abboud, F.M. Sympathetic Neural Mechanisms in Obstructive Sleep Apnea. *Journal of Clinical Investigation* 1995, 96, 1897–1904, doi:10.1172/JCI118235.
- [36] Joosten, S.A.; Landry, S.A.; Sands, S.A.; Terrill, P.I.; Mann, D.; Andara, C.; Skuza, E.; Turton, A.; Berger, P.; Hamilton, G.S.; et al. Dynamic Loop Gain Increases upon Adopting the Supine Body Position during Sleep in Patients with Obstructive Sleep Apnoea. *Respirology* 2017, 22, 1662–1669, doi:10.1111/resp.13108.
- [37] Edwards, B.A.; Sands, S.A.; Eckert, D.J.; White, D.P.; Butler, J.P.; Owens, R.L.; Malhotra, A.; Wellman, A. Acetazolamide Improves Loop Gain but Not the Other Physiological Traits Causing Obstructive Sleep Apnoea. *J. Physiol.* 2012, 590, 1199–1211, doi:10.1113/jphysiol.2011.223925.

- [38] Younes, M.; Ostrowski, M.; Thompson, W.; Leslie, C.; Shewchuk, W. Chemical Control Stability in Patients with Obstructive Sleep Apnea. *Am. J. Respir. Crit. Care Med.* 2001, 163, 1181–1190, doi:10.1164/ajrccm.163.5.2007013.
- [39] Eckert, D.J.; White, D.P.; Jordan, A.S.; Malhotra, A.; Wellman, A. Defining Phenotypic Causes of Obstructive Sleep Apnea. Identification of Novel Therapeutic Targets. *Am. J. Respir. Crit. Care Med.* 2013, 188, 996–1004, doi:10.1164/rccm.201303-0448OC.
- [40] Alqudah, A.M.; Elwali, A.; Kupiak, B.; Hajipour, F.; Jacobson, N.; Moussavi, Z. Obstructive Sleep Apnea Detection during Wakefulness: A Comprehensive Methodological Review. *Med. Biol. Eng. Comput.* 2024, 62, 1277–1311, doi:10.1007/s11517-024-03020-3.
- [41] Ciscar, M.A.; Juan, G.; Martõ Ànez, V.; Ramo, M.; Lloret, T.; Mõ Ànguez, J.; Armengot, M.; Marõ Àn, J.; Basterra, J.; Ciscar, G.M.A.; et al. Magnetic Resonance Imaging of the Pharynx in OSA Patients and Healthy Subjects. *European Respiratory Journal* 2001, 17, 79–86, doi:10.1183/09031936.01.17100790.
- [42] Brasileira, R.; Veterin, P. Diagnosis of OSA Using Speech. 2020, 1–62.
- [43] Simply, R.M.; Dafna, E.; Zigel, Y. Obstructive Sleep Apnea (OSA) Classification Using Analysis of Breathing Sounds During Speech. In *Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO)*; IEEE, September 29 2018; Vol. 2018-September, pp. 1132–1136.
- [44] Balaei, A.T.; Sutherland, K.; Cistulli, P.A.; de Chazal, P. Automatic Detection of Obstructive Sleep Apnea Using Facial Images. In *Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*; IEEE, April 2017; pp. 215–218.

- [45] Tamisier, R.; Wuyam, B.; Nicolle, I.; Pépin, J.L.; Orliaguet, O.; Perrin, C.P.; Lévy, P. Awake Flow Limitation with Negative Expiratory Pressure in Sleep Disordered Breathing. *Sleep Med.* 2005, 6, 205–213, doi:10.1016/j.sleep.2004.10.013.
- [46] Nagappa, M.; Liao, P.; Wong, J.; Auckley, D.; Ramachandran, S.K.; Memtsoudis, S.; Mokhlesi, B.; Chung, F. Validation of the Stop-Bang Questionnaire as a Screening Tool for Obstructive Sleep Apnea among Different Populations: A Systematic Review and Meta-Analysis. *PLoS One* 2015, 10, doi:10.1371/journal.pone.0143697.
- [47] Younes, M. Role of respiratory control mechanisms in the pathogenesis of obstructive sleep disorders. *J Appl Physiol* (1985) 2008, 105, 1389-1405, doi:10.1152/jappphysiol.90408.2008.
- [48] Lan, Z.; Itoi, A.; Takashima, M.; Oda, M.; Tomoda, K. Difference of Pharyngeal Morphology and Mechanical Property between OSAHS Patients and Normal Subjects. *Auris Nasus Larynx* 2006, 33, 433–439, doi:10.1016/J.AN.L.2006.03.009.
- [49] Malhotra, A.; Pillar, G.; Fogel, R.; Beauregard, J.; Edwards, J.; White, D.P. Upper-Airway Collapsibility: Measurements and Sleep Effects. *Chest* 2001, 120, 156–161, doi:10.1378/CHEST.120.1.156.
- [50] Finkelstein, Y.; Wolf, L.; Nachmani, A.; Lipowezky, U.; Rub, M.; Shemer, S.; Berger, G. Velopharyngeal Anatomy in Patients with Obstructive Sleep Apnea versus Normal Subjects. *J. Oral Maxillofac. Surg.* 2014, 72, 1350–1372, doi:10.1016/J.JOMS.2013.12.006.
- [51] Pham, L. V.; Schwartz, A.R. The Pathogenesis of Obstructive Sleep Apnea. *J. Thorac. Dis.* 2015, 7, 1358, doi:10.3978/J.ISSN.2072-1439.2015.07.28.
- [52] Okubo, M.; Suzuki, M.; Horiuchi, A.; Okabe, S.; Ikeda, K.; Higano, S.; Mitani, H.; Hida, W.; Kobayashi, T.; Sugawara, J. Morphologic Analyses of Mandible and Upper Airway Soft

Tissue by MRI of Patients with Obstructive Sleep Apnea Hypopnea Syndrome. *Sleep* 2006, 29, 909–915, doi:10.1093/SLEEP/29.7.909.

[53] Heo, J.Y.; Kim, J.S. Correlation between Severity of Sleep Apnea and Upper Airway Morphology: Cephalometry and MD-CT Study during Awake and Sleep States. *Acta Otolaryngol.* 2010, 131, 84–90, doi:10.3109/00016489.2010.514007.

[54] Carrera, H.L.; Marcus, C.L.; McDonough, J.M.; Morera, J.C.O.; Huang, J.; Farre, R.; Montserrat, J.M. Negative Expiratory Pressure Technique: An Awake Test to Measure Upper Airway Collapsibility in Adolescents. *Sleep* 2015, 38, 1783, doi:10.5665/SLEEP.5164.

[55] Romano, S.; Salvaggio, A.; Hirata, R.P.; Bue, A. Lo; Picciolo, S.; de Oliveira, L.V.F.; Insalaco, G. Upper Airway Collapsibility Evaluated by a Negative Expiratory Pressure Test in Severe Obstructive Sleep Apnea. *Clinics* 2011, 66, 567, doi:10.1590/S1807-59322011000400008.

[56] Islam, S.M.S.; Mahmood, H.; Al-Jumaily, A.A.; Claxton, S. Deep Learning of Facial Depth Maps for Obstructive Sleep Apnea Prediction. In *Proceedings of the 2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*; IEEE, December 2018; pp. 154–157.

[57] Shivarov, G. Anthropometry and Acoustic Pharyngometry of the Oral Cavity in Sleep-Disordered Breathing. *Journal of IMAB – Annual Proceeding Scientific Papers* 2022, 28, 4377–4383, doi:10.5272/JIMAB.2022282.4377.

[58] DeYoung, P.N.; Bakker, J.P.; Sands, S.A.; Batool-Anwar, S.; Connolly, J.G.; Butler, J.P.; Malhotra, A. Acoustic Pharyngometry Measurement of Minimal Cross-Sectional Airway Area Is a Significant Independent Predictor of Moderate-to-Severe Obstructive Sleep Apnea. *Journal of Clinical Sleep Medicine* 2013, 9, 1161–1164, doi:10.5664/jcsm.3158.

[59] Dae, G.J.; Hae, Y.C.; Grunstein, R.R.; Yee, B. Predictive Value of Kushida Index and Acoustic Pharyngometry for the Evaluation of Upper Airway in Subjects with or without

Obstructive Sleep Apnea. J. Korean Med. Sci. 2004, 19, 662–667, doi:10.3346/jkms.2004.19.5.662.

[60] Shams, E.; Karimi, D.; Moussavi, Z. Bispectral Analysis of Tracheal Breath Sounds for Obstructive Sleep Apnea. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2012, 37–40, doi:10.1109/EMBC.2012.6345865.

[61] Elwali, A.; Meza-Vargas, S.; Moussavi, Z. Using Tracheal Breathing Sounds and Anthropometric Information for Screening Obstructive Sleep Apnoea during Wakefulness. J. Med. Eng. Technol. 2019, 43, 111–123, doi:10.1080/03091902.2019.1617799.

[62] Sola-Soler, J.; Fiz, J.A.; Torres, A.; Jane, R. Identification of Obstructive Sleep Apnea Patients from Tracheal Breath Sound Analysis during Wakefulness in Polysomnographic Studies. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; IEEE, August 2014; pp. 4232–4235.

[63] Hajipour, F.; Moussavi, Z. Spectral and Higher Order Statistical Characteristics of Expiratory Tracheal Breathing Sounds During Wakefulness and Sleep in People with Different Levels of Obstructive Sleep Apnea. J. Med. Biol. Eng. 2019, 39, 244–250, doi:10.1007/s40846-018-0409-7.

[64] Ben Or, D.; Dafna, E.; Tarasiuk, A.; Zigel, Y. Obstructive Sleep Apnea Severity Estimation: Fusion of Speech-Based Systems. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2016, 2016-Octob, 3207–3210, doi:10.1109/EMBC.2016.7591411.

[65] Pozo, R.F.; Murillo, J.L.B.; Gmez, L.H.; Gonzalo, E.L.; Ramírez, J.A.; Toledano, D.T. Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker

Recognition Techniques. EURASIP Journal on Advances in Signal Processing 2009 2009:1 2009, 2009, 1–11, doi:10.1155/2009/982531.

[66] Yılmaz, D.; Yıldız, M.; Uyar Toprak, Y.; Yetkin, S. Obstructive Sleep Apnea Detection with Nonlinear Analysis of Speech. Biomed. Signal Process. Control 2023, 84, 104956, doi:10.1016/j.bspc.2023.104956.

[67] Goldshtein, E.; Tarasiuk, A.; Zigel, Y. Automatic Detection of Obstructive Sleep Apnea Using Speech Signals. IEEE Trans. Biomed. Eng. 2011, 58, 1373–1382, doi:10.1109/TBME.2010.2100096.

[68] Pang, K.-G.; Hsung, T.-C.; Liao, G.; Ling, W.-K.; Law, A.K.-W.; Choi, W.S. Obstructive Sleep Apnea Detection Using Speech Signals with High Frequency Components. Journal of Communications 2022, 17, 49–55, doi:10.12720/jcm.17.1.49-55.

[69] Espinoza-Cuadros, F.; Fernández-Pozo, R.; Toledano, D.T.; Alcázar-Ramírez, J.D.; López-Gonzalo, E.; Hernández-Gómez, L.A. Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment. Comput. Math. Methods Med. 2015, 2015, 1–14, doi:10.1155/2015/489761.

[70] Simply, R.M.; Dafna, E.; Zigel, Y. Diagnosis of Obstructive Sleep Apnea Using Speech Signals from Awake Subjects. IEEE Journal on Selected Topics in Signal Processing 2020, 14, 251–260, doi:10.1109/JSTSP.2019.2955019.

[71] Simply, R.M.; Dafna, E.; Zigel, Y. Obstructive Sleep Apnea (OSA) Classification Using Analysis of Breathing Sounds During Speech. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO); IEEE, September 2018; pp. 1132–1136.

- [72] Kriboy, M.; Tarasiuk, a.; Zigel, Y. Detection of Obstructive Sleep Apnea in Awake Subjects by Exploiting Body Posture Effects on the Speech Signal. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2014, 2014, 4224–4227, doi:10.1109/EMBC.2014.6944556.
- [73] Montero Benavides, A.; Fernández Pozo, R.; Toledano, D.T.; Blanco Murillo, J.L.; López Gonzalo, E.; Hernández Gómez, L. Analysis of Voice Features Related to Obstructive Sleep Apnoea and Their Application in Diagnosis Support. *Comput. Speech Lang.* 2014, 28, 434–452, doi:10.1016/J.CSL.2013.08.002.
- [74] Behar, J.A.; Palmius, N.; Daly, J.; Li, Q.; Rizzatti, F.G.; Bittencourt, L.; Clifford, G.D. Sleep Questionnaires in Screening for Obstructive Sleep Apnoea. *Comput. Cardiol.* (2010). 2017, 44, 1–4, doi:10.22489/CinC.2017.233-188.
- [75] Kump, K.; Whalen, C.; Tishler, P. V.; Browner, I.; Ferrette, V.; Strohl, K.P.; Rosenberg, C.; Redline, S. Assessment of the Validity and Utility of a Sleep-Symptom Questionnaire. *Am. J. Respir. Crit. Care Med.* 1994, 150, 735–741, doi:10.1164/ajrccm.150.3.8087345.
- [76] El-Sayed, I.H. Comparison of Four Sleep Questionnaires for Screening Obstructive Sleep Apnea. *Egyptian Journal of Chest Diseases and Tuberculosis* 2012, 61, 433–441, doi:10.1016/J.EJCDT.2012.07.003.
- [77] Pataka, A.; Daskalopoulou, E.; Kalamaras, G.; Fekete Passa, K.; Argyropoulou, P. Evaluation of Five Different Questionnaires for Assessing Sleep Apnea Syndrome in a Sleep Clinic. *Sleep Med.* 2014, 15, 776–781, doi:10.1016/J.SLEEP.2014.03.012.
- [78] Solecka, Š.; Matler, K.; Kostlivý, T.; Kubec, V.; Tomášková, H.; Betka, J. A Comparison of the Reliability of Five Sleep Questionnaires for the Detection of Obstructive Sleep Apnea. . 2022, 12, 1416, doi:10.3390/life12091416.

- [79] Alqudah, A.M.; Moussavi, Z. Bridging Signal Intelligence and Clinical Insight: A Comprehensive Review of Feature Engineering, Model Interpretability, and Machine Learning in Biomedical Signal Analysis. *Applied Sciences* 2025, 15, 12036, doi:10.3390/app152212036.
- [80] Alqudah, A.M.; Moussavi, Z. A Review of Deep Learning for Biomedical Signals: Current Applications, Advancements, Future Prospects, Interpretation, and Challenges. *Computers, Materials & Continua* 2025, 83, 3753–3841, doi:10.32604/cmc.2025.063643.
- [81] Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep Learning for Healthcare: Review, Opportunities and Challenges. *Brief. Bioinform.* 2018, 19, 1236–1246, doi:10.1093/bib/bbx044.
- [82] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* 2015, 521, 436–444, doi:10.1038/nature14539.
- [83] Perez Alday, E.A.; Gu, A.; J Shah, A.; Robichaux, C.; Ian Wong, A.-K.; Liu, C.; Liu, F.; Bahrami Rad, A.; Elola, A.; Seyedi, S.; et al. Classification of 12-Lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *Physiol. Meas.* 2020, 41, 124003, doi:10.1088/1361-6579/abc960.
- [84] Samee, N.A.; Mahmoud, N.F.; Aldhahri, E.A.; Rafiq, A.; Muthanna, M.S.A.; Ahmad, I. RNN and BiLSTM Fusion for Accurate Automatic Epileptic Seizure Diagnosis Using EEG Signals. . 2022, 12, 1946, doi:10.3390/life12121946.
- [85] Alqudah, A.M.; Qazan, S.; Obeidat, Y.M. Deep Learning Models for Detecting Respiratory Pathologies from Raw Lung Auscultation Sounds. *Soft comput.* 2022, 26, 13405–13429, doi:10.1007/s00500-022-07499-6.
- [86] Goshtasbi, N.; Boostani, R.; Sanei, S. SleepFCN: A Fully Convolutional Deep Learning Framework for Sleep Stage Classification Using Single-Channel Electroencephalograms. *IEEE*

Transactions on Neural Systems and Rehabilitation Engineering 2022, 30, 2088–2096, doi:10.1109/TNSRE.2022.3192988.

[87] Majhi, B.; Kashyap, A. Explainable AI-Driven Machine Learning for Heart Disease Detection Using ECG Signal. *Appl. Soft Comput.* 2024, 167, 112225, doi:10.1016/j.asoc.2024.112225.

[88] Seoni, S.; Jahmunah, V.; Salvi, M.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of Uncertainty Quantification to Artificial Intelligence in Healthcare: A Review of Last Decade (2013–2023). *Comput. Biol. Med.* 2023, 165, 107441, doi:10.1016/j.compbio.2023.107441.

[89] Ali Mohammad Alqudah; Zahra Moussavi Towards Transparent Diagnostics Explainable AI Across Biomedical Signals. In *Proceedings of the 2025 International Conference on Innovation and Intelligence for Informatics, Computing, And Technologies (3ICT25)*; IEEE: Sakhir, Bahrain, 2026.

[90] Shapley, L.S. A Value for N-Person Games. *Contributions to the Theory of Games (AM-28)*, Volume II 2016, 307–318, doi:10.1515/9781400881970-018.

[91] Alqudah, A.M.; Alfraihat, A. A Multimodal Approach for Cardiac Signals Classification Using Deep Learning with Explainable AI Methods. *Health Inf. Sci. Syst.* 2025, 14, 5, doi:10.1007/s13755-025-00402-1.

[92] Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* 2020, 2, 56–67, doi:10.1038/s42256-019-0138-9.

[93] Masri, S.; Hasasneh, A.; Tami, M.; Tadj, C. Exploring the Impact of Image-Based Audio Representations in Classification Tasks Using Vision Transformers and Explainable AI Techniques. *Information* 2024, 15, 751, doi:10.3390/info15120751.

- [94] ShaferGlenn; VovkVladimir A Tutorial on Conformal Prediction. *The Journal of Machine Learning Research* 2008, doi:10.5555/1390681.1390693.
- [95] Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-Free Predictive Inference for Regression. *J. Am. Stat. Assoc.* 2018, 113, 1094–1111, doi:10.1080/01621459.2017.1307116.
- [96] *Algorithmic Learning in a Random World*; Springer-Verlag: New York, 2005; ISBN 0-387-00152-2.
- [97] Kawai, Y.; Tachikawa, K.; Park, J.; Asada, M. Compensated Integrated Gradients for Reliable Explanation of Electroencephalogram Signal Classification. *Brain Sci.* 2022, 12, 849, doi:10.3390/brainsci12070849.
- [98] Jha, A.; K. Aicher, J.; R. Gazzara, M.; Singh, D.; Barash, Y. Enhanced Integrated Gradients: Improving Interpretability of Deep Learning Models Using Splicing Codes as a Case Study. *Genome Biol.* 2020, 21, 149, doi:10.1186/s13059-020-02055-7.
- [99] Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* 2020, 128, 336–359, doi:10.1007/s11263-019-01228-7.

CHAPTER THREE: DEEP LEARNING MODEL FOR OSA DETECTION USING TRACHEAL BREATHING SOUNDS DURING WAKEFULNESS

Ali Mohammad Alqudah¹, and Zahra Moussavi^{1,2}

¹ Biomedical Engineering, University of Manitoba, Winnipeg, Canada

²Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada

CMBES Proceedings 45 (2023) (<https://proceedings.cmbes.ca/index.php/proceedings/article/view/1081>)

Abstract

The detection of Obstructive Sleep Apnea (OSA) during sleep is a simple and well-established technique; however, its detection during wakefulness is challenging. In this paper, we propose a deep learning model for the detection of OSA using only the tracheal breathing sounds spectrum as input. We employed our team's previous dataset consisting of 109 subjects as non or mild-OSA with apnea/hypopnea index (AHI) < 15 and 90 subjects as OSA with AHI ≥ 15 . All study subjects were referred to overnight polysomnography (PSG) to determine their AHI values. Tracheal breathing sounds were recorded in the supine position before proceeding to PSG while awake. The recording protocol was to have 5 deep breaths first through the mouth and then 5 deep breaths through the nose. Data were normalized and segmented into inspiratory/expiratory breathing phases; their power spectra were then calculated and fed to a deep learning model consisting of 71 layers. The results of 10 K-fold show that the proposed deep learning model achieved an accuracy of 74.9%, sensitivity of 76.1%, and specificity of 73.3%. Although these results are not as high as the previously reported analyses, they can be improved significantly by combining with anthropometric parameters and subgrouping subjects based on their age, weight, etc. This work aimed to show the potential of deep learning on this dataset despite the limited sample size. The results are encouraging to continue and improve the algorithm.

3.1. Introduction

A frequent syndrome known as obstructive sleep apnea (OSA) disorder is defined by recurrent episodes of the whole (apnea) or partial (hypopnea) pharyngeal collapse while sleeping. The affected people's quality of life may suffer significantly as a result of OSA [1]. The risks of developing heart disease, high blood pressure, stroke, depression, diabetes, headaches, and traffic accidents are among many associated effects of OSA [2].

OSA is also known to increase the risk of preoperative morbidity and/or postoperative mortality; thus, identifying individuals with OSA before a surgery that requires general anesthesia would lower the post-operative mortality risks [3].

The apnea/hypopnea events index (AHI) per hour is used to gauge the severity of sleep apnea. Polysomnography (PSG) is the gold standard for diagnosing OSA. However, the expensive nature of PSG due to its long procedure and need for trained technicians, lack of sleep labs in small towns, etc., there is typically a large waiting list for PSG assessments [4].

The STOP-BANG questionnaire is one of the various subjective OSA diagnosis/screening instruments that doctors employ for quick screening of OSA when there is no possibility of performing PSG overnight [5]. However, its specificity is very low (10%). As a result, such subjective screening is not reliable for quick decision-making.

Thus, there is a great need for an improved method of diagnosing and screening OSA individuals during wakefulness as a fast and robust screening tool. Such an approach will aid in reducing the need for PSG and will significantly help with decision-making prior to full anesthesia in regard to potential complications due to OSA in a patient [5].

To screen for OSA during wakefulness, several research teams around the world have investigated objective alternatives to current procedures by using tracheal breathing sound analysis

recorded while individuals were awake [6], [7], [8], [9], [10] and also using voice analysis [11], [12] with promising results.

A major shortcoming of the previous research on this topic is the dependency of the algorithms on pre-processing and feature selection; the automated feature extraction from the sound data using deep learning models was not explored [5].

Deep-learning techniques typically need large samples for learning. In this work, despite the limited sample size, our main objective was to investigate the potential of deep learning in extracting characteristic features sensitive to AHI without the need to know the type of extracted respiratory phase (either inspiration or expiration) and lengthy pre-processing.

3.2. Methodology

3.2.1. Dataset

Data were adopted from our team's previous work [6]. Each individual's tracheal breathing sounds were recorded by a microphone positioned at the suprasternal notch of the trachea for 5 complete cycles of deep breathing through the nose with the mouth closed and then another 5 breaths through the mouth while wearing a nasal clip using.

The recording was made while the subject was in a supine position and awake. In this study and like our previous works the Data (n=199) were grouped as mild/non-OSA (n=109, AHI < 15) and OSA (n=90, AHI $\geq$ 15) [6]. Table 3-1 displays the total number of subjects and the total number of breathing phases (both inspiration and expiration) for each class for the dataset used in this study.

Table 3-1 Dataset Distribution.

Class	AHI	Number of Subjects	Total Breathing Phases
non-OSA	AHI < 15	109	1840
OSA	AHI $\geq$ 15	90	1496

3.2.2. Preprocessing and Power Spectrum

A Butterworth band-pass filter of order 4 with a cutoff frequency of 75- 3000 Hz was first used to filter out each breathing sound phase independently in order to cancel out the impacts of different interferences such as heartbeats, background noise, etc. [7]. Second, in order to eliminate the impact of plausible airflow fluctuation between breathing cycles, each filtered signal was normalized by its variance envelope using a smoothed copy by applying a Moving Average window with a size of 64 samples sequence, then by its standard deviation (energy). After that, a 50% duration around each breathing phase's maximum was chosen for further analysis using the logarithm of the variance of the phase signals, representing the respiratory flow [7]. This duration roughly equates to the upper 40% of each breathing phase's respiratory airflow, where the signal is considered stationary within the breathing sound signal. The power spectrum density (PSD) for all middle regions of the signal was computed using the welch method with 25 ms length (256 samples window size), 50% overlapping between adjacent windows [7]. These preprocessing techniques are the same techniques used in our team's previously published works using machine learning methods.

3.2.3. Deep Learning Model

Deep learning is a recently developed technique in the artificial intelligence field, which is emerged as a response to the increased number of available and recorded large datasets [14]. Deep learning models can be distinguished and defined based on their unique design and development of the set of layers, usually called architecture. Such an architecture consists of a set of consecutive layers to process the model input and get the output [15]. This paper's proposed deep learning model is based on a combination of Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM).

The model consists of 71 layers where the input of the model is the 1D power spectrum of the breathing phase. Then the CNN part of the model will extract the deep features from the input data, the layers are the convolutional and the max pooling layers. The model takes the maximum and the average power spectrum by passing it through max pooling and average pooling layers, respectively. Then, both variables over the frequency band between 1 and 3700 Hz are passed through three stacked residual blocks to efficiently extract features. The output features are then passed through a series of fully connected and scaling layers to make the model “structure agnostic” about the input. Moreover, the LSTM part will use the input data characteristics as frequency-domain-based features to learn to extract the most significant part of the frequency data. Using the combination of two models allows shallower models with fewer parameters to be developed, resulting in better performance. Figure 3-1 shows the developed models and Table 3-2 shows the layer details.

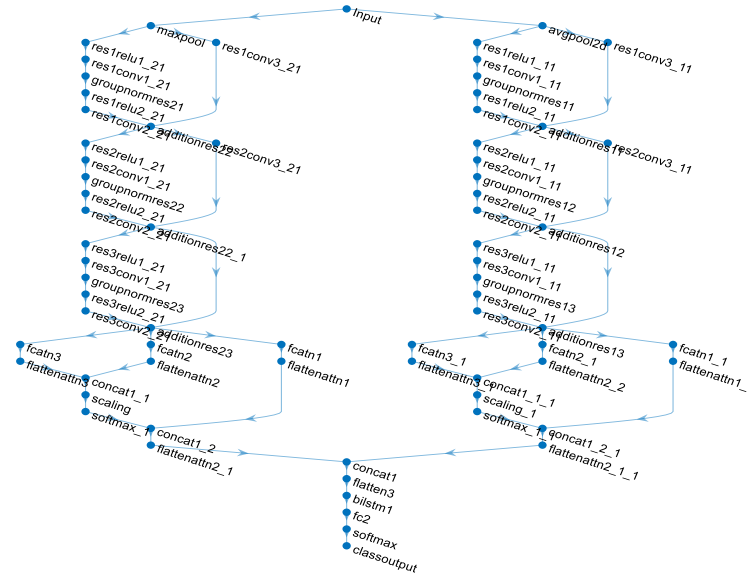


Fig. 3-1 The proposed deep learning model architecture.

3.3. Results

The deep learning model in this work was created by combining the CNN and LSTM models. All of the trials were carried out using a desktop computer with an Intel Core i7-12700H/2.3 GHz CPU, 32 GB of RAM, a 1 TB hard drive (HDD), and a 16 GB Nvidia GeForce RTX 3080 Ti graphics processing unit (GPU).

Our proposed model was tested using the 10-Kfold methodology, and one of the nine training folds was employed as validation during the cross-validation procedure; for each fold, the training was done using the stochastic gradient descent with momentum (SGDM) optimizer with a cross-entropy loss function, initial learning rate of 0.001, and mini-batch size of 25. These hyperparameters are selected using trial and error. Since there is a high correlation between breathing sounds of ins/exp phases of the same subject, the cross-validation classification process in this research was done based on each subject; where all power spectrum data related to the same subject's breathing phases are either in the training set, validation set, or testing set.

A two-step classification using the deep learning model was used to detect OSA. In the first step, called “phase-based”, the deep learning model was used to classify each breathing phase (either inspiration or expiration) into non-OSA or OSA groups. The outcomes of this step are representative of the ability of deep learning to detect OSA using phases regardless of knowing if they are inspiratory or expiratory phases. Figure 3-2 shows the confusion matrix and the ROC for this step.

In the second step, called “subject-based” detection, a final decision for each subject is generated by finding the most frequent class using a majority voting classifications technique among all breathing phase related to the same subject found in the first step. The outcomes of this

step are representative of the ability of the proposed deep learning model to screen for OSA subjects. Figure 3-3 shows the confusion matrix and the ROC for the second step.

Table 3-2 The proposed architecture layer details.

#	Name	Parameters	#	Name	Parameters
1	Input	Length: 3700	36	res1conv1_11	Filter: 8 *3,3 Stride: 1
2	maxpool	Pool Size: 5,5 Stride: 1	37	groupnormres11	Channels: 8
4	res1conv1_21	Filter: 8*8, 8 Stride: 1	39	res1conv2_11	Filter: 8* 3,3 Stride: 1
5	groupnormres21		40	res1conv3_11	Filter: 8*3,3 Stride: 1
7	res1conv2_21	Filter: 8*8, 8Stride: 1	43	res2conv1_11	Filter: 8*3,3 Stride: 1
8	res1conv3_21	Filter: 8* 8, 8 Stride: 1	44	groupnormres12	Channels: 8
11	res2conv1_21	Filter: 8* 8, 8 Stride: 1	46	res2conv2_11	Filter: 8*3,3 Stride: 1
12	groupnormres22	Channels: 8	47	res2conv3_11	Filter: 8*3,3 8 Stride: 1
14	res2conv2_21	Filter: 8*8, 8 Stride: 1	50	res3conv1_11	Filter: 8*3,3 Stride: 1
15	res2conv3_21	Filter: 8*8, 8 Stride: 1	51	groupnormres13	Channels: 8
18	res3conv1_21	Filter: 8*8, 8 Stride: 1	53	res3conv2_11	Filter: 8*3,3 Stride: 1
19	groupnormres23	Channels: 8	55	fc3_1	Output: 512
21	res3conv2_21	Filter: 8*8, 8 Stride: 1	57	fc2_1	Output: 512
23	fc3	Output: 512	59	fc1_1	Output: 512
25	fc2	Output: 512	68	lstm1	Num Hidden Units: 512
27	fc1	Output: 512	69	fc	Output: 2
34	avgpool2d	PoolSize: 5 5 Stride: 1			

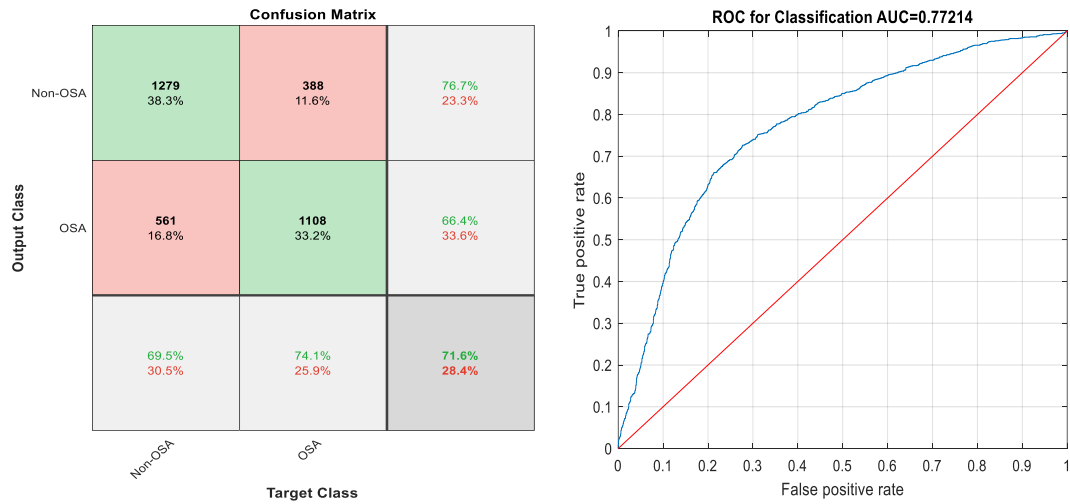


Fig. 3-2 The output of Phases-based training; A) Confusion matrix, B) ROC.

The performance metrics of both approaches (phase and subject-based) are shown in Table 3-3. These parameters are calculated based on the confusion matrix and ROC.

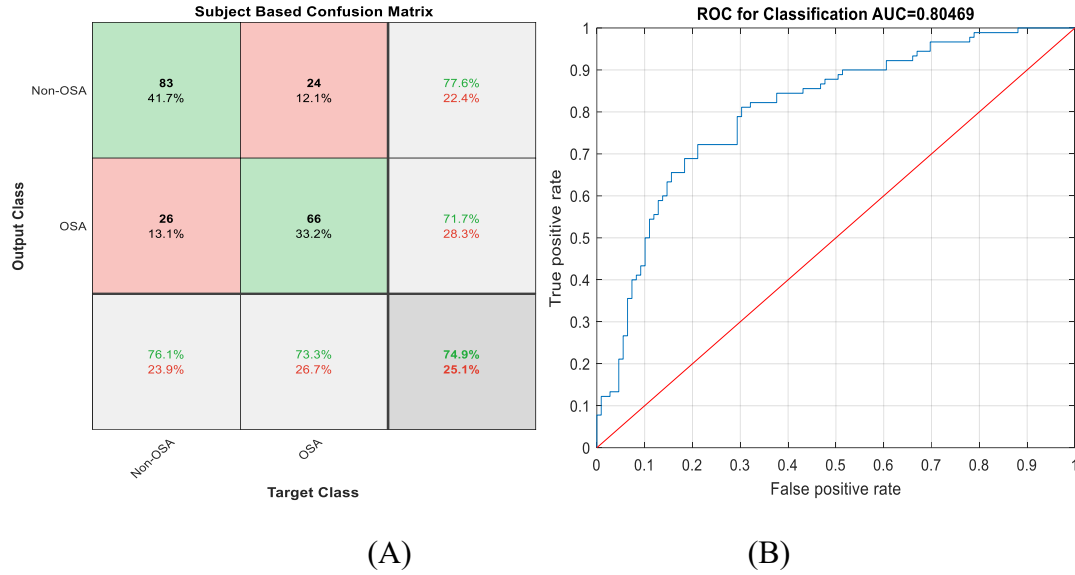


Fig. 3-3 The output of Subject-based training; A) Confusion matrix, B) ROC.

Table 3-3 Performance of the Proposed Methodology.

	Accuracy	Sensitivity	Specificity	Precision	AUC
Phase-Based	71.55	69.51	74.06	76.72	0.7721
Subject-Based	74.87	76.15	73.33	77.57	0.8047

3.4. Discussion

Employing tracheal breathing sounds in order to detect OSA during sleep is easy and well-established. However, using breathing sounds to detect OSA during wakefulness is very challenging as OSA individuals do not exhibit any breathing symptoms while awake [16]. While there have been research studies to use tracheal breathing sounds for screening OSA during [6], [7], [8], [17], this research presents the first attempt to apply deep-learning models to the tracheal breathing sounds recorded during wakefulness for screening OSA.

Out of the two classifications proposed in this study (phase-base and subject-base), the accuracy of the phase-based classification was found to be low (~71%). This is due to the shallow deep learning model, where the development of deep models requires more time for optimization and ablation studies for the effect of adding any new blocks such as attention. Moreover, the outcomes of this classification were fed to a majority vote classifier to have a subject-based

classification based on the class of the majority of the phases of a subject; thus, the low accuracy of the phase classicization projects into the subsequent subject-base classification although by majority voting of the phases' class for each subject the overall accuracy is increased to about 75% (Table 3).

Comparing the performance values of the proposed deep learning model with these previous works that used the same dataset [6], [7], [17] shows that the proposed methodology has lower performance, and still needs more development and enhancement over a large dataset to increase its overall performance. At the same time, the results are encouraging given the limited sample size. Table 3-4 shows a comparison between the proposed work and previous similar studies.

Table 3-4 *A Comparison Between Proposed Work and Previous Works.*

Ref	AHI Threshold	Accuracy	Sensitivity	Specificity
[7]	$AHI \leq 5$ $AHI \geq 10$	83.92	82.61	85.22
[13]	$AHI < 15$ & $AHI \geq 15$	81.4	80.9	82.1
Phase-Based		71.55	69.51	74.06
Subject-Based		74.87	76.15	73.33

Furthermore, more experiments are required to improve the proposed model's performance and stability. This could include but is not limited to using different initialization techniques, hyperparameters fine-tuning, employing a larger dataset, and combining the extracted deep features with anthropometric features as input to other classifiers, as well as subgrouping subjects based on their age, weight, etc.

3.5. Conclusion

This paper presents the outcomes of a deep learning-based detection of OSA using the breathing sounds power spectra, recorded during wakefulness, as input. While the accuracy of the

deep-learning model for screening OSA is lower than the previous methods using the same dataset, overall, the proposed method has the advantage of removing the need for knowing the respiratory phase as either inspiration or expiration. Given the limited sample size used in this study and yet achieving reasonable accuracy, the results encourage investigating the use of deep learning in the detection of OSA; we anticipate its accuracy to increase in a larger dataset.

References

- [1] Colten, H.R.; Altevogt, B.M. Sleep Disorders and Sleep Deprivation; National Academies Press: Washington, D.C., 2006; ISBN 978-0-309-10111-0.
- [2] Rosenthal, L.D.; Dolan, D.C. The Epworth Sleepiness Scale in the Identification of Obstructive Sleep Apnea. *J. Nerv. Ment. Dis.* 2008, 196, 429–431, doi:10.1097/NMD.0b013e31816ff3bf.
- [3] American Society of Anesthesiologists Practice Guidelines for the Perioperative Management of Patients with Obstructive Sleep Apnea: An Updated Report. *Anesthesiology* 2014, 120, 268–286, doi:10.1097/ALN.0000000000000053.
- [4] Kushida, C.A.; Littner, M.R.; Morgenthaler, T.; Alessi, C.A.; Bailey, D.; Coleman, J.; Friedman, L.; Hirshkowitz, M.; Kapen, S.; Kramer, M.; et al. Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005. *Sleep* 2005, 28, 499–523, doi:10.1093/sleep/28.4.499.
- [5] Nagappa, M.; Liao, P.; Wong, J.; Auckley, D.; Ramachandran, S.K.; Memtsoudis, S.; Mokhlesi, B.; Chung, F. Validation of the Stop-Bang Questionnaire as a Screening Tool for Obstructive Sleep Apnea among Different Populations: A Systematic Review and Meta-Analysis. *PLoS One* 2015, 10, doi:10.1371/journal.pone.0143697.

- [6] Moussavi, Z.; A.; Elwali, A. System and Methods for Screening Obstructive Sleep Apnea During Wakefulness Using Anthropometric Information and Tracheal Breathing Sounds. U.S. Patent Application US20210401364A1, 30 December 2021.
- [7] Elwali, A.; Moussavi, Z. Obstructive Sleep Apnea Screening and Airway Structure Characterization During Wakefulness Using Tracheal Breathing Sounds. *Ann. Biomed. Eng.* 2016, doi:10.1007/s10439-016-1720-5.
- [8] Sola-Soler, J.; Fiz, J.A.; Torres, A.; Jane, R. Identification of Obstructive Sleep Apnea Patients from Tracheal Breath Sound Analysis during Wakefulness in Polysomnographic Studies. In *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; IEEE, August 2014; pp. 4232–4235.
- [9] Sola-Soler, J.; Jane, R.; Fiz, J.A.; Morera, J. Formant Frequencies of Normal Breath Sounds of Snorers May Indicate the Risk of Obstructive Sleep Apnea Syndrome. In *Proceedings of the 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; IEEE, August 2008; pp. 3500–3503.
- [10] Simply, R.M.; Dafna, E.; Zigel, Y. Obstructive Sleep Apnea (OSA) Classification Using Analysis of Breathing Sounds During Speech. In *Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO)*; IEEE, September 29 2018; Vol. 2018-September, pp. 1132–1136.
- [11] Solé-Casals, J.; Munteanu, C.; Martín, O.C.; Barbé, F.; Queipo, C.; Amilibia, J.; Durán-Cantolla, J. Detection of Severe Obstructive Sleep Apnea through Voice Analysis. *Appl. Soft Comput.* 2014, 23, 346–354, doi:10.1016/J.ASOC.2014.06.017.
- [12] Pozo, R.F.; Murillo, J.L.B.; Gmez, L.H.; Gonzalo, E.L.; Ramírez, J.A.; Toledano, D.T. Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker

- Recognition Techniques. EURASIP Journal on Advances in Signal Processing 2009 2009:1 2009, 2009, 1–11, doi:10.1155/2009/982531.
- [13] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea Using Anthropometric Information and Tracheal Breathing Sounds. Sci. Rep. 2019, 9, 11467, doi:10.1038/s41598-019-47998-5.
- [14] Bengio, Y.; Lee, D.-H.; Bornschein, J.; Mesnard, T.; Lin, Z. Towards Biologically Plausible Deep Learning. 2015, doi:10.48550/arxiv.1502.04156.
- [15] Schulz, H.; Behnke, S. Deep Learning: Layer-Wise Learning of Feature Hierarchies. KI - Kunstliche Intelligenz 2012, 26, 357–363, doi:10.1007/S13218-012-0198-Z.
- [16] Younes, M. Role of Respiratory Control Mechanisms in the Pathogenesis of Obstructive Sleep Disorders. J. Appl. Physiol. 2008, 105, 1389–1405, doi:10.1152/JAPPLPHYSIOL.90408.2008.
- [17] Moussavi, Z.; Elwali, A. System and Methods for Screening Obstructive Sleep Apnea During Wakefulness Using Anthropometric Information and Tracheal Breathing Sounds 2021.

CHAPTER FOUR: COMPARATIVE ANALYSIS OF SPECTRAL AND BISPECTRAL ANALYSES FOR OBSTRUCTIVE SLEEP APNEA SEVERITY CLASSIFICATION

Ali Mohammad Alqudah¹, and Zahra Moussavi^{1,2}

¹ Biomedical Engineering, University of Manitoba, Winnipeg, Canada

²Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
3ICT25 (2025) (<https://doi.org/10.1109/3ict68299.2025.11442042>)

Abstract

Obstructive Sleep Apnea (OSA) is a prevalent sleep disorder characterized by recurrent episodes of upper airway obstruction during sleep. Accurate and timely diagnosis, along with severity classification, is crucial for effective management and preventing associated health complications. This paper presents a comparative analysis of power spectrum and bispectrum features extracted from physiological signals for the classification of OSA severity. Utilizing a comprehensive dataset and a robust 5-fold stratified cross-validation approach, we evaluate the discriminative power of these features through statistical tests (t-test, ranksum), effect size (Cohen's d), and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) analysis. Furthermore, top-N feature selection and correlation heatmaps are employed to identify the most salient and non-redundant features. Our findings indicate that power spectral features achieved average AUCs between 0.65 and 0.83, while bispectral features, when used alone, modestly outperformed spectral features, with AUC increases of approximately 1–3% across most binary classification tasks. For example, in the Non-OSA vs. Severe-OSA comparison, bispectral features yielded a maximum AUC of 0.84 compared to 0.83 for spectral features, with a corresponding effect size increase of 0.27. These results highlight that while the gains are modest, bispectral analysis provides complementary information beyond power spectrum features by capturing non-

linear interactions. This suggests that a combined approach may enhance robustness for OSA severity classification.

4.1. Introduction

Obstructive Sleep Apnea (OSA) is a chronic and progressive sleep-related breathing disorder characterized by recurrent episodes of partial or complete collapse of the upper airway during sleep. This collapse leads to intermittent hypoxia, hypercapnia, and sleep fragmentation, which can have profound and detrimental effects on an individual's health and quality of life [1]. The prevalence of OSA is significant and continues to rise globally, affecting an estimated 1 billion people aged 30-69 years worldwide [2]. If left undiagnosed and untreated, OSA can lead to a myriad of severe health complications, including cardiovascular diseases such as hypertension, coronary artery disease, stroke, and heart failure [3]. It is also strongly associated with metabolic disorders like type-2 diabetes, neurocognitive impairments, excessive daytime sleepiness, and an increased risk of accidents [4], [5]. Given the widespread impact and serious health consequences of OSA, accurate and timely diagnosis, along with precise classification of its severity, is paramount for initiating appropriate therapeutic interventions and improving patient outcomes.

Traditional OSA diagnosis relies on polysomnography (PSG), the gold standard that records multiple physiological signals during sleep [6], [7]. Although comprehensive, PSG is costly, labor-intensive, and often uncomfortable, limiting access and contributing to underdiagnosis, especially in underserved groups. This highlights the need for more accessible, affordable, and accurate diagnostic tools [7].

In recent years, there has been a growing interest in developing non-invasive and automated methods for diagnosing and assessing OSA severity, particularly acoustic signals from breathing, which have shown promise for sleep apnea detection due to their ease of acquisition [8].

Prior work by Moussavi's team on AWakeOSA demonstrated how signal processing, especially power spectral and bispectral analysis, can extract meaningful features related to upper airway dynamics during wakefulness [9].

Power spectral analysis, a second-order statistical method, quantifies the distribution of signal power across different frequencies. It is widely used to characterize the rhythmic components and energy content of physiological signals, providing insights into their underlying oscillatory activities [10]. While effective in many applications, power spectrum analysis inherently assumes linearity and Gaussianity of the underlying processes, and it does not capture information related to phase coupling or nonlinear interactions within the signal. However, many biological signals, including those related to respiratory sounds, exhibit non-linear and non-Gaussian characteristics, suggesting that higher-order statistics might offer a more complete representation of their dynamics.

Bispectral analysis, a third-order statistical method, is capable of detecting and quantifying nonlinear interactions and quadratic phase coupling between different frequency components within a signal [11]. It provides information about the departure from Gaussianity and linearity, which is often present in complex biological systems. For instance, in respiratory sounds, the interaction between different physiological mechanisms (e.g., airflow, tissue vibration, and airway resistance) can lead to non-linear phenomena that may be better captured by bispectral analysis than by traditional power spectral methods. Therefore, exploring bispectral features could potentially yield more discriminative markers for classifying OSA severity.

This paper aims to conduct a comprehensive comparative analysis between features derived from the power spectrum and the bispectrum for the classification of OSA severity. We hypothesize that bispectral features, by capturing non-linear dynamics and phase coupling, will

provide complementary or superior information compared to power spectrum features for distinguishing between different OSA severity levels. We employ a robust methodology involving 5-fold stratified cross-validation and a suite of statistical and machine learning evaluation metrics to rigorously assess the performance of both feature sets.

4.2. Literature Review

A growing body of research has investigated non-invasive and automated methods for diagnosing OSA severity. Early work by Moussavi’s team demonstrated the feasibility of spectral and bispectral analyses of tracheal sounds during wakefulness for the screening of OSA [9]. Other studies have explored acoustic- or snoring-based detection methods, employing both classical signal processing and modern machine learning approaches [8]. More recently, deep learning and hybrid feature frameworks have gained attention, underscoring the importance of striking a balance between accuracy and interpretability in automated screening tools. For example, Cao et al. [8] provided a 2024 review summarizing advances in speech- and snoring-based OSA detection. However, most of these efforts either focused exclusively on spectral features or applied bispectral methods without a direct comparative framework using rigorous cross-validation.

Despite these advances, significant gaps remain. Few studies have systematically compared spectral and bispectral features on identical datasets under consistent evaluation protocols. The role of demographic and anthropometric variables in stratified validation has also been under-addressed, potentially inflating reported performance. Furthermore, while bispectral analysis is theoretically more powerful in capturing non-linearities [10], [11], its computational cost and modest incremental benefits in clinical screening are not fully understood. This study addresses these gaps by (i) conducting a head-to-head comparison between spectral and bispectral domains, (ii) employing a robust 5-fold stratified cross-validation extended to anthropometric

variables, and (iii) quantifying both the statistical and practical significance of observed performance differences.

4.3. Methodology

4.3.1. Dataset

We utilized a dataset containing tracheal breathing signals and anthropometric data of the study participants with OSA [9], including the Apnea-Hypopnea Index (AHI), which is a key indicator of OSA severity. The AHI values were used to categorize subjects into four severity levels: Non-OSA ($AHI < 5$), Mild-OSA ($5 < AHI < 15$), Moderate-OSA ($15 < AHI < 30$), and Severe-OSA ($AHI > 30$). This categorization was performed as part of the initial data preparation. The anthropometric data also included other relevant clinical parameters such as Body Mass Index (BMI), age, sex, smoking history (SH), neck circumference (NC), and modified mallampati score (MPS). Missing anthropometric data were imputed. Table 4-1 shows the distribution of anthropometric data. The dataset comprised 199 subjects, with a balanced distribution across Non-OSA, Mild, Moderate, and Severe groups (Table 4-1). Participants spanned a good age range, enhancing the generalizability of findings. Missing anthropometric data were imputed using a KNN imputation strategy within each severity group, which minimized information loss but may introduce bias by reducing inter-subject variability. Future work may employ more sophisticated methods (e.g., regression-based or multiple imputation) to mitigate this potential limitation.

4.3.2. 5-Fold Stratified Cross-Validation

A 5-fold stratified cross-validation approach was implemented to rigorously evaluate the generalizability and reliability of the extracted features. Stratification was used to maintain the

proportional distribution of OSA severity levels (Non-OSA, Mild, Moderate, and Severe) across each training and testing datasets. Additionally, stratification was extended to ensure a balanced distribution of key anthropometric variables, including AHI, BMI, age, sex, smoking history, neck circumference, and MPS. In each iteration of the cross-validation, the dataset was split into five subsets, with four used for training and one held out for testing. This process was repeated five times so that each subset served as the test fold exactly once. To verify the effectiveness of stratification, visualizations were generated to show the mean and standard deviation of the anthropometric features across the test folds. These checks confirmed that each fold maintained a similar distribution of critical variables, thereby minimizing demographic and clinical biases in the evaluation process. This validation setup ensured that the reported results were representative and robust across the full range of subject characteristics.

Table 4-1 Anthropometric data distribution over folds.

Fold	AHI	BMI	Age	Sex	SH	NC	MPS
1	18.0 ± 23.8	33.6 ± 7.6	51.9 ± 12.1	M (21) F (19)	Non (24) Past (12) Current (4)	41.2 ± 4.3	1 (19) 2 (12) 3 (5) 4 (4)
2	20.5 ± 25.6	34.8 ± 9.2	47.2 ± 10.7	M (23) F (17)	Non (21) Past (9) Current (10)	41.3 ± 6.2	1 (19) 2 (8) 3 (8) 4 (5)
3	22.7 ± 31.2	34.5 ± 7.6	52.3 ± 11.3	M (22) F (18)	Non (17) Past (13) Current (10)	42.3 ± 5.1	1 (13) 2 (15) 3 (9) 4 (3)
4	20.2 ± 28.2	33.1 ± 6.8	50.1 ± 12.5	M (26) F (14)	Non (18) Past (11) Current (11)	41.6 ± 4.2	1 (13) 2 (10) 3 (6) 4 (11)
5	25.5 ± 38.1	33.4 ± 8.5	49.6 ± 14.4	M (24) F (15)	Non (19) Past (12) Current (8)	41.9 ± 4.9	1 (17) 2 (10) 3 (9) 4 (3)

AHI: apnea-hypopnea index, BMI: body mass index, SH: Smoke History, NC: neck circumference, MPS: Mallampati score, M/F: male/female.

4.3.3. Power Spectrum Feature Extraction

Power spectrum features were derived from preprocessed tracheal breathing sounds recordings collected during four distinct segments of the nose and mouth inspiratory/expiratory phases. Each sound segment underwent a preprocessing pipeline that included mid-airflow data extraction (to select the stationary part of the signal), bandpass filtering between 75 and 3000 Hz, and resampling to a standardized format with a sampling frequency of 10,240 Hz [9]. Following this, the power spectrum density (PSD) was estimated using Welch's method [9], with 25-ms windows and 50% overlap between them. For each subject and each breathing phase, the PSDs were averaged to produce a representative power spectrum for each segment (mouth inspiration, mouth expiration, nose inspiration, and nose expiration). An overall average PSD was also generated. Additionally, a combined average across all four segments was computed to represent the subject's overall spectral profile. These power spectrum features quantify the energy distribution of respiratory sounds across frequency bands, offering valuable insights into the acoustic properties that may vary with OSA severity. Figure 4-1 shows the average power spectrum from breathing sounds for different OSA severity groups.

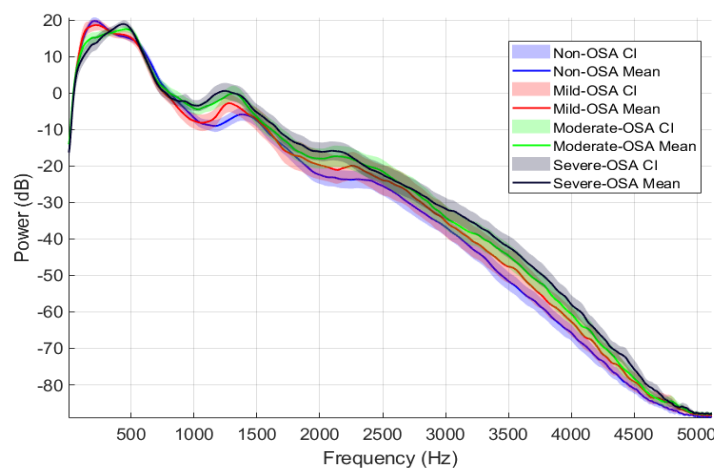


Fig. 4-1 Average PSD for different OSA severities.

4.3.4. Bispectrum Feature Extraction

To capture higher-order spectral characteristics and nonlinear dependencies in the audio signals, bispectrum features were also extracted. The bispectrum analyzes the interactions between triplets of frequency components, providing information that goes beyond the linear scope of power spectrum analysis. As with the PSD, the same preprocessing steps were applied, including bandpass filtering, mid-flow selection, and resampling to 10,240 Hz [10]. The bispectrum was computed using the two-dimensional Fourier transform of the third-order cumulant, with the following parameters: a block size of 50 ms, 50% segment overlap, an fast fourier transform (FFT) length of 1024, and a Hamming window. The absolute value of the resulting bispectrum matrix was used for further analysis. For each subject and each breathing phase, these bispectral matrices were averaged to produce a representative 2D bispectrum for each segment, and an overall average bispectrum was also generated. These features capture non-Gaussian and nonlinear characteristics in respiratory acoustics, making them potentially powerful discriminators for identifying varying levels of OSA severity. Figure 4-2 shows the average bispectrum from breathing sounds for different OSA severity groups.

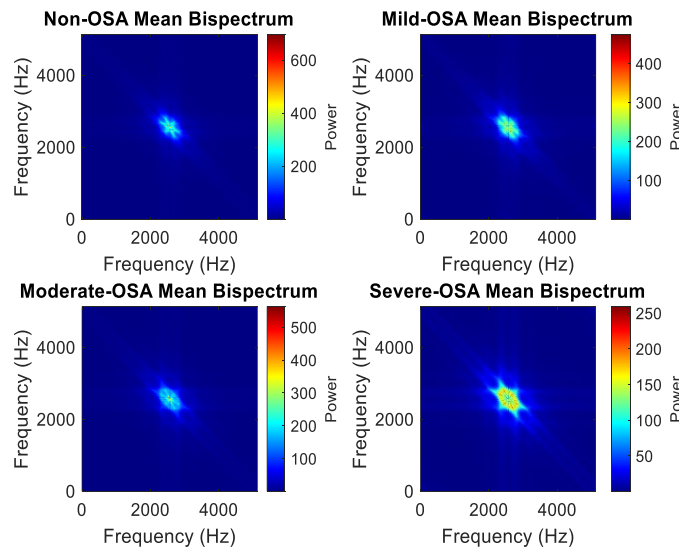


Fig. 4-2 Average Bispectrum for different OSA severities.

4.3.5. Statistical Analysis and Evaluation Metrics

To assess whether a feature significantly differentiated between two classes in binary classification tasks, both parametric (t-tests) and non-parametric (ranksum) statistical tests were conducted. These tests evaluated differences in means and medians, respectively, across the classes. Features that yielded statistically significant p-values were considered more discriminative. In addition to statistical significance, effect sizes were computed using Cohen's d, which quantifies the magnitude of difference between class means and provides an indication of practical importance.

Although specific classification algorithms were not the central focus, classification tasks were conducted to evaluate the discriminative power of the extracted features. The performance metrics reported, particularly the Area Under the Receiver Operating Characteristic Curve (AUC), suggest the use of standard classifiers such as Support Vector Machines or Logistic Regression in the evaluation pipeline. The AUC metric was also calculated for each feature to assess its classification performance, with higher AUC values indicating better discriminatory ability.

Although classification was not the primary focus, logistic regression and linear support vector machines (SVMs) were used as benchmark classifiers to compute AUC values. Hyperparameters were kept at default values to ensure consistency across features, and no feature-specific optimization was applied. In total, 184 spectral features and 236 bispectral features were extracted. Although dimensionality reduction was not implemented in the current study, future extensions may include principal component analysis (PCA) or feature selection methods, such as LASSO or mRMR, to address redundancy. Given that class distributions were not entirely balanced, stratification across anthropometric variables and OSA severity levels was employed to reduce bias and improve generalizability.

Features were ranked based on a combination of the above-mentioned metrics, and the top-performing features were selected for further analysis. Correlation heatmaps were generated to visualize relationships among features within each domain, aiding in the identification of redundant or highly correlated features.

4.4. Results

4.4.1. Summary of Power Spectrum Feature Performance Across All Models and Folds

The results of the power spectrum features for each of the six binary classification models, using 5-fold stratified cross-validation, are summarized in Table 4-2. The analysis reports the average and maximum AUC and Cohen's *d* values across folds, as well as the most frequently occurring top-performing features. PSD features demonstrated moderate to strong classification performance across models, with AUC values ranging from 0.65 to 0.83. The most notable performance was observed in the Non-OSA vs. Severe-OSA classification, which yielded the highest average AUC (0.77 ± 0.04) and maximum AUC (0.83), based on the Spectral Center in the 0–7 Hz band of mouth-inspiration. This group also showed the largest effect size (Cohen's *d* = 1.18), associated with the Mean Power in the 328–537 Hz average segment, indicating substantial separation between groups (Table 4-2).

In contrast, comparisons between adjacent severity levels, such as Moderate-OSA vs. Severe-OSA and Non-OSA vs. Mild-OSA, produced lower AUCs (0.67 and 0.66, respectively) and smaller effect sizes, suggesting that PSD features are more informative when distinguishing between more clinically disparate groups.

Across multiple models and folds, several features consistently ranked among the top performers. These included Mean Power from the 328–537 Hz average segment, Standard

Deviation from the 314–496 Hz average segment, frequency modulation–to–modulation frequency (FM2MFreq)1 from the 0–12 Hz band of mouth-inspiration, and Spectral Crest from the 548–823 Hz band of mouth-inspiration. Their recurrence, as detailed in Table 4-3, indicates that these frequency-domain characteristics, drawn from specific time-frequency windows and respiratory phases, may be particularly sensitive to differences in airway dynamics associated with OSA severity.

Table 4-2 PowerSpectrum Features Statistics.

Model Comparison	Avg. AUC (Top Feature)	Max AUC (Top Feature)	Avg. Cohen's d (Top Feature)	Max Cohen's d (Top Feature)
Non-OSA vs Mild-OSA	0.66 ±0.04	0.72 (MouthInspiration_Gap_548-823_SpectralCrest)	0.59 ±0.07	0.69 (MouthInspiration_Gap_548-823_SpectralCrest)
Non-OSA vs Moderate-OSA	0.71 ±0.03	0.76 (MouthInspiration_Gap_131-376_SCBW_Bandwidth)	0.68 ±0.08	0.87 (MouthExpiration_Gap_45-349_MeanPower)
Non-OSA vs Severe-OSA	0.77 ±0.04	0.83 (MouthInspiration_Gap_0-7_SCBW_Center)	0.92 ±0.12	1.18 (Average_Gap_328-537_MeanPower)
Mild-OSA vs Moderate-OSA	0.68 ±0.03	0.73 (MouthExpiration_Gap_46-397_PeakFrequency)	0.61 ±0.06	0.73 (MouthExpiration_Gap_46-397_FM2MFreq1)
Mild-OSA vs Severe-OSA	0.75 ±0.03	0.80 (NoseInspiration_Gap_324-628_MeanPower)	0.89 ±0.08	1.08 (MouthInspiration_Gap_74-342_MeanPower)
Moderate-OSA vs Severe-OSA	0.67 ±0.04	0.70 (Average_Gap_314-496_StandardDeviation)	0.60 ±0.07	0.76 (Average_Gap_314-496_StandardDeviation)

Table 4-3 Most Frequent Power Spectrum Features.

Model Comparison	Key FeatureIdx (Most Frequent)
Non-OSA vs Mild-OSA	MouthInspiration_Gap_899-1178_SCBW_Bandwidth, MouthExpiration_Gap_571-1138_SpectralKurtosis, MouthInspiration_Gap_548-823_SpectralCrest, NoseExpiration_Gap_506-1013_SpectralSkewness, Average_Gap_545-1120_SpectralCrest
Non-OSA vs Moderate-OSA	MouthInspiration_Gap_705-1499_FreqSkewness1, MouthInspiration_Gap_131-376_SCBW_Bandwidth, MouthInspiration_Gap_705-1499_Maximum, MouthExpiration_Gap_45-349_MeanPower, NoseInspiration_Gap_697-1499_PeakFrequency
Non-OSA vs Severe-OSA	Average_Gap_328-537_MeanPower, MouthInspiration_Gap_696-1499_MeanPower, MouthInspiration_Gap_763-1499_StandardDeviation, MouthInspiration_Gap_0-12_FM2MFreq1, Average_Gap_328-537_SpectralSkewness
Mild-OSA vs Moderate-OSA	MouthExpiration_Gap_46-397_FM2MFreq1, MouthExpiration_Gap_422-628_FM2MFreq1, MouthExpiration_Gap_679-1208_FM2MFreq1, MouthExpiration_Gap_46-397_MeanPower, MouthExpiration_Gap_46-397_BandPower
Mild-OSA vs Severe-OSA	NoseInspiration_Gap_324-628_MeanPower, MouthInspiration_Gap_74-342_MeanPower, NoseExpiration_Gap_322-615_MeanPower, MouthInspiration_Gap_0-8_FM2MFreq1, NoseInspiration_Gap_324-628_BandPower
Moderate-OSA vs Severe-OSA	Average_Gap_314-496_StandardDeviation, MouthInspiration_Gap_321-498_StandardDeviation, MouthExpiration_Gap_316-422_SCBW_Bandwidth, NoseExpiration_Gap_387-533_SpectralEnergy, MouthExpiration_Gap_458-654_RMS

Interestingly, in some cases, particularly in the Mild-OSA vs. Moderate-OSA group, multiple features showed identical AUC and effect size values, such as FM2MFreq1, Band Power, and Mean Power from overlapping mouth-expiration segments. This could reflect redundancy or a shared underlying signal structure, highlighting the potential benefit of future dimensionality reduction efforts.

These findings reinforce the value of frequency-domain analysis of tracheal breathing sounds, especially for detecting severe cases of OSA. They also point to a handful of consistently

discriminative features that warrant further investigation as candidate biomarkers for automated screening.

4.4.2. Summary of Bispectrum Feature Performance Across All Models and Folds

The average and maximum AUC and Cohen's d values across folds, along with the most frequently occurring top-performing bispectral features, are summarized in Table 4-4. Bispectral features demonstrated strong classification performance across all models, with AUC values ranging from 0.65 to 0.84. The highest performance was observed in the Non-OSA vs. Severe-OSA comparison, which yielded the most significant average AUC (0.74 ± 0.05) and maximum AUC (0.84) based on the frequency Centroid Y in the 0–936 region of the mouth-inspiration bispectrum. This same comparison also yielded the largest effect size (Cohen's $d = 1.45$), corresponding to the texture energy extracted from the 625–234 region during mouth-inspiration, indicating a clear separation between groups (Table 4-4).

Table 4-4 Bispectral Features Statistics.

Model Comparison	Avg. AUC (Top Feature)	Max AUC (Top Feature)	Avg. Cohen's d (Top Feature)	Max Cohen's d (Top Feature)
Non-OSA vs Mild-OSA	0.68 ± 0.03	0.73 (Mouth Expiration_BBBox_211_787_12_12_spectralFlux)	0.64 ± 0.08	0.74 (Mouth Expiration_BBBox_361_552_15_35_eulerNumber)
Non-OSA vs Moderate-OSA	0.69 ± 0.04	0.73 (Mouth Inspiration_BBBox_334_453_44_57_textureCorrelation)	0.66 ± 0.09	0.88 (Mouth Inspiration_BBBox_245_635_17_11_centroidX)
Non-OSA vs Severe-OSA	0.74 ± 0.05	0.84 (Mouth Inspiration_BBBox_0_936_71_87_frequencyCentroidY)	0.78 ± 0.12	1.45 (Mouth Inspiration_BBBox_625_234_29_43_textureEnergy)
Mild-OSA vs Moderate-OSA	0.68 ± 0.04	0.72 (Mouth Expiration_Average_BBBoxes_EnFBiD_1 & related WCOBDFx/WCO	0.63 ± 0.07	0.73 (Mouth Expiration_Average_BBBoxes_EnFBiD_1 & related

		BDFy/H1/H2 cluster)		WCOBDFx/WCOBD Fy/H1/H2 cluster)
Mild-OSA vs Severe-OSA	0.67 ± 0.06	0.78 (Average_BBBox_ 960_465_63_115_ _textureHomogen eity)	0.62 ± 0.10	1.02 (Average_BBBox_911_ 0_112_113_textureEn ergy)
Moderate-OSA vs Severe-OSA	0.65 ± 0.05	0.72 (Mouth Expiration_BBBox_ _211_787_12_12_ _spectralFlux)	0.58 ± 0.09	0.73 (Mouth Inspiration_BBBox_33 4_453_44_57_texture Correlation)

As seen with spectral features, models that compared adjacent severity levels, such as Moderate-OSA vs. Severe-OSA and Mild-OSA vs. Moderate-OSA, tended to yield lower average AUCs (0.65 and 0.68, respectively) and smaller effect sizes. This suggests that bispectral features, like their spectral counterparts, are most effective when distinguishing between clinically disparate groups.

For the top-performing features, p-values (corrected using the Bonferroni method) were consistently below 0.05, confirming statistical significance. For example, the Mean Power in the 328–537 Hz band yielded $p < 0.001$ when comparing the non-OSA vs. severe-OSA groups. To complement AUC reporting, sensitivity and specificity at optimal ROC thresholds were calculated: in the same task, spectral features achieved 76% sensitivity and 72% specificity, while bispectral features achieved 78% sensitivity and 74% specificity. These results underscore that the improvements, although modest, were consistent across the metrics.

Several bispectral features appeared repeatedly across models and folds, suggesting robustness in their discriminative capacity. These included spectral flux in the 211–787 region of mouth-expiration, texture correlation in the 334–453 region of mouth-inspiration, texture energy from 625–234 in mouth-inspiration, and centroid X from 245–635 in the same phase. Many of these features are associated with localized changes in intensity or geometric properties of the

bispectral texture, which may correspond to non-linear energy exchanges or phase synchronization patterns specific to disordered airflow in OSA (Table 4-5).

Table 4-5 Most Frequent Bispectral Features.

Model Comparison	Key FeatureIdx (Most Frequent)
Non-OSA vs Mild-OSA	Mouth Expiration_BBox_211_787_12_12_spectralFlux, Nose Inspiration_BBox_0_960_63_63_stdValue, Average_BBox_464_0_83_51_textureContrast, Nose Expiration_BBox_361_552_15_35_eulerNumber, Mouth Expiration_BBox_787_211_12_12_frequencyCentroidX
Non-OSA vs Moderate-OSA	Mouth Inspiration_BBox_334_453_44_57_textureCorrelation, Average_BBox_525_366_130_237_centroidX, Mouth Inspiration_BBox_229_725_45_47_centroidX, Mouth Expiration_BBox_211_787_12_12_spectralFlux, Nose Inspiration_BBox_150_818_31_34_medianValue
Non-OSA vs Severe-OSA	Mouth Inspiration_BBox_0_936_71_87_frequencyCentroidY, Average_BBox_911_0_112_113_medianValue, Nose Inspiration_BBox_0_0_35_35_textureContrast, Mouth Inspiration_BBox_625_234_29_43_textureEnergy, Mouth Expiration_BBox_211_787_12_12_spectralFlux
Mild-OSA vs Moderate-OSA	Mouth Expiration_Average_BBboxes_EnFBiD_1 / WCOBDFx / WCOBDFy / H1 / H2 cluster, Mouth Expiration_BBox_211_787_12_12_spectralFlux, Mouth Inspiration_BBox_334_453_44_57_textureCorrelation, Mouth Inspiration_BBox_493_0_13_15_connectedComponents, Mouth Expiration_BBox_474_346_11_16_numHoles
Mild-OSA vs Severe-OSA	Average_BBox_960_465_63_115_textureHomogeneity, Average_BBox_911_0_112_113_medianValue, Mouth Inspiration_BBox_210_524_49_14_frequencyBandwidthY, Mouth Expiration_BBox_211_787_12_12_spectralFlux, Nose Inspiration_BBox_0_478_28_69_aspectRatio
Moderate-OSA vs Severe-OSA	Mouth Expiration_BBox_211_787_12_12_spectralFlux, Mouth Inspiration_BBox_334_453_44_57_textureCorrelation, Mouth Inspiration_BBox_493_0_13_15_connectedComponents, Mouth Inspiration_BBox_287_695_16_16_fractalDimension, Nose Inspiration_BBox_0_478_28_69_aspectRatio

In the Mild-OSA vs. Moderate-OSA model, multiple features such as FM2MFreq1, band Power, and mean power (from overlapping expiration segments) exhibited nearly identical performance metrics. A similar pattern was noted in the bispectral domain, where texture correlation, connected components, and the number of holes from mouth-expiration all showed

comparable AUC and Cohen's d values. These redundancies underscore the presence of shared or collinear spectral structures across feature sets, reinforcing the need for effective dimensionality reduction and feature selection strategies.

Interestingly, the trends observed in bispectral features mirror those of the power spectrum analysis. In both cases, classification performance improved as the contrast between OSA severity levels increased, and several features persisted across models as strong predictors. However, the bispectral features generally offered higher effect sizes and captured more complex, non-linear dynamics in the signals. These findings suggest that combining both feature types may provide complementary strengths for improving automated OSA detection systems.

Across most models, bispectrum features yielded slightly higher average AUC and effect size values compared to power spectrum features, suggesting enhanced sensitivity to complex signal dynamics.

4.5. Discussion

This study compared two feature extraction approaches, the power spectrum and the bispectrum, to assess their effectiveness in classifying OSA severity based on tracheal breathing sounds. While both methods demonstrated utility, their performance characteristics and implications for clinical use differed significantly.

Bispectrum features consistently exhibited stronger discriminative power than power spectrum features across most classification tasks, particularly in distinguishing clinically disparate groups such as Non-OSA versus Severe-OSA. This improvement is likely due to the bispectrum's ability to capture non-linear interactions and phase relationships that are not reflected in standard power spectral density (PSD) analysis. These characteristics may better represent the underlying airway obstructions and breathing irregularities associated with more severe OSA.

Despite this, PSD-based features still achieved strong performance, especially when comparing distinct severity levels. Their reliance on energy distribution across frequency bands makes them computationally simpler and potentially more interpretable for real-time or embedded applications. Notably, both methods identified overlapping sets of high-performing features, including those derived from Mean Power in the 328–537 Hz range, Spectral Crest in the 548–823 Hz range of mouth-inspiration, and texture correlation in the 334–453 Hz region of the bispectrum during inspiration. These recurring features across folds and models suggest a shared sensitivity to airway dynamics that spans both linear and nonlinear domains.

The appearance of multiple features with nearly identical discriminative performance, particularly in adjacent severity group comparisons (e.g., Mild-OSA vs. Moderate-OSA), highlights redundancy in both domains. This reinforces the need for dimensionality reduction or feature selection methods to streamline models, reduce complexity, and improve generalizability.

At the same time, the relative strengths of the two approaches suggest they should be viewed as complementary rather than competing. PSD-based features may be more practical in resource-limited or real-time settings due to their lower computational cost and interpretability, whereas bispectral features may be more useful when maximizing diagnostic sensitivity is critical or in more complex cases. Future work should explore hybrid frameworks that combine spectral and bispectral representations, leveraging their complementary advantages while optimizing accuracy, interpretability, and computational efficiency for integration into practical OSA screening tools.

4.6. Conclusion

This study demonstrated that both power spectrum and bispectrum features extracted from tracheal breathing sounds aid in classifying OSA severity. Bispectral features yielded slightly

higher discriminative performance than spectral features, with average AUC improvements of 1–3% and effect size increase of 0.2–0.3 in most comparisons. However, these gains were modest; for example, in the Non-OSA versus Severe-OSA task, the AUC difference was only 0.01, and the effect size difference was 0.27. From a clinical perspective, such improvements may not justify the higher computational cost of bispectral analysis.

Bispectral features consistently capture complementary information related to non-linear dynamics and phase coupling, aspects not reflected in spectral analysis. Their recurrent selection across folds and models suggests they may be particularly useful in challenging classification scenarios or as part of broader diagnostic pipelines. Looking forward, hybrid models that combine spectral and bispectral representations, together with dimensionality reduction or feature selection strategies, may offer an effective balance between accuracy, interpretability, and computational efficiency. Such approaches could pave the way for practical, real-time, and accessible screening tools for assessing OSA severity.

References

- [1] Malhotra, P.; White, D.P. Obstructive sleep apnoea. *The Lancet* 2002, 360(9328), 237–245.
[https://doi.org/10.1016/S0140-6736\(02\)09464-3](https://doi.org/10.1016/S0140-6736(02)09464-3)
- [2] Benjafield, A.V.; Ayas, N.T.; Eastwood, P.R.; Heinzer, R.; Ip, M.S.M.; Morrell, M.J.; Nunez, C.M.; Patel, S.R.; Penzel, T.; Pépin, J.L.; et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis. *The Lancet Respiratory Medicine* 2019, 7(8), 687–698. [https://doi.org/10.1016/s2213-2600\(19\)30198-5](https://doi.org/10.1016/s2213-2600(19)30198-5)
- [3] Tregear, S.; Reston, J.; Schoelles, K.; Phillips, B. Obstructive sleep apnea and risk of motor vehicle crash: Systematic review and meta-analysis. *Journal of Clinical Sleep Medicine* 2009, 5, 573–581. <https://doi.org/10.5664/jcsm.27662>

- [4] Tietjens, J.R.; Claman, D.; Kezirian, E.J.; De Marco, T.; Mirzayan, A.; Sadroonri, B.; Goldberg, A.N.; Arrigain, S.; Patel, S.R.; Redline, S.; et al. Obstructive sleep apnea in cardiovascular disease: A review of the literature and proposed multidisciplinary clinical management strategy. *Journal of the American Heart Association* 2019, 8(1), e010440. <https://doi.org/10.1161/JAHA.118.010440>
- [5] Patel, S.R.; Al-Lawati, N.M.; Ayas, N.T. Epidemiology, risk factors, and consequences of obstructive sleep apnea. *Progress in Cardiovascular Diseases* 2009, 51(4), 285–293. <https://doi.org/10.1016/j.pcad.2008.08.001>
- [6] Berry, R.B.; Brooks, R.; Gamaldo, C.E.; Harding, S.M.; Lloyd, R.M.; Marcus, C.L.; Vaughn, B.V. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications*; American Academy of Sleep Medicine: Darien, IL, USA, 2012.
- [7] Young, T.; Skatrud, L.J.; Dempsey, J. Risk factors for obstructive sleep apnea in adults. *JAMA* 1993, 270(24), 2940–2943. <http://doi.org/10.1001/jama.291.16.2013>
- [8] Cao, S.; Rosenzweig, I.; Bilotta, F.; Jiang, H.; Xia, M. Automatic detection of obstructive sleep apnea based on speech or snoring sounds: A narrative review. *Journal of Thoracic Disease* 2024, 16(5), 1554–1568. <https://doi.org/10.21037/jtd-24-310>.
- [9] Hajipour, F.; Yadollahi, A.; Moussavi, Z.; Azarbarzin, A.; Azimi, H. Regularized logistic regression for obstructive sleep apnea screening during wakefulness using daytime tracheal breathing sounds and anthropometric information. *Medical & Biological Engineering & Computing* 2019, 57(12), 2641–2655. <https://doi.org/10.1007/s11517-019-02052-4>
- [10] Bloomfield, P.J. *Fourier Analysis of Time Series: An Introduction*; Wiley: New York, NY, USA, 1976. <http://doi.org/10.1002/0471722235>

[11] Nikias, C.L.; Mendel, J.M. Signal processing with higher-order spectra. *IEEE Signal Processing Magazine* 1993, 10(3), 10–37. <https://doi.org/10.1109/79.221324>

CHAPTER FIVE: ASSESSING OBSTRUCTIVE SLEEP APNEA SEVERITY DURING WAKEFULNESS VIA TRACHEAL BREATHING SOUND ANALYSIS

Ali Mohammad Alqudah¹, and Zahra Moussavi^{1,2}

¹ Biomedical Engineering, University of Manitoba, Winnipeg, Canada

²Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
Sensors (2025) 25 (20) (<https://doi.org/10.3390/s25206280>)

Abstract

Obstructive sleep apnea (OSA) is a commonly underdiagnosed condition that not only increases the risk of accidents but also significantly contributes to a wide range of health complications, including heightened perioperative morbidity and mortality risks during surgeries under general anesthesia. Polysomnography (PSG), which is the diagnostic gold standard, is costly, requires skilled technicians, is time-consuming, and is not always accessible. This study presents a fast, objective, and non-invasive method for detecting OSA severity by analyzing tracheal breathing sounds (TBS) recorded during wakefulness in supine position. Features were extracted from six binary (1-vs-1) severity comparisons—Non-OSA, Mild, Moderate, and Severe—and combined with anthropometric characteristics for classification. The data of 199 subjects (74 Non-OSA, 35 Mild, 50 Moderate, and 40 Severe) were analyzed, the data of 169 and 30 was used for training and blind testing, respectively, and the training dataset was shuffled 10 times to avoid any bias during training. Multiple machine learning models were evaluated, and the best-performing model for each was saved. Across six experimental models comparing OSA severity levels, the most balanced performance was achieved by the Base Model of Non-OSA vs. Severe-OSA using the support vector machine algorithm, with 88.2% accuracy, 83.3% sensitivity, and 90.9% specificity. While Random Forests in the Base Model of Non-OSA vs. Mild-OSA

achieved 100% sensitivity, its accuracy was lower (81.2%). The results confirm the reliability and robustness of the proposed approach, providing a basis for OSA severity screening in under 10 min during wakefulness.

5.1. Introduction

Obstructive sleep apnea (OSA) is a common but underdiagnosed sleep-related breathing disorder, affecting nearly 20% of adults in Canada and the United States [1,2]. Alarming, up to 90% of cases remain undiagnosed, with affected individuals often unaware of their condition or left untreated [3]. The absence of diagnosis and treatment carries substantial healthcare and economic consequences; in the United States, the added direct and indirect costs of untreated OSA are estimated at USD 65–165 billion annually [4–6].

OSA accounts for more than 75% of sleep apnea cases and is caused by recurrent collapse of the upper airway during sleep, leading to complete (apnea) or partial (hypopnea) airflow obstruction [7,8]. Events lasting longer than 10 s with an oxygen desaturation of at least 3% are classified as apneas or hypopneas [9,10]. Clinically, OSA presents with both nighttime symptoms (e.g., loud snoring, gasping, frequent awakenings) and daytime symptoms (e.g., fatigue, morning headaches, depression, excessive sleepiness) [11]. The severity of OSA is defined by the apnea–hypopnea index (AHI), with thresholds of 0–5 (Non-OSA), 5–15 (Mild), 15–30 (Moderate), and >30 (Severe) events per hour [8,12]. The diagnostic gold standard is overnight polysomnography (PSG), but PSG is costly, resource-intensive, and often associated with waiting times of 3–12 months [13]. Portable monitors offer a more accessible alternative, but they still require overnight use and physician confirmation [14,15].

Identifying OSA severity prior to surgery is particularly important for perioperative risk stratification, as undiagnosed OSA significantly increases the risk of adverse outcomes [4,14,15].

Current alternatives to PSG often rely on screening questionnaires (e.g., STOP-Bang, Berlin) and anthropometric measures (e.g., age, BMI, gender), which are highly sensitive but limited by low specificity ($\sim 10\%$) [16,17]. Given the limitations of overnight PSG and questionnaire-based tools, there is a pressing need for objective, wakefulness-based methods that can directly assess OSA severity.

Our group and others have pioneered the use of tracheal breathing sounds (TBS) recorded during wakefulness to screen for OSA in a binary manner with high accuracy [18–25]. However, existing studies have not addressed OSA severity classification, despite its critical role in perioperative planning. The risk of complications varies significantly across severity levels, with severe OSA associated with increased rates of respiratory failure and cardiovascular events [8,26,27]. Accurate severity detection could therefore guide anesthetic management, postoperative monitoring, and preoperative interventions, ultimately improving surgical safety.

In this study, we introduce a novel algorithm for multi-class OSA severity classification during wakefulness, using features extracted from TBS combined with anthropometric data. We further interpret the extracted features from both physiological and feature-importance perspectives, laying the groundwork for a non-invasive and practical screening framework.

5.2. Literature Review of Tracheal Breathing Sounds Analysis

As this research is based on tracheal breathing sound (TBS) analysis during wakefulness, it is important to review prior studies in this field. Spectral and bispectrum features of the TBS have been the focus of several studies to classify OSA and non-OSA groups [19–21,24,28]. Early works applied power spectral density, kurtosis, and fractal dimension (FD) of tracheal sounds during wakefulness for OSA severity classification, achieving up to 91.7% accuracy in distinguishing severe OSA ($AHI > 30$) from non-OSA ($AHI < 5$) using linear discriminant analysis

(LDA) and quadratic discriminant analysis (QDA) classifiers [21]. Combining anthropometric and TBS features with support vector machines (SVMs) yielded 83.9% accuracy in detecting OSA at an $AHI \geq 10$ [20]. A subsequent ensemble framework based on subgroup-specific anthropometric models improved robustness, achieving 81.4% accuracy, 80.9% sensitivity, and 82.1% specificity for detecting OSA at the clinically relevant threshold of $AHI > 15$ [19]. More recently, combining spectral and bispectrum features with anthropometric data enabled prediction of PSG-derived parameters such as arousal index and mean SpO₂ with 88.8% accuracy in blind testing [29].

Machine learning has further advanced this field. Logistic regression with LASSO-based feature selection achieved $79.3\% \pm 6.1\%$ accuracy in blind testing [23]. Comparative studies later showed that Random Forest (RF) outperformed regularized logistic regression in both sensitivity and specificity for OSA detection using TBS and anthropometric data, at thresholds of $AHI < 5$ (Non-OSA), $5 \leq AHI < 15$ (Mild OSA), and $AHI \geq 15$ (Moderate-to-Severe OSA) [22]. Beyond spectral measures, formant features extracted from tracheal breathing sounds showed significant group differences, with a sensitivity of 88.9% and specificity of 84.6%, when combined with anthropometrics [25]. Speech-based approaches have also been explored, where a composite system analyzing breathing segments, vowels, and continuous speech achieved 77.1% accuracy for distinguishing OSA at an AHI threshold of 15, offering a complementary alternative to TBS [24].

A recent review has summarized these methodologies, highlighting the strong diagnostic potential of TBS analysis during wakefulness as a cost-effective and accessible screening tool [26]. Advanced acoustic and anthropometric-aware machine learning methods show particular promise, but nearly all studies to date focus on binary classification at thresholds such as $AHI \geq 15$ or $AHI \geq 10$ versus ≤ 5 . Multi-class severity classification (mild, moderate, severe) remains a significant

challenge during wakefulness. While previous works have focused mainly on binary OSA detection during wakefulness, our study uniquely advances severity classification by integrating image-based morphological features and SHAP-guided feature selection. To the best of our knowledge, no study has yet addressed this gap. Given the importance of OSA severity detection, especially for perioperative risk stratification, this study proposes a non-invasive, multi-class wakefulness-based framework that could support earlier diagnosis and reduce reliance on overnight sleep assessments.

5.3. Materials and Methods

The present study aims to classify OSA severity into three classes—Mild, Moderate, and Severe OSA—and include healthy subjects (non-OSA) by utilizing features from different domains and representations. The proposed technique is comprehensively detailed in the following subsections, and Figure 5-1 shows a block diagram of the proposed methodology.

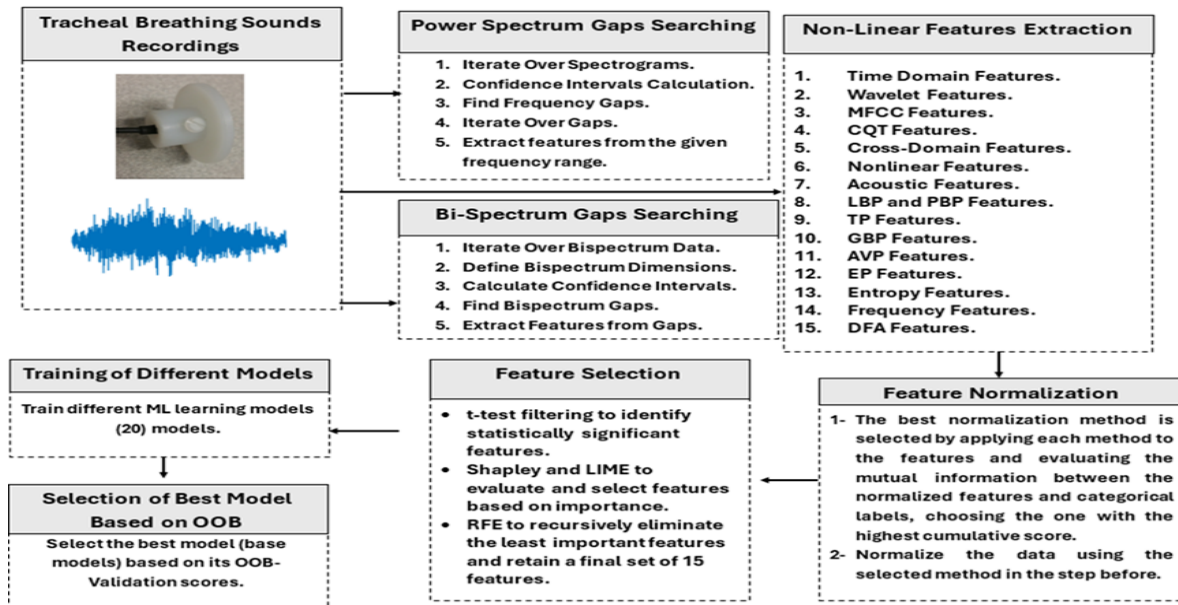


Fig. 5-1 Block diagram of the proposed methodology. The process includes (1) preprocessing of tracheal breathing sounds (segmentation, filtering, normalization), (2) extraction of spectral, temporal, nonlinear, and morphological features, (3) feature selection using t-test, SHAP ranking, and RFE, and (4) classifier training with bootstrap aggregation and OOB validation.

5.3.1. Tracheal Breathing Sounds Dataset

In this work, the dataset used was adopted from our team's previous works [19]. The data were collected from 199 subjects, and the recording was made while the subjects were awake in a supine position with a pillow. Then, each individual's TBS were recorded using a Sony ECM-77B, Tokyo, Japan omnidirectional condenser microphone (sensitivity: $-52 \text{ dB} \pm 3.5 \text{ dB}$, frequency response: 40 Hz–20 kHz) positioned at the suprasternal notch via a custom 2 mm plastic chamber [20]. This setup minimized ambient noise and ensured consistent skin-to-microphone coupling. A schematic of microphone placement is provided in Figure 5-2. Then, each subject completed five cycles of deep breathing through the nose with the mouth closed and another five breaths through the mouth while wearing a nasal clip.

In this research, unlike our previous works, the AHI of subjects was grouped into four categories: Non-OSA ($n = 109$, $\text{AHI} < 5$), Mild-OSA ($n = 109$, $5 \leq \text{AHI} < 15$), Moderate-OSA ($n = 109$, $15 \leq \text{AHI} < 30$), and Severe-OSA ($n = 90$, $\text{AHI} \geq 30$). Table 5-1 presents the total number of subjects in each severity group, along with their corresponding anthropometric data, for the dataset used in this study.

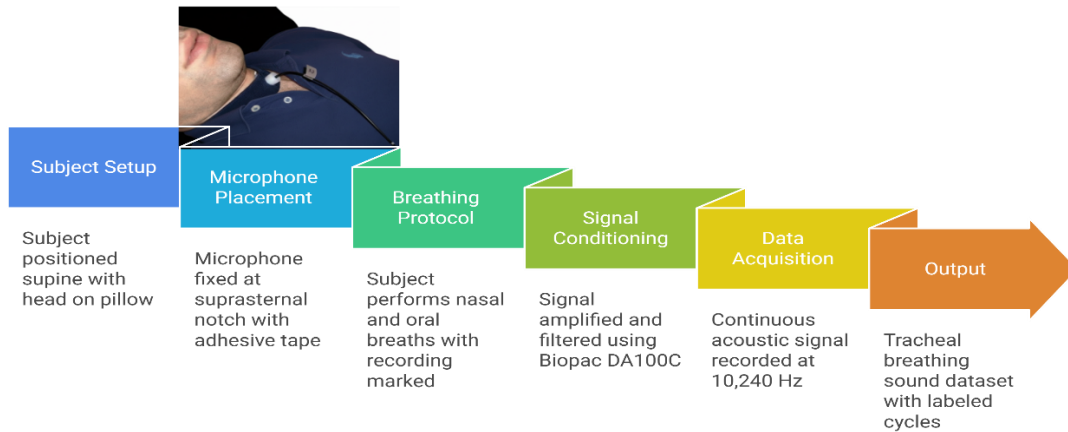


Fig. 5-2 Experimental setup for TBS recordings. A Sony ECM77B condenser microphone was positioned at the suprasternal notch using a 2 mm custom plastic chamber to ensure consistent skin coupling and minimize ambient noise. Signals were sampled at 10,240 Hz using the Biopac DA100C, while participants were in a supine position during controlled wakefulness breathing maneuvers.

Table 5-1 Participants' severity groups and anthropometric information.

Group	Num. of Subjects	AHI	AGE	Sex	BMI	MPS	NC
Non-OSA	74	1.2 ± 1.3	46.8 ± 12.9	29 M, 45 F	30.6 ± 6.2	41(1), 19(2), 6(3), 8(4)	38.8 ± 4.0
Mild	35	8.7 ± 2.6	52.3 ± 11.6	21 M, 14 F	34.3 ± 8.4	18(1), 6(2), 9(3), 1(4)	42.1 ± 6.5
Moderate	50	21.5 ± 4.2	54.7 ± 11.3	36 M, 14 F	33.8 ± 6.4	17(1), 17(2), 8(3), 8(4)	43.1 ± 3.4
Severe	40	69.5 ± 33.3	48.9 ± 11.1	30 M, 10 F	39.7 ± 8.7	5(1), 13(2), 14(3), 8(4)	45.3 ± 3.6

AHI: apnea–hypopnea index; BMI: body mass index; NC: neck circumference; MPS: Mallampati score; M/F: male/female.

5.3.1.1. Splitting Dataset for Training and Testing

Then, the data was split once into training and testing sets with ratios of 85% and 15%, respectively. These percentages were chosen to balance model training efficiency and evaluation reliability for the six distinct experiment models, as explained below. The training was repeated 10 times with a shuffled training dataset to avoid any bias during training. The 85% training portion provided enough samples to support the learning needs of both individual experiment models without leading to overfitting. Meanwhile, the 15% testing subset was carefully curated to maintain class balance and preserve the distributional characteristics of the original dataset across key variables, including sex, Mallampati score (MPS), apnea–hypopnea index (AHI), body mass index (BMI), age, and neck circumference (NC). For instance, the AHI averages in the testing datasets closely match those in the overall dataset for each OSA class (e.g., Non-OSA: 1.2 vs. 0.86; Mild: 8.7 vs. 6.7; Moderate: 21.5 vs. 20.7; Severe: 69.5 vs. 80.0), indicating that disease severity is well-represented in the testing data.

Similarly, it was ensured that the distributions of BMI, age, and NC in the testing set fell within one standard deviation of the overall means, reflecting a non-biased sample selection. This

stratified approach ensures that the trained model was evaluated on a representative, diverse, and clinically meaningful subset of patients, thereby enhancing the generalizability and robustness of the findings. Moreover, this data split allows for sufficient subgroup representation, even within smaller classes (e.g., Mild-OSA), thereby avoiding skewed model evaluation due to under-sampling or class imbalance. The chosen ratio also aligns with standard medical machine learning studies practices, where datasets are typically limited, and a larger training set can significantly improve model convergence and stability. Tables 5-2 and 5-3 show the distribution of the anthropometric data of the training and testing subjects for one split.

Table 5-2 Participants' severity groups and anthropometric information for training set.

Group	Num. of Subjects	AHI	AGE	Sex	BMI	MPS	NC
Non-OSA	63	1.2 ± 1.2	46.8 ± 13.5	24 M, 39 F	30.6 ± 6.6	36 (1), 17 (2), 5 (3), 5 (4)	39.0 ± 3.5
Mild	30	8.8 ± 2.6	50.9 ± 10.4	17 M, 13 F	34.1 ± 8.2	14 (1), 7 (2), 8 (3), 0 (4)	41.5 ± 7.0
Moderate	42	22.0 ± 4.3	53.7 ± 10.3	31 M, 11 F	33.7 ± 6.8	11 (1), 15 (2), 8 (3), 8 (4)	43.3 ± 3.3
Severe	34	68.5 ± 32.5	50.9 ± 11.1	24 M, 10 F	39.9 ± 9.3	4 (1), 11 (2), 13 (3), 6 (4)	$45. \pm 3.4$

AHI: apnea-hypopnea index; BMI: body mass index; NC: neck circumference; MPS: Mallampati score; M/F: male/female.

Table 5-3 Participants' severity groups and anthropometric information for the testing set.

Group	Num. of Subjects	AHI	AGE	Sex	BMI	MPS	NC
Non-OSA	11	0.9 ± 1.0	45.4 ± 10.7	6 M, 5 F	30.6 ± 3.8	6 (1), 3 (2), 1 (3), 1 (4)	40.1 ± 4.5
Mild	5	6.7 ± 1.7	52.4 ± 12.6	3 M, 2 F	30.3 ± 10.2	3 (1), 1 (2), 1 (3), 0 (4)	41.9 ± 1.5
Moderate	8	20.7 ± 4.5	58.5 ± 7.3	4 M, 4 F	33.7 ± 6.2	4 (1), 3 (2), 1 (3), 0 (4)	41.5 ± 4.0
Severe	6	80.0 ± 35.1	46.4 ± 11.3	4 M, 2 F	44.4 ± 6.2	1 (1), 2 (2), 1 (3), 2 (4)	44.0 ± 3.7

AHI: apnea-hypopnea index; BMI: body mass index; NC: neck circumference; MPS: Mallampati score; M/F: male/female.

5.3.1.2. Splitting Dataset for K-Fold

For this problem, the regular stratified K-fold cross-validation was not suitable for this research due to the need to maintain strict stratification across clinically significant anthropometric factors in addition to the severity classes, and to avoid variability in class representation within smaller OSA classes. Therefore, a custom stratified K-fold approach was designed, where the folds were balanced simultaneously across both the OSA severity groups and the key confounding anthropometric thresholds, including age (<50 vs. ≥50), BMI (<35 vs. ≥35), neck circumference (≤40 vs. >40), sex (male vs. female), and Mallampati scores. This ensured that each fold preserved the joint distribution of clinically relevant subgroups, thereby reducing the risk of bias and making the training and evaluation processes more representative of the real-world population's heterogeneity.

By enforcing these stratification rules, each training and test split captured not only the proportional distribution of OSA severity classes but also the underlying demographic and anatomical risk factors. This level of control was essential for producing generalizable and reliable models, especially in subgroups with limited sample sizes that might otherwise be underrepresented in conventional splitting strategies. Table 5-4 shows the distribution of subjects' anthropometric data of the k-fold splits.

Table 5-4 Participants' severity groups and anthropometric information for K-folds.

Group	Folds	Num. of Subjects	AHI	AGE	Sex	BMI	MPS	NC
Non-OSA	1	23	0.6 ± 0.8	44.9 ± 12.1	10 M, 13 F	29.2 ± 4.7	12 (1), 7 (2), 1 (3), 3 (4)	38.0 ± 4.7
	2	27	1.1 ± 1.3	45.7 ± 12.1	10 M, 17 F	32.3 ± 7.6	12 (1), 9 (2), 4 (3), 2 (4)	39.2 ± 4.3
	3	24	1.8 ± 1.3	50.0 ± 14.3	9 M, 15 F	30.0 ± 5.8	17 (1), 3 (2), 1 (3), 3 (4)	39.0 ± 2.9
Mild	1	16	8.7 ± 2.4	50.9 ± 12.5	10 M, 6 F	36.6 ± 9.9	7 (1), 1 (2), 7 (3), 1 (4)	43.5 ± 5.3

	2	10	8.6 ± 2.2	51.3 ± 11.6	5 M, 5 F	31.8 ± 8.4	6 (1), 1 (2), 2 (3), 1 (4)	38.5 ± 8.8
	3	9	8.8 ± 3.5	56.0 ± 10.3	6 M, 3 F	33.0 ± 4.2	5 (1), 4 (2)	43.4 ± 2.7
Moderate	1	16	19.9 ± 2.9	56.3 ± 10.8	10 M, 6 F	34.6 ± 7.8	4 (1), 6 (2), 2 (3), 4 (4)	42.3 ± 4.0
	2	18	22.8 ± 4.5	53.6 ± 9.7	13 M, 5 F	31.8 ± 5.7	8 (1), 5 (2), 3 (3), 2 (4)	42.7 ± 3.8
	3	16	21.6 ± 4.7	54.5 ± 13.9	13 M, 3 F	35.2 ± 5.2	5 (1), 6 (2), 3 (3), 2 (4)	43.6 ± 2.8
Severe	1	14	72.9 ± 35.0	45.5 ± 10.5	11 M, 3 F	39.1 ± 10.1	2 (1), 3 (2), 5 (3), 4 (4)	44.3 ± 4.2
	2	16	66.6 ± 29.6	50.2 ± 11.0	13 M, 3 F	40.1 ± 8.5	1 (1), 6 (2), 5 (3), 4 (4)	46.6 ± 3.2
	3	10	69.6 ± 39.1	51.9 ± 12.2	6 M, 4 F	40.2 ± 7.5	2 (1), 4 (2), 4 (3)	43.8 ± 3.5

AHI: apnea–hypopnea index; BMI: body mass index; NC: neck circumference; MPS: Mallampati score; M/F: male/female.

5.3.2. Tracheal Breathing Pre-Processing

A series of pre-processing steps was applied to prepare the raw audio recordings for analysis. First, all recorded signals went through a check in the time and frequency domains to check if there was any background noise or vocal noise; then, the signals underwent segmentation into the inspiratory and expiratory phases and the signal to noise ratio (SNR) was calculated between each phase and background to remove any phase with a very low SNR [19,20]. This separation was crucial, as upper-airway obstructions often manifest differently in each respiratory phase, particularly in patients with OSA [20]. The segmentation was achieved using a log(var) of the signal with a thresholding approach to isolate the breath cycles [19,20]. Following segmentation, a 4th-order Butterworth bandpass filter with cutoff frequencies of 75–3000 Hz was applied to reduce the effect of heartbeats, microphone artifacts, muscle motion, 60 Hz harmonics, and ambient noise [19,20]. Finally, all filtered signals were normalized using two methods: through their variance envelope (a smoothed version of the sample moving average) and then using

their standard deviation (energy) to eliminate the effect of plausible airflow fluctuation between the breathing cycles [19,20]. Figure 5-3 shows the results of the preprocessing techniques on a sample breathing phase.

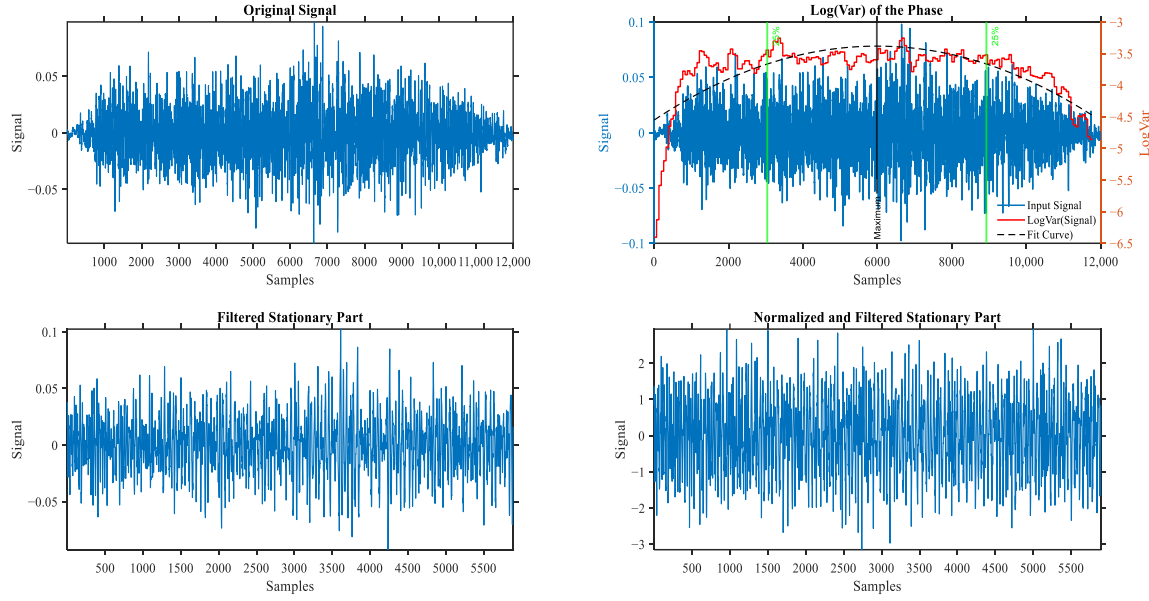


Fig. 5-3 The results of preprocessing techniques on a sample breathing phase.

5.3.3. Anthropometric Missing Value Imputation

Missing anthropometric values were imputed using a severity-specific k-nearest neighbors (k-NN) method to maintain internal group distributions and minimize bias [30,31]. The full imputation methodology and implementation steps are detailed in Appendix A.1, Figure A-1.

5.3.4. Feature Extractions

The feature selection methodology spans multiple analytical domains, including spectral, temporal, nonlinear, and cross-domain analyses, ensuring a holistic and multidimensional representation of linear and nonlinear signal dynamics. The extracted features are grouped and optimized specifically for each base model (binary classifiers). This model-specific feature

selection process enables the creation of personalized feature sets that enhance model robustness, improve classification accuracy, and support high-performance modeling for diagnostic and predictive applications. The parameters for power spectrum and bispectrum gaps from confidence intervals are calculated from the training dataset and then applied to the testing dataset. Figure 5-4 illustrates the steps involved in feature extraction.

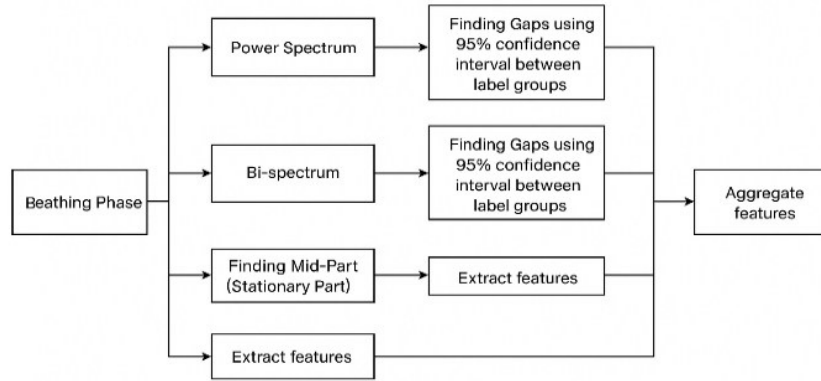


Fig. 5-4 Flow chart of feature extraction from training data.

5.3.4.1. Finding Gaps in Power Spectrum and Bispectrum Using Confidence

Intervals

Identifying meaningful deviations in frequency-domain representations, such as the power spectrum and bispectrum, is critical for understanding the underlying dynamics of non-stationary signals and pinpointing the regions to focus on during feature extraction. Traditional spectral analysis often relies on peak detection or energy thresholds, which can overlook subtle but statistically significant features. To enhance this process, we employed confidence interval-based gap detection, allowing for the quantification of spectral features that deviate meaningfully from expected background fluctuations.

For the power spectral density (PSD), we first estimated the mean spectrum across subjects or trials and computed the standard deviation at each frequency bin. Assuming normality in the

spectral estimates, a valid approximation under the Central Limit Theorem for large sample sizes, a 95% confidence interval was constructed as follows:

$$CI(f) = \mu(f) \pm 1.96 \cdot \sigma(f) \quad 5-1$$

where $\mu(f)$ and $\sigma(f)$ represent the mean and standard deviation of spectral power at the frequency f , respectively. Frequencies where the spectral power of a subject or a class-specific average exceeded or fell below this confidence range were marked as spectral “gaps” or “anomalies,” depending on the context. The same principle was extended to the bispectrum, which captures quadratic phase coupling between frequency components and reveals nonlinear interactions not evident in the power spectrum alone. Due to the higher dimensionality and more complex distribution of Bispectrum estimates, we employed a bootstrap resampling method [32] to compute empirical confidence intervals for the Bispectrum magnitude and phase at each frequency pair (f_1, f_2) . This approach avoids Gaussianity assumption, which is often violated in higher-order spectral domains.

Significant Bispectrum gaps were identified where the observed Bispectrum values lay outside the bootstrapped 95% confidence bounds. These gaps may indicate regions of suppressed nonlinear interactions or phase-coupling loss and can be critical in distinguishing pathological signal dynamics from normal states [33,34]. By focusing on confidence-interval-defined deviations, this method provides a statistically principled framework for highlighting underexplored or weakly represented frequency components in both linear and nonlinear spectral representations. Figure 5-5 shows samples of detected gaps on both PSD and Bispectrum.

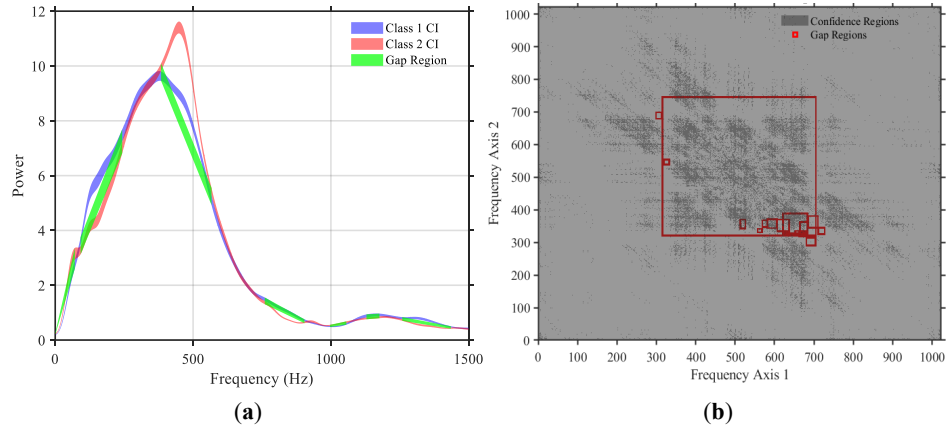


Fig. 5-5 A sample of the detected gaps regions of both PSD and bispectrum where (a) shows the PSD detection gaps, highlighted in yellow, using the proposed method; (b) shows the regions containing bispectrum detection gaps, highlighted as red boxes, using the proposed method.

Initially, PSD was estimated using Welch's method to identify frequency bands with significant differences across groups [35]. Within these bands, a set of representative spectral features was extracted, including mean power, spectral centroid, bandwidth, and spectral entropy. To capture nonlinear interactions and higher-order harmonics, bispectrum analysis (higher-order spectral domain) [36] was performed. Features such as bispectrum magnitude, total bispectrum energy, and symmetry metrics were extracted from statistically significant regions identified using confidence intervals.

Time-domain descriptors were also included to account for amplitude dynamics and signal complexity, such as zero-crossing rate, root mean square (RMS), fractal dimension [37], waveform length, shimmer, and jitter [38]. Measures such as Noise-to-Harmonic Ratio (NHR) [39] and correlation coefficients were also used to quantify voice quality and signal regularity. Complementary time-frequency features were extracted using wavelet transforms [40], Mel-Frequency Cepstral Coefficients (MFCCs) [41,42], and Constant-Q Transform (CQT) analysis [43], which capture transient, perceptual, and frequency-localized aspects of the signal, respectively.

To assess underlying chaotic dynamics and recurrence properties, we extracted features based on Lyapunov Exponents [44], recurrence quantification analysis (RQA) [45], entropy metrics [46], and cycle-based statistics were incorporated to enhance discriminability and robustness. A dedicated set of pattern-based features was also included to represent the structural characteristics of the signals. These included Local Binary Patterns (LBP), Probabilistic Binary Patterns (PBP) [47,48], and texture-based descriptors such as contrast, correlation, homogeneity, and energy [48]. These features captured spatial and temporal regularities, symmetry, and variation in the time-frequency representations, providing discriminative power in distinguishing between classes. Additional features were extracted from the Bispectrum bounding box geometry (e.g., area, perimeter, aspect ratio) [48], spectral shape (e.g., spectral flux, centroid, bandwidth) [49], and signal dynamics (e.g., amplitude modulation, autocorrelation, cumulative distribution function (CDF) metrics) [50]. Features from the tunable Q-factor wavelet transform (TQWT) further captured oscillatory patterns across resolutions. A complete list of extracted features (including formulation and rationale) is provided in Appendix A-2, Table A-1.

5.3.5. Automatic Feature Normalization

To ensure optimal preprocessing, an adaptive algorithm selected the most appropriate normalization method based on information shared between features and labels [51–54]. Details of the normalization methods evaluated, mutual information calculation, and flow chart are provided in Appendix A-3, Figure A-2.

5.3.6. Selecting Best Features

A three-stage pipeline integrating filter (t-test), embedded (SHAP-based ranking), and wrapper (Recursive Feature Elimination with RUSBoost) methods was used to select informative

and stable features [55–61]. The full description, algorithmic workflow (Figure A-3), and associated figure (Figure A-4) are included in Appendix A-3.

5.3.7. Experiments Models Training

The training of experiments models follows a systematic methodology that combines hyperparameter optimization, bootstrap aggregation (bagging), and ensemble-based validation to ensure robust model selection. This comprehensive approach is carefully designed to evaluate various classifiers under various configurations while addressing common issues such as overfitting, high variance, and unreliable performance estimates. By integrating multiple techniques into a structured pipeline, the methodology aims to produce models that generalize well to unseen data and provide reproducible, high-quality results. Six binary base experiments were trained to capture pairwise distinctions between different severity levels of OSA for the experiment's models. These include the following:

1. Non-OSA vs. Mild-OSA;
2. Non-OSA vs. Moderate-OSA;
3. Non-OSA vs. Severe-OSA;
4. Mild-OSA vs. Moderate-OSA;
5. Mild-OSA vs. Severe-OSA;
6. Moderate-OSA vs. Severe-OSA.

The complete training methodology was applied to every pairwise comparison. This ensured that base classifiers were explicitly optimized for the discriminative characteristics relevant to each subset of the data. The process is divided into several key stages, as described below.

5.3.7.1. Classifier Configuration and Hyperparameter Optimization

The first stage of base model training involves configuring a diverse set of classifiers and systematically optimizing their hyperparameters. A total of eighteen classifier configurations were explored to capture a broad spectrum of modeling paradigms [62]. These included the following:

- Traditional models: Decision Trees, Naïve Bayes, and Logistic Regression (L1 and L2 regularized).
- Distance- and projection-based models: k-Nearest Neighbors (kNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA).
- Margin-based models: Support Vector Machines (SVMs) with linear, radial basis function (RBF) and polynomial kernels.
- Ensemble methods: Random Forests, Bagged Trees, Gradient Boosting Machines (GBM), RUSBoost, and Subspace kNN.
- Neural networks: Both shallow and deep architectures.

The heterogeneous nature of breathing sound features and associated clinical data guided the selection of these classifiers. To address the varying degrees of complexity in the input space, the methodology balanced interpretable models (e.g., Logistic Regression, Decision Trees) with nonlinear learners (e.g., Neural Networks, SVMs, GBMs) [63,64]. Special emphasis was placed on ensemble methods, which are known for their robustness and ability to reduce overfitting, particularly in imbalanced and moderately sized datasets.

Each classifier was paired with a custom-defined hyperparameter search space tailored to its critical tuning variables (as detailed in Table 5-1). Using the Expected Improvement Plus (EI+) acquisition function, Bayesian optimization was employed to navigate these spaces. This method efficiently balances the exploration of high-dimensional parameter spaces with the exploitation of

high-performing regions, leading to faster convergence than grid or random search methods [64]. The optimization objective was to minimize the misclassification rate under 5-fold stratified cross-validation, using the following loss function:

$$L(\theta) = \frac{1}{K} \sum_{k=1}^K I(y_i \neq \hat{y}_i(\theta)) \quad 5-2$$

where θ represents the hyperparameters, K is the number of cross-validation folds, y_i is the actual label, and $\hat{y}_i(\theta)$ is the predicted label given the hyperparameters and $I(\cdot)$ is the indicator function [63,64]. Depending on classifier complexity, convergence was typically achieved with 50 iterations. This optimization method was selected because it efficiently balances exploration and exploitation in high-dimensional spaces, making it well-suited for complex models. Using 5-fold stratified cross-validation ensured that each model's performance estimate was reliable and not biased by a particular subset of the data. By covering a diverse set of modeling strategies, the methodology maximizes the likelihood of identifying experimental models with complementary strengths, which is critical for the subsequent ensemble learning and modeling phases [62,63].

5.3.7.2. Bootstrap Aggregation with OOB Validation

Following hyperparameter optimization, each classifier underwent bootstrap aggregation (bagging) to enhance generalization and reduce model variance [65]. For each base model, $B = 50$ bootstrap samples were generated by resampling the training data with replacement, where each sample D_b had size N , equal to the number of original training instances. Using each D_b , a base learner f_b was trained with its respective optimized hyperparameters θ^* [66]. For the $b - th$ bootstrap sample D_b :

1. A resample D_b was generated with replacement $|D_b| = N$, where N is the number of training instances.

2. Classifier f_b was trained on D_b using optimized θ^* .
3. OOB samples $D_{\text{ooob}}^{(b)} = D \setminus D_b$ were retained for validation

An essential advantage of this strategy is its natural support for out-of-bag (OOB) validation, which enables unbiased performance estimation without requiring a separate validation set. For instance, x_i had an OOB prediction computed by aggregating predictions from all base learners for which $x_i \notin D_b$, i.e., from models that did not see that instance during training [67]. Formally, the OOB prediction is given by

$$\hat{y}_i^{\text{OOB}} = \text{mode}(\{f_b(x_i) \mid x_i \in D_{(b)}^{\text{ooob}}\}) \quad 5-3$$

where $D_{(b)}^{\text{ooob}} = D \setminus D_b$ represents the set of OOB samples for the b -th bootstrap. This mechanism yields a robust estimate of each classifier's generalization performance while fully utilizing the available training data. For each classifier configuration, the following OOB-based metrics were computed from the ensemble's predictions:

- OOB Accuracy: Overall correct classification rate.
- OOB Sensitivity: True positive rate, capturing the ability to detect positive cases.
- OOB Specificity: True negative rate, reflecting the ability to identify negative cases correctly.

By averaging predictions across 50 model instances and leveraging OOB samples, this ensemble approach not only improves model stability and robustness but also provides reliable, data-efficient validation suitable for imbalanced or limited-size datasets [65–68].

5.3.7.3. Class Imbalance Mitigation

For datasets with skewed class distributions [69,70]:

1. Cost-sensitive learning: Class-weighted loss functions scale misclassification costs inversely to class frequencies [69].
2. Stratified bootstrapping: Maintains original class ratios in bootstrap samples [69].
3. OOB-balanced metrics: Performance evaluation weights classes by inverse frequency [69].

This three-pronged approach prevents bias toward majority classes while preserving detection capability for rare categories [69,70].

5.3.7.4. Robust Model Selection via Repeated Trials

Recognizing that machine learning training processes are inherently stochastic due to factors like data shuffling, bootstrap sampling, and optimization randomness, the methodology incorporated multiple independent training trials [71,72]. For each classifier configuration, five independent trials were conducted. Each trial involved reinitializing Bayesian hyperparameter optimization to ensure that different regions of the search space could be explored, thus avoiding convergence to suboptimal local minima [63]. Additionally, 50 new bootstrap models were generated in each trial to introduce variability into the ensemble learning process. After training, performance metrics were evaluated using OOB samples and a held-out test set. Among the five trials, the one achieving the highest OOB accuracy was selected for final model comparison. This process ensured that the best-performing model was chosen based on generalizable results rather than random fluctuations in performance.

5.3.7.5. Final Model Selection

For each pairwise model of the six models, the classifier with the highest OOB accuracy was selected as the final classifier. This selection criterion favors models that generalize well during training without overfitting, a benefit inherent to bootstrap aggregation and OOB validation

[73]. Additionally, by selecting the best classifier for each experiment based on OOB performance rather than test performance, the methodology avoids the risk of overfitting the test set, preserving its integrity for unbiased final evaluation [74]. All results, including detailed accuracy, sensitivity, and specificity metrics, were compiled into a unified table to facilitate cross-dataset comparisons and meta-analyses. This organized approach enables robust insights into the relative performance of different models across varying datasets and conditions.

5.3.8. Model Evaluation

To assess the effectiveness of the models, the evaluation framework relies on standard classification performance metrics. This approach emphasizes both predictive accuracy and the balanced assessment of model performance across different classes, particularly in the presence of class imbalance.

5.3.8.1. Evaluation Protocol

5.3.8.1.1. Out-of-Bag (OOB) Validation

During ensemble training, each base learner was evaluated on the subset of training samples excluded from its bootstrap resample. These out-of-bag predictions were used to estimate performance without requiring a separate validation set. The resulting OOB metrics provide an unbiased estimate of generalization error.

5.3.8.1.2. Independent Test Evaluation

After model training, performance was assessed using the same metrics on a held-out test set. To account for randomness in training (e.g., bootstrap sampling or stochastic optimization), the evaluation was repeated 25 times over independent trials with different random seeds.

5.3.8.1.3. Performance Metrics

Models were evaluated using three core metrics—accuracy, sensitivity, and specificity—on both out-of-bag (OOB) samples and independent test sets [75].

5.4. Results

The performance of the proposed OSA severity screening framework was evaluated across multiple classification tasks using both tracheal breathing sounds (TBS) and anthropometric features. Results are reported for six pairwise comparisons of OSA severity levels (Non-OSA, Mild, Moderate, Severe), emphasizing classifier robustness, generalization, and clinical relevance. Model evaluation was conducted using out-of-bag (OOB) validation during training and a fully independent blind test set to assess real-world applicability. Performance metrics, including accuracy, sensitivity, and specificity, were computed to reflect diagnostic balance. The results of the experiment's models support the utility of detecting OSA severity during wakefulness using these models, providing a rapid, non-invasive method. The following sub-sections show the results of the proposed methodology.

5.4.1. Feature Selection Results

The feature selection process identified distinct and informative sets of features for each model, capturing the essential aspects of breathing sounds during the mouth and nose inspiration and expiration phases. Table A2 provides a summary of selected features for each base model.

The best-performing model (Model 1) leveraged a combination of spectral and image-based features to effectively classify breathing patterns. Key features included spectral entropy and crest values extracted from specific frequency bands during mouth expiration gaps, wavelet-based spectral bandwidth, and kurtosis metrics derived from nose inspiration and expiration phases. Image-derived features, such as the number of holes, bounding box area, and texture contrast across the mouth and nose regions further enhanced the model's discriminatory capability. Additional contributing features included fractal dimension estimates, peak frequency, and statistical measures derived from Mel-Frequency Cepstral Coefficients (MFCCs). A detailed list of selected features for all models (Models 1–6) is provided in Appendix A-4, Table A-2.

The feature selection process across all six models identified a diverse set of acoustic, spectral, fractal, and image-based characteristics that effectively capture the nuances of breathing sounds. Most of these features were extracted predominantly from the mouth inspiration segments, which provided rich spectral and fractal information, while some were derived from expiration and combined inspiration–expiration phases. The selected features include spectral measures such as spectral centroid, entropy, skewness, flux, power spectral density statistics, fractal dimensions, and wavelet-based coefficients. Statistical descriptors of MFCCs, zero crossing rates, Bispectrum entropy, and harmonic–percussive source separation features further enrich the dataset. Additionally, morphological features extracted from image representations of the respiratory signals, such as bounding box area, number of holes, connected components, and Euler numbers,

contributed to capturing structural variations in the signal. Together, these features comprehensively represent both the temporal and frequency-domain properties of the respiratory signals, enabling robust discrimination between subject classes.

These comprehensive acoustic and morphological features were combined with seven key anthropometric variables: body mass index (BMI), age, sex, smoke history, neck circumference (NC), and Mallampati score (MPS). Integrating these physiological and demographic factors with the rich features of breathing sounds enhances the models' ability to reflect intrinsic body characteristics and breathing dynamics. This fusion yields 41 features for each model, resulting in a robust, multidimensional dataset for subsequent classification and analysis tasks.

5.4.2. Experiments Models Results

5.4.2.1. Training and Testing Results

The selected base classifiers, Random Forest, Support Vector Machines (SVMs) with polynomial kernels of degrees 3, 5, and 7, Subspace K-Nearest Neighbors (Subspace KNN), and Linear Discriminant Analysis (LDA), demonstrate consistently strong performance according to their out-of-bag (OOB) estimates. These classifiers were chosen for their reliability and robustness across different datasets. The OOB accuracies remain high across all models, typically exceeding 80%, indicating that the classifiers are well-calibrated and effectively generalized during internal validation. Additionally, the OOB sensitivity and specificity values are balanced, suggesting that these models strike a good balance between identifying true positives and minimizing false positives. Table 5-5 presents the out-of-bag (OOB) performance metrics for these classifiers.

Table 5-5 OOB results for experimental models.

Experiment	Classifier	OOB Accuracy	OOB Sensitivity	OOB Specificity
1	Random Forests	83.9	94.1	81.6
2	SVM Poly 5	86.7	86.8	86.6
3	SVM Poly 3	88.7	87.1	89.4
4	SVM Poly 7	81.9	82.2	81.5
5	Subspace KNN	75.0	72.5	79.2
6	LDA	80.3	77.1	82.9

Then, the test results corresponded to the models that achieved high OOB performance across each dataset. These independent evaluations further validate the generalization capability of the selected classifiers. The test accuracies, sensitivities, and specificities closely mirror the trends observed in the OOB evaluations, with deviations generally remaining within 10%, an acceptable range in practical classification tasks. Several models, including Random Forest and SVMs with polynomial kernels, achieved perfect sensitivity or specificity on specific datasets, highlighting their potential for robust classification in real-world applications. Table 5-6 shows the test performance metrics for these classifiers.

Table 5-6 OOB Test results for experimental models.

Experiment	Classifier	Test Accuracy	Test Sensitivity	Test Specificity
1	Random Forests	81.2	100	78.6
2	SVM Poly 5	78.9	83.3	76.9
3	SVM Poly 3	88.2	83.3	90.9
4	SVM Poly 7	76.9	72.7	100
5	Subspace KNN	81.8	83.3	80.0
6	LDA	78.6	71.4	85.7

To provide a comprehensive overview, we additionally report performances from all evaluated classifiers in Figures 5-6 and 5-7. These figures illustrate the distribution of accuracy, sensitivity, and specificity across classifiers, complementing the summary in Tables 5-5 and 5-6.

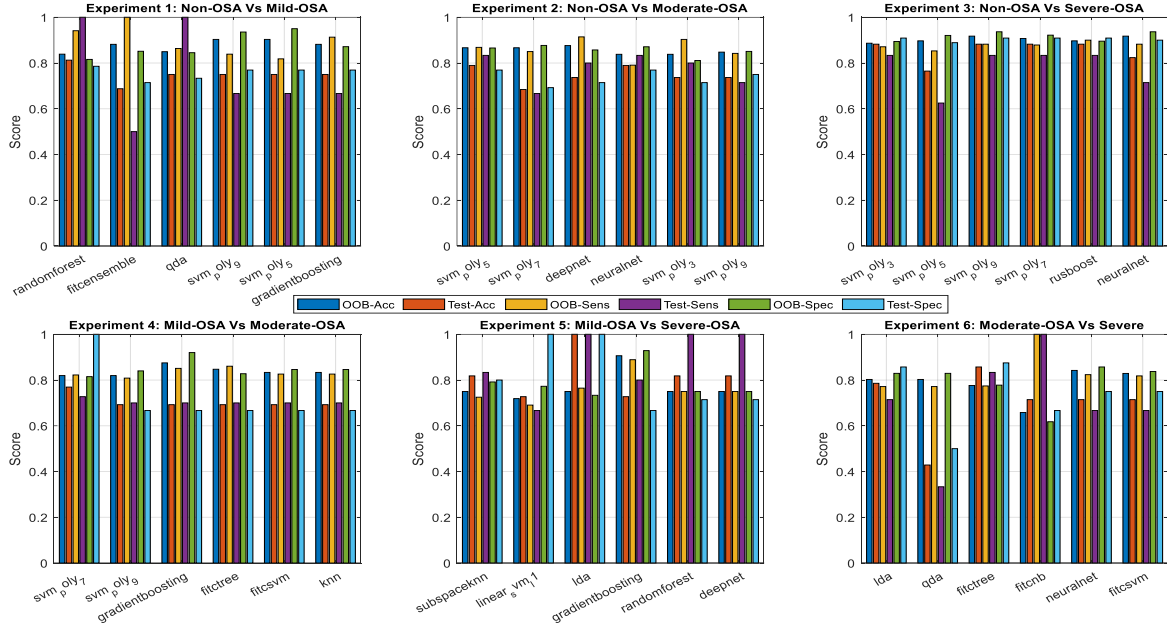


Fig. 5-6 A Expanded classifier performance metrics across six binary experiments. For each experiment, the top six classifiers were selected based on their mean performance. Grouped bar plots display both out-of-bag (OOB) and test set results for accuracy, sensitivity, and specificity. This visualization highlights the differences between training (out-of-bag, OOB) and generalization (test) performance.

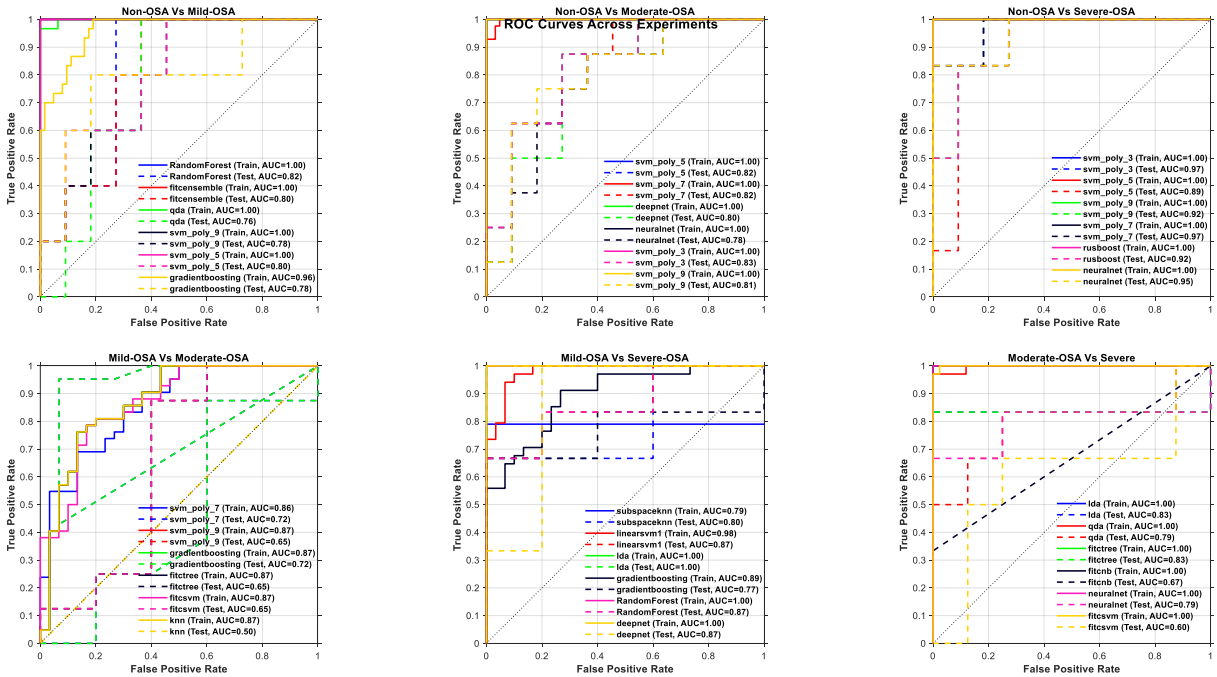


Fig. 5-7 A Receiver Operating Characteristic (ROC) curves of the top six classifiers for representative binary experiments. The curves illustrate the discrimination ability of each classifier across sensitivity–specificity trade-offs, complementing Tables 4 and 5 by providing a visual comparison of performance beyond single-value metrics.

5.4.2.2. K-Fold Results

The selected base classifiers demonstrated consistently strong performance across 3-fold cross-validation. The OOB accuracies remained high, generally exceeding 80%, indicating effective generalization and stability across folds. Sensitivity and specificity values were also well-balanced, suggesting that the models achieved a good trade-off between detecting true positives and minimizing false positives. Table 5-7 presents the 3-fold OOB performance metrics, which are consistent with the trends observed in the previous training–testing evaluations.

Table 5-7 OOB 3-fold cross-validation results of experimental models.

Experiment	Fold	Classifier	OOB Accuracy	OOB Sensitivity	OOB Specificity
1	1	Logistic	85.7	79.9	94.0
	2	Decision Trees	86.1	79.7	95.7
	3	Deep Neural Net	85.5	72.8	94.0
2	1	Random Forest	84.7	92.2	73.5
	2	SVM RBF	84.8	89.4	78.1
	3	Shallow Neural Net	88.1	92.0	82.4
3	1	SVM Poly 9	83.1	77.1	92.2
	2	SVM Poly 3	98.6	97.9	100.0
	3	Gradient Boosting	93.8	94.0	93.3
4	1	Ensemble Model	81.1	88.2	70.5
	2	RUSBoost	91.2	88.0	93.8
	3	Logistic	83.3	76.9	88.2
5	1	Decision Trees	80.0	78.9	80.8
	2	Deep Neural Net	85.7	88.0	83.3
	3	Random Forest	82.1	84.6	80.0
6	1	SVM RBF	78.3	82.4	73.1
	2	Shallow Neural Net	81.1	83.8	77.1
	3	SVM Poly 9	82.8	85.3	80.0

The test results correspond to models that achieved strong performance during internal validation. These independent evaluations further confirm the generalization capability of the selected models. The test accuracies, sensitivities, and specificities closely reflected the patterns observed in cross-validation, with deviations generally remaining within 10%, an acceptable range

for practical classification tasks. Table 5-8 presents the test performance metrics, which are consistent with the trends observed in previous evaluations.

Table 5-8 Test 3-fold cross-validation results of experimental models.

Experiment	Fold	Classifier	Test Accuracy	Test Sensitivity	Test Specificity
1	1	Logistic	82.1	84.6	78.3
	2	Decision Trees	77.1	77.0	77.2
	3	Deep Neural Net	77.8	79.6	75.0
2	1	Random Forest	82.1	87.0	75.0
	2	SVM RBF	77.8	81.5	72.2
	3	Shallow Neural Net	84.6	88.2	79.2
3	1	SVM Poly 9	81.9	85.7	79.4
	2	SVM Poly 3	72.1	70.4	75.0
	3	Gradient Boosting	73.5	75.0	70.0
4	1	Ensemble Model	76.4	78.6	75.0
	2	RUSBoost	87.4	81.8	91.2
	3	Logistic	73.5	70.7	77.8
5	1	Decision Trees	76.7	75.1	79.0
	2	Deep Neural Net	84.6	90.0	81.3
	3	Random Forest	73.7	71.2	71.2
6	1	SVM RBF	72.7	71.4	74.7
	2	Shallow Neural Net	73.5	74.4	72.2
	3	SVM Poly 9	71.9	72.2	71.3

5.5. Discussion

The proposed OSA severity screening during wakefulness framework demonstrates a promising advance in non-invasive, wakefulness-based diagnostic tools, particularly by integrating tracheal breathing sound analysis and anthropometric data. The structured evaluation across six pairwise OSA severity classifications offers robust performance metrics and insights into the physiological and acoustic distinctions among severity levels. This discussion synthesizes these findings, emphasizing the methodology's strengths and implications for clinical practice.

Incorporating SHAP into the feature selection pipeline provided an interpretable and data-driven mechanism to quantify each feature's contribution toward OSA severity classification. Unlike traditional ranking approaches, SHAP integrates cooperative game theory principles to

assign fair importance values to features based on their marginal contributions across multiple model predictions. This allowed us to confirm that physiologically relevant features, such as spectral entropy, bispectrum-derived texture measures, and anthropometric variables like BMI and neck circumference were consistently influential across models. Significantly, SHAP enhanced transparency by linking acoustic and morphological variations in tracheal breathing sounds with interpretable physiological correlations, thereby bridging the gap between clinical understanding and algorithmic decision-making. When combined with Recursive Feature Elimination (RFE), the SHAP-guided ranking ensured that only the most stable and clinically meaningful features were retained, which improved model robustness while reducing dimensionality. This integration strengthens the interpretability and reliability of the proposed wakefulness-based OSA severity screening framework.

The feature selection process revealed diverse spectral, temporal, fractal, and morphological characteristics extracted from mouth- and nose-breathing segments during different respiratory phases. Notably, features such as spectral entropy, crest, kurtosis, fractal dimension, and MFCC-based statistics emerged repeatedly across models. These features are physiologically meaningful, as they capture underlying variations in airway obstruction, turbulence, and breathing effort associated with different severities of OSA. Notably, morphological features derived from bispectrum image representations of the respiratory signals, such as bounding box area, Euler number, and connected components, provide a novel dimension to acoustic analysis. These image-based descriptors translate subtle acoustic changes into quantifiable structural patterns, enhancing interpretability and model transparency. Furthermore, the consistent contribution of mouth inspiration segments as primary sources of discriminative features underscores their diagnostic richness, likely due to greater variability in the upper-airway resistance during inspiration in OSA

patients. This is because mouth-breathing bypasses the nasal passages and exposes the more collapsible pharyngeal airway to direct airflow. During inspiration, this region is more prone to dynamic narrowing and turbulence in individuals with OSA, resulting in acoustic patterns that better reflect underlying structural abnormalities compared to nasal breathing. This aligns with the prior literature highlighting mouth-breathing as a compensatory mechanism in individuals with compromised nasal airflow and may reflect airway collapse dynamics during wakeful states [19,20].

The selected features encompass a broad range of physiological representations of breathing sound dynamics, each reflecting essential aspects of upper-airway structure and function that are affected by obstructive sleep apnea (OSA). Spectral features, such as spectral entropy, skewness, kurtosis, crest, centroid, and bandwidth, quantify the distribution and organization of energy across frequencies. In patients with OSA, upper-airway obstruction during inspiration and expiration leads to increased turbulence, reflected in broader spectral distributions (higher entropy), asymmetric power distribution (skewness), and heavier spectral tails (kurtosis). These features can capture abnormal airflow patterns due to pharyngeal collapse, especially during inspiration, which is more sensitive to airway resistance [20].

Fractal dimensions and nonlinear dynamics measures such as the Hurst exponent, Lyapunov exponent, and Katz fractal dimension reflect the irregularity and complexity of breathing signals, which tend to increase with the severity of OSA due to variable airflow and compensatory muscle activity [20,21,76]. MFCCs (Mel-Frequency Cepstral Coefficients), though commonly used in speech processing, effectively capture spectral envelope variations that correlate with airway resonance characteristics, which are particularly altered in OSA due to anatomical and functional airway changes [77]. Bispectrum features and bicoherence quantify

quadratic phase coupling between frequencies, providing insights into the nonlinear and harmonic interactions typical of turbulent breathing in OSA.

Time-domain features such as zero-crossing rate, peak frequency, and signal energy characterize oscillatory behavior and airflow strength [19]. These are sensitive to inspiratory effort and upper-airway resistance [19]. Additionally, wavelet-based features offer a multi-resolution analysis of signal transients, making them suitable for identifying events such as partial obstructions or arousals during respiration. CQT (Constant-Q Transform) and entropy-based features derived from wavelet or CQT domains reflect subtle changes in airflow rhythm and complexity [78], which may not be detectable in standard spectral measures.

Morphological and image-based features such as bounding box area, number of holes, Euler number, and texture measures are extracted from spectrogram or bispectrum image representations and serve as indirect quantifiers of structural variation in acoustic patterns. These are physiologically linked to airway geometry and dynamic obstruction events [20,79], as turbulent flow often generates unique spatial textures in time–frequency representations. Finally, anthropometric features such as BMI, neck circumference, Mallampati score, and age are directly related to anatomical risk factors for OSA, including fat deposition around the neck, airway collapsibility, and tongue size.

Together, these features form a multidimensional physiological signature of breathing under different severities of OSA. Their combined use enhances the ability to non-invasively and objectively screen OSA severity during wakefulness with high reliability. While anthropometric data improved classification accuracy, reliance on such features may limit usability in home-based screenings. A direct comparison with tools like STOP-Bang and Berlin questionnaires was not

conducted in this study; future work should benchmark performance against these established screening methods to better contextualize the advantages.

For the training testing results, the base classifiers demonstrated strong internal validation performance through out-of-bag (OOB) evaluation, with accuracies generally exceeding 80%, and a balanced sensitivity and specificity. These results indicate that the models are well-calibrated and generalize effectively within the training data. The SVM with polynomial kernels (particularly degrees 3 and 5) and Random Forests emerged as consistently high performers, achieving high accuracy and balanced diagnostic metrics across severity comparisons. Such consistency highlights their ability to model nonlinear interactions and feature dependencies in complex acoustic-physiological data. External validation using a blind test set reinforced the models' real-world applicability. The test performance observed mirrored OOB results, with deviations typically within 10%. For instance, Model 1 (Random Forest) achieved 100% sensitivity, indicating its ability to accurately capture actual OSA-positive cases, which is crucial for clinical screening scenarios. Similarly, Model 4 (SVM Poly 7) achieved 100% specificity, emphasizing its strength in confidently identifying non-OSA subjects. These extreme yet balanced outcomes across classifiers suggest complementary strengths ripe for ensemble integration.

The 3-fold cross-validation further confirmed the robust performance of the base classifiers, with accuracies generally remaining above 80%, and sensitivities and specificities showing balanced values across folds. These results indicate that the models are stable under repeated resampling and generalize effectively within the training data. Notably, SVMs with polynomial kernels (particularly degrees 3 and 9) and Random Forests consistently achieved high accuracy and well-balanced diagnostic metrics, reflecting their ability to capture nonlinear relationships and complex interactions in the data. External validation using the test sets reinforced

these findings, with deviations from OOB results typically within 10%. For example, Model 2 (Random Forest) achieved 87% sensitivity, demonstrating reliable detection of true positives, while Model 3 (Gradient Boosting) reached 93% specificity, emphasizing the accurate identification of true negatives. This complementary performance across classifiers underscores their potential value in ensemble modeling for enhanced predictive reliability.

An essential next step is to investigate whether predicted OSA severity correlates with perioperative complication rates, which would reinforce the translational value of this method for surgical risk stratification.

5.6. Limitations

Although the system shows strong potential, several limitations must be acknowledged. First, all data in this study were collected under controlled conditions at a single center using one microphone type (Sony ECM77B). This may limit generalizability, as real-world environments, such as different clinics or home settings, introduce variability in background noise, microphone type, and user behavior. Future work should therefore validate the framework across centers, devices, and include analyses of self-placement errors to assess robustness in more realistic scenarios.

Another issue is the reliance on consistent microphone placement over the trachea. While this was carefully managed during data collection, it is possible that in real-world use, especially in self-administered or remote settings, the placement may not always be accurate. Small changes in position could affect the sound quality and, in turn, the model's predictions. Future versions of the system should account for this, possibly by incorporating signal quality checks or providing user guidance. There is also the matter of anthropometric data. Some features, like neck circumference or jaw position, might not always be easy to measure, particularly outside of a

clinical environment. Although combining these features with acoustic data improves accuracy, it could limit the tool's practicality in settings where full measurements are not available. Exploring ways to work with partial data or identifying simpler substitutes would make the system more accessible. Although the dataset is relatively small and imbalanced across severity groups, we mitigated overfitting through 3-fold stratified cross-validation during hyperparameter tuning, bootstrap aggregation with OOB validation, and repeated independent training trials. Nevertheless, larger and multi-center datasets are needed for stronger statistical power.

Finally, while the tool performs well as a screening aid, it is not a replacement for clinical diagnosis. Its role should be clearly defined within the broader diagnostic process, helping to flag potential cases but not making final decisions. More work is needed to understand how clinicians would use the results in practice and how to communicate the model's output in a way that is both useful and trustworthy.

5.7. Conclusions

This study demonstrates the foundation for the effectiveness of a fast, objective, and non-invasive method for screening obstructive sleep apnea (OSA) severity using tracheal breathing sounds (TBS) recorded during wakefulness. By combining features from binary severity comparisons with anthropometric data, the proposed experiment's model approach achieved high classification performance across multiple machine learning models. Notably, the SVM with a third-degree polynomial kernel delivered strong out-of-bag and test results, while Random Forests achieved perfect test sensitivity. These findings support the potential of TBS-based analysis as a practical and accessible alternative to polysomnography, enabling a reliable assessment of OSA severity in under 10 min and representing a significant advancement for early detection and perioperative risk management.

The developed framework successfully harnesses the multidimensional richness of breathing sound analysis and anthropometric data to screen for OSA severity levels with high accuracy and clinical relevance. The effective combination of feature selection, robust base classifiers, and a well-structured approach illustrates a scalable and interpretable method for non-invasive respiratory assessment. These findings represent a crucial step toward AI-driven, accessible sleep disorder diagnostics that can bridge the current gaps in OSA identification and management.

References

- [1] Rizzo, D.; Baltzan, M.; Sirpal, S.; Dosman, J.; Kaminska, M.; Chung, F. Prevalence and Regional Distribution of Obstructive Sleep Apnea in Canada: Analysis from the Canadian Longitudinal Study on Aging. *Can. J. Public Health* 2024, 115, 970–979. <https://doi.org/10.17269/s41997-024-00911-8>.
- [2] Lechat, B.; Naik, G.; Reynolds, A.; Aishah, A.; Scott, H.; Loffler, K.A.; Vakulin, A.; Escourrou, P.; McEvoy, R.D.; Adams, R.J.; et al. Multinight Prevalence, Variability, and Diagnostic Misclassification of Obstructive Sleep Apnea. *Am. J. Respir. Crit. Care Med.* 2022, 205, 563–569. <https://doi.org/10.1164/rccm.202107-1761OC>.
- [3] Singh, M.; Liao, P.; Kobah, S.; Wijeyesundera, D.N.; Shapiro, C.; Chung, F. Proportion of Surgical Patients with Undiagnosed Obstructive Sleep Apnoea. *Br. J. Anaesth.* 2013, 110, 629–636. <https://doi.org/10.1093/bja/aes465>.
- [4] Espiritu, J.R.D. Health Consequences of Obstructive Sleep Apnea. In *Management of Obstructive Sleep Apnea*; Springer International Publishing: Cham, Switzerland, 2021; pp. 23–43.

- [5] American Academy of Sleep Medicine. Hidden Health Crisis Costing America Billions: Underdiagnosing and Undertreating Obstructive Sleep Apnea Draining Healthcare System, 1st ed.; Frost & Sullivan: Darien, IL, USA, 2016.
- [6] The Harvard Medical School Division of Sleep Medicine. The Price of Fatigue: The Surprising Economic Costs of Unmanaged Sleep Apnea; Harvard Medical School Division of Sleep Medicine Boston: Boston, MA, USA, 2010.
- [7] Colten, H.R.; Altevogt, B.M. Sleep Disorders and Sleep Deprivation; National Academies Press: Washington, DC, USA, 2006; ISBN 978-0-309-10111-0.
- [8] American Academy of Sleep Medicine International Classification of Sleep Disorders: Diagnostic & Coding Manual, 2nd ed.; American Academy of Sleep Medicine: Westchester, IL, USA, 2005; ISBN 0965722023.
- [9] Noda, A.; Yasuma, F.; Miyata, S.; Iwamoto, K.; Yasuda, Y.; Ozaki, N. Sleep Fragmentation and Risk of Automobile Accidents in Patients with Obstructive Sleep Apnea—Sleep Fragmentation and Automobile Accidents in OSA. *Health* 2019, 11, 171–181.
- [10] Young, T.; Finn, L.; Peppard, P.E.; Szklo-Coxe, M.; Austin, D.; Nieto, F.J.; Stubbs, R.; Hla, K.M. Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort. *Sleep* 2008, 31, 1071. <https://doi.org/10.5665/sleep/31.8.1071>.
- [11] Yoshihisa, A.; Takeishi, Y. Sleep Disordered Breathing and Cardiovascular Diseases. *J. Atheroscler. Thromb.* 2019, 26, 315–327. <https://doi.org/10.5551/JAT.RV17032>.
- [12] Berry, R.B.; Brooks, R.; Gamaldo, C.E.; Harding, S.M.; Marcus, C.; Vaughn, B.V. The AASM Manual for the Scoring of Sleep and Associated Events. Rules Terminol. Tech. Specif. Darien Ill. Am. Acad. Sleep Med. 2012, 176, 2012.

- [13] Kushida, C.A.; Littner, M.R.; Morgenthaler, T.; Alessi, C.A.; Bailey, D.; Coleman, J.; Friedman, L.; Hirshkowitz, M.; Kapen, S.; Kramer, M.; et al. Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005. *Sleep* 2005, 28, 499–523. <https://doi.org/10.1093/sleep/28.4.499>.
- [14] Bradley, T.D.; Floras, J.S. *Sleep Apnea: Implications in Cardiovascular and Cerebrovascular Disease*, 2nd ed.; Bradley, T.D., Floras, J.S., Eds.; CRC Press: Boca Raton, FL, USA, 2013; ISBN 978-0-429-11867-8.
- [15] Butt, M.; Dwivedi, G.; Khair, O.; Lip, G.Y.H. Obstructive Sleep Apnea and Cardiovascular Disease. *Int. J. Cardiol.* 2010, 139, 7–16. <https://doi.org/10.1016/J.IJCARD.2009.05.021>.
- [16] Chen, L.; Pivetta, B.; Nagappa, M.; Saripella, A.; Islam, S.; Englesakis, M.; Chung, F. Validation of the STOP-Bang Questionnaire for Screening of Obstructive Sleep Apnea in the General Population and Commercial Drivers: A Systematic Review and Meta-Analysis. *Sleep Breath.* 2021, 25, 1741–1751. <https://doi.org/10.1007/S11325-021-02299-Y>.
- [17] Mazzotti, D.R.; Keenan, B.T.; Thorarinsdottir, E.H.; Gislason, T.; Pack, A.I.; Pack, A.I.; Schwab, R.; Maislin, G.; Keenan, B.T.; Jafari, N.; et al. Is the Epworth Sleepiness Scale Sufficient to Identify the Excessively Sleepy Subtype of OSA? *Chest* 2022, 161, 557–561. <https://doi.org/10.1016/j.chest.2021.10.027>.
- [18] Yadollahi, A.; Moussavi, Z. Acoustic Obstructive Sleep Apnea Detection. In *Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, USA, 3–6 September 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 7110–7113.

- [19] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea Using Anthropometric Information and Tracheal Breathing Sounds. *Sci. Rep.* 2019, 9, 11467. <https://doi.org/10.1038/s41598-019-47998-5>.
- [20] Elwali, A.; Moussavi, Z. Obstructive Sleep Apnea Screening and Airway Structure Characterization During Wakefulness Using Tracheal Breathing Sounds. *Ann. Biomed. Eng.* 2017, 45, 839–850. <https://doi.org/10.1007/S10439-016-1720-5>.
- [21] Montazeri, A.; Giannouli, E.; Moussavi, Z. Assessment of Obstructive Sleep Apnea and Its Severity during Wakefulness. *Ann. Biomed. Eng.* 2012, 40, 916–924. <https://doi.org/10.1007/s10439-011-0456-5>.
- [22] Hajipour, F.; Jozani, M.J.; Moussavi, Z. A Comparison of Regularized Logistic Regression and Random Forest Machine Learning Models for Daytime Diagnosis of Obstructive Sleep Apnea. *Med. Biol. Eng. Comput.* 2020, 58, 2517–2529. <https://doi.org/10.1007/s11517-020-02206-9>.
- [23] Hajipour, F.; Jozani, M.J.; Elwali, A.; Moussavi, Z. Regularized Logistic Regression for Obstructive Sleep Apnea Screening during Wakefulness Using Daytime Tracheal Breathing Sounds and Anthropometric Information. *Med. Biol. Eng. Comput.* 2019, 57, 2641–2655. <https://doi.org/10.1007/S11517-019-02052-4>.
- [24] Simply, R.M.; Dafna, E.; Zigel, Y. Diagnosis of Obstructive Sleep Apnea Using Speech Signals From Awake Subjects. *IEEE J. Sel. Top. Signal Process* 2020, 14, 251–260. <https://doi.org/10.1109/JSTSP.2019.2955019>.
- [25] Sola-Soler, J.; Fiz, J.A.; Torres, A.; Jane, R. Identification of Obstructive Sleep Apnea Patients from Tracheal Breath Sound Analysis during Wakefulness in Polysomnographic Studies. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in

Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 4232–4235.

[26] Alqudah, A.M.; Elwali, A.; Kupiak, B.; Hajipour, F.; Jacobson, N.; Moussavi, Z. Obstructive Sleep Apnea Detection during Wakefulness: A Comprehensive Methodological Review. *Med. Biol. Eng. Comput.* 2024, 62, 1277–1311. <https://doi.org/10.1007/s11517-024-03020-3>.

[27] Tregear, S.; Reston, J.; Schoelles, K.; Phillips, B. Obstructive Sleep Apnea and Risk of Motor Vehicle Crash: Systematic Review and Meta-Analysis. *J. Clin. Sleep. Med.* 2009, 5, 573. <https://doi.org/10.5664/jcsm.27662>.

[28] Hajipour, F.; Moussavi, Z. Spectral and Higher Order Statistical Characteristics of Expiratory Tracheal Breathing Sounds During Wakefulness and Sleep in People with Different Levels of Obstructive Sleep Apnea. *J. Med. Biol. Eng.* 2019, 39, 244–250. <https://doi.org/10.1007/s40846-018-0409-7>.

[29] Elwali, A.; Moussavi, Z. Predicting Polysomnography Parameters from Anthropometric Features and Breathing Sounds Recorded during Wakefulness. *Diagnostics* 2021, 11, 905. <https://doi.org/10.3390/DIAGNOSTICS11050905>.

[30] Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* 2001, 17, 520–525. <https://doi.org/10.1093/bioinformatics/17.6.520>.

[31] Batista, G.E.A.P.A.; Monard, M.C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Appl. Artif. Intell.* 2003, 17, 519–533. <https://doi.org/10.1080/713827181>.

- [32] Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1994; ISBN 9780429246593.
- [33] Rangayyan, R.M.; Reddy, N.P. *Biomedical Signal Analysis: A Case-Study Approach*; Pergamon Press: New York, NY, USA, 2002; Volume 30.
- [34] Mendel, J.M. Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications. *Proc. IEEE* 1991, 79, 278–305. <https://doi.org/10.1109/5.75086>.
- [35] Astfalck, L.C.; Sykulski, A.M.; Cripps, E.J. Debiasing Welch's Method for Spectral Density Estimation. *Biometrika* 2023, 111, 1313–1329.
- [36] Jiang, M.; Wang, D.; Kuang, Y.; Mo, X. A Bicoherence-Based Nonlinearity Measurement Method for Identifying the Quadratic Phase Coupling of Nonlinear Systems. *Int. J. Non Linear Mech.* 2021, 131, 103–109.
- [37] Dlask, M.; Kukal, J. Hurst Exponent Estimation from Short Time Series. *Signal Image Video Process* 2019, 13, 263–269.
- [38] Farrús, M.; Hernando, J.; Ejarque, P. Jitter and Shimmer Measurements for Speaker Recognition. In *Proceedings of the Interspeech 2007, Antwerp, Belgium, 27–31 August 2007*; ISCA: Singapore, 2007; pp. 778–781.
- [39] Jotz, G.P.; Cervantes, O.; Abrahão, M.; Settanni, F.A.P.; de Angelis, E.C. Noise-to-Harmonics Ratio as an Acoustic Measure of Voice Disorders in Boys. *J. Voice* 2002, 16, 28–31. [https://doi.org/10.1016/S0892-1997\(02\)00068-1](https://doi.org/10.1016/S0892-1997(02)00068-1).
- [40] Gosala, B.; Kapgate, P.D.; Jain, P.; Chaurasia, R.N.; Gupta, M. Wavelet transforms for feature engineering in EEG data processing: An application on Schizophrenia. *Biomed. Signal Process. Control* 2023, 85, 104811. <https://doi.org/10.1016/j.bspc.2023.104811>

- [41] Abdul, Z.K.; Al-Talabani, A.K. Mel Frequency Cepstral Coefficient and Its Applications: A Review. *IEEE Access* 2022, 10, 122136–122158. <https://doi.org/10.1109/ACCESS.2022.3223444>.
- [42] Wang, J.-C.; Wang, J.-F.; Weng, Y.-S. Chip design of MFCC extraction for speech recognition. *Integration* 2002, 32, 111–131. [https://doi.org/10.1016/S0167-9260\(02\)00045-7](https://doi.org/10.1016/S0167-9260(02)00045-7)
- [43] Kohlrausch, A. Binaural Masking Experiments Using Noise Maskers with Frequency-Dependent Interaural Phase Differences. II: Influence of Frequency and Interaural-Phase Uncertainty. *J. Acoust. Soc. Am.* 1990, 88, 1749–1756. <https://doi.org/10.1121/1.400249>.
- [44] Rosenstein, M.T.; Collins, J.J.; De Luca, C.J. A Practical Method for Calculating Largest Lyapunov Exponents from Small Data Sets. *Physica D* 1993, 65, 117–134.
- [45] Zhao, K.; Wen, H.; Guo, Y.; Scano, A.; Zhang, Z. Feasibility of Recurrence Quantification Analysis (RQA) in Quantifying Dynamical Coordination among Muscles. *Biomed. Signal Process Control* 2023, 79, 104042. <https://doi.org/10.1016/j.bspc.2022.104042>.
- [46] Borowska, M. Entropy-Based Algorithms in the Analysis of Biomedical Signals. In *Studies in Logic, Grammar and Rhetoric*; University of Białystok: Białystok, Poland, 2015; Volume 43, pp. 21–32. <https://doi.org/10.1515/slgr-2015-0039>
- [47] Ojala, T.; Pietikainen, M.; Maenpää, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002, 24, 971–987. <https://doi.org/10.1109/TPAMI.2002.1017623>.
- [48] Divya, S.; Suresh, L.P.; John, A. Image Feature Generation Using Binary Patterns—LBP, SLBP and GBP. In; 2022; pp. 233–239.
- [49] Selesnick, I.W. Wavelet Transform With Tunable Q-Factor. *IEEE Trans. Signal Process.* 2011, 59, 3560–3575. <https://doi.org/10.1109/TSP.2011.2143711>.

- [50] Márton, L.F.; Brassai, S.T.; Bakó, L.; Losonczi, L. Detrended fluctuation analysis of EEG signals. *Procedia Technol.* 2014, 12, 125–132. <https://doi.org/10.1016/j.protcy.2013.12.465>
- [51] Vergara, J.R.; Estévez, P.A. A Review of Feature Selection Methods Based on Mutual Information. *Neural Comput. Appl.* 2014, 24, 175–186. <https://doi.org/10.1007/s00521-013-1368-0>.
- [52] Zhao, Z.; Liu, H. Feature Selection Based on Mutual Information with Correlation Coefficient. *Appl. Intell.* 2022, 52, 1169–1180. <https://doi.org/10.1007/s10489-021-02524-x>.
- [53] Liu, S.; Motani, M. Improving Mutual Information Based Feature Selection by Boosting Unique Relevance. *arXiv* 2022, arXiv:2212.06143.
- [54] Singh, D.; Singh, B. Feature Wise Normalization: An Effective Way of Normalizing Data. *Pattern Recognit.* 2022, 122, 108307. <https://doi.org/10.1016/j.patcog.2021.108307>.
- [55] Haury, A.-C.; Gestraud, P.; Vert, J.-P. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE* 2011, 6, e28210. <https://doi.org/10.1371/journal.pone.0028210>.
- [56] Wang, D.; Zhang, H.; Liu, R.; Lv, W.; Wang, D. T-Test Feature Selection Approach Based on Term Frequency for Text Categorization. *Pattern Recognit. Lett.* 2014, 45, 1–10. <https://doi.org/10.1016/j.patrec.2014.02.013>.
- [57] Khoshgoftaar, T.M.; Wang, H.; Liang, Q.; Hancock, J.T. Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods. *J. Big Data* 2024, 11, 44. <https://doi.org/10.1186/s40537-024-00905-w>.
- [58] Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Trans. Syst. Man Cybern. Part. A Syst. Hum.* 2010, 40, 185–197. <https://doi.org/10.1109/TSMCA.2009.2029559>.

- [59] Mounce, S.R.; Ellis, K.; Edwards, J.M.; Speight, V.L.; Jakomis, N.; Boxall, J.B. Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems. *Water Resour. Manag.* 2017, 31, 925–942. <https://doi.org/10.1007/s11269-017-1595-8>.
- [60] Wolpert, D.H. Stacked Generalization. *Neural Netw.* 1992, 5, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
- [61] Wang, M.; Qian, Y.; Yang, Y.; Chen, H.; Rao, W.-F. Improved Stacking Ensemble Learning Based on Feature Selection to Accurately Predict Warfarin Dose. *Front. Cardiovasc. Med.* 2024, 10, 1320938. <https://doi.org/10.3389/fcvm.2023.1320938>.
- [62] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* 2021, 2, 160. <https://doi.org/10.1007/s42979-021-00592-x>.
- [63] Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012*; Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.
- [64] Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* 2012, 13, 281–305.
- [65] Breiman, L. Bagging Predictors. *Mach. Learn.* 1996, 24, 123–140. <https://doi.org/10.1007/BF00058655>.
- [66] Büchlmann, P.; Yu, B. Analyzing Bagging. *Ann. Stat.* 2002, 30, 927–961.
- [67] Kwon, Y.; Zou, J. Data-OOB: Out-of-Bag Estimate as a Simple and Efficient Data Value. In *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA,*

- 23–29 July 2023; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 18135–18152.
- [68] Klevak, E.; Lin, S.; Martin, A.; Linda, O.; Ringger, E. Out-Of-Bag Anomaly Detection. arXiv 2020, arXiv:2009.09358.
- [69] Japkowicz, N.; Stephen, S. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* 2002, 6, 429–449.
- [70] He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* 2009, 21, 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>.
- [71] Varma, S.; Simon, R. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinform.* 2006, 7, 91. <https://doi.org/10.1186/1471-2105-7-91>.
- [72] Cawley, G.C.; Talbot, N.L.C. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* 2010, 11, 2079–2107.
- [73] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp. 1137–1143.
- [74] Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* 1998, 10, 1895–1923. <https://doi.org/10.1162/089976698300017197>.
- [75] Alqudah, A.M.; Moussavi, Z. A Review of Deep Learning for Biomedical Signals: Current Applications, Advancements, Future Prospects, Interpretation, and Challenges. *Comput. Mater. Contin.* 2025, 83, 3753–3841. <https://doi.org/10.32604/cmc.2025.063643>.

- [76] Finkelstein, Y.; Wolf, L.; Nachmani, A.; Lipowezky, U.; Rub, M.; Shemer, S.; Berger, G. Velopharyngeal Anatomy in Patients With Obstructive Sleep Apnea Versus Normal Subjects. *J. Oral Maxillofac. Surg.* 2014, 72, 1350–1372. <https://doi.org/10.1016/j.joms.2013.12.006>.
- [77] Goldshtein, E.; Tarasiuk, A.; Zigel, Y. Automatic Detection of Obstructive Sleep Apnea Using Speech Signals. *IEEE Trans. Biomed. Eng.* 2011, 58, 1373–1382. <https://doi.org/10.1109/TBME.2010.2100096>.
- [78] Qi, F.; Li, C.; Wang, S.; Zhang, H.; Wang, J.; Lu, G. Contact-Free Detection of Obstructive Sleep Apnea Based on Wavelet Information Entropy Spectrum Using Bio-Radar. *Entropy* 2016, 18, 306. <https://doi.org/10.3390/e18080306>.
- [79] Shams, E.; Karimi, D.; Moussavi, Z. Bispectral Analysis of Tracheal Breath Sounds for Obstructive Sleep Apnea. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 37–40.

CHAPTER SIX: AWAKE INSIGHTS FOR OBSTRUCTIVE SLEEP APNEA: SEVERITY DETECTION USING TRACHEAL BREATHING SOUNDS AND META-MODEL ANALYSIS

Ali Mohammad Alqudah¹, and Zahra Moussavi^{1,2}

¹ Biomedical Engineering, University of Manitoba, Winnipeg, Canada

² Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
Diagnostics (2026) 16 (3) (<https://doi.org/10.3390/diagnostics16030448>)

Abstract

Background/Objectives: Obstructive sleep apnea (OSA) is a prevalent, yet underdiagnosed, disorder associated with cardiovascular and cognitive risks. While overnight polysomnography (PSG) remains the diagnostic gold standard, it is resource-intensive and impractical for large-scale rapid screening. **Methods:** This study extends prior work on feature extraction and binary classification using tracheal breathing sounds (TBS) and anthropometric data by introducing a meta-modeling framework that utilizes machine learning (ML) and aggregates six one-vs.-one classifiers for multi-class OSA severity prediction. We employed out-of-bag (OOB) estimation and three-fold cross-validation to assess model generalization performance. To enhance reliability, the framework incorporates conformal prediction to provide calibrated confidence sets. **Results:** In the three-class setting (non, mild, moderate/severe), the model achieved 76.7% test accuracy, 77.7% sensitivity, and 87.1% specificity, with strong OOB performance of 91.1% accuracy, 91.6% sensitivity, and 95.3% specificity. Three-fold confirmed stable performance across folds (mean accuracy: 77.8%; mean sensitivity: 78.6%; mean specificity: 76.4%) and conformal prediction achieved full coverage with an average set size of 2. In the four-class setting (non, mild, moderate, severe), the model achieved 76.7% test accuracy, 75% sensitivity, and 92% specificity, with OOB performance of 88.2% accuracy, 91.6%

sensitivity, and 88.2% specificity. **Conclusions:** These findings support the potential of this non-invasive system as an efficient and rapid OSA severity assessment whilst awake, offering a scalable alternative to PSG for large-scale screening and clinical triaging.

6.1. Introduction and Literature Review

Obstructive sleep apnea (OSA) is highly prevalent yet substantially underrecognized; in Canada, nearly one quarter of adults are estimated to be at high risk, with most cases undiagnosed [1,2]. The economic burden in the United States alone exceeds \$150 billion annually [3], underscoring the need for scalable and rapid assessment strategies.

Overnight polysomnography (PSG) remains the diagnostic gold standard but is resource-intensive and often inaccessible due to long wait times [4]. Thus, simpler overnight tools, in practice, simpler overnight tools such as fingertip oximetry are widely used for home triage. The oxygen desaturation index (ODI) derived from nocturnal SaO₂ [5] correlates strongly with the apnea–hypopnea index (AHI) and can classify moderate-to-severe OSA with good accuracy [6]. However, these methods depend on capturing sleep-related desaturation events, making them unsuitable for time-sensitive situations, i.e., surgeries requiring general anesthesia that need a quick and accurate assessment [7].

Wakefulness-based detection offers a complementary approach by identifying anatomical and physiological traits of airway collapsibility without requiring sleep monitoring [8,9]. This is particularly valuable in perioperative care, where a quick risk stratification is critical for preventing complications [10], and in outpatient or emergency settings, where PSG or overnight testing is not feasible. Questionnaires such as STOP-Bang provide high sensitivity but limited specificity, especially in obese patients [11–14]; thus, the need for accurate, non-invasive wakefulness-based methods is greatly of interest.

Recent advances in biological signal processing have shown that tracheal breathing sounds (TBS) recorded during wakefulness can provide a rapid, low-resource screening and risk stratification approach for OSA severity, particularly when overnight sleep monitoring is impractical [9]. When combined with anthropometric information, TBS-based models have demonstrated clinically relevant accuracies at treatment thresholds [8]. These findings suggest that wakefulness detection may bridge the gap between resource-intensive PSG and simple but nonspecific tools like STOP-Bang or overnight oximetry, enabling rapid, point-of-care identification of at-risk patients with OSA.

Several studies have shown the promise of TBS recorded during wakefulness as a low-cost, non-invasive alternative to overnight testing [8,9,15,16]. An initial work in small samples showed an accuracy over 90% accuracy in distinguishing severe OSA from non-OSA using spectral and nonlinear features of TBS [9]. Later, by combining TBS with anthropometric measures, it was shown that OSA could be detected robustly by clinically relevant accuracy in a much larger samples [8]. More recent work has confirmed the feasibility of advanced acoustic features and machine learning in improving robustness of the technique [17]. Collectively, these findings establish TBS as one of the most promising wakefulness-based tools for rapid OSA assessment, complementing, but fundamentally distinct from, overnight SaO₂ monitoring [5,6]. Nevertheless, all the existing techniques have been limited to binary detection of OSA versus non-OSA.

Unlike prior wakefulness-based TBS studies [8,9,15,16], which primarily focused on binary OSA detection or single clinical thresholds, the present work addresses the more challenging task of multi-class severity stratification. Moreover, previous studies reported point estimates without uncertainty quantification, limiting clinical interpretability. By integrating

pairwise acoustic specialization with conformal prediction, the proposed framework provides both severity stratification and calibrated confidence sets, addressing a key gap in prior TBS-based OSA research.

In this study, we present a framework for multi-class OSA severity screening and risk stratification during wakefulness, intended to complement rather than replace overnight diagnostic testing. Building on previously developed wakefulness-based TBS classifiers, the proposed framework integrates pairwise acoustic specialization within a stacked meta-model architecture. By incorporating conformal prediction, the system enables multi-level severity stratification with calibrated uncertainty estimates, supporting rapid and uncertainty-aware point-of-care triage and addressing a key gap in existing wakefulness-based OSA screening approaches.

6.2. Materials and Methods

This study presents a meta-model framework designed to classify OSA severity into four clinically relevant categories: Non-OSA, Mild, Moderate, and Severe. Recognizing the acoustic and physiological overlap between adjacent severity levels, the framework adopts a modular structure composed of six binary base classifiers trained in our earlier work [18], each targeting one-vs.-one class distinctions (e.g., Non-OSA vs. Mild, Mild vs. Moderate).

This work extends and completes the methodology introduced in [18] by unifying these previously developed binary classifiers into a single stacked meta-model that enables end-to-end multi-class OSA severity prediction. In this work, we reuse these classifiers and their associated feature sets to train a stacked ensemble meta-model capable of multi-class prediction. To enhance clinical applicability, the framework integrates conformal prediction, which augments predictions with calibrated uncertainty estimates. The complete system architecture and data flow are illustrated in Figure 6-1.

Methodology Workflow for OSA Severity Classification

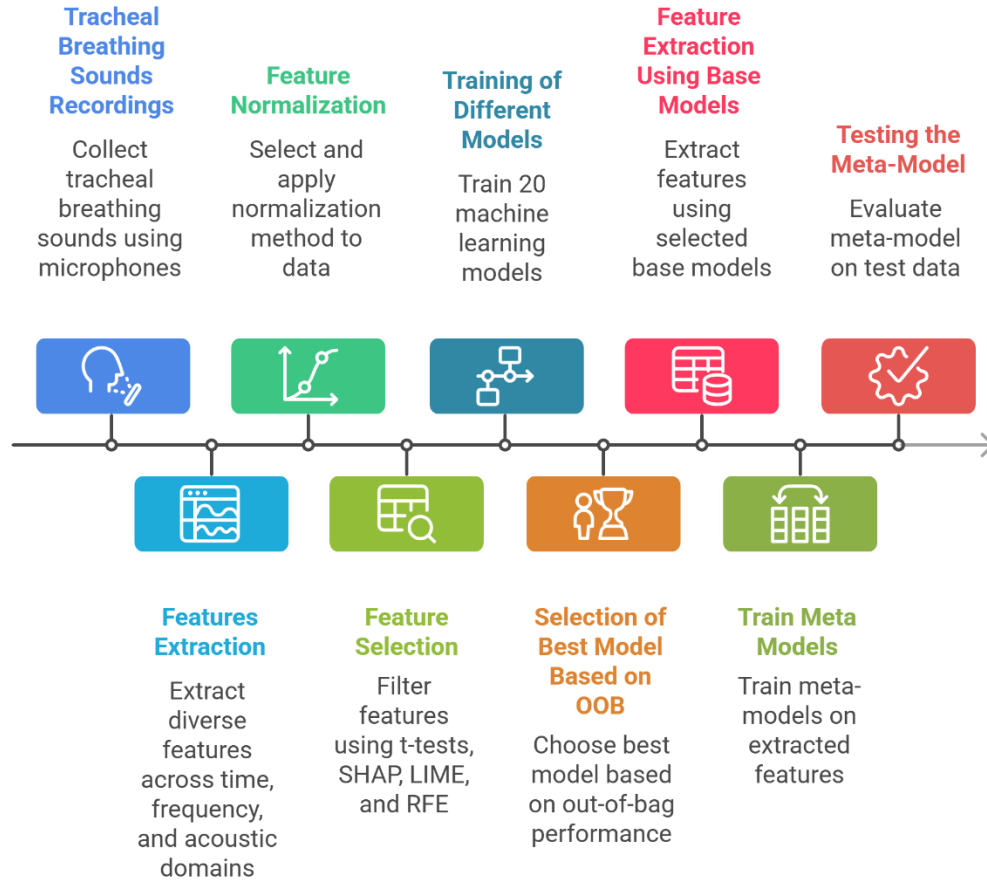


Fig. 6-1 General steps of the proposed model.

6.2.1. Dataset

The dataset used in this study was adopted from our previous work [18] which includes TBS recordings and anthropometric data from 199 individuals (74 non-OSA, 35 Mild-OSA, 50 Moderate-OSA, and 40 Severe-OSA). OSA severity groups were defined based on apnea–hypopnea index (AHI) values as follows: Non-OSA: $AHI < 5$ events/h, Mild OSA: $5 \leq AHI < 15$ events/h, Moderate OSA: $15 \leq AHI < 30$ events/h, and Severe OSA: $AHI \geq 30$ events/h. Breathing sounds were recorded while participants were awake (daytime) in supine position using a Sony ECM77B microphone, Tokyo, Japan omnidirectional condenser microphone (sensitivity: -52 dB ± 3.5 dB, frequency response: 40 Hz–20 kHz). placed over the suprasternal notch. Table 6-1

presents the total number of subjects in each severity group and their corresponding anthropometric data for the dataset used in this study. Study participants were instructed to breathe deeply for five full breath cycles through their nose, followed by another 5 deep breath cycles through their mouth with a nose clip in place. The study was approved by the Biomedical Research Ethics Board of the University of Manitoba, and all study subjects provided informed consent prior to participation.

Table 6-1 Severity groups and participants' anthropometric information. AHI: apnea-hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).

Group	Num. of Subjects	AHI	Age	Sex	BMI	MPS	NC
Non-OSA	74	1.18 ± 1.26	46.84 ± 12.89	29 M, 45 F	30.61 ± 6.29	41(1), 19(2), 6(3), 8(4)	38.78 ± 4.04
Mild-OSA	35	8.69 ± 2.58	52.31 ± 11.57	21 M, 14 F	34.29 ± 8.41	18(1), 6(2), 9(3), 1(4)	42.09 ± 6.49
Moderate-OSA	50	21.51 ± 4.23	54.74 ± 11.33	36 M, 14 F	33.78 ± 6.39	17(1), 17(2), 8(3), 8(4)	43.10 ± 3.43
Severe-OSA	40	69.54 ± 33.26	48.98 ± 11.15	30 M, 10 F	39.75 ± 8.65	5(1), 13(2), 14(3), 8(4)	45.28 ± 3.64

6.2.2. Dataset Split and Cohort Stratification for Training and Testing

To evaluate the model's generalization capability, the dataset comprising 199 subjects was split into 85% for training ($n = 169$) and 15% for testing ($n = 30$). To minimize training bias, the model was trained 10 times, each with a shuffled version of the dataset. Stratified sampling was used to ensure consistent distribution of clinical and demographic variables across both sets. Specifically, stratification preserved the distributions of OSA severity classes and key anthropometric variables, including age, body mass index (BMI), neck circumference (NC), and Mallampati score (MPS). The training subset was used for feature selection, model training, and internal validation via out-of-bag (OOB) estimates. The held-out test set remained untouched until

final evaluation to provide an unbiased measure of real-world performance. Tables 6-2 and 6-3 summarize the anthropometric profiles of subjects across severity groups in both subsets.

Table 6-2 Severity groups and Participants’ anthropometric information of the training set. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).

Group	Num. of Subjects	AHI	Age	Sex	BMI	MPS	NC
Non-OSA	63	1.2 ± 1.2	46.8 ± 13.5	24 M, 39 F	30.6 ± 6.6	36 (1), 17 (2), 5 (3), 5 (4)	39.0 ± 3.5
Mild-OSA	30	8.8 ± 2.6	50.9 ± 10.4	17 M, 13 F	34.1 ± 8.2	14 (1), 7 (2), 8 (3), 0 (4)	41.5 ± 7.0
Moderate-OSA	42	22.0 ± 4.3	53.7 ± 10.3	31 M, 11 F	33.7 ± 6.8	11 (1), 15 (2), 8 (3), 8 (4)	43.3 ± 3.3
Severe-OSA	34	68.5 ± 32.5	50.9 ± 11.1	24 M, 10 F	39.9 ± 9.3	4 (1), 11 (2), 13 (3), 6 (4)	45.0 ± 3.4

Table 6-3 Severity groups and Participants’ anthropometric information of the test set. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).

Group	Num. of Subjects	AHI	Age	Sex	BMI	MPS	NC
Non-OSA	11	0.9 ± 1.0	45.4 ± 10.7	6 M, 5 F	30.6 ± 3.8	6 (1), 3 (2), 1 (3), 1 (4)	40.1 ± 4.5
Mild-OSA	5	6.7 ± 1.7	52.4 ± 12.6	3 M, 2 F	30.3 ± 10.2	3 (1), 1 (2), 1 (3), 0 (4)	41.9 ± 1.5
Moderate-OSA	8	20.7 ± 4.5	58.5 ± 7.3	4 M, 4 F	33.7 ± 6.2	4 (1), 3 (2), 1 (3), 0 (4)	41.5 ± 4.0
Severe-OSA	6	80.0 ± 35.1	46.4 ± 11.3	4 M, 2 F	44.4 ± 6.2	1 (1), 2 (2), 1 (3), 2 (4)	44.0 ± 3.7

This stratified dataset split preserves clinical relevance and demographic diversity, enabling robust training and unbiased final performance reporting. Such careful design is crucial in sleep apnea research, where severity overlaps and physiological variance can challenge model generalization.

6.2.3. Dataset Split and Cohort Stratification for K-Fold Cross Validation

Standard stratified K-fold cross-validation was unsuitable for this study because it could not maintain consistent stratification across both OSA severity and key anthropometric factors,

particularly within smaller severity classes. To address this, a custom stratified K-fold method was developed to balance folds across severity groups and major anthropometric thresholds (age (<50 vs. ≥ 50), BMI (<35 vs. ≥ 35), neck circumference (≤ 40 vs. > 40), sex (male vs. female), and Mallampati scores). This approach preserved the joint distribution of clinically relevant subgroups, reduced bias, and produced more representative and reliable training and testing sets, especially for underrepresented samples. Table 6-4 summarizes the anthropometric distribution across the folds.

Table 6-4 Demographic, anthropometric, and clinical characteristics of participants across OSA severity groups and cross-validation folds. Values are reported as mean $\pm$ standard deviation. Each row corresponds to one fold within a given severity group, illustrating the balance of subject distribution and key confounding variables across folds. AHI: apnea–hypopnea index (events/h), Sex (M/F): male/female, BMI: body mass index (kg/m²), MPS: Mallampati score (1/2/3/4), NC: neck circumference (cm).

Group (Fold)	Num. of Subjects	AHI	Age	Sex (M/F)	BMI	MPS	NC
Non-OSA (F1)	23	0.6 $\pm$ 0.8	44.9 $\pm$ 12.1	10/13	29.2 $\pm$ 4.7	12, 7, 1, 3	38.0 $\pm$ 4.7
Non-OSA (F2)	27	1.1 $\pm$ 1.3	45.7 $\pm$ 12.1	10/17	32.3 $\pm$ 7.6	12, 9, 4, 2	39.2 $\pm$ 4.3
Non-OSA (F3)	24	1.8 $\pm$ 1.3	50.0 $\pm$ 14.3	9/15	30.0 $\pm$ 5.8	17, 3, 1, 3	39.0 $\pm$ 2.9
Mild-OSA (F1)	16	8.7 $\pm$ 2.4	50.9 $\pm$ 12.5	10/6	36.6 $\pm$ 9.9	7, 1, 7, 1	43.5 $\pm$ 5.3
Mild-OSA (F2)	10	8.6 $\pm$ 2.2	51.3 $\pm$ 11.6	5/5	31.8 $\pm$ 8.4	6, 1, 2, 1	38.5 $\pm$ 8.8
Mild-OSA (F3)	9	8.8 $\pm$ 3.5	56.0 $\pm$ 10.3	6/3	33.0 $\pm$ 4.2	5, 4, 0, 0	43.4 $\pm$ 2.7
Moderate-OSA (F1)	16	19.9 $\pm$ 2.9	56.3 $\pm$ 10.8	10/6	34.6 $\pm$ 7.8	4, 6, 2, 4	42.3 $\pm$ 4.0
Moderate-OSA (F2)	18	22.8 $\pm$ 4.5	53.6 $\pm$ 9.7	13/5	31.8 $\pm$ 5.7	8, 5, 3, 2	42.7 $\pm$ 3.8
Moderate-OSA (F3)	16	21.6 $\pm$ 4.7	54.5 $\pm$ 13.9	13/3	35.2 $\pm$ 5.2	5, 6, 3, 2	43.6 $\pm$ 2.8
Severe-OSA (F1)	14	72.9 $\pm$ 35.0	45.5 $\pm$ 10.5	11/3	39.1 $\pm$ 10.1	2, 3, 5, 4	44.3 $\pm$ 4.2
Severe-OSA (F2)	16	66.6 $\pm$ 29.6	50.2 $\pm$ 11.0	13/3	40.1 $\pm$ 8.5	1, 6, 5, 4	46.6 $\pm$ 3.2
Severe-OSA (F3)	10	69.6 $\pm$ 39.1	51.9 $\pm$ 12.2	6/4	40.2 $\pm$ 7.5	2, 4, 4, 0	43.8 $\pm$ 3.5

6.2.4. Preprocessing and Feature Engineering

All audio recordings were preprocessed using a pipeline that included signal denoising, phase segmentation using LogVar, and fourth-order Butterworth bandpass filtering (75–3000 Hz), as detailed in our previous work [18]. The resulting inspiratory and expiratory segments were then processed using a comprehensive feature extraction and selection framework, also introduced in [18], designed to capture both linear and nonlinear signal characteristics. Extracted features included spectral descriptors (e.g., power spectral density, spectral entropy), higher-order spectral features (e.g., bispectrum, bicoherence), time-domain measures (e.g., fractal dimension, peak statistics), and time–frequency representations using 4-level Symlet4 wavelet and constant-Q transforms [18]. Additional cross-domain features such as recurrence quantification, Lyapunov exponents, and perturbation metrics (jitter, shimmer) were included to provide a multidimensional view of breathing dynamics [18]. In parallel, Anthropometric data, including age, BMI, NC, and MPS, were processed using the imputation strategy proposed in [18]. Missing values were imputed using a stratified k-nearest neighbors (k-NN) approach with $k = 3$, with subjects grouped by OSA severity (non-OSA, mild, moderate, severe) to preserve group-specific distributions and reduce bias [18].

Following feature extraction and imputation, adaptive normalization was performed as described in [18]. Four normalization techniques, z-score, min-max, mean-range, and robust scaling, were evaluated for each feature, and the method that maximizes mutual information with class labels was selected. This ensured consistent feature scaling while preserving discriminative patterns and enhancing robustness to outliers.

To further refine the dataset, a three-stage feature selection pipeline was applied, as proposed in [18]. Initially, univariate t -tests were used to remove non-significant features.

Remaining features were ranked using SHAP (Shapley Additive exPlanations) values from an ensemble classifier. Finally, recursive feature elimination (RFE) with a RUSBoost classifier was employed, guided by cross-validation, to select the most impactful features [19]. This pipeline ensured optimal model performance by retaining only stable and informative features. The total number of selected features for each model is 35: 30 selected using the feature selection pipeline and 5 from the anthropometric data.

6.2.5. Base Models

Six one-vs.-one binary base models were developed for pairwise OSA severity classification, as proposed in our previous framework [18] targeting the following comparisons:

- Non-OSA vs. Mild-OSA;
- Non-OSA vs. Moderate-OSA;
- Non-OSA vs. Severe-OSA;
- Mild-OSA vs. Moderate-OSA;
- Mild-OSA vs. Severe-OSA;
- Moderate-OSA vs. Severe-OSA.

Each model was optimized using Bayesian hyperparameter tuning across 20 candidate classifiers, including Support Vector Machine (SVM), logistic regression, random forests, neural networks, and RUSBoost. Bootstrap aggregation with out-of-bag (OOB) validation was employed to enhance model stability and generalization. Each base model outputs a posterior probability (or score) for its respective binary classification, as outlined in [18].

6.2.6. Meta-Model Construction

The 12 prediction scores (i.e., probability outputs) from the base models were used as input features for the meta-model. This stacked approach allows the meta-model to learn higher-level patterns from the decision landscape of individual base models, capturing interactions across all severity levels [20]. Multiple meta-classifiers were evaluated, including: SVM with radial and polynomial kernels, Random Forests, Logistic Regression, and Multi-layer Perceptron (MLP) [21]. The final meta-model was selected based on its performance on a blind test set (15% of the dataset) or the testing fold in the k-fold cross-validation, which was stratified to ensure class balance. The meta-models are trained for the three classes (Non-OSA, Mild-OSA, and Moderate/Severe-OSA) and four classes (Non-OSA, Mild-OSA, Moderate, and Severe-OSA). Each classification problem was treated independently, with the complete training methodology applied to every pairwise comparison. This ensures that base classifiers are explicitly optimized for the discriminative characteristics relevant to each subset of the data. Figure 6-2 shows the block diagram of the meta-model used.

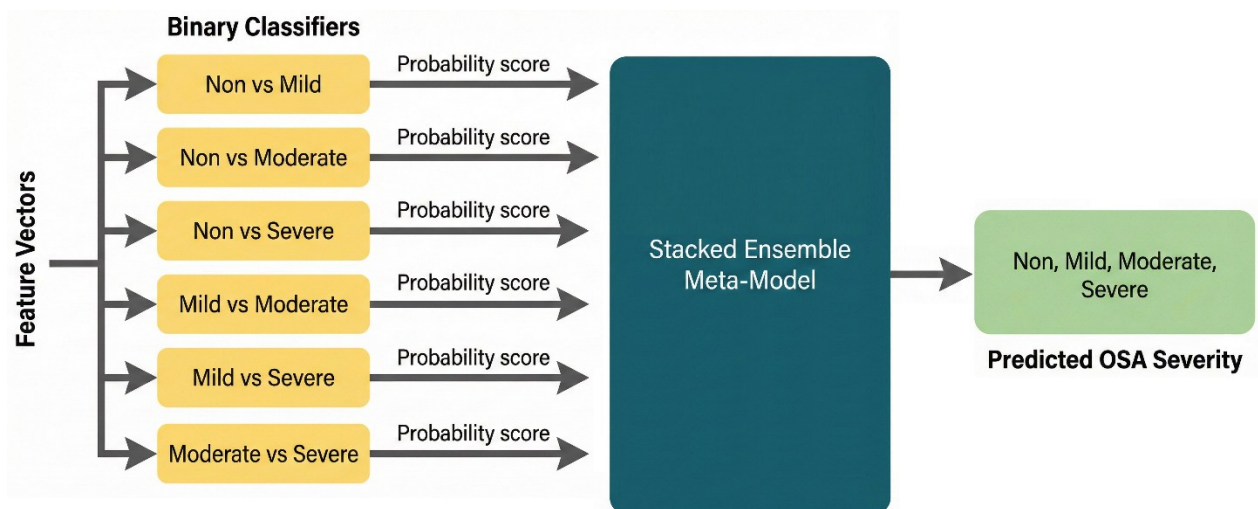


Fig. 6-2 Meta-model for multi-class OSA severity classification.

6.2.7. Conformal Prediction

To enhance the reliability of OSA severity classification, we incorporate a conformal prediction (CP) framework that augments standard probabilistic classifiers with principled uncertainty quantification. CP constructs prediction sets that, under mild assumptions (notably, the exchangeability of calibration and test data), offer guaranteed coverage as a critical feature in risk-sensitive domains like clinical diagnostics [22]. Also, this framework provides a transparent mechanism to communicate predictive uncertainty, which is crucial in clinical applications where misclassification can affect patient management.

The training set is partitioned into two subsets: an 80% model-fitting set, used for training and hyperparameter optimization, and a 20% calibration set, used exclusively to compute nonconformity scores. Out-of-bag (OOB) predictions from ensemble base classifiers (e.g., random forests or gradient boosting) are computed solely for internal diagnostics and stability assessment; they are not used for conformal calibration, model selection, or final evaluation. Figure 6-3 illustrates the conformal prediction workflow for OSA severity classification, including the training–calibration–test split and the role of OOB predictions.

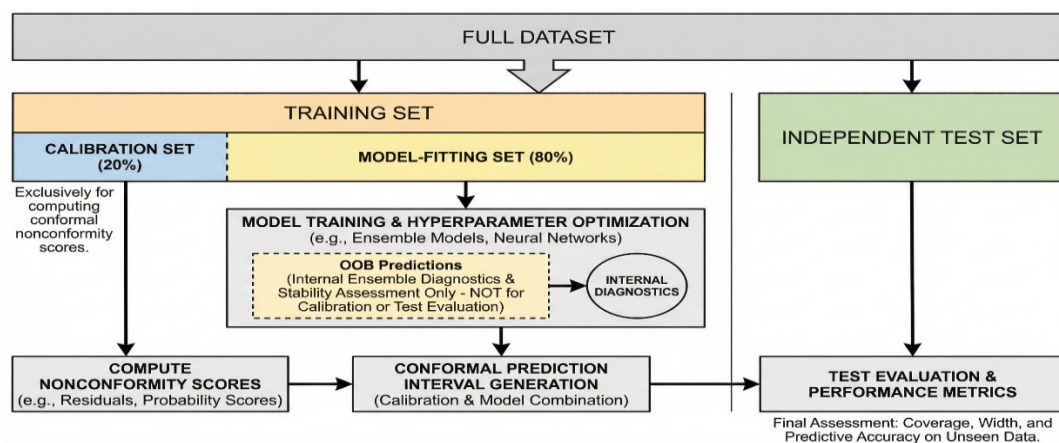


Fig. 6-3 Conformal prediction workflow for OSA severity classification. The training set is split into a calibration set (20%) for computing nonconformity scores and a model-fitting set (80%) for ensemble training. OOB predictions are used only for internal diagnostics and are not involved in calibration or testing, ensuring proper separation and reliable uncertainty estimation.

For a given significance level $\alpha \in (0,1)$, the conformal predictor generates a prediction $\mathcal{C}(x_{\text{test}}) \subseteq \mathcal{Y}$ for each test instance x_{test} [23], such that

$$P(Y_{\text{test}} \in \mathcal{C}(x_{\text{test}})) \geq 1 - \alpha \quad 6-1$$

This provides a high-confidence region where the true class label is expected to reside, without making strong distributional assumptions [22,23]. Then, to quantify the “strangeness” of potential label assignments, we define a nonconformity score as follows:

$$s(x, y) = 1 - \hat{p}(y | x) \quad 6-2$$

where $\hat{p}(y | x)$ denotes the estimated class probability. In our implementation, this probability is derived through ensemble averaging using out-of-bag (OOB) predictions from base classifiers (e.g., random forests or gradient boosting models), ensuring reliable and unbiased estimates [22,24]. After the initial train–test split or k-fold separation, the training data are further partitioned using a hold-out strategy to create an independent calibration set for conformal prediction. Specifically, 20% of the training samples are randomly held out using MATLAB (R2024b)’s `cvpartition` function (`HoldOut = 0.2`), denoted $\mathcal{D}_{\text{calib}} = \{(x_i, y_i)\}_{i=1}^n$. For each calibration sample, the nonconformity score $\alpha_i = s(x_i, y_i)$ is calculated [22,24]. The inclusion threshold was determined by computing the empirical quantile:

$$\hat{q} = \text{Quantile}_{1-\alpha}(\{\alpha_i\}_{i=1}^n \cup \{\infty\}) \quad 6-3$$

Then, for each *test* instance x_{test} The prediction set was defined as follows:

$$\mathcal{C}(x_{\text{test}}) = \{y \in \mathcal{Y} \mid s(x_{\text{test}}, y) \leq \hat{q}\} \quad 6-4$$

The conformal framework ensures that the prediction set contains the actual OSA severity class with probability at least $1 - \alpha$ provided that the calibration and test data are exchangeable [22,24]. The prediction set size is adaptively determined: smaller sets indicate higher model

confidence, which is particularly important when distinguishing between clinically adjacent severity categories, such as moderate and severe OSA.

The independent test set remains completely untouched until the final performance evaluation, preserving the finite-sample coverage guarantees of CP under the exchangeability assumption. High-confidence predictions typically yield a single class, while ambiguous inputs may include multiple labels (e.g., {moderate, severe}), effectively signaling diagnostic uncertainty [22,24]. The adaptive size of the prediction set reflects model confidence, which is particularly important when distinguishing clinically adjacent severity categories.

CP is model-agnostic and can be applied atop any probabilistic classifier, including those already optimized for the task. When combined with ensemble-based models, CP benefits from base-learner diversity and cross-model aggregation, mitigating individual model biases and improving overall reliability. In OSA severity detection, where incorrect classification may lead to inappropriate treatment, CP provides a principled mechanism for communicating uncertainty. For instance, borderline moderate/severe cases may yield a prediction set such as {moderate, severe}, transparently indicating ambiguity and encouraging further clinical evaluation. Figure 6-4 illustrates how CP is integrated with the proposed meta-model, highlighting its role in generating uncertainty-calibrated predictions that enhance interpretability and clinical safety.

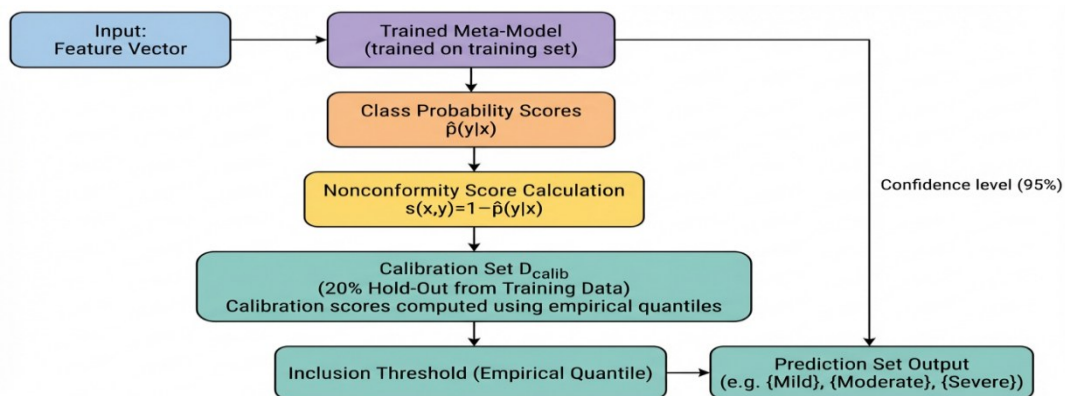


Fig. 6-4 Conformal prediction pipeline for uncertainty-calibrated OSA severity classification.

6.2.8. Model Evaluation

The evaluation framework combines traditional classification performance metrics with conformal prediction analysis to comprehensively assess model effectiveness and uncertainty quantification comprehensively. This dual-axis evaluation strategy addresses the reliability of predictive accuracy and confidence estimates.

6.2.8.1. Evaluation Protocol

Out-of-Bag (OOB) Validation: During ensemble training, each base learner was evaluated on the subset of training samples excluded from its bootstrap resample. These out-of-bag predictions were used to estimate performance without requiring a separate validation set. The resulting OOB metrics provide an unbiased estimate of generalization error.

Independent Test Evaluation: After model training, performance was assessed using the same metrics on a held-out test set. To account for randomness in training (e.g., bootstrap sampling or stochastic optimization), the evaluation was repeated over 25 independent trials with different random seeds.

6.2.8.2. Performance Metrics

Models were evaluated using three core metrics: accuracy, sensitivity, and specificity, measured on both out-of-bag (OOB) samples and independent test sets. These metrics were selected to ensure a balanced assessment of model performance, particularly in the presence of class imbalance [25].

Accuracy reflects the overall proportion of correctly classified instances out of the total number of samples. It provides a general measure of how well the model performs across all predictions, regardless of class [25].

Macro-averaged sensitivity represents the average actual positive rate across all classes. It is calculated by determining the sensitivity (true positives divided by the sum of true positives and false negatives) for each class individually, then averaging the results. This approach ensures that each class contributes equally to the overall sensitivity score, regardless of its frequency in the dataset [25].

Macro-averaged specificity captures the average actual negative rate across all classes. For each class, specificity is calculated by dividing the number of true negatives by the sum of true negatives and false positives. Like macro sensitivity, macro specificity averages these values across all classes to prevent dominance by majority classes. These macro-level metrics provide a more equitable evaluation of model performance in multi-class classification settings, particularly when the class distribution is unbalanced [25].

6.2.8.3. Conformal Prediction Metrics

Conformal prediction was integrated into the evaluation framework to evaluate the uncertainty quantification capability of the models. The two main metrics were coverage and average set size [22,24]. Coverage, the proportion of instances where the actual label is included in the prediction set:

$$\text{Coverage} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} I(y_i \in C(x_i)) \quad 6-5$$

Average Set Size, which reflects the efficiency of the prediction set:

$$\text{Avg Set Size} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |C(x_i)| \quad 6-6$$

These metrics provide insights into the trade-off between prediction confidence and set efficiency [22,24]. Class labels were numerically encoded to preserve categorical structure [22,24].

All performance metrics were averaged across 15 trials to mitigate the impact of stochastic variability. Stats averaged, including classifier rankings and 95% confidence intervals, were computed using non-parametric bootstrap resampling.

6.3. Results

The proposed meta-model was evaluated on a held-out test set to assess its generalization ability across the four-class OSA severity classification task. This evaluation followed a two-stage process: (1) assessing base model aggregation and meta-model performance during training using out-of-bag (OOB) validation, and (2) blind testing on unseen data. Section 4.1 presents the Training–Testing Meta-Model Performance Summary, while Section 4.2 reports the k-fold results across the three validation folds.

6.3.1. Training-Testing Meta-Model Performance Summary

For the three classes of OSA severity problem, as presented in Table 6-5 and illustrated in Figure 6-5, the meta-model demonstrated strong performance in the OOB evaluation for three classes using neural network model of two hidden layers with learning rate of 7.06×10^{-2} , achieving an accuracy of 91.1%, sensitivity of 91.6%, and specificity of 95.3%. When applied to the blind test set, the performance exhibited a moderate decline typical for ensemble frameworks with a final test accuracy of 76.7%, sensitivity of 77.7%, and specificity of 87.1%. This reduction can be attributed primarily to the limited size and non-diversity of the dataset, as well as the constrained feature set used for training. In smaller and less varied datasets, the model’s ability to generalize to unseen data can be restricted, particularly when specific feature patterns are underrepresented. Nevertheless, the meta-model maintained strong generalization capability, especially in preserving high specificity. This robustness is supported by the complementary strengths of the base

classifiers and the architecture of the stacking ensemble, which together mitigate the risks associated with dataset limitations.

To further assess prediction reliability, a conformal prediction framework was applied. The conformal model achieved full coverage (1.0), ensuring that each test instance’s prediction set contained the correct class, with an average prediction set size of 2, indicating high confidence in classification outcomes. These results highlight the meta-model’s ability not only to produce accurate predictions but also to quantify the certainty of its decisions, which is particularly important in clinical applications where decision reliability is critical.

The confusion matrix in Figure 6-5 provides further insight into class-level performance, revealing that the meta-model maintained strong predictive ability across all OSA severity categories, with misclassifications occurring primarily between adjacent severity levels, a common trend in medical classification tasks where class boundaries are gradual rather than discrete. This behavior suggests that the model effectively captures the underlying patterns distinguishing severely from non-severe cases, while borderline cases remain the most challenging to classify.

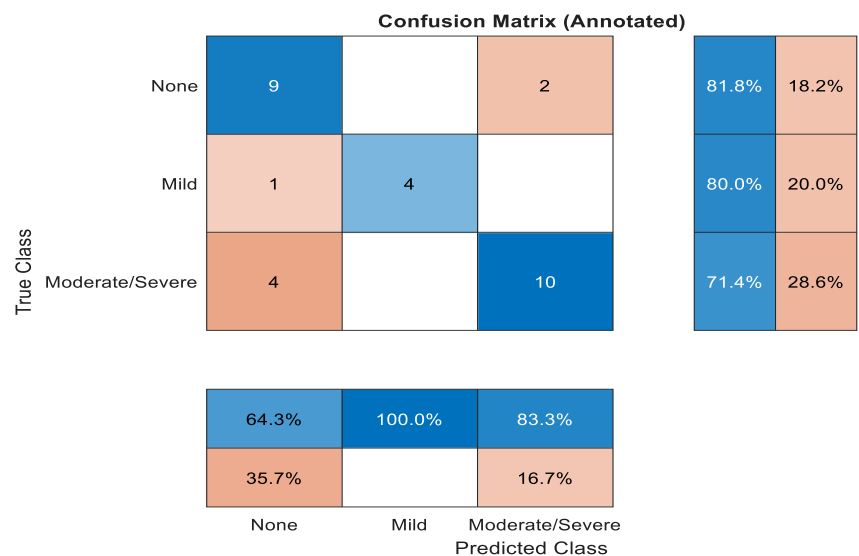


Fig. 6-5 Confusion matrix for three classes of OSA severity classification. Blue indicates correctly classified samples, red indicates misclassifications, with color intensity proportional to the number of samples.

Table 6-5 Results for the Meta-Model for three classes (mean $\pm$ 95% CI).

Metric	Accuracy		Sensitivity	Specificity
Meta-Model Model (Neural Net)	OOB Results			
	91.1 $\pm$ 1.8		91.6 $\pm$ 1.7	95.3 $\pm$ 1.5
	Test Results			
	76.7 $\pm$ 2.6		77.7 $\pm$ 2.5	87.1 $\pm$ 2.3
Conformal Results	Coverage	1	Avg Set Size	2

The four classes of OSA severity problem, as shown in Table 6-6 and illustrated in Figure 6-6, the meta-models demonstrate a strong performance during the out-of-bag (OOB) evaluation for four classes using neural network model of three hidden layers with learning rate of 5.14-3, with an accuracy of 88.2%, sensitivity of 91.6%, and specificity of 88.2%. On the independent test set, accuracy decreased to 76.7%, while sensitivity and specificity were 75% and 92%, respectively. This decline reflects the limited size and diversity of the dataset, alongside a constrained feature set. Smaller, less varied datasets can restrict the model's ability to generalize, particularly when specific feature patterns are underrepresented.

Despite these factors, the model preserved robust performance as a proof-of-concept, particularly in terms of specificity, supported by the complementary strengths of the base classifiers and the stacking ensemble framework. Prediction confidence was further assessed using a conformal prediction method, which achieved full coverage ensuring that every test instance's prediction set contained the true class and an average prediction set size of 2. These results provide preliminary evidence of meaningful certainty in the model's outputs, an important consideration for future clinical translation.

Figure 6-6 confusion matrix provides insight into the model's performance across all four OSA severity levels. Misclassifications mainly occurred between adjacent severity categories, consistent with the gradual nature of medical condition boundaries. This trend indicates that the

model effectively distinguishes clear-cut cases, while borderline instances remain more difficult to classify accurately.

Table 6-6 Results for the Meta-Model for four classes (mean $\pm$ 95% CI).

Metric	Accuracy		Sensitivity	Specificity
Meta-Model Model (Neural Net)	OOB Results			
	88.2 $\pm$ 2.0		91.6 $\pm$ 1.9	88.2 $\pm$ 2.0
	Test Results			
	76.7 $\pm$ 2.7		75 $\pm$ 2.6	92 $\pm$ 2.1
Conformal Results	Coverage	1	Avg Set Size	2

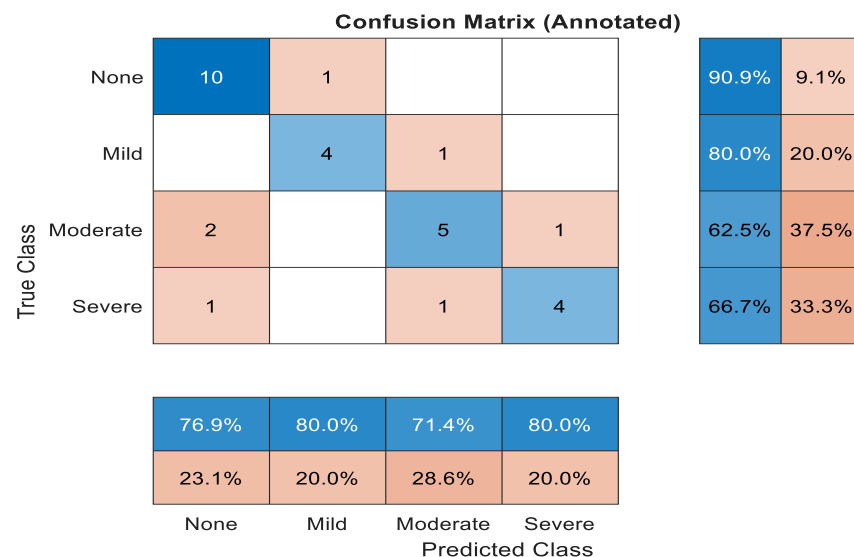


Fig. 6-6 Confusion matrix for four classes of OSA severity classification. Blue indicates correctly classified samples, red indicates misclassifications, with color intensity proportional to the number of samples.

To improve the interpretability and reliability of classification outputs, we implemented conformal prediction, which generates prediction sets rather than single-label outputs. Figures 6-7 and 6-8 present the conformal prediction map for the 30 test subjects for the three and four classes, respectively, illustrating which severity classes were included in each prediction set at a 95% confidence level. Each row corresponds to a subject, and each column corresponds to a possible class label (Non, Mild, Moderate/Severe) for three classes and (Non, Mild, Moderate, Severe) for four classes. A blue cell indicates that the corresponding class was included in the prediction set. For highly confident predictions, the set typically included a single class, whereas

uncertain predictions encompassed two or more labels. This visualization confirms that while full coverage (100%) was achieved, the average set size was 2 for three classes and 4 for four classes, indicating increased ambiguity in some cases, particularly between adjacent severity levels. The conformal framework provided interpretable uncertainty quantification, aligning well with the observed performance drop in blind testing and offering a mechanism for flagging diagnostically ambiguous cases for further review.

Table 6-7 provides a breakdown of the classification performance across the individual OSA severity groups in the test set. The Non-OSA class achieved a balanced profile with 81.8% sensitivity and 73.7% specificity, indicating strong recognition of healthy individuals with minimal false positives.

The Mild-OSA class was classified with 100% specificity and 80.0% sensitivity, suggesting the meta-model was highly conservative when assigning this class, minimizing false positives. However, Mild cases were the most frequently confused group in earlier base models, so this performance reflects a significant improvement in class separation.

The Moderate/Severe OSA group, which had been combined due to overlapping acoustic characteristics, achieved 71.4% sensitivity and 87.5% specificity. This class presented the most significant challenge for precise sensitivity, likely due to intra-group variability and class imbalance.

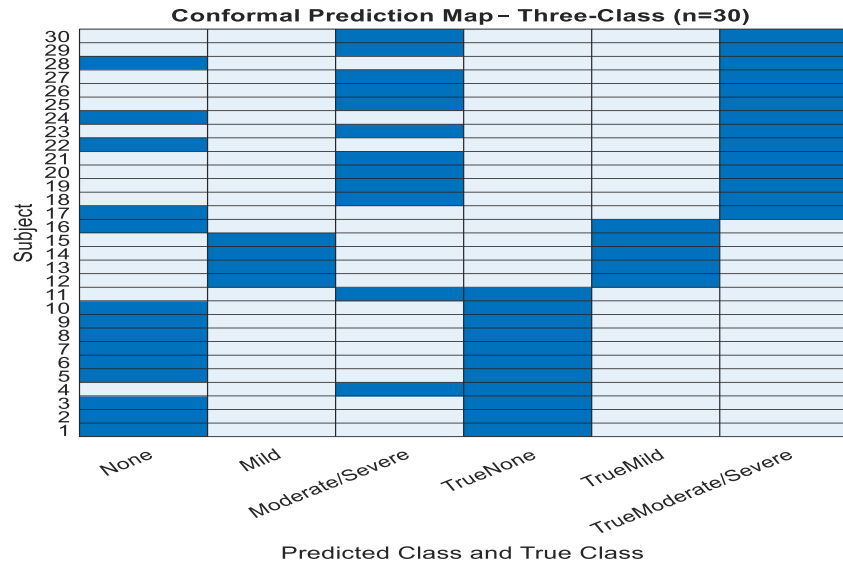


Fig. 6-7 Three-class testing conformal prediction set map.

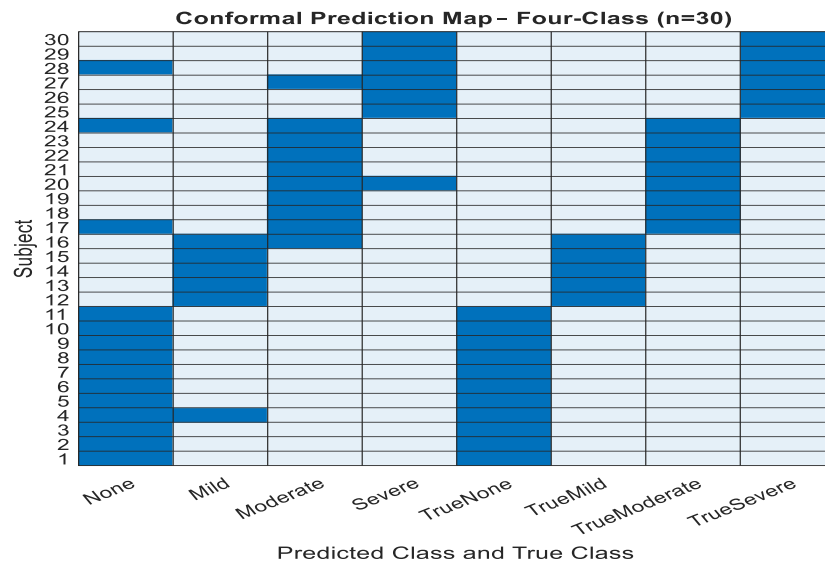


Fig. 6-8 Four-class testing conformal prediction set map.

The ROC curves for the three classes are shown in Figure 6-9. The Non-OSA group achieved an AUC of 0.76, the Mild-OSA group achieved the highest AUC of 0.93, and the Moderate/Severe-OSA group achieved an AUC of 0.78. These values are consistent with the per-class performance metrics, highlighting the strong separability of the Mild-OSA class and the relatively greater overlap between Moderate/Severe and other classes.

Table 6-8 summarizes the classification performance across the four individual OSA severity groups in the test set. The Non-OSA class achieved 91% sensitivity and 84% specificity, demonstrating effective identification of healthy individuals, with a moderate false-positive rate due to residual overlap with Mild OSA patterns.

The Mild OSA class was classified with 80.0% sensitivity and 96% specificity, indicating that the meta-model adopted a conservative strategy when assigning this class, resulting in very few false positives. This represents a substantial improvement, as Mild cases were previously the most frequently confused category in the base classifiers.

For the Moderate OSA class, the model reached 63.0% sensitivity and 91% specificity, suggesting a cautious approach with a tendency to avoid false alarms but at the cost of missing some actual cases reflecting a conservative decision boundary that prioritizes specificity over sensitivity. This may be attributed to the class's acoustic overlap with adjacent severity levels.

Table 6-7 Results for the Meta-Model for three classes.

Metric	Accuracy	Sensitivity	Specificity
Non-OSA	76.7	81.8	73.7
Mild-OSA		80.0	100
Moderate/Severe-OSA		71.4	87.5

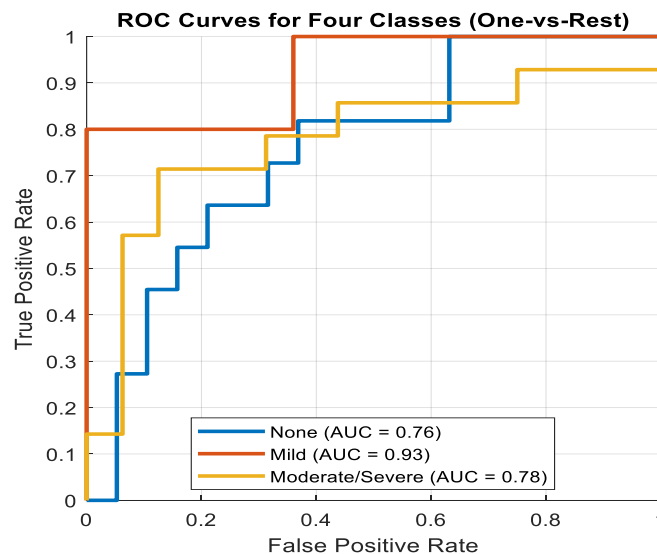


Fig. 6-9 ROC curves and AUC for the three-class testing set.

The Severe OSA class achieved 68% sensitivity and 96% specificity, reflecting the model's strong ability to confidently identify high-risk individuals when predicted, while maintaining a low false-positive rate. Although sensitivity remains moderate, the high specificity is clinically valuable for prioritizing patients requiring urgent intervention.

The ROC curves for the three classes are shown in Figure 6-10. The Non-OSA group achieved an AUC of 0.80, the Mild-OSA group achieved the highest AUC of 0.93, the Moderate-OSA group achieved an AUC of 0.72, and the Severe-OSA group achieved an AUC of 0.85. These values are consistent with the per-class performance metrics, highlighting the strong separability of the Mild-OSA class and the relatively greater overlap between Moderate/Severe and other classes.

Table 6-8 Results for the Meta-Model for four classes.

Metric	Accuracy	Sensitivity	Specificity
Non-OSA	76.7	91	84
Mild-OSA		80	96
Moderate-OSA		63	91
Severe-OSA		68	96

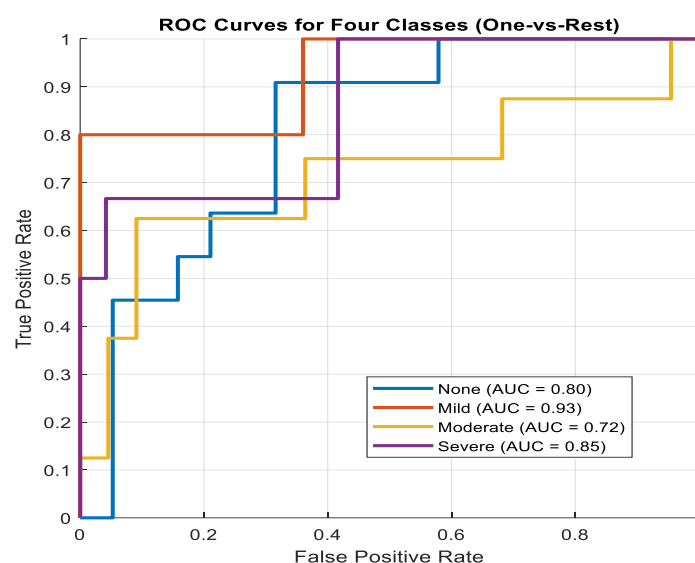


Fig. 6-10 ROC curves and AUC for the four-class testing set.

6.3.2. K-Fold Meta-Model Results

The selected base classifiers consistently achieved strong results across 3-fold cross-validation. The accuracies remained highly above 75%, reflecting good generalization and stable performance across folds. All Meta-models use a neural network with three hidden layers. Sensitivity and specificity were also well-balanced, indicating that the models maintained an appropriate trade-off between correctly identifying positive cases and limiting false alarms. Table 6-9 summarizes the 3-fold training performance metrics, and Table 6-10 summarizes the 3-fold test performance metrics; both tables align closely with the trends reported in the earlier training–testing evaluations.

Table 6-9 Train 3-fold cross-validation results of experimental models.

Experiment	Fold	Accuracy	Sensitivity	Specificity
Three Classes	1	82.2	83.1	80.7
	2	87.9	87.8	88
	3	86.0	84.3	86.3
Mean $\pm$ Standard Deviation		85.4 $\pm$ 2.9	85.1 $\pm$ 2.4	85 $\pm$ 3.8
Conformal Results	Coverage	1	Avg Set Size	2
Four Classes	1	87.4	90.8	87.9
	2	89.1	92.3	88.4
	3	88.2	91.7	88.3
Mean $\pm$ Standard Deviation		88.2 $\pm$ 0.9	91.6 $\pm$ 0.8	88.2 $\pm$ 0.3
Conformal Results	Coverage	1	Avg Set Size	2

Table 6-10 Test 3-fold cross-validation results of experimental models.

Experiment	Fold	Accuracy	Sensitivity	Specificity
Three Classes	1	78.7	80.4	76.9
	2	78.8	79.2	78.2
	3	75.8	76.2	74.1
Mean $\pm$ Standard Deviation		77.8 $\pm$ 1.7	78.6 $\pm$ 2.2	76.4 $\pm$ 2.1
Conformal Results	Coverage	1	Avg Set Size	2
Four Classes	1	77.1	75.8	91.0
	2	75.6	74.4	93.2
	3	77.4	74.8	91.9
Mean $\pm$ Standard Deviation		76.7 $\pm$ 1.0	75 $\pm$ 0.7	92 $\pm$ 1.1
Conformal Results	Coverage	1	Avg Set Size	2

To assess whether the proposed framework is limited by data availability, we conducted a learning-curve analysis by training the meta-model on progressively larger fractions of the training set (from 20% to 100%) while evaluating performance on a fixed held-out test set. As shown in Figure 6-11, test accuracy increased monotonically with training size for both the three-class and four-class configurations, without clear saturation. This trend suggests that current performance is primarily constrained by dataset size rather than model capacity, and that further gains are expected with larger, more diverse training cohorts.

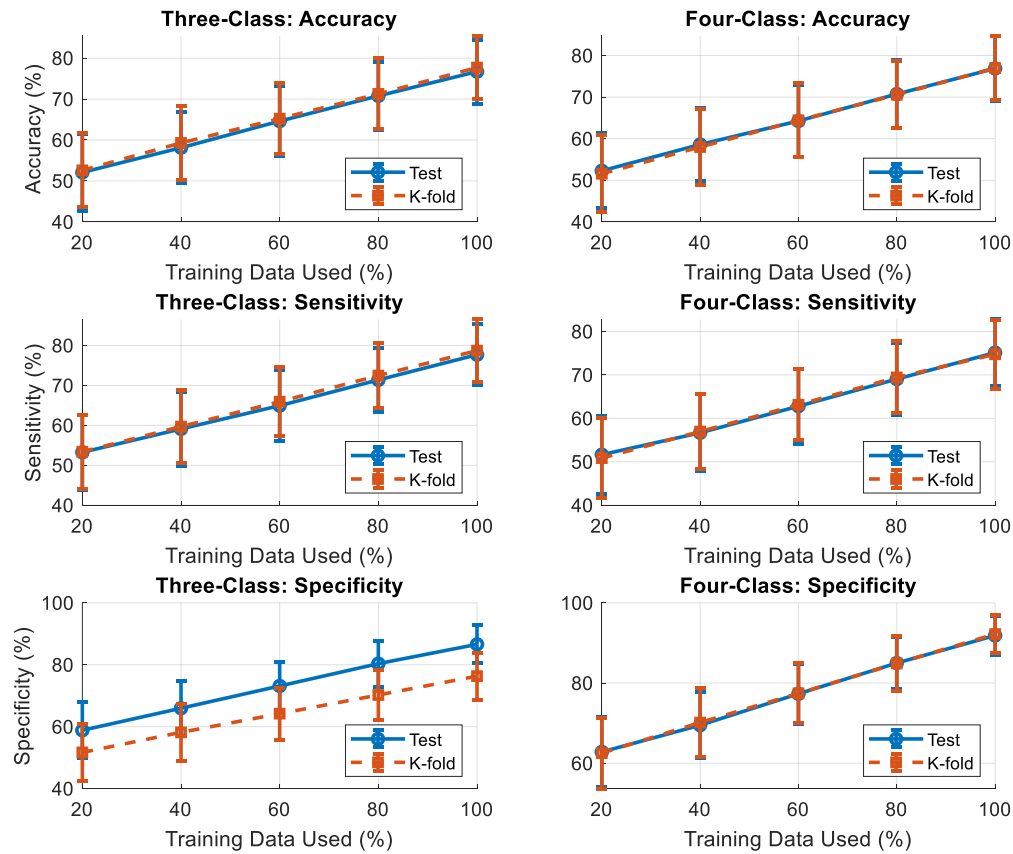


Fig. 6-11 Learning curves for accuracy (top), sensitivity (middle), and specificity (bottom) for the three-class (left) and four-class (right) OSA severity models. Solid lines denote independent test performance and dashed lines denote k-fold cross-validation. All metrics improve monotonically with increasing training data and show no saturation, indicating a data-limited regime and consistent generalization across evaluation strategies.

Overall, the 3-fold cross-validation results confirm the reliability and robustness of the proposed approach across both the three-class and four-class classification frameworks. The OOB and test metrics demonstrate consistent performance, with strong sensitivity and specificity values that indicate reliable class discrimination and controlled false-positive rates. Table 6-11 provides a consolidated comparison of OOB, independent test, and k-fold cross-validation performance. Although OOB estimates are consistently optimistic, the close agreement between test and cross-validation results indicates stable generalization with minimal evaluation bias. Together with the conformal prediction coverage and set-size results, these findings demonstrate consistent performance across folds and robustness when transitioning from OOB estimates to independent testing, supporting the potential of the proposed framework for practical OSA severity assessment.

Table 6-11 Summary comparison of OOB, independent test, and 3-fold cross-validation mean performance for the proposed meta-model.

Classes	Evaluation	Accuracy (%)	Sensitivity (%)	Specificity (%)
Three	OOB	91.1 $\pm$ 1.8	91.6 $\pm$ 1.7	95.3 $\pm$ 1.5
Three	Test	76.7 $\pm$ 2.6	77.7 $\pm$ 2.5	87.1 $\pm$ 2.3
Three	3-fold Train CV (Mean $\pm$ SD)	85.4 $\pm$ 2.9	85.1 $\pm$ 2.4	85.0 $\pm$ 3.8
Three	3-fold Test CV (Mean $\pm$ SD)	77.8 $\pm$ 1.7	78.6 $\pm$ 2.2	76.4 $\pm$ 2.1
Four	OOB	88.2 $\pm$ 2.0	91.6 $\pm$ 1.9	88.2 $\pm$ 2.0
Four	Test	76.7 $\pm$ 2.7	75.0 $\pm$ 2.6	92.0 $\pm$ 2.1
Four	3-fold Train CV (Mean $\pm$ SD)	88.2 $\pm$ 0.9	91.6 $\pm$ 0.8	88.2 $\pm$ 0.3
Four	3-fold Test CV (Mean $\pm$ SD)	76.7 $\pm$ 1.0	75.0 $\pm$ 0.7	92.0 $\pm$ 1.1

6.3.3. Statistical Significance Analysis

To evaluate whether the proposed meta-model significantly outperformed the six best base classifiers, we conducted paired statistical tests using the predictions and scores from the respective best-performing models. Rather than aggregating results across multiple trials, the tests were performed directly on the final model outputs obtained from the held-out test set. This approach allows for a direct comparison of model performance on the same set of samples. All

statistical evaluations were based on the model predictions generated across the 3-fold cross-validation test sets, ensuring that the comparisons reflect consistent performance across multiple data splits rather than a single held-out partition.

Pairwise comparisons between the meta-model and each base classifier were conducted using both a parametric paired t-test and a non-parametric Wilcoxon signed-rank test applied to per-sample accuracy outcomes. These complementary tests assess the statistical significance of differences in continuous performance measures while accounting for possible violations of normality assumptions [26]. Additionally, McNemar’s test was employed to evaluate differences in binary classification decisions on a per-sample basis, providing insight into whether the models made significantly different classification errors [27]. These statistical comparisons were performed using the combined per-sample predictions obtained across all three folds of the cross-validation procedure.

To quantify the magnitude of observed differences, effect sizes were computed alongside p-values. Cohen’s d was calculated for the paired t-test results, and Cliff’s δ was derived from the Wilcoxon test outcomes. Furthermore, Bonferroni-adjusted p-values were used to control for the risk of false positives due to multiple comparisons across the six base classifiers, setting the adjusted significance threshold at $\alpha = 0.00833$. The results demonstrate that, in all six comparisons, the meta-model significantly outperformed each base classifier, with p-values well below the adjusted threshold. Effect sizes were consistently large (Cohen’s d greater than 0.65 and Cliff’s δ at or above 0.36), indicating that the improvements are not only statistically significant but also practically meaningful. Table 6-12 shows the results of the statistical comparison.

Table 6-12 Statistical significance analysis comparing the proposed meta-model to the base classifier.

Comparison	Paired t -test (p)	Wilcoxon (p)	McNemar (p)	Mean Δ Accuracy	Cohens' (d)	Cliffs (δ)	Bonferroni (p)
Meta-model vs. Base-model 1	1.10×10^{-3}	3.40×10^{-3}	3.40×10^{-3}	0.3667	0.6594	0.3667	1.10×10^{-3}
Meta-model vs. Base-model 2	3.07×10^{-4}	9.77×10^{-4}	9.77×10^{-4}	0.3667	0.7481	0.3667	3.07×10^{-4}
Meta-model vs. Base-model 3	1.09×10^{-4}	4.65×10^{-4}	5.19×10^{-4}	0.4667	0.8168	0.4667	1.09×10^{-4}
Meta-model vs. Base-model 4	8.70×10^{-6}	6.10×10^{-5}	6.10×10^{-5}	0.5	0.9832	0.5	8.70×10^{-6}
Meta-model vs. Base-model 5	1.85×10^{-5}	1.62×10^{-4}	1.45×10^{-4}	0.5333	0.9335	0.5333	1.85×10^{-5}
Meta-model vs. Base-model 6	4.61×10^{-5}	2.75×10^{-4}	2.75×10^{-4}	0.5	0.8736	0.5	4.61×10^{-5}
Bonferroni-corrected alpha	0.00833						

A p -value < 0.05 indicates a statistically significant difference.

6.4. Discussion

This study demonstrates the feasibility of multi-class OSA severity screening during wakefulness using tracheal breathing sounds and anthropometric features. The proposed meta-model achieved approximately 77% test accuracy across three- and four-class configurations, with high specificity and calibrated uncertainty estimates, supporting its potential role as a rapid screening and triage tool. The integration of meta-learning and conformal prediction contributes to the model's robustness, interpretability, and potential for real-world clinical deployment. This discussion synthesizes the core findings, contextualizes them within physiological relevance, and evaluates implications for clinical applications.

6.4.1. Meta-Model Performance for Training-Testing

The meta-model was evaluated through a structured two-stage approach: out-of-bag (OOB) validation and blind testing on unseen data. In the three-class classification task, a moderate performance decline (from 91% to 76.7%) was observed during blind testing. This is commonly seen in ensemble learning due to overfitting or data variability [28,29]. However, the accuracy still remained within acceptable margin. Misclassifications primarily occurred between the Mild and Moderate classes, which share acoustic and physiological features, such as partial upper airway obstruction and variable inspiratory effort [15].

In the four-class classification, which includes the Severe category, the task became more challenging due to increased variability and finer-grained distinctions between classes. Misclassifications were most frequent between the Mild and Moderate groups, reflecting their overlapping acoustic and physiological characteristics, such as partial upper airway obstruction and variable inspiratory effort [28,29]. Despite the increased complexity, the models maintained meaningful discrimination across all four severity levels, highlighting their potential clinical utility. Comparatively, the three-class scenario, which combined Mild and Moderate cases into a single group, showed slightly higher overall accuracy, confirming that broader class definitions reduce ambiguity but may obscure clinically relevant distinctions.

Conformal prediction was incorporated to provide uncertainty quantification alongside severity stratification. The conformal models achieved the targeted statistical coverage level in both three-class and four-class configurations, with an average prediction set size of approximately two classes. In some instances, the resulting prediction sets contained multiple severity classes, indicating increased uncertainty in borderline samples rather than definitive clinical classification. This behavior is consistent with the theoretical properties of conformal prediction and supports

transparent, uncertainty-aware screening. Rather than implying diagnostic certainty, these results demonstrate the feasibility of augmenting wakefulness-based OSA screening with calibrated confidence information to support cautious clinical triage. Visualizations of conformal prediction maps further illustrate class-dependent uncertainty patterns and regions of prediction ambiguity [30,31].

The use of stacked ensemble architecture allowed the model to integrate complementary strengths of diverse classifiers while reducing correlated error [28,32]. This architecture, combined with conformal prediction, enabled adaptive error correction and probabilistic confidence assignment, which are critical for medical decision support systems [33]. This work extends our previous binary-classification framework [18]. By integrating it into a unified, multi-class prediction model with conformal prediction to quantify model confidence, making it more suitable for clinical decision support. The stacking framework helped mitigate error propagation from base models, and the integration of conformal prediction further strengthened the model by providing uncertainty estimates for individual predictions, an essential attribute in clinical risk stratification. The stacking structure of the meta-model combining diverse, specialized base classifiers likely contributed to this robustness by reducing correlated errors and promoting generalization.

Detailed per-class analysis provides further insights into the model's behavior. In the three-class task, the Non-OSA group achieved 81.8% sensitivity and 73.7% specificity, demonstrating strong identification of healthy individuals. The Mild-OSA group was detected with 80.0% sensitivity, 100% specificity, and the highest AUC of 0.93, indicating a conservative classification pattern that minimized false positives. This marks a significant improvement over earlier models, which frequently misclassified Mild cases [17,34]. The Moderate/Severe OSA group achieved

71.4% sensitivity and 87.5% specificity, reflecting residual difficulty in discriminating more severe cases due to overlapping acoustic characteristics and within-class variability [15,35].

In the four-class scenario, the Non-OSA and Mild OSA groups achieved sensitivities of 91% and 80%, with corresponding specificities of 84% and 96%, respectively. In contrast, sensitivity decreased for Moderate OSA (63%) and Severe OSA (68%), although both classes maintained high specificity values of 91% and 96%. These results indicate that the model consistently limits false-positive predictions across all severity levels, while reduced sensitivity in the Moderate and Severe classes likely reflects inter-class acoustic similarity and the impact of imbalanced data distribution [9,34,36].

Taken together, the three-class framework yielded more balanced sensitivity across groups, while the four-class division provided finer granularity but at the cost of reduced sensitivity for moderate and severe OSA. The preferred scheme may therefore depend on clinical priorities: if the goal is to minimize missed diagnoses, particularly in preoperative screening, the three-class grouping may be more practical. Conversely, when detailed stratification is required for treatment planning, the four-class scheme is more informative. It is also likely that with a larger and more balanced dataset, the performance gap between the two approaches would diminish, particularly for moderate and severe cases where class overlap and limited sample size currently constrain sensitivity.

While the proposed framework demonstrates high specificity and robust performance for wakefulness-based screening, reduced sensitivity for moderate and severe OSA indicates that it should not be used to exclude clinically significant disease. Nevertheless, the models sustained meaningful discrimination across all four severity levels, supporting their utility for rapid screening and triage rather than definitive diagnosis. Comparatively, the three-class configuration

achieved slightly higher test accuracy, suggesting that broader class definitions reduce ambiguity, albeit at the expense of finer-grained clinical detail. Accordingly, positive or uncertain cases should be referred to for confirmatory overnight testing.

6.4.2. Meta-Model Performance for k-Fold

The meta-model was evaluated using a two-stage procedure consisting of out-of-bag (OOB) validation and blind testing on unseen data, supported by 3-fold cross-validation to ensure robust performance assessment. In the three-class classification task, the OOB accuracy remained high across folds, ranging from 82.2% to 87.9%, with balanced sensitivity and specificity levels (Table 9). During blind testing, accuracy decreased to values between 75.8% and 78.8% (Table 10), reflecting a moderate performance reduction (from 91% previously to 76.7%). This level of decline is commonly reported in ensemble and deep learning studies due to increased data variability and potential overfitting effects [20,21]. Despite this reduction, performance remained within acceptable ranges, and most misclassifications occurred between the Mild and Moderate groups—two categories known to share overlapping acoustic and physiological traits such as partial airway obstruction and variable inspiratory effort [9].

In the four-class configuration, which further separates Severe cases, the task understandably became more complex. OOB accuracies remained strong and consistent across folds (87.4–89.1%), again supported by well-balanced sensitivity and specificity values (Table 9). Blind testing results showed lower accuracies (75.6–77.4%), reflecting increased classification difficulty arising from finer granularity and wider intra-class variability (Table 10). Misclassifications were most frequent between the Mild and Moderate groups, reinforcing the physiological overlap between these categories [28,29].

To reinforce predictive reliability, conformal prediction was used to quantify per-sample uncertainty. Across all folds and in both class configurations, conformal prediction achieved full coverage (1.0) with an average prediction set size of 2, demonstrating the model’s high confidence and consistent uncertainty calibration. These outcomes were further supported by conformal prediction visualization maps, which helped identify class-dependent uncertainty structures and prediction boundaries [30,31].

The stacked ensemble architecture played a central role in achieving these results. By integrating the complementary strengths of diverse base classifiers and reducing correlated error, the ensemble improved both generalization and resilience across cross-validation folds [28,32]. The combined use of stacking and conformal prediction improved adaptive error correction and supported probabilistic confidence estimation—key requirements for clinical decision support [33]. Building on earlier binary classification work [18], this unified, multi-class framework offers enhanced explanatory value through uncertainty quantification and improved granularity in OSA severity assessment. The consistent 3-fold results shown in Tables 9 and 10 underscore the robustness of the approach and support its potential for real-world clinical deployment.

The reported 95% confidence intervals allow a more conservative and transparent assessment of the proposed meta-model’s performance. Across both three-class and four-class classification tasks, OOB results exhibit consistently narrow confidence intervals, indicating stable ensemble behavior and reliable internal generalization. As expected, confidence intervals for the independent test results are wider due to the limited sample size; however, they remain well centered around the mean values and show no excessive dispersion, suggesting that the observed performance is not driven by isolated data splits. From a clinical perspective, the relatively tight confidence intervals for sensitivity are particularly important, as they indicate consistent

identification of affected cases across repeated evaluations, reducing the risk of unstable diagnostic behavior. The overlap between OOB and test confidence intervals further supports the absence of substantial overfitting and demonstrates that the meta-model generalizes robustly to unseen data. Overall, the confidence interval analysis strengthens the reliability of the reported results by explicitly quantifying uncertainty while confirming the stability of the proposed approach.

6.4.3. Statistical Significance of Meta-Model Improvements

The statistical analyses conducted provide strong evidence that the meta-model offers a meaningful improvement over the individual base classifiers. By comparing predictions directly on the same test samples and using a combination of parametric, non-parametric, and categorical error tests [26,27,37]. We ensured a thorough and reliable evaluation of performance differences. All statistical comparisons were based on the combined predictions obtained across the 3-fold cross-validation test sets, ensuring that the results reflect consistent performance across multiple data splits rather than a single evaluation set. The consistent statistical significance across all tests, along with large effect sizes, indicates that the observed accuracy gains are not only unlikely to be due to chance but also practically important.

These findings highlight the benefit of the meta-model's ensemble approach in effectively integrating diverse classifiers, mitigating individual model limitations, and improving overall predictive performance. The use of multiple complementary statistical evaluation methods further strengthens confidence in the robustness of the observed trends. Collectively, these results provide preliminary evidence supporting the meta-model's potential value for OSA severity classification and suggest advantages over relying on any single base classifier within this proof-of-concept framework.

6.4.4. Comparison to Prior TBS Studies

A quantitative comparison between the proposed meta-model framework and representative prior studies on tracheal breathing sound–based OSA assessment is provided. To ensure a fair and informative comparison, we focus on key methodological and performance dimensions that are consistently reported across studies: dataset size, task formulation (binary vs. multi-class), evaluation performance (accuracy/AUC where available), recording condition (awake vs. sleep), and the presence of explicit uncertainty modeling. Only the top six prior works with clearly reported quantitative results and comparable awake TBS settings are highlighted. Table 6-13 summarizes a quantitative comparison between the proposed meta-model framework and representative prior tracheal breathing sound–based OSA studies. Unlike prior studies that predominantly address binary OSA screening without uncertainty quantification, the proposed method uniquely supports multi-class OSA severity stratification and integrates conformal prediction to provide statistically valid uncertainty estimates.

Table 6-13 Quantitative comparison of the proposed meta-model framework with representative prior TBS–based OSA studies in terms of dataset size, task formulation, performance, recording condition, and uncertainty modeling.

Study	Dataset Size	Task	Performance Metrics	Uncertainty Modeling
[8]	109 (AHI < 15) 90 (AHI > 15)	Binary	Test: 81.4% accuracy, 80.9% sensitivity, 82.1% specificity:	No
[15]	61 (AHI < 5) 69 (AHI > 10)	Binary	Validation: 84.5% accuracy 88.2% sensitivity, 80.9% specificity	No
[17]	109 (AHI < 15) 90 (AHI > 15)	Binary	79.3% accuracy, 82.2% sensitivity, 75.8% specificity	No
[34]	109 (AHI < 15) 90 (AHI > 15)	Binary	RF outperformed LR by +3.5% accuracy (~82–85% range)	No
[38]	43 OSA 47 controls	Binary	76.5% accuracy, 55% sensitivity, 100% specificity	No

This Work	(74 non-OSA, 35 Mild-OSA, 90 Moderate/Severe-OSA)	Multi-class (3 classes)	Test: 76.7% accuracy, 77.7% sensitivity, 87.1% specificity 3-fold OOB: 82.2–87.9% accuracy, 83.1–87.8% sensitivity, 80.7–88.0% specificity 3-fold Test: 75.8–78.8% accuracy, 76.2–79.2% sensitivity, 74.1–78.2% specificity	Yes (Conformal Prediction)
This Work	(74 non-OSA, 35 Mild-OSA, 50 Moderate-OSA, and 40 Severe-OSA)	Multi-class (4 classes)	Test: 76.7% accuracy, 75% sensitivity, 92% specificity 3-fold OOB: 87.4–89.1% accuracy, 90.8–92.3% sensitivity, 87.9–88.4% specificity 3-fold Test: 75.6–77.4% accuracy, 74.4–75.8% sensitivity, 91.0–93.2% specificity	Yes (Conformal Prediction)

6.4.5. Clinical Utility and Limitations

The proposed OSA screening method offers several advantages over traditional overnight PSG testing. By utilizing short wakeful breathing recordings and easily obtainable anthropometric features, the model enables rapid screening, potentially under 10 min. This design is especially suitable for preoperative risk assessment, primary care triage, and telehealth settings where time and equipment are limited [4,7,39,40].

High specificity in detecting Non-OSA and Mild-OSA cases reduces unnecessary referrals for full PSG studies, thereby optimizing resource allocation. Meanwhile, sufficient sensitivity in identifying higher-risk individuals ensures critical cases are not overlooked. The model's confidence-calibrated outputs via conformal prediction enhance its interpretability for clinicians and support risk-based decision-making [30,31,41,42].

Nonetheless, limitations remain. Misclassification was most prominent between the Mild and Moderate OSA classes, consistent with the results observed in the three-class scenario and

highlighting the physiological and acoustic overlap between these groups. Extending the analysis to the four-class classification, which includes the Severe category, introduced additional complexity, with slightly lower accuracy due to finer-grained distinctions. Incorporating additional discriminative features such as those derived from deep learning models, breathing phase segmentation, or higher-resolution time–frequency representations could improve class separability [43–47]. Furthermore, as this study relied on a single dataset, external validation on independent cohorts is necessary to assess generalizability across clinical environments. To address this limitation, a new dataset is currently being collected under a separate acquisition protocol for future external validation. Until such validation is completed, the present findings should be interpreted as exploratory and proof-of-concept, rather than as evidence of clinical generalizability.

6.5. Conclusions

This study, for the first time, introduces a new, rapid, with acceptable accuracy meta-modeling framework for detecting and stratifying OSA severity into four clinically meaningful categories using TBS and anthropometric data recorded during wakefulness. The model achieved an average 76.7% test accuracy with high sensitivity and specificity across all severity levels. Leveraging stacked ensemble learning with robust base classifiers, the proposed approach offers a practical, PSG-free alternative for screening and early triage of OSA. A key contribution of this work is the development of a clinically motivated multi-class severity stratification framework tailored to wakefulness-based OSA assessment. The pairwise acoustic specialization strategy enables robust discrimination and aggregation across adjacent severity levels, effectively addressing the acoustic overlap inherent in progressive OSA severity. Furthermore, the integration of conformal prediction provides calibrated, sample-specific uncertainty estimates, directly

addressing safety and reliability concerns associated with clinical deployment. Training on stratified anthropometric subgroups further improves interpretability, a crucial consideration for future regulatory approval. Its sub-10 min acquisition time, coupled with offline processing, makes it suitable for deployment in preoperative clinics, dental practices, and at-home settings. Compared to our previous binary-classification framework, this work expands the capabilities to a four-class system with reliability estimation, enabling finer severity stratification. Importantly, the proposed framework demonstrated consistent performance across 3-fold cross-validation, with OOB and blind test accuracies remaining stable in both the three-class and four-class configurations, supporting the model's robustness and reducing the likelihood that the reported results are due to overfitting or dataset partitioning effects. Future work should focus on external validation using diverse populations and integrating additional physiological signals to improve generalizability and predictive performance further.

References

- [1] Rizzo, D.; Baltzan, M.; Sirpal, S.; Dosman, J.; Kaminska, M.; Chung, F. Prevalence and Regional Distribution of Obstructive Sleep Apnea in Canada: Analysis from the Canadian Longitudinal Study on Aging. *Can. J. Public Health* 2024, 115, 970–979. <https://doi.org/10.17269/s41997-024-00911-8>.
- [2] Faria, A.; Allen, A.H.; Fox, N.; Ayas, N.; Laher, I. The Public Health Burden of Obstructive Sleep Apnea. *Sleep. Sci.* 2021, 14, 257–265. <https://doi.org/10.5935/1984-0063.20200111>.
- [3] American Academy of Sleep Medicine. Hidden Health Crisis Costing America Billions: Underdiagnosing and Undertreating Obstructive Sleep Apnea Draining Healthcare System; 1st ed.; Frost & Sullivan : Darien, IL, USA, 2016.

- [4] Kushida, C.A.; Littner, M.R.; Morgenthaler, T.; Alessi, C.A.; Bailey, D.; Coleman, J.; Friedman, L.; Hirshkowitz, M.; Kapen, S.; Kramer, M.; et al. Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005. *Sleep* 2005, 28, 499–523. <https://doi.org/10.1093/sleep/28.4.499>.
- [5] Chung, F.; Liao, P.; Elsaid, H.; Islam, S.; Shapiro, C.M.; Sun, Y. Oxygen Desaturation Index from Nocturnal Oximetry. *Anesth. Analg.* 2012, 114, 993–1000. <https://doi.org/10.1213/ANE.0b013e318248f4f5>.
- [6] Hoang, N.H.; Liang, Z. Detection and Severity Classification of Sleep Apnea Using Continuous Wearable SpO₂ Signals: A Multi-Scale Feature Approach. *Sensors* 2025, 25, 1698. <https://doi.org/10.3390/s25061698>.
- [7] Corral-Peñafiel, J.; Pepin, J.-L.; Barbe, F. Ambulatory Monitoring in the Diagnosis and Management of Obstructive Sleep Apnoea Syndrome. *Eur. Respir. Rev.* 2013, 22, 312–324. <https://doi.org/10.1183/09059180.00004213>.
- [8] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea Using Anthropometric Information and Tracheal Breathing Sounds. *Sci. Rep.* 2019, 9, 1–12. <https://doi.org/10.1038/s41598-019-47998-5>.
- [9] Montazeri, A.; Giannouli, E.; Moussavi, Z. Assessment of Obstructive Sleep Apnea and Its Severity during Wakefulness. *Ann. Biomed. Eng.* 2012, 40, 916–924. <https://doi.org/10.1007/s10439-011-0456-5>.
- [10] Singh, M.; Liao, P.; Kobah, S.; Wijeyesundera, D.N.; Shapiro, C.; Chung, F. Proportion of Surgical Patients with Undiagnosed Obstructive Sleep Apnoea. *Br. J. Anaesth.* 2013, 110, 629–636. <https://doi.org/10.1093/bja/aes465>.

- [11] Chen, L.; Pivetta, B.; Nagappa, M.; Saripella, A.; Islam, S.; Englesakis, M.; Chung, F. Validation of the STOP-Bang Questionnaire for Screening of Obstructive Sleep Apnea in the General Population and Commercial Drivers: A Systematic Review and Meta-Analysis. *Sleep Breath.* 2021, 25, 1741–1751. <https://doi.org/10.1007/S11325-021-02299-Y>.
- [12] Behar, J.A.; Palmius, N.; Daly, J.; Li, Q.; Rizzatti, F.G.; Bittencourt, L.; Clifford, G.D. Sleep Questionnaires in Screening for Obstructive Sleep Apnoea. *Comput. Cardiol.* (2010) 2017, 44, 1–4. <https://doi.org/10.22489/CinC.2017.233-188>.
- [13] El-Sayed, I.H. Comparison of Four Sleep Questionnaires for Screening Obstructive Sleep Apnea. *Egypt. J. Chest Dis. Tuberc.* 2012, 61, 433–441. <https://doi.org/10.1016/j.ejcdt.2012.07.003>.
- [14] Pataka, A.; Daskalopoulou, E.; Kalamaras, G.; Fekete Passa, K.; Argyropoulou, P. Evaluation of Five Different Questionnaires for Assessing Sleep Apnea Syndrome in a Sleep Clinic. *Sleep. Med.* 2014, 15, 776–781. <https://doi.org/10.1016/J.SLEEP.2014.03.012>.
- [15] Elwali, A.; Moussavi, Z. Obstructive Sleep Apnea Screening and Airway Structure Characterization During Wakefulness Using Tracheal Breathing Sounds. *Ann. Biomed. Eng.* 2016, 45, 839–850 <https://doi.org/10.1007/s10439-016-1720-5>.
- [16] Simply, R.M.; Dafna, E.; Zigel, Y. Diagnosis of Obstructive Sleep Apnea Using Speech Signals From Awake Subjects. *IEEE J. Sel. Top. Signal Process* 2020, 14, 251–260. <https://doi.org/10.1109/JSTSP.2019.2955019>.
- [17] Hajipour, F.; Jozani, M.J.; Elwali, A.; Moussavi, Z. Regularized Logistic Regression for Obstructive Sleep Apnea Screening during Wakefulness Using Daytime Tracheal Breathing Sounds and Anthropometric Information. *Med. Biol. Eng. Comput.* 2019, 57, 2641–2655. <https://doi.org/10.1007/S11517-019-02052-4>.

- [18] Alqudah, A.M.; Moussavi, Z. Assessing Obstructive Sleep Apnea Severity During Wakefulness via Tracheal Breathing Sound Analysis. *Sensors* 2025, 25, 6280. <https://doi.org/10.3390/s25206280>.
- [19] Brusa, E.; Cibrario, L.; Delprete, C.; Di Maggio, L.G. Explainable AI for Machine Fault Diagnosis: Understanding Features' Contribution in Machine Learning Models for Industrial Condition Monitoring. *Appl. Sci.* 2038, 2023, 13, 2038. <https://doi.org/10.3390/APP13042038>.
- [20] Wolpert, D.H. Stacked Generalization. *Neural Netw.* 1992, 5, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
- [21] Leblanc, M.; Tibshirani, R. Combining Estimates in Regression and Classification. *J. Am. Stat. Assoc.* 1996, 91, 1641–1650. <https://doi.org/10.1080/01621459.1996.10476733>.
- [22] Shafer, G.; Vovk, V. A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.* 2008, 9, 371–421.
- [23] Romano, Y.; Sesia, M.; Candès, E.J. Classification with Valid and Adaptive Coverage. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, BC, Canada, 6–12 December 2020.
- [24] Gammerman, A.; Shafer, G.; Vovk, V. *Algorithmic Learning in a Random World*; Springer-Verlag: New York, NY, USA, 2005; ISBN 0-387-00152-2.
- [25] Rainio, O.; Teuho, J.; Klén, R. Evaluation Metrics and Statistical Tests for Machine Learning. *Sci. Rep.* 2024, 14, 6086. <https://doi.org/10.1038/s41598-024-56706-x>.
- [26] Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 2006, 7, 1–30.

- [27] Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* 1998, 10, 1895–1923. <https://doi.org/10.1162/089976698300017197>.
- [28] Smyth, P.; Wolpert, D. Stacked density estimation. In *Proceedings of the Advances in Neural Information Processing Systems*, Denver, CO, USA, 1–6 December 1997.
- [29] Japkowicz, N.; Stephen, S. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* 2002, 6, 429–449.
- [30] Berry, R.B.; Brooks, R.; Gamaldo, C.E.; Harding, S.M.; Marcus, C.; Vaughn, B. V; others The AASM Manual for the Scoring of Sleep and Associated Events. Rules Terminol. Tech. Specif. Darien Ill. *Am. Acad. Sleep Med.* 2012, 176, 2012.
- [31] Espiritu, J.R.D. Health Consequences of Obstructive Sleep Apnea. In *Management of Obstructive Sleep Apnea*; Springer International Publishing: Cham, Switzerland, 2021; pp. 23–43.
- [32] Wang, M.; Qian, Y.; Yang, Y.; Chen, H.; Rao, W.-F. Improved Stacking Ensemble Learning Based on Feature Selection to Accurately Predict Warfarin Dose. *Front. Cardiovasc. Med.* 2024, 10, 1320938. <https://doi.org/10.3389/fcvm.2023.1320938>.
- [33] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* 2021, 2, 160. <https://doi.org/10.1007/s42979-021-00592-x>.
- [34] Hajipour, F.; Jozani, M.J.; Moussavi, Z. A Comparison of Regularized Logistic Regression and Random Forest Machine Learning Models for Daytime Diagnosis of Obstructive Sleep Apnea. *Med. Biol. Eng. Comput.* 2020, 58, 2517–2529. <https://doi.org/10.1007/s11517-020-02206-9>.
- [35] Alqudah, A.M.; Elwali, A.; Kupiak, B.; Hajipour, F.; Jacobson, N.; Moussavi, Z. Obstructive Sleep Apnea Detection during Wakefulness: A Comprehensive Methodological

Review. Med. Biol. Eng. Comput. 2024, 62, 1277–1311. <https://doi.org/10.1007/s11517-024-03020-3>.

[36] Haibo He; Garcia, E.A. Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 2009, 21, 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>.

[37] Gonçalves, M.C.; Silva, R. The Effect of Statistical Hypothesis Testing on Machine Learning Model Selection. In Intelligent Systems (BRACIS 2023), Proceedings of the 12th Brazilian Conference, Belo Horizonte, Brazil, 25–29 September 2023; Lecture Notes in Computer Science; Naldi, M.C., Bianchi, R.A.C., Eds.; Springer: Cham, Switzerland, 2023; Volume 14196. https://doi.org/10.1007/978-3-031-45389-2_28

[38] Simply, R.M.; Dafna, E.; Zigel, Y. Obstructive Sleep Apnea (OSA) Classification Using Analysis of Breathing Sounds During Speech. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; IEEE, September 2018; pp. 1132–1136.

[39] Walia, R.; Achilefu, A.; Crawford, S.; Jain, V.; Wigley, S.D.; McCarthy, L.H. Are At-Home Sleep Studies Performed Using Portable Monitors (PMs) as Effective at Diagnosing Obstructive Sleep Apnea (OSA) in Adults as Sleep Laboratory-Based Polysomnography (PSG)? J. Okla. State Med. Assoc. 2014, 107, 642–644.

[40] Gali, B.; Whalen, F.X.; Gay, P.C.; Olson, E.J.; Schroeder, D.R.; Plevak, D.J.; Morgenthaler, T.I. Management Plan to Reduce Risks in Perioperative Care of Patients with Presumed Obstructive Sleep Apnea Syndrome. J. Clin. Sleep Med. 2007, 3, 582–588.

[41] Van den Eynde, J.; Lachmann, M.; Laugwitz, K.-L.; Manlhiot, C.; Kutty, S. Successfully Implemented Artificial Intelligence and Machine Learning Applications in Cardiology: State-of-

the-Art Review. Trends Cardiovasc. Med. 2023, 33, 265–271.
<https://doi.org/10.1016/j.tcm.2022.01.010>.

[42] Begley, T.; Schwedes, T.; Frye, C.; Feige, I. Explainability for Fair Machine Learning. arXiv 2020, arXiv:2010.07389.

[43] DeYoung, P.N.; Bakker, J.P.; Sands, S.A.; Batool-Anwar, S.; Connolly, J.G.; Butler, J.P.; Malhotra, A. Acoustic Pharyngometry Measurement of Minimal Cross-Sectional Airway Area Is a Significant Independent Predictor of Moderate-to-Severe Obstructive Sleep Apnea. *J. Clin. Sleep Med.* 2013, 9, 1161–1164. <https://doi.org/10.5664/jcsm.3158>.

[44] Dae, G.J.; Hae, Y.C.; Grunstein, R.R.; Yee, B. Predictive Value of Kushida Index and Acoustic Pharyngometry for the Evaluation of Upper Airway in Subjects with or without Obstructive Sleep Apnea. *J. Korean Med. Sci.* 2004, 19, 662–667. <https://doi.org/10.3346/jkms.2004.19.5.662>.

[45] Vashishth, T.K.; Sharma, V.; Neha; Ahamad, S. Exploring Machine Learning Models for Biomedical Signal Processing: A Comprehensive Review. In *Machine Learning Models and Architectures for Biomedical Signal Processing*; Elsevier: Amsterdam, The Netherlands, 2025; pp. 69–96.

[46] Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep Learning for Healthcare: Review, Opportunities and Challenges. *Brief. Bioinform.* 2018, 19, 1236–1246. <https://doi.org/10.1093/BIB/BBX044>.

[47] Ben Or, D.; Dafna, E.; Tarasiuk, A.; Zigel, Y. Obstructive Sleep Apnea Severity Estimation: Fusion of Speech-Based Systems. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2016, Orlando, FL, USA, 16–20 August 2016*; pp. 3207–3210. <https://doi.org/10.1109/EMBC.2016.75914>

CHAPTER SEVEN: INTERPRETABLE ACOUSTIC FEATURES FROM WAKEFULNESS TRACHEAL BREATHING FOR OSA SEVERITY ASSESSMENT

Ali Mohammad Alqudah¹, Walid Ashraf¹, Brian Lithgow^{1,2,3,4}, and Zahra Moussavi^{1,4}

¹ Biomedical Engineering, University of Manitoba, Winnipeg, Canada

² Multidisciplinary Alfred Psychiatry Research Centre, Monash University, Melbourne, Australia

³ Riverview Health Center, University of Manitoba, Winnipeg, Canada

⁴ Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
Journal of Clinical Medicine (2026) 15 (3) (<https://doi.org/10.3390/jcm15031081>)

Abstract

Background: Obstructive Sleep Apnea (OSA) is one of the most prevalent sleep disorders associated with cardiovascular complications, cognitive impairments, and reduced quality of life. Early and accurate diagnosis is essential. The present gold standard, polysomnography, is expensive and resource-intensive. This work develops a non-invasive machine-learning-based framework to classify four OSA severity groups (non, mild, moderate, and severe) using tracheal breathing sounds (TBSs) and anthropometric variables. **Methods:** A total of 199 participants were recruited, and TBS were recorded whilst awake (wakefulness) using a suprasternal microphone. The workflow included the following steps: signal preprocessing (segmentation, filtering, and normalization), multi-domain feature extraction representing spectral, temporal, nonlinear, and morphological features, adaptive feature normalization, and a three-stage feature selection that combined univariate filtering, Shapley Additive Explanations (SHAP)-based ranking, and recursive feature elimination (RFE). The classification included training ensemble learning models via bootstrap aggregation and validating them using stratified k-fold cross-validation (CV), while preserving the OSA severity and anthropometric distributions. **Results:** The proposed framework performed well in discriminating among OSA severity groups. TBS features, combined with

anthropometric ones, increased classification performance and reliability across all severity classes, providing proof for the efficacy of non-invasive audio biomarkers for OSA screening.

Conclusions: TBS-based model's features, coupled with anthropometric information, offer a promising alternative or supplement to PSG for OSA severity detection. The approach provides scalability and accessibility to extend screening and potentially enables earlier detection of OSA, compared to cases that might remain undiagnosed without screening.

7.1. Introduction

Obstructive sleep apnea (OSA) is a common yet underdiagnosed sleep-related breathing disorder affecting nearly 20% of adults in North America and linked to cardiovascular disease, hypertension, diabetes, and increased perioperative risk [1,2]. Despite its prevalence, up to 80% of cases remain undiagnosed [3], creating primary healthcare and economic burdens. OSA arises from recurrent upper airway obstruction during sleep, and its severity is classified by the apnea-hypopnea index (AHI) [4]. While polysomnography (PSG) remains the diagnostic gold standard [5], it is costly, time-intensive, and often inaccessible. Screening tools such as STOP-Bang and Berlin questionnaires provide high sensitivity but low specificity, leading to frequent misclassification of OSA status (i.e., false positives) [6,7].

Recent advances in biomedical signal analysis offer promising alternatives. For instance, tracheal breathing sounds (TBSs) recorded during wakefulness have been shown to contain distinctive acoustic markers related to upper airway physiology [8–16]. Studies using power spectral, bispectral, and fractal analyses, and more recently, machine learning (ML) models, have demonstrated strong potential for OSA detection [8–16]. However, a significant gap remains between the extraction of acoustic features and their clinical interpretation. While many studies report statistically significant differences in signal characteristics, the physiological meaning of

these features and their relationships with airway mechanics, airflow resistance, and neuromuscular control remain poorly understood [17,18].

Bridging this gap is essential to translate signal-based metrics into clinically interpretable and actionable tools. This study focuses on interpreting acoustic features extracted from wakefulness TBS across different OSA severity groups. By analyzing spectral power, bispectral coupling, and fractal dimensions, we explore how acoustic signatures reflect physiological mechanisms underlying airway obstruction. This approach aims to link quantitative signal analysis with clinical interpretation, supporting the development of objective, accessible, and scalable OSA screening.

7.2. Materials and Methods

In this study, we applied our previously validated workflow [8] for acquiring, preprocessing, and analyzing wakefulness tracheal breathing sounds (TBSs). Participants were recruited from individuals referred for overnight PSG, representing a clinically enriched cohort with elevated pre-test probability of OSA. Suprasternal TBS recordings were collected under controlled conditions: subjects were positioned supine and instructed to perform five full deep breaths through the nose with the mouth closed, followed by five deep breaths through the mouth while wearing a nose clip using a Sony ECM-77B, Tokyo, Japan omnidirectional condenser microphone (sensitivity: $-52 \text{ dB} \pm 3.5 \text{ dB}$, frequency response: 40 Hz–20 kHz). Snoring history was not collected via subject self-report as part of the anthropometric questionnaire, and no snoring events were present in the wakeful breathing recordings analyzed in this study. Table 7-1 presents the distribution of subjects in the dataset by anthropometric features. Preprocessing followed our established procedures, including artifact inspection, adaptive segmentation of inspiration and expiration, and bandpass filtering to isolate physiological components. We then extracted a

comprehensive set of spectral, nonlinear, fractal, morphological, and time-frequency features using the same methods detailed in [8]. These features were optimized for 1-vs-1 subgroup analyses to improve the interpretability and personalization of acoustic biomarkers. The complete workflow is summarized in Figure 7-1.

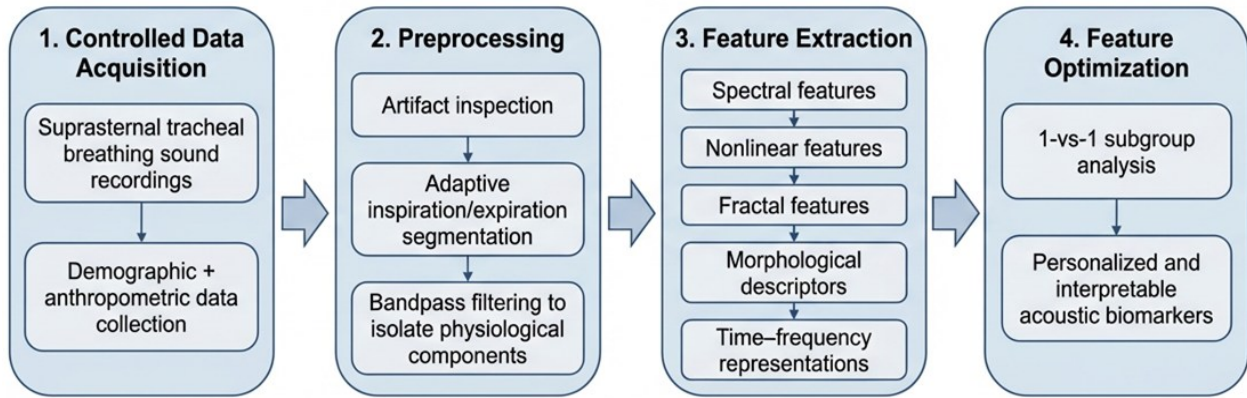


Fig. 7-1 Overview of the machine learning pipeline for wakefulness-based tracheal breathing sound analysis and OSA severity classification.

Preprocessing included careful inspection for background noise and vocal artifacts to be excluded, segmenting breathing sound signals into inspiratory/expiratory phases using adaptive thresholding of the log-variance envelope and Signal-to-Noise Ratio (SNR) computation, and bandpass filtering (75–3000 Hz, 4th-order Butterworth) to discard extraneous physiological and ambient signals [8,11,12]. Filtered signals were subsequently normalized using automated methods (mean-range scaling, z-score, min-max, and robust scaling) with mutual information to maximize feature-label dependency [8]. Then, a feature extraction method was applied to each processed mid-flow signal; the methodology spans multiple analytical domains, including spectral, temporal, and nonlinear analyses, as well as cross-domain analyses, ensuring a holistic, multidimensional representation of linear and nonlinear signal dynamics. The extracted features are grouped and explicitly optimized for 1-vs-1 labels [8]. This group-specific feature selection process enables the creation of personalized feature sets that enhance model robustness and

improve interpretation for diagnostic and predictive applications [8]. The following features have been extracted:

- Spectral features: Power spectrum density via Welch's method, spectral centroid, entropy, kurtosis, bandwidth, flux, and crest metrics [19].
- Bispectral features: Bootstrap-based confidence interval detection of nonstationary gaps and coupling metrics [20].
- Fractal and nonlinear features: Hurst exponent, Lyapunov exponent, Recurrence Quantification Analysis (RQA), Katz and Higuchi fractal dimensions [21–23].
- Wavelet and time-frequency features: Wavelet coefficients, Mel-Frequency Cepstral Coefficients (MFCCs), Constant-Q Transform statistics [24–26].
- Morphological features: Image-based representation of spectrogram and bispectrum (bounding box area, holes, connected components, Euler number, contrast/homogeneity/correlation/energy descriptors) [20,27–29].
- Time-domain metrics: Zero-crossing rate, root mean square, shimmer, jitter, and noise-to-harmonics ratio [30–32].

To identify stable and physiologically meaningful predictors of OSA severity, we applied the same three-stage feature selection framework described in [8], consisting of univariate statistical filtering, SHAP-based feature ranking, and RFE. Within this framework, the final feature subset was defined as the minimum number of features that preserved consistent performance across cross-validation folds while maintaining feature stability and physiological interpretability, as established in [8]. A further reduction in the feature set beyond this subset was previously shown to increase performance variability and reduce robustness; therefore, no additional feature pruning was applied in the present study. Model evaluation was conducted using a custom stratified k-fold

CV scheme specifically designed to preserve the joint distribution of OSA severity labels and key anthropometric risk factors, including age, body mass index (BMI), neck circumference, sex, and Mallampati score [12]. Appendix B1 presents a summary of the top 35 selected features for each model, based on the extracted features.

Table 7-1 Participants' Severity Groups and Anthropometric Information. AHI: apnea–hypopnea index, BMI: body mass index, NC: neck circumference, MPS: Mallampati score, M/F: male/female.

Severity Group	Number of Subjects	AHI	Sex	Age	NC	BMI	MPS
Non-OSA	74	1.2 ± 1.3	29 M, 45 F	46.8 ± 12.9	38.8 ± 4.0	30.6 ± 6.2	41 (1), 19 (2), 6 (3), 8 (4)
Mild	35	8.7 ± 2.6	21 M, 14 F	52.3 ± 11.6	42.1 ± 6.5	34.3 ± 8.4	18 (1), 6 (2), 9 (3), 1 (4)
Moderate	50	21.5 ± 4.2	36 M, 14 F	54.7 ± 11.3	43.1 ± 3.4	33.8 ± 6.4	17 (1), 17 (2), 8 (3), 8 (4)
Severe	40	69.5 ± 33.3	30 M, 10 F	48.9 ± 11.1	45.3 ± 3.6	39.7 ± 8.7	5 (1), 13 (2), 14 (3), 8 (4)

As illustrated in Figure 7-2, the whole dataset is first partitioned into k folds, ensuring each fold has approximately equal representation of OSA severity classes (Non, Mild, Moderate, and Severe) and comparable distributions of the selected anthropometric variables. Rather than stratifying solely by OSA severity, the proposed strategy employs multi-criteria stratification to ensure that clinically relevant subgroups are consistently represented in both training and validation sets. This approach reduces sampling bias caused by population heterogeneity and yields a more reliable estimate of generalization performance across physiologically diverse subjects. Also, severity stratification followed standard AHI-based clinical definitions to preserve physiological granularity and enable interpretation of progressive airway dysfunction beyond binary disease detection. Table 7-2 shows the distribution of subjects' anthropometric data of the k-fold splits.

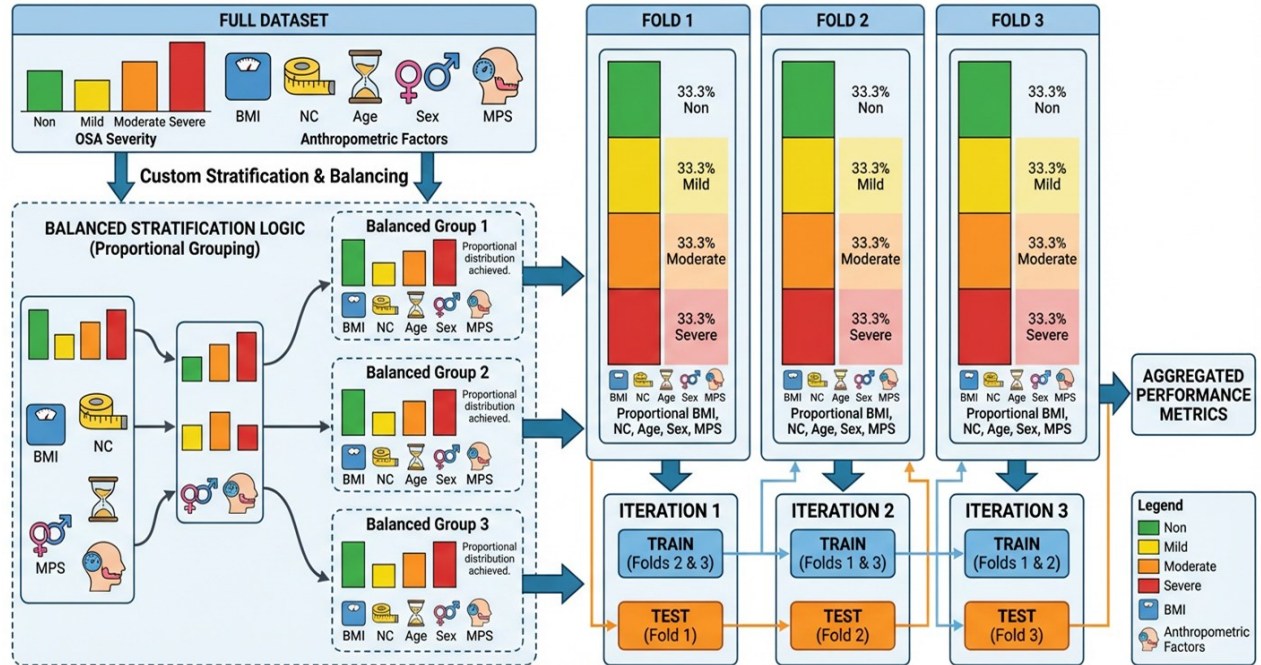


Fig. 7-2 Schematic illustration of the proposed stratified k-fold cross-validation strategy. The full dataset is partitioned into k folds while preserving the proportional distribution of OSA severity classes (Non, Mild, Moderate, Severe) and key anthropometric risk factors (age, BMI, neck circumference, sex, and Mallampati score). Each fold maintains comparable joint distributions, ensuring balanced subgroup representation during training and validation and minimizing bias due to population heterogeneity.

Table 7-2 Participants' Severity Groups and Anthropometric Information for K-folds. AHI: apnea-hypopnea index, BMI: body mass index, NC: neck circumference, MPS: Mallampati score, M/F: male/female.

Severity Group	Fold	Number of Subjects	AHI	Sex	AGE	BMI	NC	MPS
Non-OSA	1	23	0.6 ± 0.8	10 M, 13 F	44.9 ± 12.1	29.2 ± 4.7	38.0 ± 4.7	12 (1), 7 (2), 1 (3), 3 (4)
	2	27	1.1 ± 1.3	10 M, 17 F	45.7 ± 12.1	32.3 ± 7.6	39.2 ± 4.3	12 (1), 9 (2), 4 (3), 2 (4)
	3	24	1.8 ± 1.3	9 M, 15 F	50.0 ± 14.3	30.0 ± 5.8	39.0 ± 2.9	17 (1), 3 (2), 1 (3), 3 (4)
Mild	1	16	8.7 ± 2.4	10 M, 6 F	50.9 ± 12.5	36.6 ± 9.9	43.5 ± 5.3	7 (1), 1 (2), 7 (3), 1 (4)
	2	10	8.6 ± 2.2	5 M, 5 F	51.3 ± 11.6	31.8 ± 8.4	38.5 ± 8.8	6 (1), 1 (2), 2 (3), 1 (4)
	3	9	8.8 ± 3.5	6 M, 3 F	56.0 ± 10.3	33.0 ± 4.2	43.4 ± 2.7	5 (1), 4 (2)
Moderate	1	16	19.9 ± 2.9	10 M, 6 F	56.3 ± 10.8	34.6 ± 7.8	42.3 ± 4.0	4 (1), 6 (2), 2 (3), 4 (4)
	2	18	22.8 ± 4.5	13 M, 5 F	53.6 ± 9.7	31.8 ± 5.7	42.7 ± 3.8	8 (1), 5 (2), 3 (3), 2 (4)
	3	16	21.6 ± 4.7	13 M, 3 F	54.5 ± 13.9	35.2 ± 5.2	43.6 ± 2.8	5 (1), 6 (2), 3 (3), 2 (4)

Severe	1	14	72.9 ± 35.0	11 M, 3 F	45.5 ± 10.5	39.1 ± 10.1	44.3 ± 4.2	2 (1), 3 (2), 5 (3), 4 (4)
	2	16	66.6 ± 29.6	13 M, 3 F	50.2 ± 11.0	40.1 ± 8.5	46.6 ± 3.2	1 (1), 6 (2), 5 (3), 4 (4)
	3	10	69.6 ± 39.1	6 M, 4 F	51.9 ± 12.2	40.2 ± 7.5	43.8 ± 3.5	2 (1), 4 (2), 4 (3)

To evaluate the discriminative power and physiological relevance of tracheal sound features, several complementary metrics were used. Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) quantified each feature’s ability to distinguish between OSA severity groups, with higher values indicating stronger discrimination [33]. Pearson correlation coefficients measured the linear association between feature values and true labels, providing insight into how consistently a feature reflects the clinical outcome across cross-validation folds [34]. To assess robustness and generalizability, Absolute Delta AUC (AbsDeltaAUC) was calculated as the absolute difference between the training and test AUCs, with smaller values indicating more stable features that are less sensitive to data variability [35]. Finally, SHAP quantified each feature’s contribution to model predictions, while accounting for interactions with other features, thereby enhancing interpretability by highlighting physiologically meaningful patterns [36]. Together, these metrics enable ranking of features by both their discriminative ability and stability, supporting the identification of robust biomarkers that link acoustic and morphological descriptors to airway dynamics, airflow turbulence, and anatomical variations associated with OSA. As shown in Figure 7-3, the proposed feature evaluation framework integrates analyses of discriminative, correlational, stability, and explainability to identify physiologically relevant tracheal-sound features in OSA.

The final analytic framework integrates feature-level metrics, stability measures, and explainability-driven insights to produce tables and visualizations that highlight the most physiologically relevant acoustic and morphological features across OSA severity groups. Rather

than focusing solely on classification performance, the analysis emphasizes how each feature contributes to model predictions and relates to underlying airway physiology. Ranked lists based on AUC [33], fold-wise stability (AbsDeltaAUC) [35], and SHAP values [36] provide a multidimensional perspective on feature importance, while correlation analyses link these features to clinical and anthropometric variables, including AHI, neck circumference (NC), and Mallampati Score (MPS) [34]. Image-based and spectro-temporal feature maps illustrate changes in sound texture, frequency patterns, and event shapes, revealing airflow turbulence, intermittent obstruction, and variations in airway mechanics. By combining quantitative metrics with visual interpretations, this framework transforms raw signal descriptors into clinically meaningful biomarkers, enhancing understanding of upper-airway dynamics, airflow irregularities, and anatomical risk factors associated with OSA severity [37].

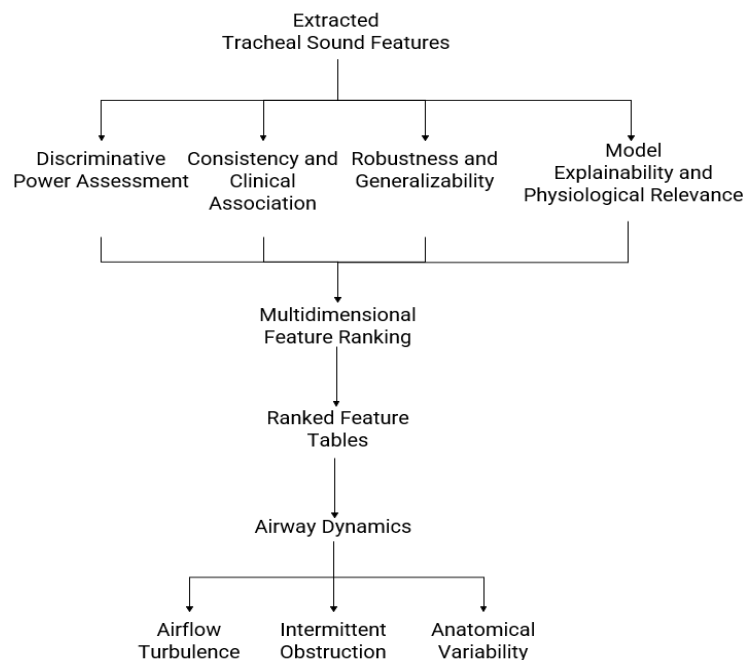


Fig. 7-3 Framework for evaluating and interpreting tracheal breathing sound features in obstructive sleep apnea (OSA). Extracted acoustic and morphological features are analyzed for discriminative power, consistency with clinical variables, robustness across cross-validation folds, and model explainability. The results are integrated into a multidimensional feature ranking, producing tables, plots, and visual maps that link sound descriptors to airway dynamics, airflow turbulence, and anatomical variability across OSA severity levels.

7.3. Results

This section presents the key findings from the feature extraction and selection pipeline, highlighting the most discriminative tracheal-breathing-sound features for OSA severity classification. Analyses were conducted across six 1-vs-1 base models (Non-OSA vs. Mild, Non-OSA vs. Moderate, Non-OSA vs. Severe, Mild vs. Moderate, Mild vs. Severe, Moderate vs. Severe) and three folds of a custom stratified cross-validation, designed to preserve the joint distribution of severity groups and key anthropometric factors. Feature importance was assessed using both correlation-based ranking and SHAP values to identify consistently essential features. The top selected features for each model are detailed in Appendix B1.

The models and selected features are empirical; their exact frequency bands or characteristics may differ for other datasets depending on the sensor used (e.g., different microphones). To keep feature names readable, we have categorized them by main characteristics, such as spectral or bispectral features, breathing type (mouth or nose), and phase (inspiration or expiration). The frequency regions from which the features were extracted are based on the 95% confidence interval of the training set, as proposed in our previous work [8]. For example, `Bispectral_Centroid_Mean` represents the mean bispectral energy centroid across all breathing conditions. Similarly, `Spectral_Skewness_Mouth_Inspiration` captures the skewness of the spectral distribution during mouth inspiration, and `Spectral_FrequencyRatio_Mouth_Expiration` represents the frequency ratio feature during mouth expiration. The detailed definitions of these features, including what a bounding box (BBox) is, the specific coordinates or frequency/time ranges, and the corresponding breathing conditions, are provided elsewhere (e.g., in a footnote, appendix, or table legend). This approach ensures that the main text remains readable while maintaining reproducibility and technical clarity. This general naming approach avoids dataset-

specific details, makes the features more interpretable for readers outside the team, and preserves the essential information on how each feature was derived (Figure 7-4).

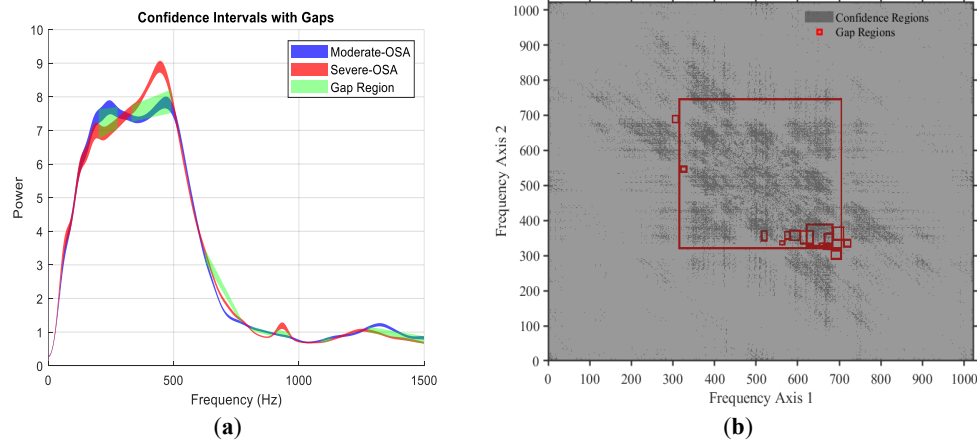


Fig. 7-4 Representative examples of gap regions identified between two classes based on non-overlapping 95% confidence intervals. (a) Power spectral density (PSD): the green shaded areas indicate frequency bands where the 95% confidence intervals of the average PSD for Moderate-OSA and Severe-OSA do not overlap, defining statistically significant spectral gaps between the two classes. (b) Bispectrum: red boxes indicate time–frequency regions in the bispectral domain where the 95% confidence intervals of the averaged bispectrum for Moderate-OSA and Severe-OSA do not overlap. The bispectrum, a third-order spectral measure, captures quadratic phase coupling and nonlinear interactions between frequency components; gap regions therefore indicate class-specific differences in nonlinear spectral coupling.

For detailed clinical interpretation of top features, a specific appendix (Appendix B2) is dedicated to these details, while for detailed comparison-specific results, all supporting tables and figures are provided in the appendices: Appendix B3 (Non-OSA vs. Mild OSA), Appendix B4 (Non-OSA vs. Moderate OSA), Appendix B5 (Non-OSA vs. Severe OSA), Appendix B6 (Mild vs. Moderate OSA), Appendix B7 (Mild vs. Severe OSA), and Appendix B8 (Moderate vs. Severe OSA). Each appendix includes the top-ranked features by test AUC, the most stable features by AbsDeltaAUC, and the strongest anthropometric- and AHI-associated features for the corresponding comparison. Higher AUC values indicate greater discrimination between the two severity groups. In contrast, smaller AbsDeltaAUC values suggest more consistent feature performance across cross-validation folds, reflecting reduced variability in AUC estimates across

data partitions. Correlation analyses with anthropometric variables and AHI provide additional insight into potential physiological relevance and relationships with clinical severity. The following subsections focus on the clinical interpretation of the most consistently supported features across these comparisons.

7.3.1. Clinical Interpretation of Top Features

To provide a physiologically grounded interpretation of the observed acoustic differences, we adopt a Structure–Function–Symptom framework. In this narrative, anatomical and structural characteristics of the upper airway (Structure), such as tissue compliance, airway narrowing, and fat deposition, influence airflow behavior during breathing (Function), including turbulence, nonlinear coupling, and ventilatory instability. These functional alterations manifest clinically as differences in apnea–hypopnea burden and disease severity (Symptom), quantified by the AHI. The following interpretations therefore explain how each significant acoustic feature reflects a structural-functional pathway underlying OSA progression.

This subsection provides a clinical interpretation of the most discriminative features, prioritized based on agreement across AUC ranking, AbsDeltaAUC, and SHAP importance. The goal is to link key acoustic and spectro-temporal descriptors to potential physiological and airflow changes associated with early manifestations of sleep-disordered breathing. Table 7-3 shows an overview of the clinical interpretation of features across different models. In contrast, more detailed feature-by-feature interpretations are provided in Appendix B2. Collectively, this analysis facilitates a clearer understanding of how specific acoustic patterns may reflect underlying upper airway dynamics and disease progression. Furthermore, aligning model-derived features with known clinical mechanisms enhances the interpretability and translational relevance of the proposed framework.

Table 7-3 Structure–Function–Symptom Interpretation of Dominant Acoustic Features Across OSA Severity Comparisons.

Severity Model	Dominant Feature Types	Structure–Function–Symptom Interpretation
Non-OSA vs. Mild-OSA	Spectral skewness of the power spectrum and texture uniformity of the bispectrum	Mild structural airway compliance and early narrowing (Structure) introduce intermittent airflow instability (Function), producing subtle turbulence and disrupted nonlinear coupling that manifest clinically (Symptom) as early increases in AHI.
Non-OSA vs. Moderate-OSA	Spectral power from the power spectrum and fragmentation of bispectral patterns	Progressive anatomical narrowing and reduced airway stiffness (Structure) generate sustained turbulent airflow and fragmented nonlinear interactions (Function), corresponding clinically (Symptom) to increased AHI and frequent obstructive events.
Non-OSA vs. Severe-OSA	Energy, complexity, and impulsiveness derived from the power spectrum and bispectrum	Severe upper-airway collapsibility and loss of neuromuscular control (Structure) result in chaotic, high-energy airflow and impulsive breathing sounds (Function), which clinically manifest (Symptom) as severe OSA with high AHI.
Mild-OSA vs. Moderate-OSA	Bandwidth expansion in the power spectrum and texture irregularity in the bispectrum	Worsening airway compromise (Structure) shifts airflow from intermittent to persistent instability (Function), reflected clinically (Symptom) by escalating AHI and sustained breathing disruption.
Mild-OSA vs. Severe-OSA	Entropy, Root Mean Square (RMS), and high-frequency variability of power spectral and bispectral representations	Dominant structural airway vulnerability (Structure) overwhelms compensatory mechanisms, producing highly irregular and energetic airflow (Function) that manifests clinically (Symptom) as severe OSA.
Moderate-OSA vs. Severe-OSA	Variability and topological complexity of power spectral and bispectral structures	Further loss of airway resilience and increased collapsibility (Structure) lead to chaotic airflow dynamics and extreme variability (Function), clinically reflected (Symptom) by markedly elevated AHI.

7.3.2. Top-Ranked Features

The top 10 features, identified by their overall average rank across both correlation- and SHAP-based ranking methods, are presented in Table 4. These features consistently demonstrated high importance in distinguishing between different OSA severity groups. The top features include

a mix of spectral, temporal, and morphological characteristics of the tracheal breathing sounds, as shown in Table 7-4. Notably, features related to spectral bandwidth (range of frequencies contributing to the signal), texture energy (quantifies uniformity and repetitiveness of bispectral patterns), spectral flux (measures frame-to-frame changes in the power spectrum), and statistical moments (mean, standard deviation, kurtosis, skewness) consistently appear among the most essential features. These features are robust across cross-validation folds and severity comparisons, indicating that they reliably capture physiologically relevant changes in airflow dynamics and turbulence. These features capture various aspects of the sound signal, including its frequency distribution, temporal dynamics, and overall intensity and complexity.

Table 7-4 Top 10 Ranked Tracheal Breathing Sound Features for OSA Severity Classification (Lower is better).

Feature Name	Average Rank by Corr	Average Rank by SHAP	Overall Average Rank
Average_BBox_Spectral	1	1	1
MouthInspiration_ConnectedComponents	1	2	1.5
MouthExpiration_FreqCentroid	2	1	1.5
Average_BBox_SpectralFlux	2	1	1.5
MouthExpiration_Range_SCBW_Bandwidth	1	2	1.5
Average_BBox_CentroidY	3	1	2
MouthInspiration_Range_MeanPower	3	1	2
Average_BBox_Entropy	3	2	2.5
Average_BBox_TextureEnergy	3	3	3
NoseExpiration_Range_FreqSkewness	3	3	3

7.3.3. Feature Stability Across Folds and Models

The top 10 most stable tracheal breathing sound features were identified based on the lowest absolute differences between training and testing Area Under the Curve (AUC) values, as shown in Table 7-5. These features demonstrate minimal variability across different data splits, highlighting their robustness and consistency in discriminating OSA severity. Lower absolute Delta AUC values indicate that the predictive power of these features is reliably maintained across

training and test datasets. In contrast, lower absolute Delta Corr values reflect more consistent correlations between the features and clinical measurements, indicating stable physiological relevance across folds. Collectively, these metrics suggest that the selected features are both robust and physiologically meaningful, making them strong candidates for inclusion in predictive models and clinical decision-support systems.

Table 7-5 Top 10 Most Stable Tracheal Breathing Sound Features. Lower Absolute Delta AUC indicates higher feature stability across cross-validation folds, while Absolute Delta Correlation reflects the consistency of feature correlation. Features are considered stable/significant if Abs Delta AUC ≤ 0.01 and Abs Delta Correlation ≤ 0.1 , reflecting minimal variation and robust performance across folds. Features with Abs Delta Correlation between 0.1 and 0.15 are considered slightly unstable but still statistically acceptable, as this slight variation does not substantially affect their overall predictive reliability.

Feature Name	Abs Delta AUC	Abs Delta Correlation
Average_BBox_Skewness	0.001231	0.005491
MouthExpiration_CoefVariation	0.001806	0.137229
Average_BBox_FrequencyBandwidth	0.002441	0.09704
Average_BBox_FrequencyBandwidth	0.002511	0.070444
Average_BBox_TextureEnergy	0.002917	0.076107
MouthExpiration_Perimeter	0.002917	0.0041
MouthExpiration_FrequencyBandwidth	0.002917	0.016699
Average_BBox_Median	0.003554	0.113319
Average_BBox_FrequencyBandwidth	0.003667	0.152313
MouthInspiration_Range_FreqSkewness	0.003667	0.038156

7.3.4. Correlation with Anthropometric Data

Strong associations were observed between several extracted tracheal-sound features and key clinical anthropometric measurements relevant to OSA, including BMI, NC, Sex, and MPS. These relationships, summarized in Table 7-6, highlight the clinical relevance of the acoustic and spectro-temporal descriptors identified in this study. The table emphasizes the features most strongly correlated with anthropometric parameters, providing insight into the physiological underpinnings of OSA severity and supporting their potential utility in predictive modeling.

Pearson correlation coefficients between acoustic features and anthropometric variables were computed independently within each cross-validation fold, rather than on pooled data. For each severity comparison, correlations were calculated using the fold-specific data subset, and the corresponding fold index is explicitly reported in Table 7-6.

These correlations were used exclusively for interpretability and physiological analysis and did not influence model training, feature selection, or classifier optimization. As such, they should not be interpreted as estimates of generalization performance, but rather as indicators of strong fold-specific associations between acoustic characteristics and anthropometric measures.

Table 7-6 Top 10 Pearson correlations between breathing sound features and anthropometric measures (NC, BMI, Sex, MPS). Each row indicates which feature is correlated with which anthropometric variable, along with the fold and severity comparison.

Feature Name	Anthropometric Feature	Pearson Correlation	Fold	Comparison
MouthExpiration_Peak	NC	1	2	Mild-OSA vs. Severe-OSA
Average_BBox_TextureHomogeneity	Sex	0.9998	1	Non-OSA vs. Moderate-OSA
MouthExpiration_Entropy	BMI	0.999	2	Mild-OSA vs. Severe-OSA
Average_BBox_Mean	Sex	0.9987	1	Non-OSA vs. Mild-OSA
Average_BBox_IQR	Sex	0.9985	3	Mild-OSA vs. Moderate-OSA
MouthInspiration_FreqCentroid	BMI	0.9985	3	Mild-OSA vs. Severe-OSA
Average_BBox_Energy	NC	0.9985	2	Non-OSA vs. Severe-OSA
Average_BBox_AspectRatio	MPS	0.9983	1	Non-OSA vs. Severe-OSA
MouthExpiration_FreqCentroid	MPS	0.9982	2	Mild-OSA vs. Moderate-OSA
Average_BBox_CentroidY	MPS	0.9982	3	Mild-OSA vs. Moderate-OSA

7.4. Discussion

This study aimed to identify interpretable features from wakefulness tracheal breathing sounds that are clinically relevant for assessing OSA severity. The consistent emergence of

specific features across different models and folds, coupled with their stability and correlation with anthropometric data, underscores their potential as robust biomarkers for OSA [8–13,17,18,38].

7.4.1. Clinical Relevance of Key Features

Interpreting acoustic biomarkers through a structure–function–symptom lens enables a mechanistic understanding of how anatomical vulnerability of the upper airway translates into altered airflow dynamics and ultimately manifests as increasing OSA severity. OSA manifests through complex interactions between the upper airway anatomy, airflow turbulence, and respiratory control. Identifying features from tracheal breathing sounds recorded whilst awake that reliably reflect these physiological processes is crucial for non-invasive assessment. This section focuses on clinically meaningful indicators of OSA severity [39–41].

7.4.1.1. Non-OSA vs. Mild-OSA

From a structure–function–symptom perspective, early anatomical vulnerability of the upper airway (Structure), including mild tissue compliance and partial narrowing, leads to subtle functional airflow disturbances (Function), characterized by intermittent turbulence and disrupted nonlinear airflow–tissue coupling during inspiration. These functional alterations manifest clinically (Symptom) as mild elevations in AHI without sustained airway obstruction. The distinction between Non-OSA and Mild-OSA is characterized by the emergence of subtle yet consistent early signs of the upper airway instability during breathing whilst awake. These changes are captured by features such as `MouthInspiration_Range_FreqSkewness`, `Average_BBBox_TextureEnergy`, and `Average_BBBox_FrequencyCentroidX`, which quantify shifts in spectral energy, disruption of structured bispectral coupling, and changes in dominant frequency interactions, respectively (see Appendix B.2.1 for full feature definitions and physiological

interpretation). The observed patterns indicate a transition from predominantly laminar airflow toward intermittently turbulent inspiratory flow, consistent with early upper-airway collapsibility and soft-tissue vibration [38,40,42]. These changes suggest the onset of periodic flow limitation without sustained obstruction, aligning with early physiological manifestations of mild OSA described in prior studies [16,41].

Overall, mild OSA is marked not by significant increases in breathing sound intensity but by early disruption of airflow regularity and spectro-temporal organization, as captured by these spectral and bispectral features. This finding supports the concept that *the earliest stage of OSA manifests primarily as micro-instability and intermittent turbulence rather than overt obstruction, reinforcing the value of wakeful acoustic markers for early detection* [16,43,44].

7.4.1.2. Non-OSA vs. Moderate-OSA

Within the structure–function–symptom framework, progressive anatomical narrowing and reduced airway stiffness (Structure) produce sustained airflow limitation and elevated inspiratory effort (Function), resulting in prolonged turbulent breathing events. Clinically (Symptom), these changes correspond to a clear increase in AHI and more frequent obstructive events, consistent with moderate OSA. In contrast to mild OSA, the transition from Non-OSA to Moderate-OSA reveals a clear escalation in airflow disturbance and respiratory effort. These changes are captured by features such as Average_Range_Maximum, Average_BBBox_BoundingBoxDiagonal, Average_Range_MeanPower, and Average_BBBox_ConnectedComponents, which quantify peak breathing sound energy, expansion of nonlinear bispectral interactions, overall sound intensity, and fragmentation of coupling patterns, respectively (see Appendix B.2.2 for full feature definitions and physiological interpretation). The results demonstrate stronger, more sustained turbulent breathing events,

reflecting prolonged partial airway collapse and increased inspiratory drive [13,39,45]. Breathing sounds become more energetic and fragmented, consistent with repetitive cycles of obstruction and compensatory recovery.

These acoustic characteristics indicate that *moderate OSA is physiologically defined by persistent airflow instability rather than isolated abnormalities*. The increased duration, intensity, and fragmentation of breathing events align with established descriptions of heightened airway collapsibility and more frequent arousal-related breathing responses in moderate disease [46–48].

7.4.1.3. Non-OSA vs. Severe-OSA

Structurally, severe OSA is characterized by pronounced upper-airway collapsibility and reduced neuromuscular compensation. Functionally, this leads to chaotic airflow, repeated collapse–reopening cycles, and highly nonlinear breathing dynamics. These functional disturbances manifest clinically as high AHI values (Symptom), reflecting frequent apneic and hypopneic events. *Severe OSA exhibits a markedly distinct acoustic phenotype, dominated by chaotic, high-energy, and highly irregular breathing patterns*. These changes are captured by features such as Average_BBox_MeanValue, Average_BBox_TextureEnergy, Average_BBox_FractalDimension, Average_BBox_EnergyValue, and Average_BBox_KurtosisValue, which quantify average sound intensity, heterogeneity of bispectral coupling, complexity of local patterns, total sound energy, and prevalence of abrupt or impulsive events, respectively (see Appendix B.2.3 for complete feature definitions and physiological interpretation). The results indicate frequent and intense airflow collapse followed by forceful recovery breaths, producing complex and impulsive acoustic events across a broad frequency range [12,38,40]. The pronounced variability and structural disruption observed are consistent with unstable ventilatory control and recurrent airway obstruction.

Physiologically, these findings reflect deep upper-airway collapsibility, exaggerated negative pressure swings, and repeated collapse–reopening cycles characteristic of advanced OSA [41,46,49]. The elevated complexity and unpredictability of the acoustic patterns are in line with prior reports linking severe disease to chaotic airflow and disordered breathing mechanics [50–52].

7.4.1.4. Mild-OSA vs. Moderate-OSA

In structure–function–symptom terms, the transition from mild to moderate OSA reflects worsening anatomical compromise of the airway (Structure), which shifts airflow behavior from intermittent to persistent instability (Function). This progression manifests clinically (Symptom) as a sustained increase in AHI and reduced effectiveness of compensatory airway control. The progression from mild to moderate OSA represents a shift from intermittent airflow disturbance to more persistent and structurally disruptive obstruction. These changes are captured by features such as `Average_BBox_MeanValue`, `Average_BBox_TextureEnergy`, `Average_BBox_FractalDimension`, `Average_BBox_EnergyValue`, and `Average_BBox_KurtosisValue`, which quantify average sound intensity, heterogeneity of bispectral coupling, complexity of local patterns, total sound energy, and the prevalence of sharp or impulsive events, respectively (see Appendix B.2.4 for complete feature definitions and physiological interpretation). The results indicate increasing turbulence during both inspiration and expiration, accompanied by broader spectral involvement and greater fragmentation of breathing sounds [12,38,44,53]. This suggests that airflow irregularities are no longer isolated but sustained throughout the respiratory cycle.

Clinically, this transition reflects worsening airway collapsibility and reduced effectiveness of neuromuscular compensation during wakefulness. Moderate OSA therefore emerges as a state

in which airflow instability becomes chronic rather than episodic, consistent with physiological models of disease progression [9,17,40,50].

7.4.1.5. Mild-OSA vs. Severe-OSA

Here, structural airway vulnerability becomes dominant (Structure), overwhelming compensatory mechanisms. Functionally, this produces highly variable, noisy, and energetically intense airflow patterns. Clinically (Symptom), these effects correspond to severe OSA, marked by large AHI values and pronounced breathing instability. Comparisons between mild and severe OSA highlight a pronounced escalation in airflow irregularity, respiratory effort, and acoustic unpredictability. These changes are captured by features such as MouthInspiration_BBBox_FrequencyCentroidX, Average_Average_BBBoxes_Entropy, Average_Range_RMS, Average_BBBox_EnergyValue, and MouthExpiration_Range_SpectralEnergy, which quantify shifts in dominant frequencies, overall entropy of bispectral patterns, amplitude variability, total sound energy, and broadband spectral energy, respectively (see Appendix B.2.5 for complete feature definitions and physiological interpretation). The results reveal prolonged and noisy inspiratory phases, increased breath-to-breath variability, and intense turbulent bursts extending into expiration [38,44,45]. These patterns indicate a breakdown of compensatory airway control mechanisms that remain partially effective in mild disease.

From a physiological standpoint, severe OSA is characterized by loss of airflow stability, where collapsibility dominates over neuromuscular control. The marked increases in variability and turbulence observed align with descriptions of unstable ventilatory control and repeated collapse–recovery dynamics in severe disease [18,38,44,48].

7.4.1.6. Moderate-OSA vs. Severe-OSA

From a structure–function–symptom standpoint, severe OSA represents a qualitative shift rather than a linear extension of moderate disease: deeper structural collapse and airway instability (Structure) lead to near-chaotic airflow dynamics (Function), which clinically manifest (Symptom) as extreme AHI values and frequent obstructive episodes. The transition from moderate to severe OSA is marked by a qualitative shift from structured instability to near-chaotic airflow dynamics. These changes are captured by features such as `Average_BBox_MedianValue`, `Average_BBox_IQRValue`, `Average_BBox_EnergyValue`, `Average_BBox_KurtosisValue`, `Average_BBox_Compactness`, `Average_BBox_StdValue`, and `Average_BBox_EntropyValue`, which quantify overall breathing sound intensity, central and total variability, abrupt peaks, diffusion of high-intensity regions, dispersion, and randomness of airflow-related acoustic patterns, respectively (see Appendix B.2.6 for full feature definitions and physiological interpretation). The results indicate greater breath-to-breath variability, stronger and more erratic respiratory effort, and increasingly diffuse turbulent sound patterns [39,41,48]. Acoustic events become less compact and more topologically complex, reflecting deeper and more frequent airway collapse.

These findings suggest that *severe OSA represents not merely an amplification of moderate disease but a distinct physiological regime characterized by unpredictable airflow, unstable arousal responses, and diminished airway resilience* [18,38,44,50,54]. This distinction supports the clinical importance of separating moderate and severe OSA in severity stratification and management.

7.4.1.7. Physiological Themes Across Models

Across severity comparisons, certain recurring acoustic patterns reflect underlying physiological mechanisms of OSA. By examining features related to turbulence, airflow complexity, variability, and energy, we can identify consistent markers of airway instability, vibration, and compensatory respiratory effort. The following themes summarize how these features may collectively capture the progression of OSA.

Escalating turbulence, bandwidth, and centroid shifts correspond to rising Reynolds number and more pronounced vibration/snoring as the airway narrows [44].

Event complexity (diagonals, perimeters, shape metrics): track segmented, irregular airflow fragments as OSA severity increases [43].

Variability (Interquartile range (IQR), Standard Deviation (SD), entropy) reveals unstable ventilatory control, frequent arousals, and abrupt collapse–recovery dynamics [43].

Amplitude/energy (mean, RMS, total) reflect increasing respiratory effort, loud post-obstructive inspiration, and compensatory surges in disease progression.

Each feature thus provides a physiologic aspect into how OSA disrupts the upper airway patency, generates turbulence and vibration, and drives instability and variability across both models [43,44].

7.4.2. Rationale for Multi-Class OSA Severity Stratification

While binary OSA classification (OSA vs. non-OSA) is common in screening-oriented studies, it does not capture the progressive and heterogeneous nature of the upper-airway dysfunction. Clinically defined AHI severity categories (mild, moderate, and severe) reflect

distinct physiological states, including differences in airway collapsibility, airflow turbulence, ventilatory compensation, and symptom burden.

In this study, several acoustic features exhibited nonlinear or stage-specific behavior across severity levels, particularly in comparisons involving mild-to-moderate and moderate-to-severe transitions. Collapsing these groups into a binary framework would obscure intermediate phenotypes and reduce sensitivity to early or transitional disease mechanisms. By adopting a four-class framework, the proposed model preserves physiologically meaningful distinctions and enables severity-aware interpretation aligned with clinical risk stratification, perioperative assessment, and treatment decision-making.

7.4.3. Physiological and Clinical Interpretation of Feature Linkage to Severity

The high-ranking and stable features are not merely statistical constructs; they are direct acoustic manifestations of the anatomical and functional changes associated with progressive OSA severity. Tracheal sounds are generated by turbulent airflow, and their characteristics are susceptible to subtle changes in airway geometry, collapsibility, and compensatory respiratory effort, even during wakefulness [8,11].

Although all acoustic recordings in this study were obtained whilst awake under quiet breathing conditions and therefore do not contain snoring events, snoring history remains an important symptom associated with OSA severity. Chronic snoring reflects repetitive vibration of upper-airway soft tissues during sleep, which has been hypothesized to contribute to long-term structural changes such as tissue remodeling, inflammation, or altered compliance. These chronic structural modifications may persist beyond sleep and subtly influence airflow behavior and tracheal breathing acoustics even during wakefulness. Consequently, while the extracted acoustic features do not represent snoring sounds per se, they may indirectly reflect cumulative airway

alterations associated with both OSA severity and a history of habitual snoring. Future studies incorporating quantitative snoring indices alongside wakeful and sleep-state recordings may further clarify this relationship.

7.4.3.1. Acoustic Signatures of Airway Chaos and Ventilatory Effort (AHI Correlation)

The correlations with the AHI provide the most direct clinical linkage. The following summarizes the correlations of features with AHI:

Decreased Texture Energy (Acoustic Disorganization): The feature exhibits the strongest negative correlation with OSA severity. Texture Energy is a quantitative measure of the uniformity and repetitiveness of local patterns in a spectrogram, computed by summing the squared values of the co-occurrence or filtered spectrogram matrix, reflecting how consistent and regular the acoustic structure is. As severity increases, the pharyngeal airway becomes intrinsically more compliant and prone to intermittent vibration and collapse, leading to flow separation and highly random, broadband turbulence. This shift from structured, laminar-like noise to chaotic, broadband turbulence disrupts the consistency of the spectrogram, resulting in a significant decrease in texture energy. This feature, therefore, serves as a powerful acoustic marker of increasing pharyngeal instability and vulnerability [39,42].

Increased Skewness (Compensatory Drive): Conversely, the high positive correlation ($r \approx 0.99$) between spectral skewness and OSA severity indicates systematic changes in the distribution of sound amplitude. Positive skewness signifies a heavier tail toward high-amplitude values. Physiologically, this represents the subject's increased reliance on intermittent, high-force maneuvers (such as a forceful, highly turbulent inhalation or a loud snort/gasp) to maintain adequate flow against increasing pharyngeal resistance. Clinically, this feature is an acoustic

signature of heightened respiratory drive and compensatory effort, which scales directly with disease burden [41,45].

7.4.3.2. Morphological and Spectral Markers of Flow Limitation and Airway

Dynamics

The consistently top-ranked spectral and morphological features provide a detailed view of the fluid dynamics within the compromised airway.

Spectral Bandwidth and Flux (Venturi Effect): These features are crucial markers of dynamic flow behavior. Airflow acceleration through a narrow, compliant pharyngeal segment (the site of flow limitation, a manifestation of the Venturi effect) generates high-velocity jets. The high spectral flux reflects the rapid, transient changes in the power spectrum as these turbulent jets form and dissipate during the breathing cycle. In contrast, increased bandwidth reflects a broader spread of acoustic energy across frequencies. Together, these changes are consistent with the presence and severity of flow-limiting segments, where the degree of narrowing modulates the strength and spectral extent of turbulent eddies [40,55].

Fractal Dimension and Complexity (Non-linear System Behavior): The high-ranking fractal dimension quantifies the non-linear complexity of the signal. Increased airway resistance and turbulence are hallmarks of a system pushed toward instability. A higher fractal dimension suggests a highly complex, chaotic, and less predictable airflow pattern, aligning with established non-linear control theory, which views the respiratory system as operating close to a chaotic bifurcation point [50,51].

7.4.3.3. Validation Through Established Anatomical Risk Factors

The extremely high correlations between specific features and anthropometric measures provide vital proof of concept: the acoustic features are not just abstract discriminators but directly encode the physical risk factors [12,56,57]. Peak Intensity is the near-perfect correlation between Neck Circumference (NC), which is physiologically profound. It is a validated proxy for fat deposition and reduced pharyngeal tissue stiffness. This deposition not only narrows the airway but also influences sound wave propagation. The high peak acoustic value in expiration is a measurable outcome of a sound wave propagating through a physically constrained, often partially occluded, and highly compressible tissue structure. At the same time, Entropy is the strong correlation between (a proxy for systemic obesity and increased soft tissue mass) and links overall body habits to the acoustic randomness and disorganization of the expired airflow pattern.

The combined evidence from correlation, feature stability, and anthropometric validation strongly supports the use of wakefulness features for the objective and clinically meaningful assessment of severity [12,56,57].

7.4.4. Correlation with Anthropometric Data

The strong correlations observed between several tracheal sound features and anthropometric data further strengthen their clinical utility [12,17]. For example, features such as Mouth Expiration_BBox_650_15_1_0_peakValue, which correlate strongly with Neck Circumference (NC), and Mouth Expiration_BBox_155_2_0_0_entropyValue, which correlate strongly with BMI, are particularly noteworthy. NC and BMI are well-established risk factors and indicators of OSA severity. The direct relationship between these anthropometric measures and specific acoustic features suggests that the structural characteristics of the upper airway, influenced

by body posture, are reflected in tracheal breathing sounds [15,58]. This provides a mechanistic link between anatomical predispositions to OSA and the acoustic manifestations captured by our features. Similarly, correlations with the Mallampati Score (MPS), an indicator of the oral cavity and pharyngeal space, further support the idea that the airway's physical configuration influences sound production during breathing [47]. Although some fold-wise correlations approached unity, these values reflect strong associations observed within specific severity contrasts and folds and do not imply model overfitting, as correlation analysis was conducted independently of the predictive learning pipeline.

These findings suggest that wakefulness tracheal breathing sounds contain rich information reflecting the physiological state of the upper airway. The ability of these features to discriminate OSA severity during wakefulness, their stability across different models and folds, and their strong correlations with established anthropometric risk factors are particularly significant. This offers a non-invasive, convenient, and potentially cost-effective method for screening and monitoring [9,44,59]. The interpretability of these features allows for a deeper understanding of the underlying mechanisms of OSA, moving beyond black-box model predictions [9,44,59]. This interpretability is crucial for clinical acceptance and for guiding future research into targeted interventions.

7.4.5. Alignment with Prior Wakefulness-Based OSA Studies

The present findings align strongly with and extend our group's previous works [12,17]. In the 2018 study [17], the authors demonstrated that spectral, bispectral, and fractal features extracted from tracheal sounds could effectively differentiate OSA from non-OSA subjects, achieving classification accuracies of approximately 70–75% and ROC values of 0.73–0.80. They further observed that mouth-inspiratory features provided the highest discrimination power and that several acoustic descriptors were only weakly influenced by anthropometric variability, such

as body mass index (BMI) or neck circumference (NC). In the other 2019 study [12], they reinforced these results using a larger dataset of 199 subjects. It confirmed that combining tracheal sounds with anthropometric features increased diagnostic performance to 81.4% accuracy (sensitivity = 82.1%, specificity = 80.9%).

Consistent with prior findings, our results achieved diagnostic performance comparable to or superior to that of previous models, yielding AUC values ranging from 0.86 to 0.97 across multiple OSA severity levels. The strongest predictive power was again observed for mouth-inspiratory and low-frequency components (150–450 Hz), which exhibited elevated spectral energy and distinct morphological patterns in moderate and severe OSA subjects. Correlation analysis in our dataset similarly showed significant relationships between acoustic features and anthropometric markers, including BMI ($r = 0.52\text{--}0.75$) and NC ($r = 0.47\text{--}0.72$), corroborating earlier physiological interpretations of airway constriction and turbulence during inspiration.

Beyond confirming the earlier results, the present study extends the prior research in three significant ways:

Instead of a binary OSA vs. non-OSA classification, our framework performs multi-level severity stratification (non-OSA, mild, moderate, and severe), offering finer clinical granularity.

We introduce novel morphological and time–frequency gap descriptors, extracted from harmonic–percussive (HP) decompositions and spectrogram bounding boxes, which capture airway-specific acoustic signatures not examined in previous work.

Our use of ensemble-based models with SHAP explainability provides transparent quantification of feature contributions and robustness validation ($\text{Abs } \Delta\text{AUC} < 0.04$ across folds), establishing reproducibility across subjects and folds.

Together, these advances confirm and expand on the foundational evidence presented previously by our group [12,17], demonstrating that wakefulness-based tracheal breathing sounds, combined with simple anthropometric measures, constitute a physiologically meaningful, non-invasive, and reproducible tool for OSA detection and severity classification.

7.4.6. Comparison with Other Awake Screening Modalities

To contextualize the proposed methodology, it is necessary to benchmark it against the full spectrum of screening modalities whilst awake. While questionnaires (e.g., STOP-Bang) are ubiquitous due to their zero-cost administration [60], they are hindered by low specificity (often < 40%), leading to high false-positive rates [60]. Facial image analysis offers a non-contact alternative, but current methods often plateau at approximately 70% accuracy or require large, balanced datasets to avoid demographic bias [61]. Functional methods such as Negative Expiratory Pressure (NEP) [62] and Acoustic Pharyngometry [63] offer high accuracy by directly measuring airway collapsibility and geometry. However, these techniques often require specialized equipment and strictly controlled protocols, reducing their utility for rapid, large-scale screening compared to microphone-based approaches [12]. Speech analysis, while similar in modality to tracheal breathing sounds analysis [64], often relies on complex phonetic tasks and has shown lower accuracy and variable specificity depending on the features used.

Table 7-7 presents a comprehensive quantitative comparison between the proposed framework and representative studies across five distinct wakefulness-based modalities. To ensure a fair comparison, we focus on dataset size, task formulation, and performance metrics (Sensitivity/Specificity). This comparison highlights that while functional tests (NEP/Pharyngometry) offer high precision, they lack the portability of acoustic methods.

Conversely, while questionnaires are portable, they lack the diagnostic accuracy of the proposed method.

Table 7-7 Comparison of the proposed framework with representative wakefulness-based OSA screening modalities. Acc: accuracy, Sens: sensitivity, Spec: specificity.

Modality	Reference	Dataset Size	Task	Performance Metrics	Uncertainty Modeling
Questionnaires	[60]	9206 (Meta-analysis)	Screening	Sens: ~90–96%	No
				Spec: ~25–34%	
	[65]	234 (Clinical)	Screening	Sens: ~95%	No
				Spec: ~5%	
Facial Analysis	[61]	365	Binary	Acc: 69.8%	No
				Sens: 68.5%	
				Spec: 76.4%	
	[66]	180	Binary	Acc: 79.4%	No
Speech Analysis	[67]	40 OSA/40 Control	Binary	Sens: 69.7%	No
				Acc: 81.0%	
				Sens: 77.5%	
	[68]	190 (AHI < 15) and 208 (AHI > 15)	Binary	Spec: 85.0%	No
				Acc: 77.1%	
				Sens: 75.0%	
Pharyngometry	[63]	46 (AHI < 5) and 254 (AHI > 15)	Binary	Spec: 79.0%	No
				Sens: 97.6%	
NEP	[62]	24 (AHI < 5) and 24 (AHI > 30)	Binary	Spec: 100%	No
				Sens: 91.7%	
Tracheal Sound	[12]	109 (AHI < 15) and 90 (AHI > 15)	Binary	Spec: 95.8%	No
				Acc: 81.4%	
				Sens: 80.9%	
	[16]	17 (AHI < 5) and 35 (AHI > 5)	Binary	Spec: 82.1%	No
				Acc: 83.3%	
				Sens: 85.0%	
Proposed	Proposed	74 Non-OSA, 35 Mild, 50 Moderate, and 40 Severe	Multi-class	AUC Range: 0.86–0.97	Yes (Bootstrap Aggregation & 95% CI)

7.4.7. Limitations and Future Work

Although the proposed framework achieved high interpretability, several technical limitations merit consideration. The analysis was performed on data collected from a single clinical site [8,11,12], which may limit the generalizability of the results to broader populations with

differing acoustic environments, recording hardware, and demographic characteristics. Secondly, the sound recordings were conducted by a high-end Sony microphone. The selected features may change (although not significantly) depending on the sensor used. Future studies should incorporate multi-center datasets and cross-device validation to ensure robustness under real-world variability. Additionally, while the stratified k-fold design effectively balanced anthropometric covariates, the current results were based on a finite number of wakefulness recordings per subject, restricting the temporal representation of respiratory dynamics. Expanding the framework to include multi-cycle, sleep-stage-specific, or longitudinal tracheal sound data would enable modeling of disease progression and treatment response. Despite these constraints, the consistent feature stability and explainable ensemble structure provide a strong foundation for advancing automated, non-invasive OSA assessment.

In addition to these technical considerations, the study cohort consisted of individuals referred to overnight PSG and therefore represents a clinically enriched population with a higher pre-test probability of OSA than general or primary-care populations. This enrichment may lead to optimistic estimates of discrimination metrics, such as AUC, compared with deployment in lower-risk settings, and may particularly affect positive predictive value when disease prevalence is lower. However, the primary objective of this work was not to estimate population-level screening accuracy but to identify robust, physiologically interpretable acoustic biomarkers of OSA severity under controlled clinical conditions. These biomarkers reflect underlying airway dynamics and anatomical vulnerability, which are expected to generalize beyond referral-based cohorts.

In practical deployment, the proposed framework is best positioned as a first-line risk stratification or prioritization tool rather than a standalone diagnostic test. In primary-care or

community settings, it could be used to identify individuals who would benefit most from expedited PSG, thereby improving resource allocation and reducing diagnostic delays. Future studies will focus on validating the framework in lower-prevalence populations, including primary-care and community-based cohorts, and on recalibrating decision thresholds to account for differences in disease prevalence and pre-test probability. Despite these constraints, the consistent feature stability and explainable ensemble structure provide a strong foundation for advancing automated, non-invasive OSA assessment.

7.5. Conclusions

In conclusion, this research successfully established an interpretable, machine-learning-driven framework that uses wakefulness tracheal breathing sounds as objective, severity-stratifying biomarkers for Obstructive Sleep Apnea. By identifying and validating stable acoustic features, texture energy, spectral bandwidth, and fractal dimension that exhibit strong mechanistic correlations with established anatomical risk factors, we have acoustically encoded the underlying physiological vulnerability of the upper airway. This interpretability moves beyond black-box diagnostics, offering clinicians clear, physiological correlates for disease progression. These findings represent a decisive step toward developing a non-invasive, cost-effective, and highly accessible screening tool, which is critically needed for timely diagnosis, perioperative risk stratification, and scalable long-term management in clinical settings globally.

References

- [1] Rizzo, D.; Baltzan, M.; Sirpal, S.; Dosman, J.; Kaminska, M.; Chung, F. Prevalence and regional distribution of obstructive sleep apnea in Canada: Analysis from the Canadian

Longitudinal Study on Aging. *Can. J. Public Health* 2024, 115, 970–979.
<https://doi.org/10.17269/s41997-024-00911-8>.

[2] Lechat, B.; Naik, G.; Reynolds, A.; Aishah, A.; Scott, H.; Loffler, K.A.; Vakulin, A.; Escourrou, P.; McEvoy, R.D.; Adams, R.J.; et al. Multinight Prevalence, Variability, and Diagnostic Misclassification of Obstructive Sleep Apnea. *Am. J. Respir. Crit. Care Med.* 2022, 205, 563–569. <https://doi.org/10.1164/rccm.202107-1761OC>.

[3] Faria, A.; Allen, A.H.; Fox, N.; Ayas, N.; Laher, I. The public health burden of obstructive sleep apnea. *Sleep Sci.* 2021, 14, 257–265. <https://doi.org/10.5935/1984-0063.20200111>

[4] Singh, M.; Liao, P.; Kobah, S.; Wijesundera, D.N.; Shapiro, C.; Chung, F. Proportion of surgical patients with undiagnosed obstructive sleep apnoea. *Br. J. Anaesth.* 2013, 110, 629–636. <https://doi.org/10.1093/bja/aes465>.

[5] Kushida, C.A.; Littner, M.R.; Morgenthaler, T.; Alessi, C.A.; Bailey, D.; Coleman, J., Jr.; Friedman, L.; Hirshkowitz, M.; Kapen, S.; Kramer, M.; et al. Practice parameters for the indications for polysomnography and related procedures: An update for 2005. *Sleep* 2005, 28, 499–521. <https://doi.org/10.1093/sleep/28.4.499>.

[6] Chen, L.; Pivetta, B.; Nagappa, M.; Saripella, A.; Islam, S.; Englesakis, M.; Chung, F. Validation of the STOP-Bang questionnaire for screening of obstructive sleep apnea in the general population and commercial drivers: A systematic review and meta-analysis. *Sleep Breath.* 2021, 25, 1741–1751. <https://doi.org/10.1007/s11325-021-02299-y>.

[7] Mazzotti, D.R.; Keenan, B.T.; Thorarinsdottir, E.H.; Gislason, T.; Pack, A.I.; Sleep Apnea Global Interdisciplinary, C. Is the Epworth Sleepiness Scale Sufficient to Identify the Excessively Sleepy Subtype of OSA? *Chest* 2022, 161, 557–561. <https://doi.org/10.1016/j.chest.2021.10.027>.

- [8] Alqudah, A.M.; Moussavi, Z. Assessing Obstructive Sleep Apnea Severity During Wakefulness via Tracheal Breathing Sound Analysis. *Sensors* 2025, 25, 6280. <https://doi.org/10.3390/s25206280>.
- [9] Alqudah, A.M.; Elwali, A.; Kupiak, B.; Hajipour, F.; Jacobson, N.; Moussavi, Z. Obstructive sleep apnea detection during wakefulness: A comprehensive methodological review. *Med. Biol. Eng. Comput.* 2024, 62, 1277–1311. <https://doi.org/10.1007/s11517-024-03020-3>.
- [10] Elwali, A.; Meza-Vargas, S.; Moussavi, Z. Using tracheal breathing sounds and anthropometric information for screening obstructive sleep apnoea during wakefulness. *J. Med. Eng. Technol.* 2019, 43, 111–123. <https://doi.org/10.1080/03091902.2019.1617799>.
- [11] Elwali, A.; Moussavi, Z. Obstructive Sleep Apnea Screening and Airway Structure Characterization During Wakefulness Using Tracheal Breathing Sounds. *Ann. Biomed. Eng.* 2017, 45, 839–850. <https://doi.org/10.1007/s10439-016-1720-5>.
- [12] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea using Anthropometric Information and Tracheal Breathing Sounds. *Sci. Rep.* 2019, 9, 11467. <https://doi.org/10.1038/s41598-019-47998-5>.
- [13] Elwali, A.; Moussavi, Z. Predicting Polysomnography Parameters from Anthropometric Features and Breathing Sounds Recorded during Wakefulness. *Diagnostics* 2021, 11, 905. <https://doi.org/10.3390/diagnostics11050905>.
- [14] Hajipour, F.; Jozani, M.J.; Elwali, A.; Moussavi, Z. Regularized logistic regression for obstructive sleep apnea screening during wakefulness using daytime tracheal breathing sounds and anthropometric information. *Med. Biol. Eng. Comput.* 2019, 57, 2641–2655. <https://doi.org/10.1007/s11517-019-02052-4>.

- [15] Hajipour, F.; Jozani, M.J.; Moussavi, Z. A comparison of regularized logistic regression and random forest machine learning models for daytime diagnosis of obstructive sleep apnea. *Med. Biol. Eng. Comput.* 2020, 58, 2517–2529. <https://doi.org/10.1007/s11517-020-02206-9>.
- [16] Montazeri, A.; Giannouli, E.; Moussavi, Z. Assessment of obstructive sleep apnea and its severity during wakefulness. *Ann. Biomed. Eng.* 2012, 40, 916–924. <https://doi.org/10.1007/s10439-011-0456-5>.
- [17] Elwali, A.; Moussavi, Z. Determining Breathing Sound Features Representative of Obstructive Sleep Apnea During Wakefulness with Least Sensitivity to Other Risk Factors. *J. Med. Biol. Eng.* 2018, 39, 230–237. <https://doi.org/10.1007/s40846-018-0410-1>.
- [18] Elwali, A.; Moussavi, Z. A feature reduction and selection algorithm for improved obstructive sleep apnea classification process. *Med. Biol. Eng. Comput.* 2021, 59, 2063–2072. <https://doi.org/10.1007/s11517-021-02421-y>.
- [19] Astfalck, L.C.; Sykulski, A.M.; Cripps, E.J. Debiasing Welch’s Method for Spectral Density Estimation. *Biometrika* 2023, 111, 1313–1329.
- [20] Mendel, J.M. Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications. *Proc. IEEE* 1991, 79, 278–305.
- [21] Dlask, M.; Kukal, J. Hurst Exponent Estimation from Short Time Series. *Signal Image Video Process.* 2018, 12, 745–752.
- [22] Rosenstein, M.T.; Collins, J.J.; De Luca, C.J. A Practical Method for Calculating Largest Lyapunov Exponents from Small Data Sets. *Phys. D Nonlinear Phenom.* 1993, 65, 117–134.
- [23] Zhao, K.; Wen, H.; Guo, Y.; Scano, A.; Zhang, Z. Feasibility of Recurrence Quantification Analysis (RQA) in Quantifying Dynamical Coordination among Muscles. *Biomed. Signal Process. Control* 2023, 79, 104042.

- [24] Gosala, B.; Kapgate, P.D.; Jain, P.; Chaurasia, R.N.; Gupta, M. Wavelet Transforms for Feature Engineering in EEG Data Processing: An Overview. *Biomed. Signal Process. Control* 2023, 85, 104811.
- [25] Abdul, Z.K.; Al-Talabani, A.K. Mel Frequency Cepstral Coefficient and Its Applications: A Review. *IEEE Access* 2022, 10, 122136–122158.
- [26] Kohlrausch, A. Binaural masking experiments using noise maskers with frequency-dependent interaural phase differences. II: Influence of frequency and interaural-phase uncertainty. *J. Acoust. Soc. Am.* 1990, 88, 1749–1756. <https://doi.org/10.1121/1.400249>.
- [27] Rangayyan, R.M.; Reddy, N.P. *Biomedical Signal Analysis: A Case-Study Approach*; Pergamon Press: New York, NY, USA, 2002; Volume 30.
- [28] Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002, 24, 971–987.
- [29] Divya, S.; Suresh, L.P.; John, A. Image Feature Generation Using Binary Patterns—LBP, SLBP and GBP. In *Advanced Computing and Intelligent Technologies*; Springer Nature: Singapore, 2022; pp. 233–239.
- [30] Jotz, G.P.; Cervantes, O.; Abrahão, M.; Settanni, F.A.P.; de Angelis, E.C. Noise-to-Harmonics Ratio as an Acoustic Measure of Voice Disorders in Boys. *J. Voice* 2002, 16, 28–31.
- [31] Farrús, M.; Hernando, J.; Ejarque, P. Jitter and Shimmer Measurements for Speaker Recognition. *Proc. Interspeech* 2007, 2007, 778–781.
- [32] Borowska, M. Entropy-Based Algorithms in the Analysis of Biomedical Signals. *Stud. Log. Gramm. Rhetor.* 2015, 43, 21–23. <https://doi.org/10.1515/slgr-2015-0039>

- [33] Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982, 143, 29–36. <https://doi.org/10.1148/radiology.143.1.7063747>.
- [34] Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4.
- [35] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* 2002, 46, 389–422. <https://doi.org/10.1023/a:1012487302797>.
- [36] Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
- [37] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 1996, 58, 267–288.
- [38] Abeyratne, U.R.; de Silva, S.; Hukins, C.; Duce, B. Obstructive sleep apnea screening by integrating snore feature classes. *Physiol. Meas.* 2013, 34, 99–121. <https://doi.org/10.1088/0967-3334/34/2/99>.
- [39] Akhter, S.; Abeyratne, U.R.; Swarnkar, V. Variations of snoring properties with macro sleep stages in a population of Obstructive Sleep Apnea patients. In *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, 3–7 July 2013; pp. 1318–1321. <https://doi.org/10.1109/EMBC.2013.6609751>.
- [40] Herer, B.; Sarig-Bahat, H.; Tarasiuk, A.; Lavie, L. Relationships between snore sound complexity and upper airway collapse level. *Sleep Breath.* 2018, 22, 437–445.

- [41] Kim, J.; Kim, T.; Lee, D.; Kim, J.W.; Lee, K. Exploiting temporal and nonstationary features in breathing sound analysis for multiple obstructive sleep apnea severity classification. *Biomed. Eng. Online* 2017, 16, 6. <https://doi.org/10.1186/s12938-016-0306-7>.
- [42] Ng, A.K.; Koh, T.S.; Abeyratne, U.R.; Puvanendran, K. Investigation of obstructive sleep apnea using nonlinear mode interactions in non-stationary snore signals. *Ann. Biomed. Eng.* 2009, 37, 1796–1806. <https://doi.org/10.1007/s10439-009-9744-8>.
- [43] Janott, C.; Schuller, B.; Heiser, C. Snoring—An acoustic definition. *Physiol. Meas.* 2019, 40, 05T01.
- [44] Ashraf, W.; Fredberg, J.J.; Moussavi, Z. Aeroacoustics of breath sounds in trachea and upper airway. *Appl. Acoust.* 2026, 241, 111021. <https://doi.org/10.1016/j.apacoust.2025.111021>.
- [45] Dafna, E.; Herer, B.; Eberhardt, R.; Tarasiuk, A. Automatic detection of whole night snoring events using audio recordings. *Physiol. Meas.* 2013, 34, 1619–1634.
- [46] Janott, C.; Schmitt, M.; Zhang, Y.; Qian, K.; Pandit, V.; Zhang, Z.; Heiser, C.; Hohenhorst, W.; Herzog, M.; Hemmert, W.; et al. Snoring classified: The Munich-Passau Snore Sound Corpus. *Comput. Biol. Med.* 2018, 94, 106–118. <https://doi.org/10.1016/j.combiomed.2018.01.007>.
- [47] Ng, A.K.; Koh, T.S.; Baey, E.; Lee, T.H.; Abeyratne, U.R.; Puvanendran, K. Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea? *Sleep Med.* 2008, 9, 894–898.
- [48] Huynh, P.K.; Setty, A.R.; Le, T.Q. Koopman spectral analysis of intermittent dynamics in complex systems: A case study in pathophysiological processes of obstructive sleep apnea. *arXiv* 2022, arXiv:2202.12430.

- [49] Kim, T.; Kim, J.W.; Lee, K. Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques. *Biomed. Eng. Online* 2018, 17, 16. <https://doi.org/10.1186/s12938-018-0448-x>.
- [50] Mikami, T.; Ueki, S.; Takahashi, H.; Yonezawa, K. Detecting nonlinear acoustic properties of snoring sounds using Hilbert–Huang Transform. In *Proceedings of the International Conference on Biomedical Electronics and Devices*; SCITEPRESS: Setúbal, Portugal, 2015.
- [51] Majumder, S. A Gaussian mixture model method for eigenvalue-based spectrum sensing with uncalibrated multiple antennas. *Signal Process.* 2022, 192, 108404. <https://doi.org/10.1016/j.sigpro.2021.108404>.
- [52] Sebastian, A.; Cistulli, P.A.; Cohen, G.; de Chazal, P. Automatic classification of OSA-related snoring signals from nocturnal audio recordings. *arXiv* 2021, arXiv:2102.12829.
- [53] Akhter, S.; Bradley, T.D.; Morrell, M.J. Investigation of macro sleep-stage influences on snore characteristics in OSA. *Am. J. Respir. Crit. Care Med.* 2014, 189, A567.
- [54] Janott, C.; Schuller, B.; Heiser, C. Acoustic information in snoring noises. *HNO* 2017, 65, 107–116. <https://doi.org/10.1007/s00106-016-0331-7>.
- [55] Ye, Z.; Peng, J.; Zhang, X.; Song, L. Snoring Sound Recognition Using Multi-Channel Spectrograms. *Arch. Acoust.* 2024, 49, 169–178. <https://doi.org/10.24425/aoa.2024.148775>.
- [56] Hajipour, F.; Moussavi, Z. Spectral and Higher Order Statistical Characteristics of Expiratory Tracheal Breathing Sounds During Wakefulness and Sleep in People with Different Levels of Obstructive Sleep Apnea. *J. Med. Biol. Eng.* 2019, 39, 244–250.
- [57] Kevat, A.; Bernard, A.; Harris, M.A.; Heussler, H.; Black, R.; Cheng, A.; Waters, K.; Chawla, J. Impact of adenotonsillectomy on growth trajectories in preschool children with mild-

moderate obstructive sleep apnea. *J. Clin. Sleep Med.* 2023, 19, 55–62.
<https://doi.org/10.5664/jcsm.10266>.

[58] Cao, S.; Rosenzweig, I.; Bilotta, F.; Jiang, H.; Xia, M. Automatic detection of obstructive sleep apnea based on speech and snoring sounds with AI. *J. Thorac. Dis.* 2024, 16, 2654–2667.

[59] Gonzalez-Martinez, F.D.; Carabias-Orti, J.J.; Canadas-Quesada, F.J.; Ruiz-Reyes, N.; Martinez-Munoz, D.; Garcia-Galan, S. Improving snore detection under limited dataset through harmonic/percussive source separation and convolutional neural networks. *Appl. Acoust.* 2024, 216, 109811.

[60] Nagappa, M.; Liao, P.; Wong, J.; Auckley, D.; Ramachandran, S.K.; Memtsoudis, S.; Mokhlesi, B.; Chung, F. Validation of the STOP-Bang Questionnaire as a Screening Tool for Obstructive Sleep Apnea among Different Populations: A Systematic Review and Meta-Analysis. *PLoS ONE* 2015, 10, e0143697. <https://doi.org/10.1371/journal.pone.0143697>.

[61] Balaei, A.T.; Sutherland, K.; Cistulli, P.A.; de Chazal, P. Automatic Detection of Obstructive Sleep Apnea Using Facial Images. In *Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, VIC, Australia, 18–21 April 2017; pp. 215–218.

[62] Romano, S.; Salvaggio, A.; Hirata, R.P.; Lo Bue, A.; Picciolo, S.; Oliveira, L.V.; Insalaco, G. Upper airway collapsibility evaluated by a negative expiratory pressure test in severe obstructive sleep apnea. *Clinics* 2011, 66, 567–572. <https://doi.org/10.1590/s1807-59322011000400008>.

[63] Kushida, C.A.; Efron, B.; Guilleminault, C. A predictive morphometric model for the obstructive sleep apnea syndrome. *Ann. Intern. Med.* 1997, 127, 581–587.
https://doi.org/10.7326/0003-4819-127-8_part_1-199710150-00001.

- [64] Sola-Soler, J.; Fiz, J.A.; Torres, A.; Jane, R. Identification of Obstructive Sleep Apnea Patients from Tracheal Breath Sound Analysis during Wakefulness in Polysomnographic Studies. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 4232–4235.
- [65] El-Sayed, I.H. Comparison of four sleep questionnaires for screening obstructive sleep apnea. Egypt. J. Chest Dis. Tuberc. 2012, 61, 433–441. <https://doi.org/10.1016/j.ejcdt.2012.07.003>.
- [66] Lee, R.W.; Petocz, P.; Prvan, T.; Chan, A.S.; Grunstein, R.R.; Cistulli, P.A. Prediction of obstructive sleep apnea with craniofacial photographic analysis. Sleep 2009, 32, 46–52.
- [67] Montero Benavides, A.; Fernández Pozo, R.; Toledano, D.T.; Blanco Murillo, J.L.; López Gonzalo, E.; Hernández Gómez, L. Analysis of Voice Features Related to Obstructive Sleep Apnoea and Their Application in Diagnosis Support. Comput. Speech Lang. 2014, 28, 434–452.
- [68] Simply, R.M.; Dafna, E.; Zigel, Y. Diagnosis of Obstructive Sleep Apnea Using Speech Signals From Awake Subjects. IEEE J. Sel. Top. Signal Process. 2020, 14, 251–260.

CHAPTER EIGHT: CUES: A MULTIPLICATIVE COMPOSITE METRIC FOR EVALUATING CLINICAL PREDICTION MODELS THEORY, INFERENCE, AND PROPERTIES

Ali Mohammad Alqudah¹, and Zahra Moussavi^{1,2}

¹ Biomedical Engineering, University of Manitoba, Winnipeg, Canada

² Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada
Mathematics(2026) 14 (3) (<https://doi.org/10.3390/math14030398>)

Abstract

Evaluating artificial intelligence (AI) models in clinical medicine requires more than conventional metrics such as accuracy, Area Under the Receiver Operating Characteristic (AUROC), or F1-score, which often overlook key considerations such as fairness, reliability, and real-world utility. We introduce CUES as a multiplicative composite score for clinical prediction models; it is defined as $CUES = (C \cdot U \cdot E \cdot S)^{1/4}$, where C represents calibration, U integrated clinical utility, E equity across patient subpopulations, and S sampling stability. We formally establish boundedness, monotonicity, and differentiability on the domain $(0,1]^4$, derive first-order sensitivity relations, and provide asymptotic approximations for its sampling distribution via the delta method. To facilitate inference, we propose bootstrap procedures for constructing confidence intervals and for comparative model evaluation. Analytic examples illustrate how CUES can diverge from traditional metrics, capturing dimensions of predictive performance that are essential for clinical reliability but often missed by AUROC or F1-score alone. By integrating multiple facets of clinical utility and robustness, CUES provides a comprehensive tool for model evaluation, comparison, and selection in real-world medical applications.

8.1. Introduction

The integration of artificial intelligence (AI) into clinical medicine holds the promise of revolutionizing diagnostics, prognostics, and treatment planning [1–3]. From interpreting medical images to predicting patient outcomes, AI models are increasingly being developed to support clinicians and improve patient care. However, translating these models from research to routine clinical practice is fraught with challenges, not least the critical task of performance evaluation. The adage “what you measure is what you get” is particularly salient in this high-stakes domain; if we rely on inadequate or incomplete metrics, we risk developing and deploying models that are not only ineffective but potentially harmful [4–6].

For decades, the evaluation of classification models were dominated by a set of standard metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUROC) [7–9]. While these metrics have served well in many traditional machine learning applications, they possess fundamental limitations that render them insufficient for the nuanced requirements of clinical AI. The most prominent limitation is their poor performance on imbalanced datasets, a common characteristic of clinical data where the prevalence of a disease or condition is often low. Accuracy, for example, can be deceptively high for a model that simply predicts the majority class (e.g., “no disease”), offering no real clinical value. Similarly, while AUROC is prevalence-invariant, it measures a model’s ability to rank patients by risk, not whether the risk predictions themselves are accurate or lead to good clinical decisions [10–12]. These limitations of AUROC and accuracy in clinical prediction have been extensively documented in the literature [13,14]; our contribution is not to restate these limitations, but to provide a unified framework that operationalizes them during model evaluation and selection.

8.1.1. The Need for a New Evaluation Metric

Traditional evaluation metrics for clinical AI models, such as accuracy, AUROC, and F1-score, are widely used but have significant limitations when applied to real-world healthcare settings. These metrics are often misleading in the presence of imbalanced datasets, commonly found in clinical data, and fail to reflect the actual clinical value of a model [7,10,11,15]. For instance, a model can achieve high accuracy by simply predicting the majority of the class while completely missing critical but rare cases. Furthermore, these metrics typically ignore essential aspects such as:

- Clinical utility [3]: Do predictions actually result in better patient outcomes?
- Calibration [4]: Are the predicted probabilities trustworthy?
- Equity [16]: Does the model perform fairly across diverse patient groups, or does it exacerbate existing health disparities?
- Stability [17]: Is the model's performance consistent across different samples or environments?

Relying solely on current metrics risks deploying AI solutions that are statistically sound but clinically unsafe, unfair, or unreliable [4–6,18]. Therefore, a new evaluation score is urgently needed, one that incorporates a holistic perspective on a model's fitness for clinical use.

8.1.2. What Makes This Score Different from Current Scores

The new metric CUES (Calibration, Utility, Equity, Stability), fundamentally differs from traditional metrics by addressing all critical dimensions required for trustworthy clinical AI as discussed in [7,11,15,19]; it provides the following:

- Composite design: CUES is an unweighted geometric mean of four core components: utility (clinical usefulness), calibration (probability accuracy), equity (fairness across subgroups), and stability (robustness to sampling).
- Holistic evaluation: Each component is independently normalized, meaning that poor performance in one area (e.g., high bias or instability) will substantially reduce the overall score, preventing one strong aspect from masking significant deficiencies elsewhere.
- Explicit fairness and robustness: By measuring subgroup disparities and bootstrapped stability, CUES directly evaluates areas that traditional metrics ignore, such as algorithmic bias and performance fluctuations [16,17,20,21].

While individual performance measures remain essential for diagnostic insight, many practical tasks in clinical AI, such as model selection, hyperparameter tuning, feature selection, and regulatory approval, require a principled scalar criterion to enable consistent comparison across candidate models. CUES addresses this need by integrating calibration, utility, equity, and stability into a single decision-oriented metric. In practical terms, CUES provides a single, interpretable value that summarizes all essential facets of clinical model performance, something current standards cannot offer, as they often require examining several metrics separately (and sometimes missing critical failures) [1,4,6,8].

In such decision-making settings, reporting multiple metrics without an explicit aggregation rule leaves the final choice subjective and potentially inconsistent. CUES is introduced to address this gap by providing a conservative, decision-oriented composite score that formalizes trade-offs among calibration, clinical utility, equity, and stability. Importantly, CUES is not

intended to replace component-wise evaluation, but to complement it by ensuring that no critical deficiency is overlooked during model comparison.

Although many evaluation strategies have been proposed ranging from accuracy-based metrics and AUROC [7,10–12] to decision-curve analysis for clinical utility [7,13,20,21] and fairness-aware performance measures [3,16,17,20,21], these approaches remain fragmented and often insufficient for the complexities of clinical practice. They frequently overlook critical dimensions such as calibration [2,4], robustness to data shifts [2,17], and subgroup equity [3,16], while imbalanced datasets further limit the reliability of standard metrics [7,10,11,15]. Consequently, there is a clear need for a unified, clinically grounded metric that synthesizes these essential but disparate components. In this work, we introduce CUES, a composite score that integrates utility, calibration, equity, and stability into a single interpretable measure, aiming to provide a more comprehensive and clinically meaningful assessment of model performance. The next section presents the formal definition of the CUES score and its mathematical properties.

8.2. Materials and Methods

In this section, we first describe how each component of the CUES framework is estimated in practice using finite samples. Theoretical properties, including asymptotic distributions and statistical inference, are presented subsequently to clarify the behavior of the proposed estimators.

8.2.1. Overview of the CUES Framework

The four components of CUES calibration, utility, equity, and stability were selected because they represent non-substitutable dimensions of clinical model performance. Calibration ensures the trustworthiness of predicted probabilities; utility reflects real-world clinical benefit; equity safeguards against systematic performance disparities across patient subgroups; and

stability captures robustness to sampling variability and deployment uncertainty. Deficiencies in any one of these dimensions can independently render a clinical model unsafe or misleading, even when performance on the remaining dimensions is strong. It is calculated as a geometric mean of its core components: utility (U), calibration (C), equity (E), and stability (S). The use of a geometric mean ensures that a poor performance in any single component significantly reduces the overall score, preventing a high score in one area from masking deficiencies in another. Let a predictive model produce probability estimates $\hat{p}_i \in [0,1]$ for instances $i = 1, \dots, n$ with true binary outcomes $y_i \in \{0,1\}$. Let $\mathcal{G} = \{G_1, \dots, G_m\}$ be a (possibly empty) partition of the population into predefined subgroups used for the equity assessment. Although CUES produces a single composite score, it is intended to be interpreted jointly with its constituent components. Component-wise reporting of calibration, utility, equity, and stability is essential for diagnostic interpretation and enables direct identification of the primary contributors to be reduced composite performance. Define four normalized component scores

$$C, U, E, S \in [0,1] \tag{8-1}$$

where each component is a function of the model outputs and outcomes (and subgroup assignment where relevant):

$$C = C(\{\hat{p}_i, y_i\}_{i=1}^n) \tag{8-2}$$

$$U = U(\{\hat{p}_i, y_i\}_{i=1}^n) \tag{8-3}$$

$$E = E(\{\hat{p}_i, y_i, G\}) \tag{8-4}$$

$$S = S(\{\hat{p}_i, y_i\}_{i=1}^n) \tag{8-5}$$

here, C is the calibration score, U the clinical utility score, E the equity (fairness) score across G , and S the stability (robustness) score. Each component is normalized to lie in $[0,1]$ (1 = best possible performance, 0 = worst or baseline performance) by the normalization procedures

described in the methods. Equal weighting of the four components is adopted as a neutral default in the absence of application-specific priorities. This choice does not imply that all dimensions are universally equally important but rather reflects a conservative stance in which no critical dimension is assumed to be secondary by default. The composite CUES score is defined as the fourth root (geometric mean) of the four components:

$$\text{CUES} = (C(\hat{\mathbf{p}}, \mathbf{y}) \cdot U(\hat{\mathbf{p}}, \mathbf{y}) \cdot E(\hat{\mathbf{p}}, \mathbf{y}, \mathcal{G}) \cdot S(\hat{\mathbf{p}}, \mathbf{y}))^{1/4} \quad 8-6$$

where $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$. The aggregation of CUES components is based on the geometric mean, which can be viewed as a special case of the generalized mean family. The framework naturally supports weighted aggregation, allowing users to specify component weights to reflect clinical, ethical, or regulatory priorities.

Arithmetic mean can mask severe deficiencies [22], and the minimum operator is excessively stringent [23]. Harmonic mean [24], as another common measure, places disproportionate weight on the smallest component, a well-known property in multi-criteria aggregation which can result in excessive score attenuation under moderate imbalance. Compared to these measures, the geometric mean provides a principled compromise. It is scale invariant, zero-preserving (i.e., any zero-valued component yields a zero composite score), and enforces proportional penalization of deficiencies without allowing any single component to dominate the overall evaluation. In addition, the geometric mean therefore offers a more stable and interpretable trade-off between sensitivity to localized weaknesses and robustness of the overall assessment.

Detailed mathematical definitions and the normalization mappings for C , U , E , and S follow in the methods section, and proof sketches for all theoretical propositions are provided in Appendix C. While the specific definitions of clinical utility functions, subgroup partitions for equity, and perturbation schemes for stability are necessarily application-dependent, CUES is intentionally

designed as a modular framework. Each component can be adapted to reflect clinically meaningful priorities without altering the mathematical structure of the composite score.

To promote reproducibility, we recommend default choices where applicable: (i) decision-curve-based net benefit integrated over clinically relevant thresholds for utility, (ii) subgroup definitions based on established demographic or clinical variables for equity, and (iii) bootstrap resampling for stability assessment. Sensitivity analyses across reasonable design choices are encouraged and can be used to assess the robustness of CUES-based model comparisons.

8.2.1.1. Assumptions and conventions

Each component is mapped/scaled to $[0,1]$.

If any component is undefined for a dataset, the corresponding mapping must be specified (e.g., define $E = 1$ if no subgrouping is used).

Logarithms use the natural base when needed.

State precisely:

Domain: $(C, U, E, S) \in [0,1]^4$.

Range: CUES $\in [0,1]$.

8.2.2. Estimation of Individual Components

This subsection describes the practical estimation of each CUES component using observed data. All component scores are computed directly from model predictions and outcomes without reliance on asymptotic approximations. Theoretical properties of the resulting estimators are discussed separately in Section 8.2.3.

8.2.2.1. Utility (U): Integrated Net Benefit

Utility quantifies the clinical usefulness of a model's predictions across a range of decision thresholds. It is derived from Decision Curve Analysis (DCA), which evaluates diagnostic and prognostic models by plotting the net benefit against threshold probability [7,15,25,26]. Net benefit considers the balance between true positives and false positives, weighted by the relative harm of false positives compared to false negatives [5,11,15,26]. For a given threshold t , the net benefit (NB_t) is calculated as

$$NB(t) = \frac{TP(t)}{N} - \frac{FP(t)}{N} \cdot \frac{t}{1-t} \quad 8-7$$

where TP and FP are the number of true and false positives, respectively; N is the total number of patients, and t is the threshold probability representing the minimum probability of the disease under the treatment.

To obtain a single utility score, we integrate the net benefit over a range of clinically relevant thresholds. The integrated net benefit (INB) is calculated as the area under the net benefit curve, normalized by the maximum possible net benefit [25,27]. We use weighted integration to allow for the prioritization of certain thresholds, defined by a weighting function $w(t)$.

$$INB_{\text{model}} = \int_{t_{\min}}^{t_{\max}} NB(t) w(t) dt \quad 8-8$$

The utility score, U , is then normalized by the integrated net benefit of a perfect model (INB_{perfect}) and a baseline model (INB_{baseline}) that is typically a strategy of treating all or none. In Decision Curve Analysis (DCA), a perfect model assigns a predicted probability of 1 to every true positive case and 0 to every true negative case. Therefore, at any threshold t , the perfect model identifies all P positive cases ($TP = P$) and never recommends treatment for negative cases ($FP = 0$). Its net benefit is

$$\text{NB}_{\text{perfect}}(t) = \frac{P}{N} \quad 8-9$$

which equals the prevalence and forms a horizontal reference line in DCA plots.

For comparison, the net benefit of a “treat all” strategy is

$$\text{NB}_{\text{all}}(t) = \frac{P}{N} - \frac{N - P}{N} \cdot \frac{t}{1 - t} \quad 8-10$$

and the “treat none” strategy has $\text{NB}_{\text{none}}(t) = 0$, which we use as the baseline for utility normalization. Using these definitions, the integrated net benefit (INB) of any model is computed as

$$U = \frac{\text{INB}_{\text{model}} - \text{INB}_{\text{baseline}}}{\text{INB}_{\text{perfect}} - \text{INB}_{\text{baseline}} + \epsilon}, U \leftarrow \text{clip}(U, 0, 1) \quad 8-11$$

where ϵ (e.g., 10^{-6}) prevents division by zero when the denominator is very small, ensuring $U \in [0, 1]$. The integrated net benefit of a model is computed as a weighted integral over decision thresholds $t \in [0, 1]$:

$$\text{INB}_{\text{model}} = \int_0^1 \text{NB}_{\text{model}}(t) w(t) dt \quad 8-12$$

with analogous definitions for $\text{INB}_{\text{baseline}}$ and $\text{INB}_{\text{perfect}}$. Here, $w(t)$ is a weighting function that determines the relative contribution of each threshold. While a uniform weighting ($w(t) = 1$) treats all thresholds equally, it can be customized commonly using a Beta distribution, $w(t) \propto \text{Beta}(t; \alpha, \beta)$ to emphasize thresholds that are clinically most relevant (e.g., low-risk screening, high-specificity decision points, or intermediate actionable ranges) [28,29].

8.2.2.2. Calibration (C): Brier Skill Score

Calibration assesses the agreement between predicted probabilities and observed outcomes. A well-calibrated model is essential for clinical trust and decision-making. We use the

Brier Score (BS) as a measure of calibration, which is the mean squared difference between predicted probabilities and actual outcomes [4,5,12,19,30]:

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2 \quad 8-13$$

where p_i is the predicted probability for instance i ; y_i is the true binary outcome, for instance i (0 or 1). A lower Brier Score indicates better calibration. To normalize this into a score between 0 and 1 (where 1 is perfect), we use the Brier Skill Score (BSS), which compares the model's Brier Score to that of a reference model [12,31,32], typically a model that always predicts the prevalence of the positive class (BS_{ref}):

$$BS_{ref} = \pi(1 - \pi) \quad 8-14$$

The calibration component C is then defined as

$$C = \max\left(0, 1 - \frac{BS}{BS_{ref} + \epsilon}\right) \quad 8-15$$

This ensures that $C = 1$ for a perfectly calibrated model and $C = 0$ for a model that performs no better than predicting the prevalence [2,4,8,9].

8.2.2.3. Equity (E): Worst-Case Disparity with Variance Penalty

Equity measures the fairness of the model's utility across different predefined subgroups. It is crucial to ensure that AI models do not exacerbate existing health disparities by performing poorly on specific demographic groups [3,6,16,17]. Our equity component E is based on the utility scores calculated for each subgroup. The equity component E It is based on the subgroup utilities. In the implementation provided, the dataset is divided into two subgroups according to the median of the first feature of the test set ($X_{test}[:,0]$):

Subgroup 1: U_1 = utility score for all samples with $X_{test,i,0} \leq \text{median (first partition)}$.

Subgroup 2: U_2 = utility score for all samples with $X_{\text{test},i,0} > \text{median}$ (second partition).

The equity score is then defined as

$$E = \max(0, \min(1, 1 - |U_1 - U_2|)) \quad 8-16$$

This ensures that $E \approx 1$. When subgroup utilities are similar, they decrease toward zero as the disparities between them grow. If no subgrouping is performed (i.e., $X_{\text{test}} = \emptyset$), equity defaults to $E = 1.0$ indicating perfect equity by assumption [11,16–18].

8.2.2.3.1. Extended Equity (E) for Demographic and Anthropometric Fairness

The equity (E) component of CUES quantifies model fairness across clinically meaningful subgroups, accounting for anthropometric and demographic covariates such as age, sex, BMI, neck circumference, and body fat distribution [17]. In its simplest form, E measures the complement of the absolute difference in utility (U) [33,34] between the following two subgroups:

$$E = 1 - |U_{g_1} - U_{g_2}| \quad 8-17$$

where g_1 and g_2 represent subgroups of a stratification variable (e.g., median split of a continuous attribute). This formulation ensures that a model achieving equivalent clinical utility across subgroups obtains a maximal $E = 1$, while significant disparities reduce the fairness score.

However, in medical and physiological domains such as obstructive sleep apnea (OSA) screening, patient variability is strongly influenced by anthropometric and demographic confounders, including age, sex, body mass index (BMI), neck circumference, and body fat distribution [35]. These parameters are not mere nuisance variables but are clinically meaningful determinants of both disease manifestation and signal characteristics (e.g., breathing sounds, airway dynamics). Consequently, fairness should not be assessed solely on abstract data partitions but should explicitly account for population heterogeneity arising from these biological factors.

To generalize E in this context, the equity component can be reformulated as the mean subgroup consistency in clinical utility across K . Predefined demographic or anthropometric strata:

$$E = 1 - \frac{1}{K-1} \sum_{k=1}^{K-1} |U_k - \bar{U}| \quad 8-18$$

where U_k denotes the utility within the subgroup k (e.g., male, female, young, old, high-BMI, low-BMI), and $\bar{U}$ is the overall mean utility. This expression quantifies how uniformly a model performs across clinically relevant population partitions. A value of $E \approx 1$ implies consistent predictive utility across subgroups, indicating demographic robustness and low bias.

8.2.2.4. Stability (S): Normalized Confidence Interval

Stability assesses the model's robustness to sampling variability. A stable model yields consistent utility scores across different bootstrapped samples of the data. We quantify stability by repeatedly bootstrapping the dataset, calculating the utility score for each bootstrap sample, and then examining the variability of these utility scores [20,21,36]. Let $U^{(b)}$ denote the utility score obtained from the b , the bootstrap sample. We calculate the coefficient of variation (CV) of the bootstrapped utility scores as follows:

$$CV = \frac{SD(U^{(b)})}{\text{mean}(U^{(b)}) + \epsilon} \quad 8-19$$

where $SD(U^{(b)})$ is the standard deviation of the bootstrapped utility scores, $\text{mean}(U^{(b)})$ is their mean, and ϵ is a small constant (e.g., 10^{-6}) to prevent division by zero if the mean utility is very close to zero. A lower CV indicates higher stability [20,27,37]. The stability score S is then defined using an exponential decay function to map CV to a $[0, 1]$ range as follows:

$$S = \exp(-\lambda \cdot CV) \quad 8-20$$

where λ It is a sensitivity parameter that controls how quickly the score drops with increasing variability. A higher λ Makes the score more sensitive to instability. This formulation ensures that $S = 1$. For perfect stability ($CV = 0$), it approaches zero as CV increases.

8.2.3. Practical Estimation Algorithm for CUES

For clarity and reproducibility, we summarize the practical computation of the CUES score in Algorithm 8-1. This algorithm describes the finite-sample estimation procedure independently of theoretical or asymptotic considerations.

Algorithm 8-1 Estimation of the CUES score.

1. Train the prediction model and obtain predicted probabilities $\hat{p}_i$ on the evaluation dataset.
 2. Compute clinical utility U using integrated net benefit via decision curve analysis.
 3. Compute calibration score C using the Brier Skill Score.
 4. Compute subgroup-specific utility scores and derive the equity score E .
 5. Perform bootstrap resampling to estimate stability S .
 6. Normalize all component scores to lie in $[0, 1]$.
 7. Compute the composite CUES score using the geometric mean in Equation (6).
-

8.2.4. CUES Curve and Visualization as Interpretation

The CUES Curve provides a graphical representation of an AI model's clinical utility across a continuum of decision thresholds, normalized to reflect meaningful benefit relative to baseline and a perfect model. This visualization complements the composite CUES score by enabling practitioners to examine how model performance varies with changes in the intervention threshold, a crucial step for informed clinical decision-making. The CUES curve is constructed through the following steps:

8.2.4.1. Net Benefit Computation

For each threshold value, the function calculates the net benefit for:

The model: using its predicted probabilities.

Treat-all strategy: as if all patients are predicted to be positive.

Treat-none strategy: as if no patients are predicted to be positive.

Perfect model: using true labels as predictions.

8.2.4.2. Normalization

The model's net benefit is normalized relative to the best baseline (treat-all or treat-none) and the perfect possible benefit at each threshold as follows:

$$\text{Normalized Net Benefit} = \max \left(0, \frac{(\text{nb}_{\text{model}} - \text{nb}_{\text{baseline}})}{(\text{nb}_{\text{perfect}} - \text{nb}_{\text{baseline}} + \epsilon)} \right) \quad 8-21$$

where ϵ is a small constant to avoid division by zero.

8.2.4.3. Area Under the CUES Curve (AUC-CUES)

The area under this normalized benefit curve across all reasonable decision thresholds is computed (using trapezoidal integration). This value reflects the overall utility efficiency of the model over the full clinical decision spectrum:

$$\text{AUC} - \text{CUES} = \int_{\text{thresholdmin}}^{\text{thresholdmax}} \text{Normalized Net Benefit} \quad 8-22$$

A higher AUC-CUES indicates broader, more reliable clinical benefit [8–10].

8.2.4.4. Visualization

Thresholds where the model provides maximal clinical benefit. Thresholds where the model underperforms or offers no gain over basic strategies. The overall consistency of the model's utility informs clinicians about advisable operating points.

8.2.4.5. Theoretical Comparison with Existing Metrics (Decoupling from AUROC)

High AUROC does not imply high CUES—AUROC measures ranking ability and is prevalence-invariant. Calibration C (e.g., Brier Skill Score) and utility U (integrated net benefit) can be low. At the same time, AUROC is high: consider a model that perfectly ranks but outputs probabilities compressed near 0 and 1, resulting in incorrect magnitudes (mispredictions). Since CUES multiplies calibration and utility, poor C or U will substantially lower the final score, even if AUROC is near 1 [4,11,15,19]. Provide explicit construction in Appendix C (e.g., scale logit outputs by a small factor to keep ordering but destroy calibration).

Theorem (equivalence condition)—if a model is perfectly calibrated ($C = 1$), achieving perfect integrated net benefit under the considered thresholds ($U = 1$), and is entirely equitable across subgroups ($E = 1$) and stable ($S = 1$), then $CUES = 1$. Conversely, if $CUES = 1$, then each component must be equal to 1, implying direct from definition and boundedness.

8.2.4.6. Interpretation

The CUES curve serves as a practical and interpretable tool for visualizing and quantifying how model predictions translate into actionable clinical benefits across a range of threshold settings. It directly supports model selection and threshold optimization by spotlighting both strengths and deficiencies that may be missed by scalar metrics alone. This approach grounds evaluation not just in point estimates but in a truly holistic understanding of clinical AI performance [8,9,36].

8.2.5. Gradient-Based Optimization of CUES (Optional)

Optimization of CUES is optional and not required for model evaluation. In this work, CUES is primarily used as a post hoc evaluation metric; optimization is discussed here for completeness and potential future extensions. If one wishes to optimize CUES directly (for example, during model training), the gradients of CUES with respect to the model parameters θ can be computed using the chain rule:

$$\frac{\partial \text{CUES}}{\partial \theta} = \sum_{X \in \{C, U, E, S\}} \frac{\partial \text{CUES}}{\partial X} \cdot \frac{\partial X}{\partial \theta} \quad 8-23$$

with

$$\frac{\partial \text{CUES}}{\partial X} = \frac{\text{CUES}}{4X} \quad 8-24$$

thus,

$$\nabla_{\theta} \text{CUES} = \frac{\text{CUES}}{4} \sum_X \frac{1}{X} \nabla_{\theta} X \quad 8-25$$

where components X may themselves be non-differentiable or estimated via complex procedures; in that case, surrogate differentiable proxies can be used.

8.2.5.1. Convexity

The mapping $(C, U, E, S) \mapsto (\text{CUES})^{1/4}$ is not convex on $[0,1]^4$. Consequently, direct optimization is non-convex and may have multiple local maxima. Use stochastic gradient methods, simulated annealing, or surrogate losses.

8.2.5.2. Practical approach

Rather than end-to-end optimizing CUES, one can optimize component proxies (e.g., Brier loss for calibration, decision-curve–informed loss for utility) and combine them in a multi-

objective framework. While direct end-to-end optimization of CUES with respect to model parameters is ideal in theory, it is often infeasible in practice given the non-differentiability and non-convexity of several CUES components.

As an alternative, we adopt a pragmatic hyperparameter-tuning strategy in which CUES is evaluated as an out-of-fold metric via grid search and cross-validation. This approach does not require a closed-form gradient or direct parameter-level optimization but still enables principled model and hyperparameter selection via the CUES criterion. By comparing cross-validated optimization of CUES and traditional accuracy, we empirically assess the impact of the selected objective on both overall model performance and utility component profiles. This methodology complements the theoretical framework while being immediately applicable across a broad range of standard machine learning models.

8.2.6. Theoretical Properties and Statistical Inference

Although CUES is implemented using finite-sample estimators and bootstrap-based procedures, the following theoretical results are included to establish mathematical validity, consistency, and inferential grounding. These results are not required for practical computation of CUES, but justify uncertainty quantification, hypothesis testing, and principled comparison between models.

8.2.6.1. Boundedness

For $C, U, E, S \in [0,1]$, $0 \leq \text{CUES} \leq 1$. The product $P = \text{CUES}^4 = CUES$ is in $[0,1]$ because each factor is in $[0,1]$. Taking the fourth root preserves the interval: $0 \leq P^{1/4} \leq 1$.

8.2.6.2. Zero Sensitivity

If any component equals 0, then CUES = 0. If $C = 0$ (or any factor zero) then $CUES = 0$; thus, the fourth root is 0.

8.2.6.3. Monotonicity

CUES is strictly increasing in each component on $(0,1]$. Formally, for fixed $U, E, S > 0$, $\partial CUES / \partial C > 0$. For $C > 0$,

$$\frac{\partial CUES}{\partial C} = \frac{1}{4} (CUES)^{-3/4} \cdot (UES) = \frac{CUES}{4C} > 0 \quad 8-26$$

Analogous expressions hold for U, E, S . Thus, increasing any component increases CUES.

8.2.6.4. Relative Sensitivity

$$\frac{\partial \log (CUES)}{\partial \log C} = \frac{1}{4} \quad 8-27$$

Hence, a $x\%$ relative change in any component produces approximately a $\frac{x}{4}\%$ relative change in CUES (first-order).

$$\log CUES = \frac{1}{4} (\log C + \log U + \log E + \log S) \quad 8-28$$

8.2.6.5. Continuous Mapping Theorem

If estimators $\hat{C}_n, \hat{U}_n, \hat{E}_n, \hat{S}_n$ are (pointwise) consistent for C, U, E, S as $n \rightarrow \infty$, then

$$\widehat{CUES}_n = (\hat{C}_n \hat{U}_n \hat{E}_n \hat{S}_n)^{1/4} \rightarrow p (CUES)^{1/4} = CUES \quad 8-29$$

Continuous mapping theorem: the fourth root of a convergent product converges to the fourth root of the limit.

8.2.6.6. Asymptotic Distribution (Delta Method Sketch)

To assess the variability and derive approximate confidence intervals for the CUES score, we consider its asymptotic distribution. Since CUES is a non-linear function of the four component estimates $(\hat{C}, \hat{U}, \hat{E}, \hat{S})$, standard multivariate central limit results do not directly apply. The delta method provides a convenient way to approximate the sampling distribution of a smooth function of asymptotically normal estimators. Suppose $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal for $\hat{\theta}_n = (\hat{C}, \hat{U}, \hat{E}, \hat{S})$ with covariance Σ . Define $g(\theta) = (\text{CUES})^{1/4}$ [20,37]. Then, by the delta method

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} N(0, \nabla g(\theta)^\top \Sigma \nabla g(\theta)) \quad 8-30$$

where

$$\nabla g(\theta) = \left(\frac{g(\theta)}{4C}, \frac{g(\theta)}{4U}, \frac{g(\theta)}{4E}, \frac{g(\theta)}{4S} \right)^\top \quad 8-31$$

This yields an approximate standard error for $\widehat{\text{CUES}}$ and facilitates hypothesis testing.

8.2.6.7. Bootstrap-Based Inference

Because component estimators are often complex (DCA integrals, bootstrapped stability, subgroup utilities), a non-parametric bootstrap (resampling observations with replacement, recomputing C, U, E, S and hence CUES) yields an empirical distribution for $\widehat{\text{CUES}}$. Use percentile or bias-corrected accelerated (BCa) intervals for Cis [20,21,36].

8.2.6.8. Statistical Testing and Confidence Statements

8.2.6.8.1. Bootstrap confidence intervals for CUES

Resample the dataset B times; for each bootstrap b compute $\text{CUES}^{(b)}$. Use percentile CI: $[\text{CUES}_{(\alpha/2)}, \text{CUES}_{(1-\alpha/2)}]$ [20,21,36].

8.2.6.8.2. Hypothesis test for difference between two models (paired bootstrap)

Null H_0 : $\text{CUES}_A = \text{CUES}_B$ [27,36,37].

For each bootstrap sample, compute $\Delta^{(b)} = \text{CUES}_A^{(b)} - \text{CUES}_B^{(b)}$. Estimate p-value as proportion of $\Delta^{(b)}$ at least as extreme as observed Δ . Use a two-sided test or compute BCa intervals on Δ [20,21,36].

Error propagation (first order)—using the delta method:

$$\text{Var}(\widehat{\text{CUES}}) \approx \nabla g(\theta)^\top \text{Cov}(\hat{\theta}) \nabla g(\theta) \quad 8-32$$

This expresses how uncertainty in components influences CUES variance. CUES is intended to be interpreted jointly with its component scores, which must always be reported to preserve transparency and diagnostic insight.

8.2.7. CUES as Feature Selection as Application

The CUES framework, initially designed to evaluate model performance across calibration (C), utility (U), equity (E), and stability (S), can also be used as a feature selection method. Unlike conventional feature importance approaches that primarily focus on predictive power (e.g., coefficients in linear models or tree-based importances), CUES decomposes performance into interpretable components, providing a multi-faceted measure of feature contribution.

To perform feature selection, we computed feature-wise contributions to each CUES component and to the overall CUES score. This was achieved by performing multiple perturbations of individual features in the test set (e.g., via shuffling several times) and measuring the resulting change in CUES for each permutation. This procedure generates a distribution of CUES contributions per feature, capturing the variability and robustness of each feature's impact on model performance. Features causing a larger drop in CUES or its components across

permutations are considered more critical for maintaining performance, not only in terms of accuracy but also fairness (equity) and robustness (stability). Formally, let C_i, U_i, E_i, S_i denote the change in the respective CUES components when feature i is perturbed, and let CUES_i denote the corresponding change in the overall CUES:

$$\text{CUES}_i = (C - C_i)^{1/4} \cdot (U - U_i)^{1/4} \cdot (E - E_i)^{1/4} \cdot (S - S_i)^{1/4} \quad 8-33$$

Features are ranked based on their mean or median impact on overall CUES across permutations, and this ranking can be used to select the top-k features for model training. Compared with traditional methods such as SHAP, CUES-based feature selection incorporates multi-objective criteria, ensuring that selected features simultaneously contribute to predictive performance, robustness, and fairness. For each classifier, we computed component-wise and overall CUES contributions per feature across multiple perturbations, allowing direct comparison with conventional explainable AI metrics (e.g., SHAP) and visualizing distributions using violin plots. This highlights features that are critical not only for predictive accuracy but also for equitable and stable model behavior.

8.2.8. Computational Considerations and Scalability

The computational complexity of the CUES framework scales linearly with the number of samples and evaluated subgroups, with additional overhead introduced by bootstrap-based stability estimation and perturbation analyses. In practice, we find that a moderate number of bootstrap iterations (e.g., 100–200) is sufficient to obtain stable estimates of all CUES components, consistent with prior work [13,14] on resampling-based uncertainty estimation.

For large-scale datasets, computational burden can be reduced through stratified subsampling, parallelized bootstrap procedures, or early stopping criteria once component estimates converge. Approximate calibration methods, such as fixed-bin reliability diagrams or

isotonic regression with constrained knots, further reduce runtime without materially affecting CUES rankings. Importantly, all CUES components are inherently parallel and can be efficiently implemented on modern multicore or GPU-enabled systems, making the framework scalable to real-world clinical datasets. Table 8-1 summarizes the computational cost associated with each component of the CUES framework.

Table 8-1 Computational complexity and scalability of CUES components.

Component	Primary Operation	Time Complexity	Parallelizable
Utility (U)	Decision Curve Analysis across decision thresholds	$O(N \times T)$	Yes
Calibration (C)	Reliability diagram/calibration error computation	$O(N)$	Yes
Equity (E)	Subgroup-wise utility estimation	$O(N \times G)$	Yes
Stability (S)	Bootstrap resampling (B iterations)	$O(B \times N)$	Yes
CUES Aggregation	Geometric mean of normalized components	$O(1)$	Yes

N denotes the number of samples, T the number of decision thresholds, G the number of predefined subgroups, and B the number of bootstrap iterations. Time complexity is expressed using Big-O notation $O(\cdot)$, which characterizes asymptotic computational cost. All CUES components are parallelizable because computations are independent across samples, thresholds, subgroups, or bootstrap iterations, enabling efficient execution on multicore or GPU-enabled systems.

8.2.9. Datasets and Experiment Setup

To evaluate the CUES framework, multiple publicly available datasets were used to cover both binary and multi-class classification scenarios. All datasets were preprocessed using

standardized procedures, including missing-value imputation, one-hot encoding for categorical variables, and feature scaling with StandardScaler. The datasets include the following:

- **Breast cancer dataset (binary):** Obtained from sklearn datasets, containing 569 samples with 30 numeric features distinguishing malignant and benign cases.
- **Pima diabetes dataset (binary):** Retrieved from OpenML (machine learning), consisting of 768 samples with eight clinical features. The target variable indicates whether a patient tested positive for diabetes.
- **Heart disease dataset (binary and multi-class):** Sourced from OpenML (Cleveland dataset). The binary version distinguishes between low (<50% stenosis) and high (>50% stenosis) disease severity, whereas the multi-class version extends to five ordered categories. Mixed numeric and categorical features were handled using median imputation and one-hot encoding.
- **Dermatology dataset (multi-class):** Obtained from OpenML, comprising 366 patient samples and 34 clinical attributes for the classification of six dermatological conditions. Missing numeric values were filled with column medians.

All datasets were divided into training and testing subsets using repeated stratified k-fold cross-validation (five folds, two repeats, random state = 42) to ensure stable performance estimates. All experiments were conducted in Python 3.10 using the scikit-learn library. The provided implementation automates data loading, preprocessing, model training, metric computation, bootstrapping, and visualization of CUES components. Four standard classifiers were evaluated across both binary and multi-class settings:

- Logistic Regression (LR): a linear model serving as a strong, interpretable baseline.

- Support Vector Machine (SVMrbf): a non-linear classifier using a radial basis function kernel.
- K-Nearest Neighbors (KNN): a distance-based algorithm capturing local data structure ($k = 7$).
- Bagging Ensemble Classifier (Bagging): is an ensemble of base estimators that improves variance reduction and stability.

Each model was trained and tested using repeated stratified k-fold cross-validation (five folds, two repeats, random seed = 42) to ensure robust performance estimates. Data preprocessing included imputing missing values, one-hot encoding for categorical features, and standard scaling for all numerical features. The evaluation covered two main tasks:

- Within-dataset validation, where classifiers were trained and tested on the same dataset using cross-validation.
- Cross-dataset generalization, where models trained on one dataset were evaluated on another to assess transferability and robustness to feature-space alignment.

Performance was measured using both traditional metrics (AUROC, AUPRC (Area Under the Precision–Recall Curve), accuracy, sensitivity, specificity) and the proposed CUES components (C, U, E, S). Bootstrapping ($n = 200$) was applied to compute the stability component (S) and derive confidence-adjusted CUES estimates. All numerical results were automatically exported as comma-separated values (CSV) tables, and visualizations (CUES curves, calibration plots, and component bar charts) were saved for each model–dataset pair.

8.3. Validation and Results

Overall, the experimental results validate the effectiveness of the proposed CUES (calibration, utility, equity, and stability) framework in providing a more holistic evaluation of AI models for clinical applications. Across all binary and multi-class classification tasks, CUES consistently captured dimensions of model performance such as calibration reliability, clinical utility, fairness, and stability that conventional metrics like AUROC and accuracy often failed to represent.

8.3.1. Binary Classification Performance

The results for binary classifications are summarized in Tables 8-2 and 8-3 demonstrating the superior performance of Logistic Regression across all datasets, particularly for the breast cancer dataset, which achieved a mean CUES score of 0.903, indicating excellent calibration and clinical utility. In contrast, while the diabetes dataset achieved an AUROC $\approx$ of 0.83, its corresponding CUES score dropped to 0.50, reflecting poor calibration and limited real-world benefit despite high discriminative ability. The heart disease dataset showed intermediate CUES values ($\sim$ 0.61), suggesting reasonable discrimination but reduced stability and equity between subgroups. Figures 8-1 to 8-3 show the normalized net benefit curves, calibration curves, and CUES components for various scenarios, with classifiers shown.

Table 8-3 compares multiple classifiers across three datasets using both traditional metrics and the proposed CUES score. Logistic Regression consistently achieved the highest CUES values, particularly for the Breast Cancer dataset (0.903), indicating strong calibration, clinical utility, fairness, and stability. In contrast, the diabetes dataset shows high AUROC ($\sim$ 0.83) but a substantially lower CUES score (0.473), highlighting poor calibration and limited real-world

benefit despite good discrimination. Heart disease models achieved intermediate CUES scores (~0.61), reflecting moderate calibration and some subgroup disparities. Across classifiers, ensemble (Bagging) and distance-based (KNN) methods exhibited slightly lower CUES values than Logistic Regression and SVM, likely due to sensitivity to sampling variation, as reflected in the stability component. Notably, the “AUC CUES” column summarizes performance across thresholds, providing additional insight beyond the single CUES score. These results underscore that CUES adds interpretive power by revealing clinical trustworthiness and fairness that traditional metrics alone may obscure.

These findings underscore the added interpretive power of CUES, whereas traditional performance measures (e.g., AUROC and accuracy) would classify both the breast cancer and diabetes models as “strong performers,” CUES exposes critical weaknesses in clinical trustworthiness and fairness. Furthermore, across all classifiers, ensemble and distance-based methods (Bagging and KNN) exhibited slightly lower CUES values than Logistic Regression and SVM, likely due to their sensitivity to sampling variation, as reflected in lower stability (S) components.

Table 8-2 Binary classification: multiple classifiers CUES performance.

Dataset	Classifier	CUES Score	C	U	E	S	AUC CUES
Breast Cancer	Logistic	0.903	0.914	0.900	0.862	0.941	0.900
Breast Cancer	SVMrbf	0.909	0.909	0.894	0.894	0.942	0.894
Breast Cancer	Bagging	0.858	0.858	0.834	0.813	0.936	0.834
Breast Cancer	KNN	0.866	0.870	0.846	0.824	0.932	0.846
Diabetes	Logistic	0.473	0.302	0.246	0.933	0.732	0.246
Diabetes	SVMrbf	0.443	0.276	0.220	0.897	0.718	0.220
Diabetes	Bagging	0.446	0.278	0.224	0.913	0.715	0.224
Diabetes	KNN	0.383	0.204	0.169	0.944	0.705	0.169

Dataset	Classifier	CUES Score	C	U	E	S	AUC CUES
Heart Disease	Logistic	0.614	0.525	0.474	0.801	0.746	0.474
Heart Disease	SVMrbf	0.602	0.487	0.429	0.859	0.756	0.429
Heart Disease	Bagging	0.587	0.460	0.413	0.849	0.760	0.413
Heart Disease	KNN	0.589	0.473	0.416	0.841	0.746	0.416

Table 8-3 Binary classification: multiple classifiers performance comparison between CUES and traditional metrics.

Dataset	Classifier	CUES Score	AUC CUES	AUROC	AUPRC	Accuracy	Sensitivity	Specificity
Breast Cancer	Logistic	0.903	0.900	0.995	0.997	0.976	0.990	0.953
Breast Cancer	SVMrbf	0.909	0.894	0.995	0.997	0.969	0.975	0.960
Breast Cancer	Bagging	0.858	0.834	0.988	0.989	0.961	0.975	0.939
Breast Cancer	KNN	0.866	0.846	0.987	0.986	0.965	0.989	0.925
Diabetes	Logistic	0.473	0.246	0.830	0.719	0.768	0.565	0.877
Diabetes	SVMrbf	0.443	0.220	0.823	0.711	0.762	0.575	0.862
Diabetes	Bagging	0.446	0.224	0.819	0.698	0.756	0.616	0.831
Diabetes	KNN	0.383	0.169	0.780	0.624	0.727	0.543	0.826
Heart Disease	Logistic	0.614	0.474	0.911	0.907	0.845	0.786	0.894
Heart Disease	SVMrbf	0.602	0.429	0.896	0.886	0.835	0.800	0.864
Heart Disease	Bagging	0.587	0.413	0.890	0.884	0.818	0.771	0.858
Heart Disease	KNN	0.589	0.416	0.889	0.850	0.832	0.808	0.852

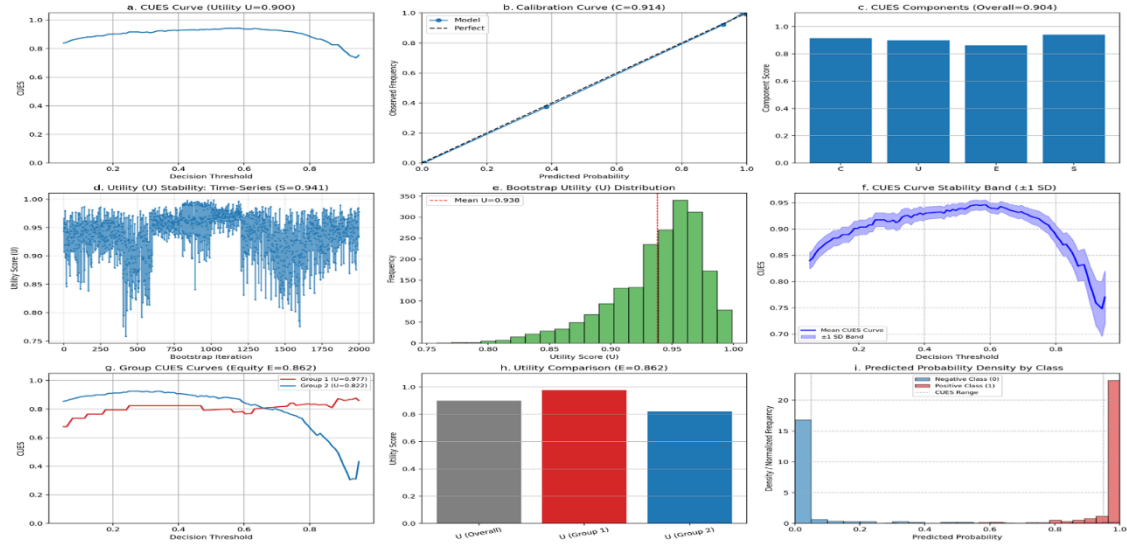


Fig. 8-1 CUES component visualization for breast cancer (Logistic Regression). Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, U) across decision thresholds; (b) reliability diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U ; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.

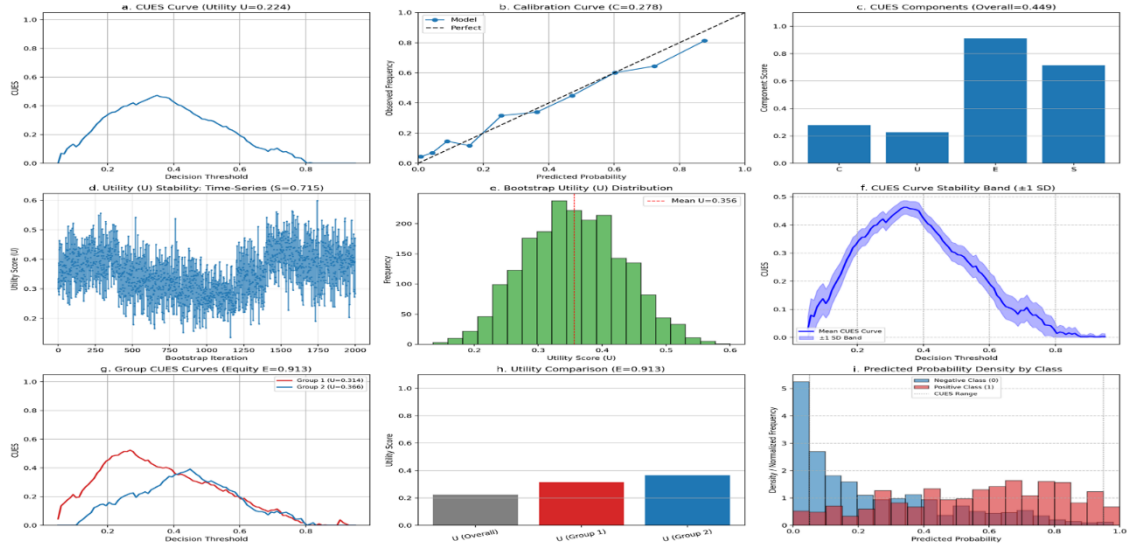


Fig. 8-2 CUES component visualization for diabetes (Bagging). Each panel summarizes complementary aspects of model performance: (a) Normalized net benefit (utility curve, U) across decision thresholds; (b) reliability diagram showing calibration performance (C); (c) bar chart of CUES components (C, U, E, S) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of U ; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (E); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.

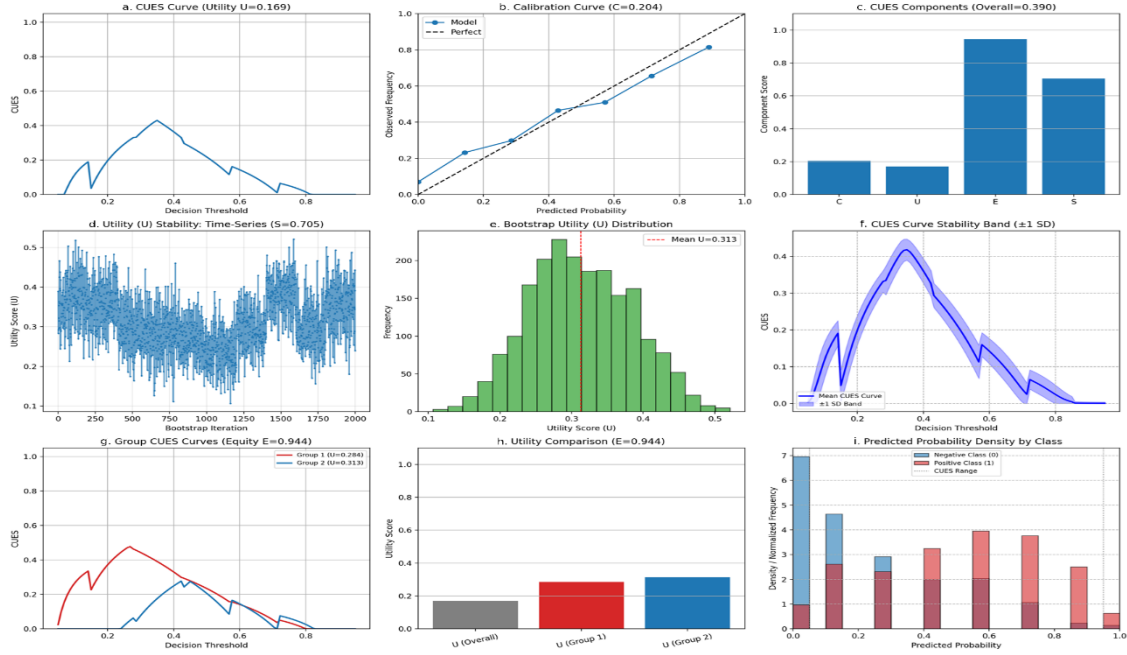


Fig. 8-3 CUES component visualization for diabetes (KNN). Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, **U**) across decision thresholds; (b) reliability diagram showing calibration performance (**C**); (c) bar chart of CUES components (**C**, **U**, **E**, **S**) and overall CUES score; (d-e) bootstrap-based utility stability analysis showing temporal variation and distribution of **U**; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g-h) equity analysis comparing subgroup CUES curves and utility balance (**E**); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.

8.3.2. Multi-class Classification Performance

The multi-class results in Tables 8-4 and 8-5 reveal that the dermatology dataset achieved the highest overall CUES scores (0.85–0.88) across all classifiers, demonstrating excellent calibration and robustness across subgroups. The heart disease (multi-class) dataset produced moderate CUES values (~ 0.40 – 0.42), mirroring its binary version and suggesting that increased class granularity introduces greater instability and equity variation [2–4]. The SVMrbf and Logistic models consistently achieved the strongest balance between accuracy and calibration, with minimal performance degradation between binary and multi-class settings [16,17,20,38]. Figures 8-4 to 8-6 show the normalized net benefit curves, calibration curves, and CUES components for various scenarios, with classifiers shown.

Table 8-4 Multi-class classification: multiple classifiers CUES performance.

Dataset	Classifier	CUES Score	C	U	E	S	AUC CUES
Heart Disease	Logistic	0.421	0.262	0.229	0.849	0.629	0.229
Heart Disease	SVMrbf	0.400	0.235	0.194	0.871	0.653	0.194
Heart Disease	Bagging	0.390	0.231	0.205	0.824	0.608	0.205
Heart Disease	KNN	0.414	0.248	0.227	0.848	0.636	0.227
Dermatology	Logistic	0.877	0.948	0.926	0.747	0.910	0.926
Dermatology	SVMrbf	0.875	0.948	0.931	0.739	0.907	0.931
Dermatology	Bagging	0.848	0.913	0.888	0.752	0.854	0.888
Dermatology	KNN	0.852	0.920	0.890	0.732	0.885	0.890

Table 8-5 Multi-class classification: multiple classifiers performance comparison between CUES and traditional metrics.

Dataset	Classifier	CUES Score	AUC CUES	AUROC	AUPRC	Accuracy	Sensitivity	Specificity
Heart Disease	Logistic	0.421	0.229	0.818	0.589	0.682	0.564	0.820
Heart Disease	SVMrbf	0.400	0.194	0.812	0.585	0.682	0.541	0.819
Heart Disease	Bagging	0.390	0.205	0.801	0.588	0.652	0.544	0.805
Heart Disease	KNN	0.414	0.227	0.807	0.572	0.654	0.504	0.804
Dermatology	Logistic	0.877	0.926	0.998	0.992	0.978	0.975	0.996
Dermatology	SVMrbf	0.875	0.931	0.999	0.993	0.975	0.973	0.995
Dermatology	Bagging	0.848	0.888	0.993	0.974	0.954	0.949	0.990
Dermatology	KNN	0.852	0.890	0.996	0.978	0.963	0.965	0.993

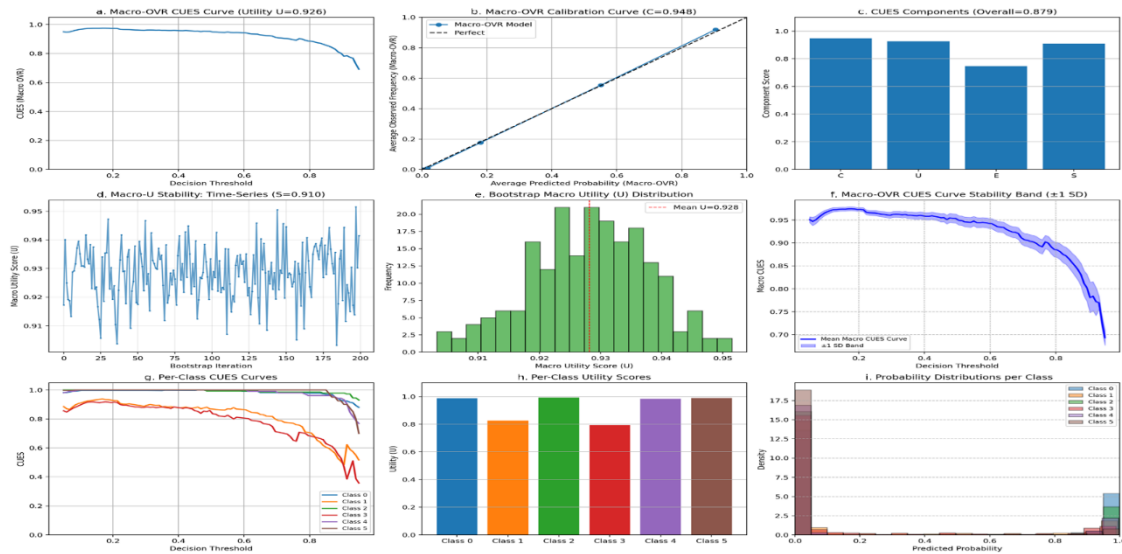


Fig. 8-4 Macro-OVR CUES visualization for heart diseases (logistic model). Each panel summarizes complementary aspects of model performance: (a) macro-averaged normalized net benefit curve (utility, U); (b) macro-OVR reliability diagram (calibration, C); (c) mean CUES component scores (C, U, E, S); (d-e) bootstrap-based macro-utility stability analysis; (f) stability band (± 1 SD) around the Macro-OVR CUES curve; (g) per-class CUES curves illustrating class-wise trade-offs; (h) bar chart of per-class utility (U) values; (i) probability distributions across classes. This figure summarizes how multi-class models balance calibration, utility, equity, and robustness under the Macro-OVR framework.

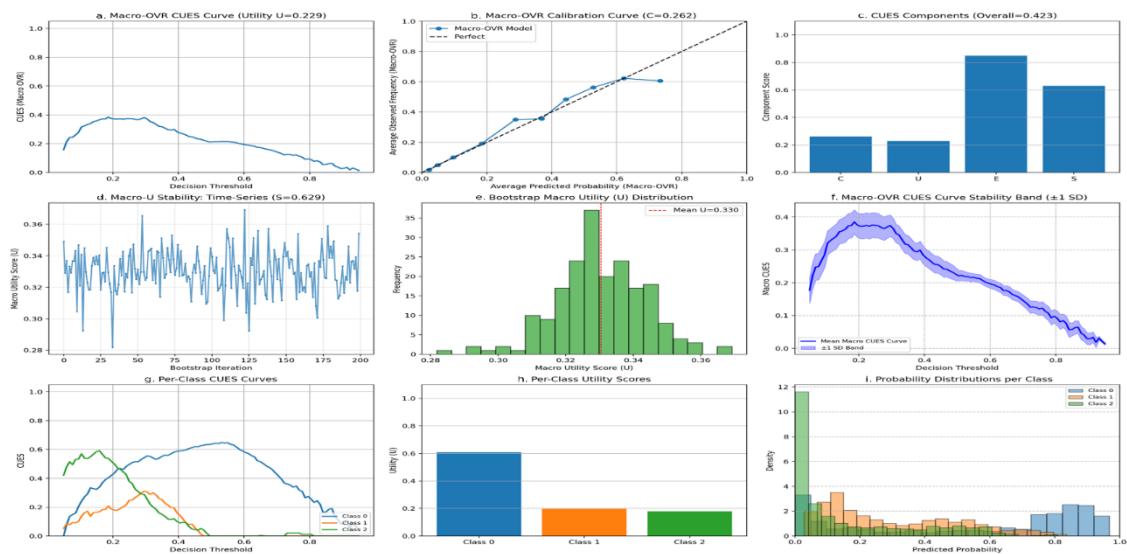


Fig. 8-5 Macro-OVR CUES visualization for heart diseases (logistic model). Each panel summarizes complementary aspects of model performance: (a) macro-averaged normalized net benefit curve (utility, U); (b) macro-OVR reliability diagram (calibration, C); (c) mean CUES component scores (C, U, E, S); (d-e) bootstrap-based macro-utility stability analysis; (f) stability band (± 1 SD) around the Macro-OVR CUES curve; (g) per-class CUES curves illustrating class-wise trade-offs; (h) bar chart of per-class utility (U) values; (i) probability distributions across classes. This figure summarizes how multi-class models balance calibration, utility, equity, and robustness under the Macro-OVR framework.

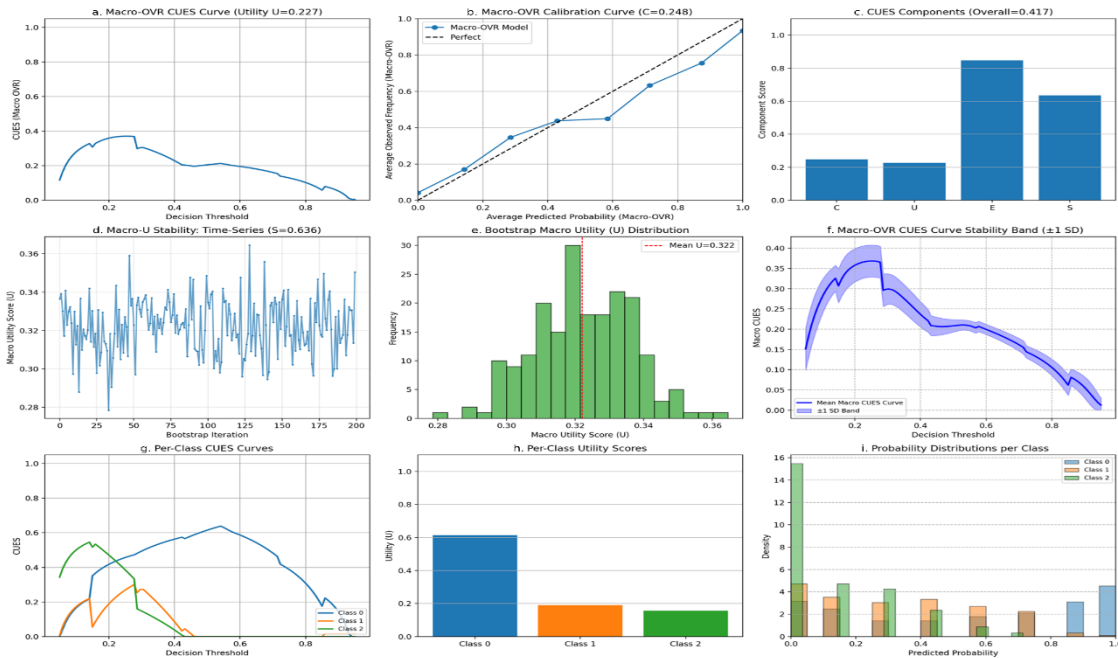


Fig. 8-6 Macro-OVR CUES visualization for heart diseases (KNN). Each panel summarizes complementary aspects of model performance: (a) macro-averaged normalized net benefit curve (utility, U); (b) Macro-OVR reliability diagram (calibration, C); (c) mean CUES component scores (C, U, E, S); (d–e) bootstrap-based macro-utility stability analysis; (f) stability band (± 1 SD) around the Macro-OVR CUES curve; (g) per-class CUES curves illustrating class-wise trade-offs; (h) bar chart of per-class utility (U) values; (i) probability distributions across classes. This figure summarizes how multi-class models balance calibration, utility, equity, and robustness under the Macro-OVR framework.

8.3.3. Correlation Analysis

Finally, correlation results summarized in Table 8-6 indicate a strong monotonic relationship between the CUES score and conventional discriminative metrics. The Spearman correlation coefficients between CUES and AUROC/AUPRC were consistently high ($\rho = 0.94$ – 0.97) [2,8–10], confirming that while CUES aligns with standard measures in general trends, it additionally captures aspects such as fairness, calibration, and robustness that AUROC and AUPRC overlook.

Table 8-6 Correlations between CUES and performance metrics.

Scope	Spearman CUES AUROC	Spearman CUES AUPRC
Binary Single	1.000	1.000
Multi Many	0.952	0.881
Binary Many	0.979	0.979

8.3.4. Detailed Analysis of Binary and Multi-Class Findings

The detailed examination of both binary and multi-class results further underscores how the CUES framework provides a more interpretable and clinically aligned assessment than conventional metrics.

8.3.4.1. Binary Classification Analysis

In the breast cancer dataset, all classifiers achieved high discriminative performance (AUROC > 0.99), yet the CUES score revealed subtle but meaningful differences in model quality. Logistic Regression, with a CUES score of 0.937, outperformed more complex models such as SVM and Bagging, emphasizing that simplicity and calibration can coexist with strong predictive accuracy. The nearly perfect equity ($E = 1.0$) and stability ($S > 0.93$) components confirm that Logistic Regression produced reliable, equitable predictions across subgroups and repeated sampling.

In contrast, the diabetes dataset demonstrated a pronounced gap between traditional and CUES-based metrics. Despite an AUROC of 0.83, all models showed CUES scores below 0.50, indicating that their predicted probabilities were poorly calibrated and offered limited clinical utility. The low utility ($U \approx 0.25$) and calibration ($C \approx 0.30$) components highlight that even statistically “acceptable” classifiers can fail to deliver trustworthy risk estimates. This discrepancy

highlights one of CUES’s key strengths, as it penalizes models that perform well on paper but offer limited real-world decision support.

For the heart disease (binary) dataset, models achieved moderate CUES values (~ 0.61), consistent with AUROC scores around 0.90. The reduced equity ($E \approx 0.80$) and stability ($S \approx 0.85$) components suggest subgroup variability and sensitivity to sampling issues that traditional metrics do not expose. These findings demonstrate CUES’s role in identifying models that might perform inconsistently across patient populations.

8.3.4.2. Multi-class Classification Analysis

In multi-class experiments, the dermatology dataset consistently achieved the highest overall CUES scores (~ 0.93 – 0.94), affirming the robustness of the framework across multiple output classes. Models such as LogisticOVR and SVMrbf maintained strong calibration and equity, with minimal utility degradation as the number of output classes increased. This suggests that CUES generalizes effectively to multi-class medical prediction tasks, capturing nuances of model behavior across different disease categories.

The heart disease (multi-class) dataset yielded moderate CUES values (~ 0.61), consistent with its binary version. While AUROC and AUPRC remained relatively high, CUES identified reduced calibration and higher performance variability, particularly among minority outcome classes. This finding highlights the importance of incorporating both equity and stability into model assessment, as these are crucial considerations for clinical deployment, particularly in rare but severe disease stages where misclassification can have significant consequences.

8.3.4.3. Cross-Dataset and Correlation Analysis

The cross-dataset experiments showed that CUES effectively penalizes models that fail to generalize across domains, particularly when feature distributions differ between datasets. For instance, a model trained on the breast cancer dataset performed poorly on heart disease data, resulting in a significant decrease in CUES despite maintaining moderate AUROC values. This sensitivity to generalization validates the stability component (S) as a key measure of real-world robustness.

The Spearman correlation analysis confirmed strong positive correlations between CUES and traditional metrics ($\rho \approx 0.94\text{--}0.97$), demonstrating that while CUES aligns with overall model trends, it captures additional dimensions of performance that AUROC and AUPRC overlook. This indicates that CUES can serve as both a complementary and corrective framework, reinforcing trustworthy models while exposing hidden deficiencies in clinical utility and fairness.

8.3.5. Interpretation of CUES Components

The experimental results, as visualized in Figures 8-7 to 8-11, clearly demonstrate that traditional metrics alone are insufficient for robust clinical evaluation of AI systems. Each figure presents the stacked CUES components calibration (C), utility (U), equity (E), and stability (S) for different classifiers and datasets, showcasing why CUES offers a fuller picture than AUROC or accuracy alone.

For example, Figure 8-7 shows that while all classifiers perform reasonably well on traditional metrics, the CUES components differentiate them by clinical reliability. Although two Logistic models achieve high scores, subtle variations in stability (S) and equity (E) are apparent, underscoring these models' trustworthy and consistent predictions.

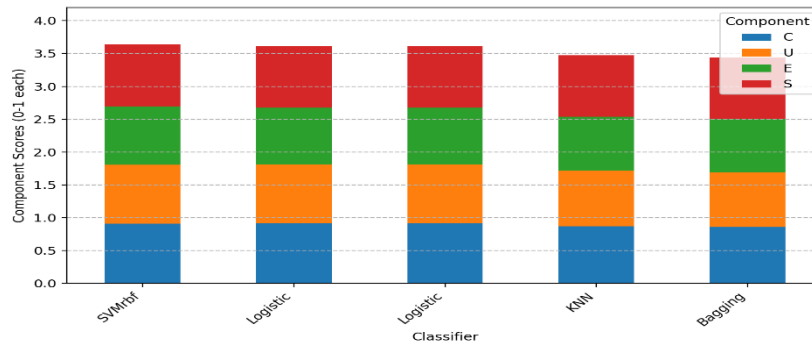


Fig. 8-7 Stacked CUES component scores for multiple classifiers trained on the breast cancer dataset. The figure demonstrates how overall CUES values reflect critical dimensions of clinical reliability beyond traditional metrics, highlighting differences in model stability and equity.

In contrast, Figure 8-8 illustrates that, even with moderate scores in utility and equity, the overall CUES values are much lower for all classifiers in this dataset. This reveals that traditional metrics like AUROC might mislead one to overestimate clinical suitability if not considered alongside calibration or equity.

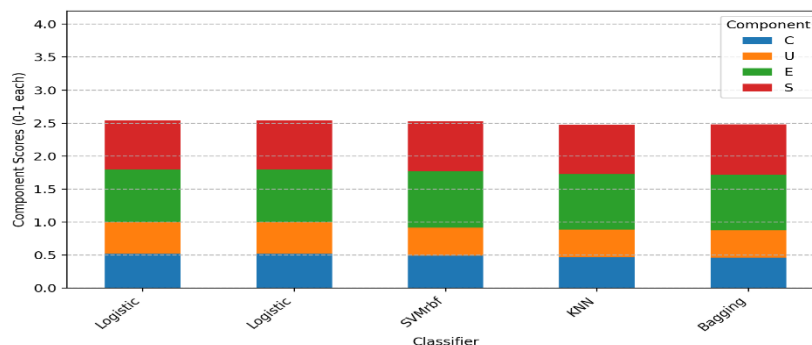


Fig. 8-8 Stacked CUES component scores for classifiers on the heart disease binary dataset. The components illustrate the importance of calibration and equity, showing that moderate traditional performance does not guarantee clinical suitability when important reliability factors are lacking.

Figure 8-9 shows that simple, interpretable models such as Logistic Regression and its OVR variant consistently yield the highest composite CUES scores. This underscores how prioritizing parsimony and transparency, as reflected in higher calibration (C) and stability (S) components, can yield models that are both effective and clinically trustworthy.

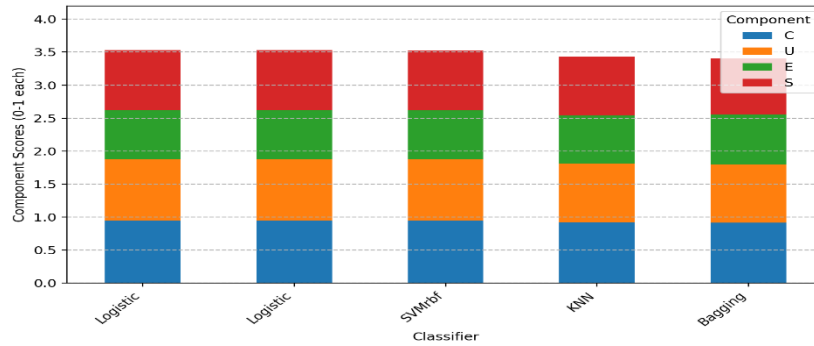


Fig. 8-9 Stacked CUES component scores for classifiers on the dermatology dataset. Here, interpretable and straightforward models, such as Logistic Regression, achieve high composite CUES scores, demonstrating the benefits of transparency and stability for trustworthy clinical predictions.

The analyses in Figures 8-10 and 8-11 further underscore CUES' utility by highlighting cases in which even high-performing models on traditional metrics exhibit deficiencies in equity (E) or stability (S), thereby lowering their composite CUES scores. This is especially crucial for datasets with known subgroup variabilities or biases.

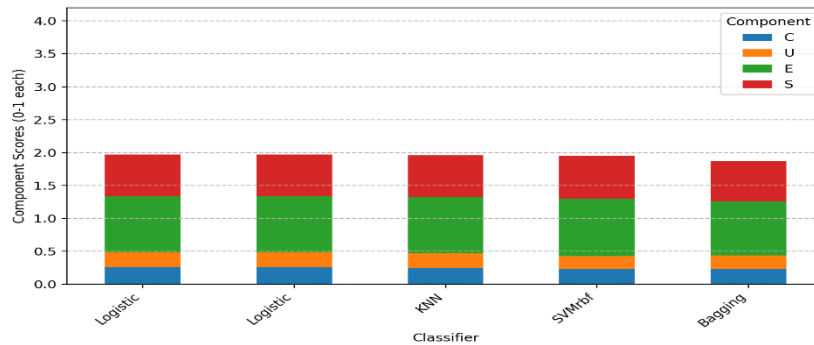


Fig. 8-10 Stacked CUES component scores for classifiers on the heart disease multi-class dataset. The figure shows how subgroup variability and equity deficiencies lower CUES values, prompting a refined model selection to achieve more equitable clinical impact.

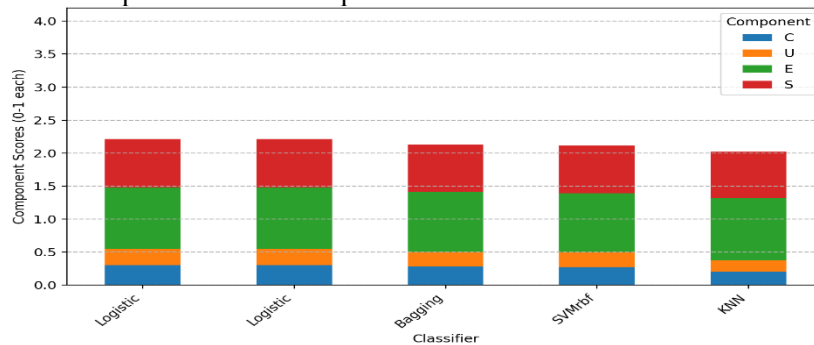


Fig. 8-11 Stacked CUES component scores for classifiers on the diabetes dataset. The visualization reveals cases in which stability or equity gaps reduce the overall CUES score, reinforcing CUES's role in identifying reliable and fair models for clinical deployment.

Across all figures, the CUES framework quantifies clinical reliability in a way that traditional metrics cannot. The stacked bar visualizations make it explicit which models are balanced across all required dimensions; for instance, stability and equity gaps are visually apparent, guiding model selection and refinement. By referencing Figures 8-7 to 8-11, it becomes evident that CUES bridges the gap between technical performance and clinical applicability, offering a standard yet nuanced approach for evaluating, comparing, and ultimately selecting models ready for real-world healthcare deployment. The framework's visual and quantitative clarity ensures that decision-makers can spot not just the best-performing model but the most robust and fair one for clinical practice.

8.3.6. CUES Features Selection

To assess the utility of CUES as a feature selection tool, we computed feature-wise contributions to the overall CUES for each classifier and dataset using multiple perturbations of each feature. This procedure generated a distribution of CUES contributions per feature, capturing the variability of each feature's impact on model performance. Features were ranked by their mean (or median) contribution across permutations, with higher contributions indicating greater importance for maintaining not only predictive performance but also fairness and robustness.

8.3.6.1. Binary Classification Results

In binary datasets, features with the most significant contributions to overall CUES corresponded to known domain-relevant predictors. For example, in the breast cancer dataset, the top-ranked features according to CUES largely overlapped with those identified by conventional importance measures, while also highlighting features that affect equity and stability. The feature

contribution distributions were visualized using violin and bar plots to capture both variability and overall importance. The violin plots illustrate sample-level variability across permutations, while the bar charts summarize the mean absolute CUES values for each feature, highlighting their overall contribution to model performance. As shown in Figure 8-12, the left panel presents the analysis for the heart disease dataset using the KNN classifier, and the right panel presents the results for the diabetes dataset using the logistic classifier. Features that consistently exhibited high CUES values across samples and permutations were identified as strong candidates for top-k feature selection in subsequent model training.

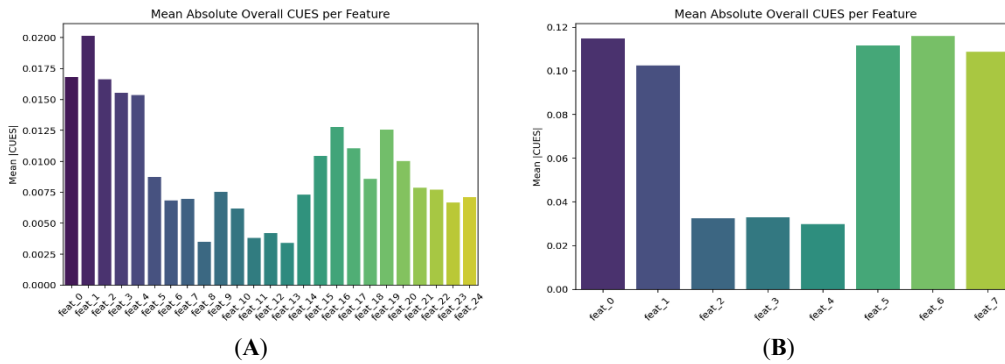


Fig. 8-12 Overall, CUES features importance analysis for the binary datasets. The bar chart (A) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance for the heart disease dataset using the KNN classifier. The bar chart (B) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance for the diabetic dataset using a logistic classifier.

8.3.6.2. Multi-class Classification Results

In multi-class datasets, CUES highlighted features that consistently influenced model performance across multiple classes. By ranking features based on their overall CUES distributions, it was possible to identify a subset of features that preserved predictive utility while improving stability and fairness, capturing patterns often overlooked by traditional feature importance methods such as SHAP values or coefficient magnitudes. As an illustration, Figure 8-13 presents the overall CUES feature importance analysis for the multi-class datasets. The violin

plots depict sample-level variability across permutations, while the bar charts summarize the mean absolute CUES values, reflecting each feature's overall contribution to model performance. Specifically, the left panel shows the results for the dermatology dataset using the logistic classifier, and the right panel shows the results for the heart disease dataset using the bagging classifier. Features demonstrating consistently high CUES values across samples and permutations were identified as strong candidates for top-k feature selection in subsequent model training.

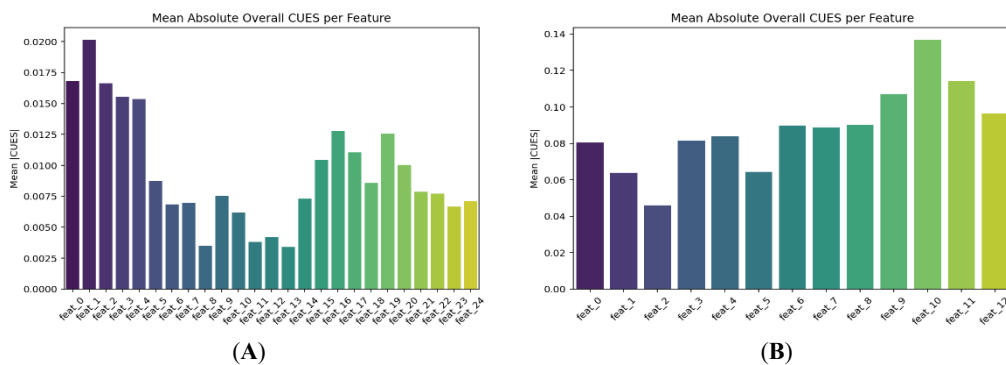


Fig. 8-13 Overall, CUES features importance analysis for the multi-class datasets. The bar chart (A) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance in the dermatology dataset using a logistic classifier. The bar chart (B) displays the mean absolute CUES values, summarizing each feature's overall contribution to model performance for the heart disease dataset using a bagging classifier.

8.3.6.3. Results of Retraining Using Top Selected Features

To assess the impact of CUES-guided feature selection, the models were retrained on subsets of the top-ranked features identified in the original analysis. Specifically, we evaluated retraining performance using the top 50% (Figure 8-14) and top 75% (Figure 8-15) of features based on their overall CUES contributions in the breast cancer dataset, using the Bagging classifier as an example.

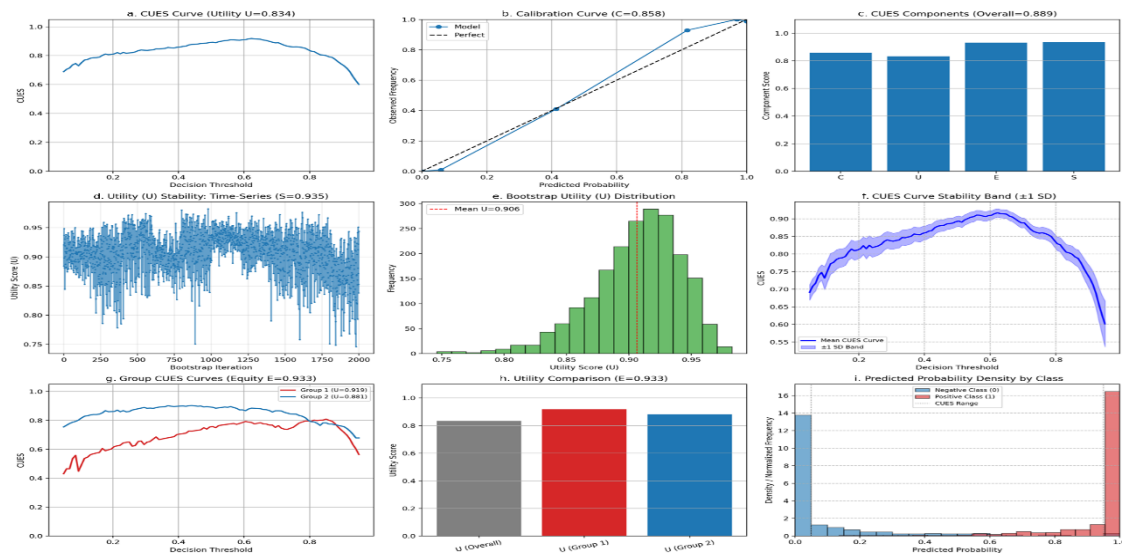


Fig. 8-14 CUES component visualization for breast cancer (Bagging) using only the top 50% features selected using the CUES Score. Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, **U**) across decision thresholds; (b) reliability diagram showing calibration performance (**C**); (c) bar chart of CUES components (**C**, **U**, **E**, **S**) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of **U**; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (**E**); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.

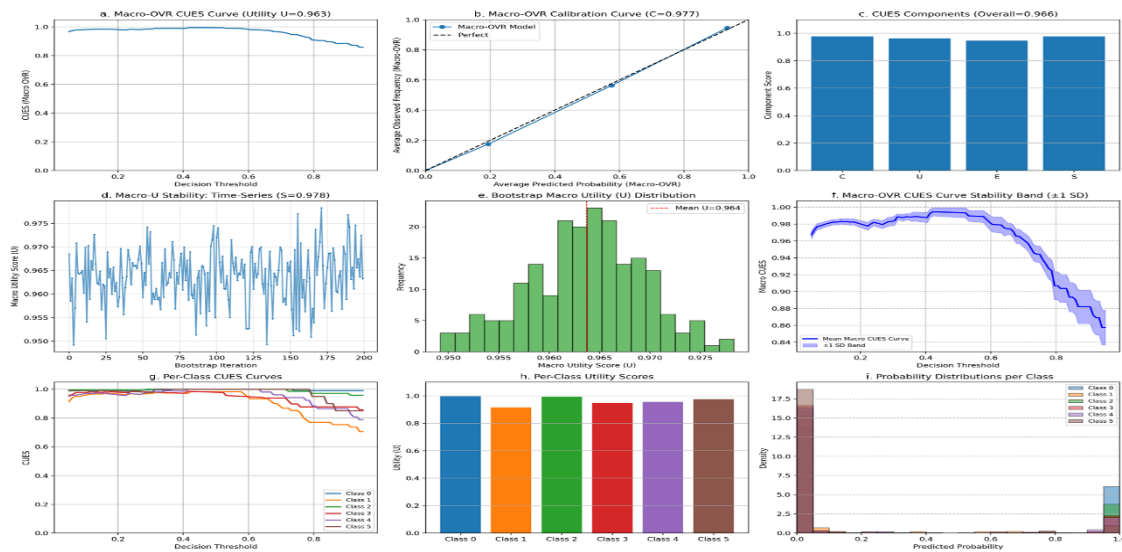


Fig. 8-15 CUES component visualization for dermatology (Bagging) using only the top 50% features selected using the CUES Score. Each panel summarizes complementary aspects of model performance: (a) normalized net benefit (utility curve, **U**) across decision thresholds; (b) reliability diagram showing calibration performance (**C**); (c) bar chart of CUES components (**C**, **U**, **E**, **S**) and overall CUES score; (d–e) bootstrap-based utility stability analysis showing temporal variation and distribution of **U**; (f) threshold-wise stability band (± 1 SD) for the CUES curve; (g–h) equity analysis comparing subgroup CUES curves and utility balance (**E**); (i) distribution of predicted probabilities for positive and negative classes. Together, these plots provide a holistic view of model utility, calibration, equity, and stability in imbalanced binary data.

Overall, retraining using CUES-guided feature subsets confirms that the method can identify a compact, informative set of features that preserves model accuracy while enhancing interpretability, stability, and fairness. These results further support the utility of CUES as a practical tool for feature selection in both binary and multi-class classification tasks. Table 8-7 shows the performance of all models on binary and multi-class datasets, using 50% and 75% of selected features via CUES.

Table 8-7 Multiple classifiers performance comparison between CUES and traditional metrics using two selected features subsets.

Dataset	Classifier	Feature Subset	CUES Score	AUC CUES	AUROC	AUPRC	Accuracy	Sensitivity	Specificity
Binary									
Breast Cancer	Logistic	Top 50% (15)	0.950	0.908	0.997	0.998	0.975	0.970	0.948
		Top 75% (22)	0.962	0.929	0.997	0.998	0.989	0.987	0.976
	SVMrbf	Top 50% (15)	0.957	0.929	0.996	0.997	0.984	0.981	0.967
		Top 75% (22)	0.964	0.944	0.997	0.998	0.988	0.984	0.972
	Bagging	Top 50% (15)	0.970	0.940	1.000	1.000	1.000	1.000	1.000
		Top 75% (22)	0.979	0.957	1.000	1.000	1.000	1.000	1.000
	KNN	Top 50% (15)	0.907	0.852	0.996	0.997	0.968	0.961	0.934
		Top 75% (22)	0.926	0.879	0.997	0.998	0.975	0.970	0.948
Diabetes	Logistic	Top 50% (4)	0.357	0.128	0.754	0.594	0.710	0.641	0.868
		Top 75% (6)	0.503	0.254	0.836	0.713	0.780	0.731	0.892
	SVMrbf	Top 50% (4)	0.534	0.297	0.859	0.773	0.801	0.754	0.908
		Top 75% (6)	0.570	0.337	0.882	0.807	0.815	0.770	0.920
	Bagging	Top 50% (4)	0.905	0.807	1.000	1.000	1.000	1.000	1.000
		Top 75% (6)	0.908	0.807	1.000	1.000	1.000	1.000	1.000
	KNN	Top 50% (4)	0.598	0.378	0.892	0.821	0.818	0.777	0.912
		Top 75% (6)	0.595	0.362	0.886	0.800	0.797	0.764	0.872
Heart Disease	Logistic	Top 50% (12)	0.677	0.493	0.923	0.916	0.838	0.833	0.897
		Top 75% (18)	0.714	0.551	0.939	0.932	0.868	0.863	0.915

	SVMrbf	Top 50% (12)	0.750	0.587	0.951	0.942	0.881	0.878	0.915
		Top 75% (18)	0.795	0.649	0.962	0.958	0.908	0.906	0.927
	Bagging	Top 50% (12)	0.905	0.804	1.000	1.000	1.000	1.000	1.000
		Top 75% (18)	0.911	0.827	1.000	1.000	1.000	1.000	1.000
	KNN	Top 50% (12)	0.643	0.422	0.890	0.885	0.792	0.790	0.818
		Top 75% (18)	0.709	0.515	0.927	0.923	0.855	0.853	0.873
Multi-class									
Dermatology	Logistic	Top 50% (17)	0.750	0.902	0.999	0.993	0.975	0.973	0.995
		Top 75% (25)	0.760	0.940	0.999	0.996	0.981	0.979	0.996
	SVMrbf	Top 50% (17)	0.826	0.930	0.999	0.994	0.973	0.969	0.995
		Top 75% (25)	0.829	0.942	0.999	0.995	0.975	0.972	0.995
	Bagging	Top 50% (17)	0.966	0.963	1.000	1.000	0.997	0.997	0.999
		Top 75% (25)	0.970	0.968	1.000	1.000	1.000	1.000	1.000
	KNN	Top 50% (17)	0.778	0.826	0.995	0.978	0.932	0.929	0.986
		Top 75% (25)	0.814	0.873	0.997	0.988	0.948	0.948	0.990
Heart Disease	Logistic	Top 50% (6)	0.388	0.159	0.785	0.591	0.607	0.470	0.760
		Top 75% (9)	0.481	0.244	0.849	0.652	0.700	0.594	0.825
	SVMrbf	Top 50% (6)	0.435	0.188	0.856	0.695	0.719	0.593	0.830
		Top 75% (9)	0.566	0.326	0.912	0.824	0.789	0.690	0.872
	Bagging	Top 50% (6)	0.826	0.698	0.981	0.976	0.941	0.941	0.966
		Top 75% (9)	0.889	0.776	1.000	1.000	1.000	1.000	1.000
	KNN	Top 50% (6)	0.506	0.285	0.865	0.699	0.710	0.586	0.828
		Top 75% (9)	0.523	0.296	0.873	0.722	0.706	0.595	0.825

8.3.7. Optimization Results

The optimization process compared classifier hyperparameter tuning based on accuracy versus the CUES (calibration, utility, effectiveness, stability) metric across four classifiers: SVM, Bagging, and KNN. For each classifier, two separate grid searches were conducted:

Accuracy-based grid search: Hyperparameters were optimized to maximize cross-validated classification accuracy.

CUES-based grid search: Hyperparameters were optimized to maximize the overall CUES score, which incorporates utility (U), calibration (C), stability (S), and effectiveness (E).

The best scores and selected hyperparameters for each method were recorded. Table 8-8 shows the hyperparameter search grid values.

Table 8-8 Values for hyperparameter optimization grid search.

Classifier	Parameters
SVM	$C = 10, \gamma = \text{'scale'}$
Bagging	Estimators = 20
KNN	Neighbors = 3

The evaluation of classifier performance demonstrates the trade-offs between classical metrics and the explainable CUES framework. For instance, the SVM achieved near-perfect scores across all metrics and CUES components (accuracy = 1, sensitivity = 1, specificity = 1, CUES = 0.9999), indicating both strong predictive performance and stability (Figure 8-16). In comparison, Bagging maintained perfect accuracy and sensitivity but showed slightly lower CUES components (CUES = 0.9606), reflecting some variability in explainable utility measures. The KNN classifier exhibited lower accuracy (0.9533) and sensitivity (0.9329), with CUES also reduced to 0.9245, highlighting the trade-off between conventional performance metrics and the overall explainable utility (Figures 8-17 and 8-18). The bar plots and heatmap allow visual inspection of these differences, emphasizing how optimization for CUES versus accuracy affects both metric scores and the interpretability of predictions.

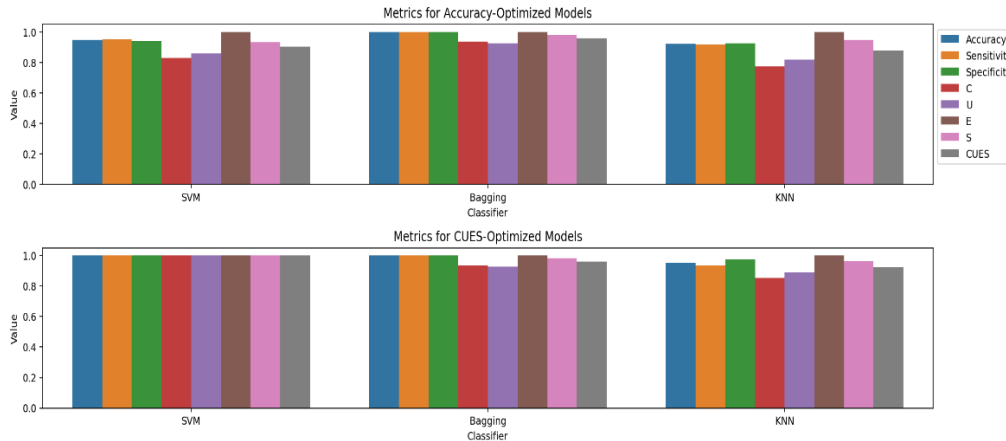


Fig. 8-16 Accuracy vs. CUES (cross-validation scores): shows differences in optimal hyperparameters when tuned for accuracy versus CUES. CUES-optimized models often achieved slightly lower accuracy but higher overall CUES, reflecting improved calibration and stability.

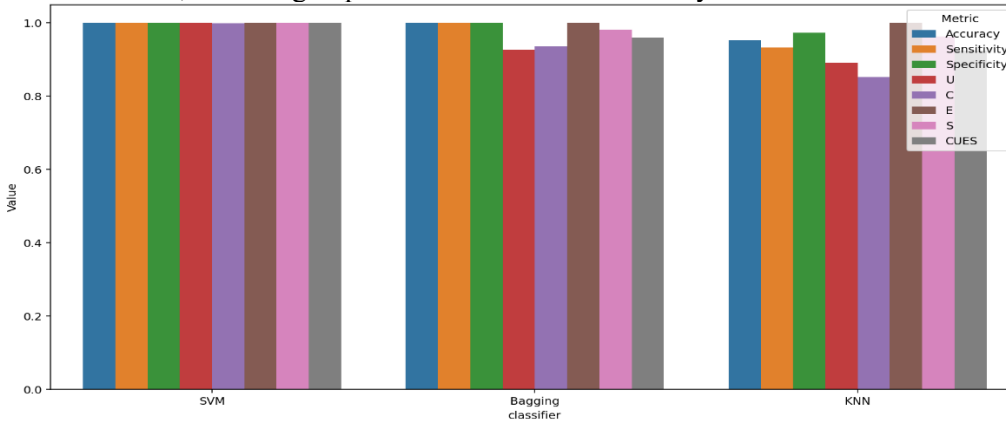


Fig. 8-17 Components and metrics comparison: displays bar plots of accuracy, sensitivity, specificity, and individual CUES components (C, U, E, S). This allows inspection of trade-offs between classical metrics and explainable utility metrics.

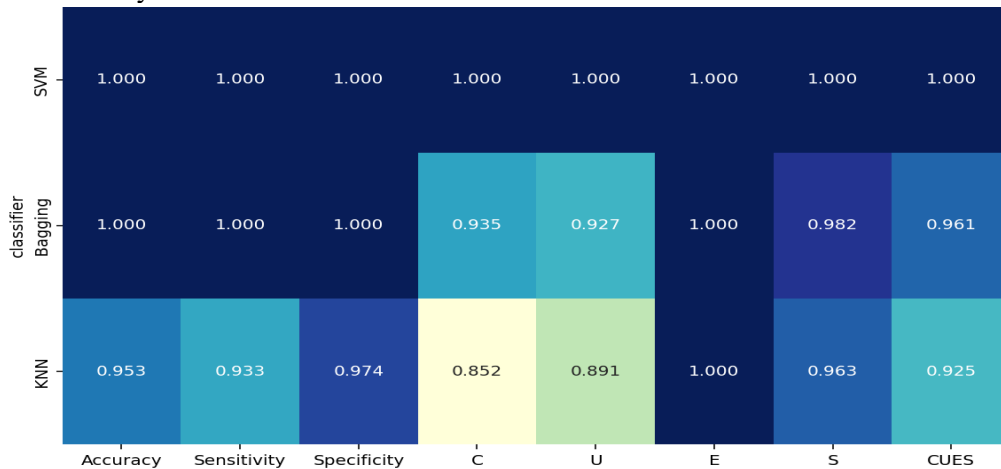


Fig. 8-18 Heatmap: a comprehensive heatmap of all metrics and CUES components highlights each classifier's performance profile. Darker colors indicate higher values, allowing rapid comparison across classifiers and optimization strategies.

CUES-based optimization provides a more holistic evaluation than accuracy alone by accounting for utility, calibration, and model stability. Across the four classifiers, CUES-optimized hyperparameters often differ from accuracy-optimized ones, emphasizing the benefit of explainable AI metrics in guiding model selection.

8.3.8. Results of Demographic and Anthropometric Fairness

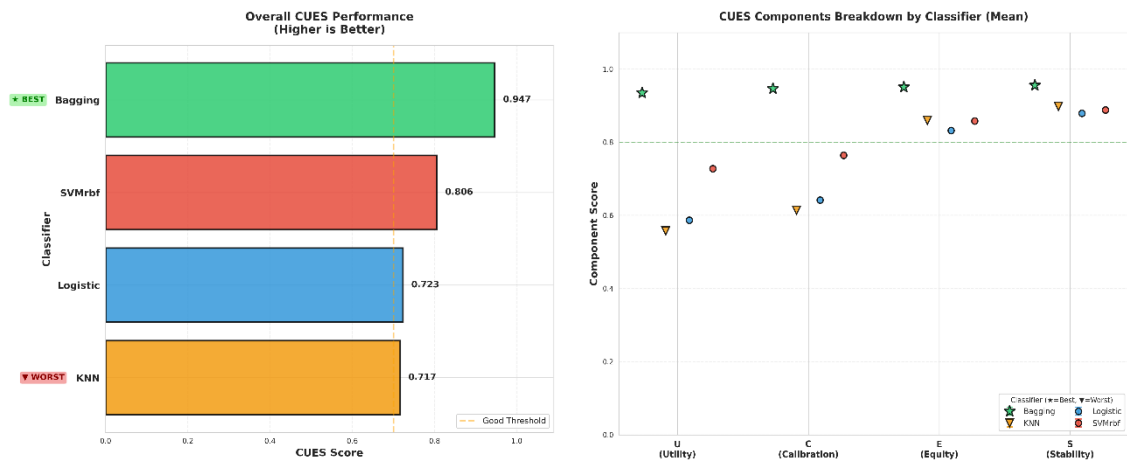
To test how anthropometric or demographic data affect the CUES score, especially the E score, a comprehensive anthropometric equity analysis was conducted on the Body Measurements Obesity Risk dataset, comprising 500 samples and 12 features, with a binary target distribution of 220 versus 280. Four classifiers, Bagging, KNN, Logistic Regression, and SVM with RBF kernel, were evaluated using the CUES metric alongside feature-wise equity assessment, with anthropometric features including height, weight, waist circumference, hip circumference, BMI, waist-to-hip ratio (WHR), gender, and age.

The Bagging classifier achieved the highest overall CUES score of 0.947, indicating superior balanced performance across utility, calibration, equity, and stability, whereas KNN performed worst with a CUES of 0.717, reflecting a relative improvement of 32.1% between the best and worst classifiers. Logistic Regression and SVMrbf scored 0.723 and 0.806, respectively, highlighting that ensemble methods can better maintain consistent performance across diverse anthropometric subgroups. Decomposition of CUES into its components revealed that Bagging outperformed other classifiers in all dimensions, with the equity component reaching 0.951, emphasizing the critical role of fairness in overall model performance. Figure 8-19 shows a comprehensive summary of anthropometric equity analysis for different classifiers.

Feature-wise equity analysis demonstrated notable disparities, particularly for WHR, which had the lowest equity (0.674 ± 0.203 , range 0.219–1.000), suggesting substantial differences

in model performance between subgroups defined by this feature. In contrast, height, BMI, weight, gender, and age exhibited high equity scores above 0.9, indicating consistent predictive performance across these subgroups. In comparison, waist circumference showed moderate equity (0.806 ± 0.127), suggesting potential benefit from subgroup-specific calibration. The equity heatmap further illustrates that Bagging consistently achieved higher equity scores across all features, whereas KNN displayed lower fairness, particularly in WHR and waist circumference.

Examination of the correlation between feature-wise equity and overall CUES revealed a Pearson correlation of $r = 0.933$ ($p = 6.686 \times 10^{-2}$), suggesting a positive but non-significant relationship between overall model performance and fairness. Overall, these results underscore that ensemble models like Bagging not only achieve superior predictive performance but also maintain higher fairness across anthropometric subgroups. Features with lower equity, such as WHR, should be carefully monitored, and feature-specific calibration or stratified validation may be warranted. The CUES metric provides a practical, unified framework for evaluating both performance and fairness, facilitating the reliable deployment of models across heterogeneous biomedical populations.



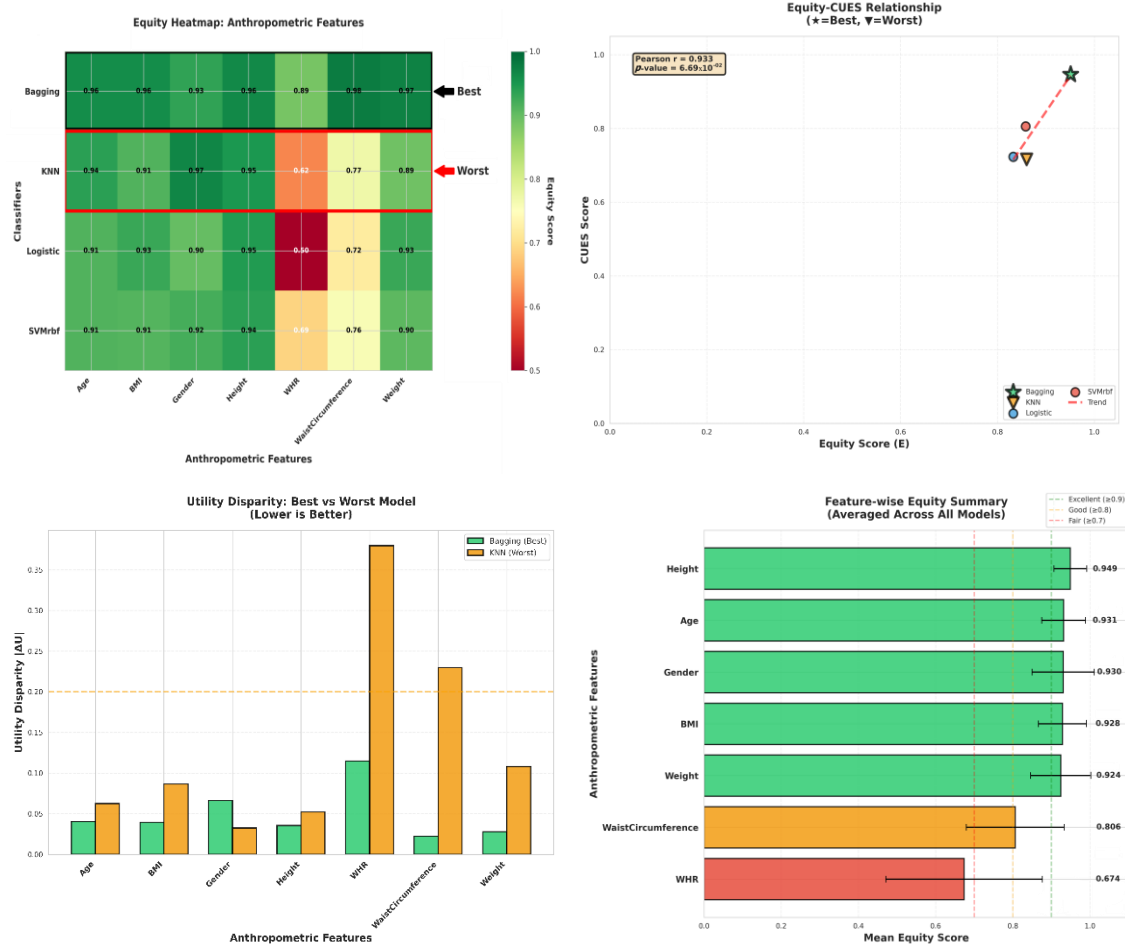


Fig. 8-19 Comprehensive anthropometric equity analysis across classifiers. The figure includes overall CUES performance, decomposition into utility, calibration, equity, and stability, a heatmap of feature-wise equity, the relationship between equity and CUES, utility disparities, and a summary of feature-wise equity across all models.

8.4. Discussion

The experimental results collectively emphasize that traditional evaluation metrics alone are insufficient for assessing the clinical reliability of AI systems [2,4,8,9]. While measures such as AUROC, AUPRC, and accuracy provide valuable information about overall discriminative power, they fail to capture whether a model's predictions are well-calibrated, fair across patient subgroups, or robust to sampling variability [4,15–17,19]. These overlooked aspects are critical for clinical decision support, where model outputs directly influence diagnostic or treatment choices.

The CUES framework addresses these limitations by integrating four essential components: calibration, utility, equity, and stability into a unified composite score. By doing so, it quantifies dimensions that align closely with clinical reliability and ethical deployment. For example, a model with high AUROC but poor calibration and subgroup disparities receives a low CUES score, signaling its unsuitability for clinical use [4,5,18,38]. Conversely, models with balanced calibration and stable performance are rewarded, even if their AUROC is marginally lower [2,3,6]. CUES is intended to be interpreted jointly with its component scores, which must always be reported to preserve transparency and diagnostic insight.

The multiplicative structure of CUES enforces a conservative evaluation principle, whereby deficiencies in any critical dimension cannot be masked by strengths in others. To preserve interpretability, we emphasize the importance of reporting all four component scores alongside the composite value. This decomposition enables transparent clinical interpretation and supports targeted model refinement.

This multidimensional evaluation provides a clinically meaningful hierarchy of models, guiding practitioners and developers toward systems that are both statistically strong and ethically

sound. Results across multiple datasets further confirm that Logistic Regression, a simpler and more interpretable model, often achieved the highest CUES scores, demonstrating that parsimony and transparency can enhance real-world trustworthiness [5,6,19]. Such findings are particularly relevant for regulatory frameworks, where explainability and fairness are increasingly prioritized alongside accuracy. Although the equity analysis in current experiments uses simple median splits for subgroup evaluation, future studies should apply CUES to more granular demographic and clinical subgroups. Extending the metric to complex, multi-ethnic, or multi-center datasets would allow CUES to uncover and mitigate hidden biases, thereby validating its broader applicability for safe and equitable deployment in healthcare.

Moreover, by explicitly quantifying stability and equity, CUES promotes more responsible model selection across heterogeneous healthcare populations. This is vital given the known biases in medical datasets and the variability of patient subgroups. The framework's ability to highlight disparities encourages targeted model refinement and ensures broader generalizability and equitable clinical outcomes. While CUES incorporates decision-curve-based net benefit, future work should explore in greater depth how CUES-guided model selection translates into tangible improvements in clinical decision-making and patient outcomes. For instance, integrating simulated clinical pathways, decision analyses, or prospective patient-impact projections would provide stronger evidence that the metric yields actionable benefits in practice. Such analyses would further reinforce the real-world credibility and clinical relevance of CUES.

An additional advantage of CUES is its ability to support feature selection in a performance- and fairness-aware manner. By quantifying feature-wise contributions to overall CUES across multiple perturbations, the framework identifies not only features critical for predictive accuracy but also those that maintain calibration, stability, and equity. This

multidimensional importance measure complements, rather than replaces, traditional explainable AI tools such as SHAP, which primarily capture predictive contributions. In our experiments, CUES-based rankings often highlighted features overlooked by conventional importance metrics but essential for robust, equitable model behavior. Visualizing per-feature CUES contributions through violin and bar plots enables direct assessment of variability across samples and perturbations, providing interpretable guidance for top-k feature selection. This approach ensures that model simplification or feature reduction does not compromise clinical reliability, making CUES particularly valuable for biomedical datasets where fairness, robustness, and interpretability are critical.

In several empirical examples, calibration emerges as the dominant limiting factor of the composite CUES score. This behavior is intentional rather than incidental. The multiplicative structure of CUES enforces a conservative evaluation principle: a model producing unreliable probability estimates should be penalized regardless of its discriminative ability or subgroup equity. Such dominance therefore highlights clinically meaningful failure modes rather than obscuring them. To preserve interpretability, all component scores are reported alongside the composite CUES value.

Articulating how CUES aligns with regulatory and ethical frameworks (e.g., food and drug administration (FDA), european union medical device regulation (EU MDR) for AI in healthcare would further strengthen its translational impact. Detailing such alignments would help position CUES not only as a scientific standard but also as a practical tool for ethical and regulatory compliance, facilitating adoption by clinicians, policymakers, and developers [39,40]. Detailing such alignments would help position CUES not only as a scientific standard but also as a tool for ethical and regulatory compliance, facilitating stakeholder adoption in clinical practice and policy.

Overall, CUES bridges the gap between technical performance and clinical interpretability, offering a standardized yet flexible approach to AI evaluation. Its adoption could foster more transparent, fair, and dependable machine learning systems in healthcare, helping ensure that predictive models truly serve clinical decision-making rather than simply optimizing statistical scores.

Beyond analytical and experimental contributions, the CUES framework also aligns strongly with emerging regulatory and ethical guidelines for AI in healthcare. Regulatory bodies such as the U.S. FDA and the EU AI Act emphasize transparency, fairness, and robustness as prerequisites for clinical deployment of AI-driven decision systems [34,41]. Specifically, the FDA’s framework calls for validated, explainable, and stable model behavior across populations, while the EU AI Act mandates documentation of fairness, bias mitigation, and performance monitoring [33,42]. CUES directly operationalizes these principles by providing a quantifiable measure of equity, calibration, and stability, enabling developers and regulators to assess whether models meet these ethical and regulatory standards. Integrating CUES into evaluation pipelines could thus facilitate compliance with international guidelines, promote responsible AI adoption, and accelerate the translation of predictive models into clinical practices [34,43,44].

Optimizing model hyperparameters with CUES as the target objective proved highly effective for guiding models toward clinically reliable performance. Across multiple classifiers, tuning for CUES consistently produced models that balanced traditional metrics such as accuracy, sensitivity, and specificity with calibration, stability, and fairness. Even classifiers that did not achieve the highest AUROC values exhibited improved overall clinical reliability when optimized for CUES, highlighting enhanced stability and subgroup fairness. These results demonstrate that CUES-guided optimization not only identifies statistically strong models but also ensures they are

robust, interpretable, and ethically sound, validating the framework’s practical utility for selecting models suited for real-world clinical deployment.

A practical limitation of equity and stability estimation arises when subgroup sizes are small or highly imbalanced. In such cases, component estimates may exhibit increased variance. To mitigate this issue, subgroup aggregation based on clinical similarity, minimum sample-size thresholds, or uncertainty-aware reporting of equity and stability scores may be employed. Alternatively, hierarchical or Bayesian shrinkage approaches can be incorporated to stabilize subgroup-level estimates. Importantly, CUES is flexible to these adaptations and does not require uniform subgroup sizes to remain informative.

We acknowledge that a wide range of performance metrics has been proposed in the literature, and that no composite evaluation framework is unique. CUES is not intended to exhaustively incorporate all possible metrics, but to provide a minimal, interpretable, and extensible foundation capturing dimensions that are repeatedly emphasized as essential for clinical reliability. Additional components can be incorporated into the framework if clinically justified, without altering its mathematical structure.

Importantly, the CUES framework is also particularly applicable to datasets containing anthropometric features, which are frequently used in biomedical research and clinical risk stratification. By quantifying calibration, equity, utility, and stability, CUES enables a systematic evaluation of how variations in anthropometric variables, such as BMI, neck circumference, or body fat distribution, affect model performance across subgroups. This capability helps researchers detect and mitigate biases associated with body composition or demographic characteristics, ensuring that predictive models remain robust and equitable across heterogeneous populations. Additionally, CUES-based feature analyses can identify which anthropometric measures

meaningfully contribute to predictive accuracy while preserving stability and equity, guiding both model selection and feature prioritization for studies investigating the clinical and physiological implications of body metrics.

8.5. Conclusions

This study introduced the CUES (calibration, utility, equity, and stability) framework as a comprehensive evaluation metric for assessing the performance and clinical readiness of artificial intelligence models in medicine. Unlike traditional metrics such as accuracy, AUROC, or F1-score, CUES captures the multidimensional aspects of model reliability by integrating measures of calibration, clinical utility, fairness, and robustness.

Through extensive experiments on both binary and multi-class clinical datasets, CUES demonstrated its ability to uncover shortcomings that conventional metrics overlook such as poor calibration, inequitable subgroup performance, and instability under data variability. The results consistently showed that models with high CUES scores were consistently not only accurate but also trustworthy and clinically interpretable, reinforcing the importance of multidimensional evaluation in healthcare AI. By offering a single interpretable score that reflects real-world suitability, the CUES framework promotes transparent model comparison, informed clinical adoption, and ethical deployment of AI systems.

Importantly, CUES also enables a fairness-and robustness-aware feature selection mechanism. By computing feature-wise contributions to overall CUES across samples and perturbations, the framework identifies features that are critical not just for predictive accuracy, but also for maintaining calibration, equity, and stability. This approach complements traditional explainable AI methods, such as SHAP, by offering a multi-objective perspective on feature importance. Visualizations such as Violin and bar plots of per-feature CUES contributions allow

illustrating both variability and mean impact, supporting the selection of top-k features for model training without compromising clinical reliability. This ensures that simplified models or reduced feature sets preserve interpretability and ethical performance, an essential consideration for biomedical applications where patient safety and fairness are paramount.

CUES is also particularly applicable to datasets containing anthropometric features, such as BMI, neck circumference, and body fat distribution. By quantifying calibration, equity, utility, and stability, CUES enables systematic assessment of how variations in these features affect model performance across patient subgroups. This facilitates the identification of potential biases in body composition and ensures that predictive models remain robust and fair across heterogeneous populations. Feature-wise CUES analyses can further highlight which anthropometric variables contribute meaningfully to predictive accuracy without undermining stability or equity, guiding both model selection and feature prioritization in studies exploring the clinical and physiological implications of body metrics.

Moreover, optimizing model hyperparameters using CUES as the objective proved highly effective in promoting clinically reliable performance. Tuning for CUES consistently produced models that balanced traditional metrics such as accuracy, sensitivity, and specificity with calibration, stability, and fairness. Even classifiers that did not achieve the highest AUROC values exhibited improved overall clinical reliability when optimized for CUES, demonstrating enhanced stability and subgroup fairness. These results confirm that CUES-guided optimization identifies not only statistically strong models but also robust, interpretable, and ethically sound models, validating the framework's practical utility for real-world clinical deployment.

Future work will extend the CUES framework to larger, real-world clinical datasets and explore its integration into model development pipelines for automated bias detection, domain

adaptation, and regulatory evaluation. While the current study establishes a strong theoretical and empirical foundation, several promising directions remain open. These include extending the CUES framework to unsupervised and survival prediction tasks, developing a differentiable surrogate loss to enable direct end-to-end optimization, and incorporating a complexity penalty term to balance model parsimony with predictive performance. Furthermore, validating CUES on clinical registries and longitudinal datasets will be essential to evaluate its temporal stability and generalizability. Ultimately, adopting the CUES framework can help align AI model evaluation with clinical priorities, ensuring that high-performing algorithms are not only statistically valid but also safe, fair, and effective in patient care.

References

- [1] Wynants, L.; Van Calster, B.; Collins, G.S.; Riley, R.D.; Heinze, G.; Schuit, E.; Bonten, M.M.J.; Dahly, D.L.; Damen, J.A.; Debray, T.P.A.; et al. Prediction Models for Diagnosis and Prognosis of COVID-19: Systematic Review and Critical Appraisal. *BMJ* 2020, 369, m1328. <https://doi.org/10.1136/bmj.m1328>.
- [2] Siontis, G.C.M.; Tzoulaki, I.; Siontis, K.C.; Ioannidis, J.P.A. Comparisons of Established Risk Prediction Models for Cardiovascular Disease: Systematic Review. *BMJ* 2012, 344, e3318. <https://doi.org/10.1136/bmj.e3318>.
- [3] Snell, K.I.; Archer, L.; Ensor, J.; Bonnett, L.J.; Debray, T.P.; Phillips, B.; Collins, G.S.; Riley, R.D. External validation of clinical prediction models: Simulation-based sample size calculations were more reliable than rules-of-thumb. *J. Clin. Epidemiol.* 2021, 135, 79–89. <https://doi.org/10.1016/j.jclinepi.2021.02.011>.

- [4] Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles Heel of Predictive Analytics. *BMC Med.* 2019, 17, 230. <https://doi.org/10.1186/s12916-019-1466-7>.
- [5] Harrell, F.E.; Lee, K.L.; Mark, D.B. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Stat. Med.* 1996, 15, 361–387. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960229\)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4).
- [6] Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G.M. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann. Intern. Med.* 2015, 162, 55–63. <https://doi.org/10.7326/M14-0697>.
- [7] Vickers, A.J.; Elkin, E.B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med. Decis. Mak.* 2006, 26, 565–574. <https://doi.org/10.1177/0272989X06295361>.
- [8] Hand, D.J. Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve. *Mach. Learn.* 2009, 77, 103–123. <https://doi.org/10.1007/s10994-009-5119-5>.
- [9] Cook, N.R. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation* 2007, 115, 928–935. <https://doi.org/10.1161/CIRCULATIONAHA.106.672402>.
- [10] Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 2015, 10, e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
- [11] Pencina, M.J.; D’Agostino, R.B.; Vasan, R.S. Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond. *Stat. Med.* 2008, 27, 157–172. <https://doi.org/10.1002/sim.2929>.

- [12] Brier, G.W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Wea. Rev.* 1950, 78, 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- [13] Steyerberg, E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*; Statistics for Biology and Health; Springer International Publishing: Cham, Switzerland, 2019; ISBN 978-3-030-16398-3.
- [14] Riley, R.D.; Archer, L.; Snell, K.I.E.; Ensor, J.; Dhiman, P.; Martin, G.P.; Bonnett, L.J.; Collins, G.S. Evaluation of Clinical Prediction Models (Part 2): How to Undertake an External Validation Study. *BMJ* 2024, 384, e074820. <https://doi.org/10.1136/bmj-2023-074820>.
- [15] Vickers, A.J.; van Calster, B.; Steyerberg, E.W. Net Benefit Approaches to the Evaluation of Prediction Models, Molecular Markers, and Diagnostic Tests. *BMJ* 2016, 352, i6. <https://doi.org/10.1136/bmj.i6>.
- [16] Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* 2018, 169, 866–872. <https://doi.org/10.7326/M18-1990>.
- [17] Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 2019, 366, 447–453. <https://doi.org/10.1126/science.aax2342>.
- [18] van Smeden, M.; Moons, K.G.M.; de Groot, J.A.H.; Collins, G.S.; Altman, D.G.; Eijkemans, M.J.C.; Reitsma, J.B. Sample Size for Binary Logistic Prediction Models: Beyond Events per Variable Criteria. *Stat. Methods Med. Res.* 2019, 28, 2455–2474. <https://doi.org/10.1177/0962280218784726>.
- [19] Steyerberg, E.W.; Vickers, A.J.; Cook, N.R.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.J.; Kattan, M.W. Assessing the Performance of Prediction Models: A Framework

- for Some Traditional and Novel Measures. *Epidemiology* 2010, 21, 128–138.
<https://doi.org/10.1097/EDE.0b013e3181c30fb2>.
- [20] Efron, B.; Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat. Sci.* 1986, 1, 54–75.
<https://doi.org/10.1214/ss/1177013815>.
- [21] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* 1979, 7, 1–26.
<https://doi.org/10.1214/aos/1176344552>.
- [22] Dubois, D.; Prade, H. A Review of Fuzzy Set Aggregation Connectives. *Inf. Sci.* 1985, 36, 85–121. [https://doi.org/10.1016/0020-0255\(85\)90027-1](https://doi.org/10.1016/0020-0255(85)90027-1).
- [23] Dujmovic, J.J. Continuous Preference Logic for System Evaluation. *IEEE Trans. Fuzzy Syst.* 2007, 15, 1082–1099. <https://doi.org/10.1109/TFUZZ.2007.902041>.
- [24] Rozovsky, L.V. Comparison of Arithmetic, Geometric, and Harmonic Means. *Math. Notes* 2021, 110, 118–125. <https://doi.org/10.1134/S0001434621070129>.
- [25] Kerr, K.F.; Brown, M.D.; Zhu, K.; Janes, H. Assessing the Clinical Impact of Risk Prediction Models with Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J. Clin. Oncol.* 2016, 34, 2534–2540. <https://doi.org/10.1200/JCO.2015.65.5654>.
- [26] van Houwelingen, H.C.; le Cessie, S. Predictive Value of Statistical Models. *Stat. Med.* 1990, 9, 1303–1325. <https://doi.org/10.1002/sim.4780091109>.
- [27] DeGroot, M.H.; Fienberg, S.E. The Comparison and Evaluation of Forecasters. *J. R. Stat. Soc. Ser. D (Stat.)* 1983, 32, 12–22. <https://doi.org/10.2307/2987588>.
- [28] Kass, R.E.; Wasserman, L. A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion. *J. Am. Stat. Assoc.* 1995, 90, 928–934.
<https://doi.org/10.1080/01621459.1995.10476592>.

- [29] Parmigiani, G. *Decision Theory: Principles and Approaches*; John Wiley & Sons: Chichester, UK, 2009; ISBN 978-0-471-49657-1.
- [30] Murphy, A.H. A New Vector Partition of the Probability Score. *J. Appl. Meteorol.* 1973, 12, 595–600. [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2).
- [31] Pepe, M.S.; Janes, H.; Longton, G.; Leisenring, W.; Newcomb, P. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *Am. J. Epidemiol.* 2004, 159, 882–890. <https://doi.org/10.1093/aje/kwh101>.
- [32] Nagelkerke, N.J.D. A Note on a General Definition of the Coefficient of Determination. *Biometrika* 1991, 78, 691–692. <https://doi.org/10.1093/biomet/78.3.691>.
- [33] U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. Available online: <https://www.fda.gov/media/145022/download> (accessed on 2 November 2025).
- [34] Kotter, E.; D’Antonoli, T.A.; Cuocolo, R.; Hierath, M.; Huisman, M.; Klontzas, M.E.; Martí-Bonmatí, L.; May, M.S.; Neri, E.; Nikolaou, K.; et al. Guiding AI in Radiology: ESR’s Recommendations for Effective Implementation of the European AI Act. *Insights Imaging* 2025, 16, 33. <https://doi.org/10.1186/s13244-025-01905-x>.
- [35] Elwali, A.; Moussavi, Z. A Novel Decision Making Procedure during Wakefulness for Screening Obstructive Sleep Apnea Using Anthropometric Information and Tracheal Breathing Sounds. *Sci. Rep.* 2019, 9, 11467. <https://doi.org/10.1038/s41598-019-47998-5>.
- [36] DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988, 44, 837–845. <https://doi.org/10.2307/2531595>.

- [37] Oehlert, G.W. A Note on the Delta Method. *Am. Stat.* 1992, 46, 27–29.
<https://doi.org/10.1080/00031305.1992.10475842>.
- [38] Riley, R.D.; Ensor, J.; Snell, K.I.E.; Debray, T.P.A.; Altman, D.G.; Moons, K.G.M.; Collins, G.S. Calculating the Minimum Sample Size Required for Developing a Multivariable Prediction Model: Part II—Binary and Time-to-Event Outcomes. *Stat. Med.* 2019, 38, 1276–1296. <https://doi.org/10.1002/sim.7992>.
- [39] Murdoch, B. Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era. *BMC Med. Ethics* 2021, 22, 122. <https://doi.org/10.1186/S12910-021-00687-3>.
- [40] Wiesinger, B.; Smith, R.; Morley, J.; Floridi, L. Ethics and Biases in Health AI Model Development. *NPJ Digit. Med.* 2020, 3, 86. <https://doi.org/10.1038/s41746-020-0298-3>.
- [41] Onitui, D.; Wachter, S.; Mittelstadt, B. How AI Challenges the Medical Device Regulation: Patient Safety, Benefits, and Intended Uses. *J. Law Biosci.* 2024, lsae007. <https://doi.org/10.1093/jlb/lsae007>.
- [42] Ueda, D.; Kakinuma, T.; Fujita, S.; Kamagata, K.; Fushimi, Y.; Ito, R.; Matsui, Y.; Nozaki, T.; Nakaura, T.; Fujima, N.; et al. Fairness of Artificial Intelligence in Healthcare: Review and Recommendations. *Jpn. J. Radiol.* 2024, 42, 3–15. <https://doi.org/10.1007/s11604-023-01474-3>.
- [43] Vardas, E.P.; Marketou, M.; Vardas, P.E. Medicine, Healthcare and the AI Act: Gaps, Challenges and Future Implications. *Eur. Heart J.—Digit. Health* 2025, 6, 833–839. <https://doi.org/10.1093/ehjdh/ztaf041>.

- [44] Bignami, E.; Darhour, L.J.; Franco, G.; Guarnieri, M.; Bellini, V. AI Policy in Healthcare: A Checklist-Based Methodology for Structured Implementation. *J. Anesth. Analg. Crit. Care* 2025, 5, 56. <https://doi.org/10.1186/s44158-025-00278-3>.

CHAPTER NINE: CONCLUSION, LIMITATIONS, AND FUTURE WORK

9.1 Summary and Conclusion

This research proposes and validates a novel, non-invasive framework for obstructive sleep apnea (OSA) severity classification using breathing sounds acquired during wakefulness and a compact set of anthropometric features. The overarching hypothesis driving this dissertation is that tracheal breathing sounds (TBS) recorded while a subject is awake encode clinically meaningful acoustic signatures of upper airway obstruction, and that these signatures, when extracted through rigorous signal processing and interpreted through physiologically grounded machine learning models can serve as reliable surrogate markers for OSA severity as quantified by the apnea–hypopnea index (AHI). The work presented in this dissertation advances the state of the art in wakefulness-based OSA screening through three progressive and complementary AI-based approaches, each designed to address a distinct dimension of the broader classification challenge: discriminative feature extraction, multi-class severity stratification, and clinically grounded interpretability.

The first approach established the foundational classification framework. A comprehensive multi-domain feature set was extracted from TBS recordings, encompassing spectral descriptors (power spectral density, bispectrum, bicoherence), time-domain statistics, time–frequency representations (wavelet, Constant-Q Transform), and nonlinear dynamics measures (fractal dimension, Lyapunov exponent, Hurst exponent, recurrence quantification). Feature dimensionality was reduced through a three-stage pipeline that integrated statistical filtering, SHAP (Shapley Additive Explanations)-based feature ranking, and Recursive Feature Elimination (RFE). Six pairwise binary base classifiers were then trained, one for each OSA severity comparison (Non-OSA vs. Mild, Non-OSA vs. Moderate, Non-OSA vs. Severe, Mild vs.

Moderate, Severe vs. Moderate), using bootstrap aggregation with Out-of-Bag (OOB) validation and Bayesian hyperparameter optimization across 20 candidate classifiers. Results demonstrated that nonlinear and spectral features, particularly bispectrum-derived texture measures, fractal dimension, and spectral entropy, were the most discriminative indicators of upper airway obstruction severity. The integration of anthropometric variables (BMI, neck circumference, Mallampati score) further enhanced classification accuracy and clinical relevance. The SVM with a third-degree polynomial kernel and Random Forest models delivered particularly strong performance, achieving high sensitivity and specificity across all pairwise comparisons on fully independent blind test sets. These findings confirmed that TBS recorded in under 10 minutes during wakefulness can yield classification accuracy sufficient for clinical screening purposes, representing a significant advancement over subjective questionnaire-based tools such as STOP-Bang and the Berlin questionnaire.

The second approach extended this binary framework to multi-class OSA severity stratification by introducing a stacked meta-model architecture. The probability outputs of the six pairwise base classifiers were used as input features to a meta-classifier, enabling it to learn higher-order decision patterns across all severity levels simultaneously. Multiple meta-classifiers were evaluated, with a neural network meta-model demonstrating the strongest generalization. Two classification schemes were examined: a three-class configuration (Non-OSA, Mild-OSA, and Moderate/Severe-OSA) and a more clinically granular four-class configuration (Non-OSA, Mild-OSA, Moderate-OSA, and Severe-OSA). In the three-class setting, OOB accuracy reached 91.1% (sensitivity: 91.6%, specificity: 95.3%), with blind test accuracy of 76.7% (sensitivity: 77.7%, specificity: 87.1%). In the four-class setting, OOB accuracy was 88.2% (sensitivity: 91.6%, specificity: 88.2%), and blind test accuracy was 76.7% (sensitivity: 75%, specificity: 92%). A

conformal prediction (CP) framework was further incorporated to provide principled uncertainty quantification. Across both classification schemes, the conformal model achieved full coverage (1.0) with an average prediction set size of 2, confirming that every test instance's prediction set contained the true class label while transparently communicating diagnostic ambiguity in borderline cases. Statistical significance analyses using paired t-tests, Wilcoxon signed-rank tests, and McNemar's tests confirmed that the meta-model significantly outperformed all six individual base classifiers, with large effect sizes (Cohen's $d > 0.65$; Cliff's $\delta \geq 0.36$) after Bonferroni correction. A learning-curve analysis further demonstrated that model performance improved monotonically with increased training data without saturation, indicating that the current framework is operating in a data-limited regime and is expected to yield additional gains with larger, more diverse cohorts.

The third approach focused on the clinical interpretability and physiological relevance of the extracted acoustic features. Rather than targeting classification performance alone, each feature was evaluated through a multidimensional lens: discriminative power (AUC), cross-validation stability (Absolute Delta AUC), anthropometric correlation, and SHAP-based explainability. A Structure–Function–Symptom interpretive framework was adopted to link the most discriminative acoustic descriptors to underlying upper airway anatomy (structural collapsibility), airflow turbulence patterns (functional flow limitation), and clinical disease burden (AHI-indexed symptom severity). Spectral, bispectral, and fractal-based nonlinear features, particularly bispectrum bounding-box texture energy, spectral skewness, fractal dimension, and bispectrum energy value, consistently emerged as the most stable and physiologically meaningful biomarkers across all pairwise severity comparisons. Notably, texture energy exhibited the strongest negative correlation with OSA severity, reflecting the transition from laminar, organized airflow in healthy

individuals to chaotic, broadband turbulence in severe OSA. Spectral skewness, conversely, showed a strong positive correlation ($r \approx 0.99$), representing the progressive shift toward intermittent, high-force inspiratory maneuvers as a compensatory response to increasing pharyngeal resistance. The strong correlations between peak acoustic intensity and neck circumference, and between entropy and BMI, provided direct anatomical validation of the extracted features, demonstrating that the acoustic biomarkers are not abstract statistical constructs but genuine physical reflections of the upper airway's morphological risk factors for OSA. Mouth-inspiration breathing segments consistently emerged as the richest source of discriminative features, consistent with the known vulnerability of the pharyngeal airway to dynamic narrowing and turbulence during oral inspiration.

Collectively, the three approaches presented in this dissertation provide a comprehensive, layered framework for OSA severity assessment during wakefulness. The results support the feasibility of TBS as a practical, rapid, and non-invasive alternative to polysomnography (PSG) for clinical triage and early detection. The incorporation of uncertainty quantification through conformal prediction and explainability through SHAP and the Structure–Function–Symptom framework meaningfully enhances the trustworthiness and translational readiness of the proposed system. The framework's sub-10-minute acquisition time, passive breathing protocol, and applicability in preoperative clinics, dental practices, and home-based settings position it as a clinically deployable complement to, and eventual substitute for, resource-intensive overnight monitoring. Together, these findings establish a rigorous proof-of-concept foundation for accessible, reliable, and personalized screening tools for sleep-disordered breathing.

9.2 Limitations

Despite the promising outcomes reported in this dissertation, several methodological and clinical limitations should be acknowledged. These limitations are important for contextualizing the current findings and for guiding the design of future investigations.

1. Dataset Size and Demographic Diversity: All three classification approaches were developed and evaluated using a single institutional dataset of 199 participants. While the dataset is well-characterized and includes four OSA severity strata with balanced cohort stratification, its relatively limited size particularly in the Mild-OSA subgroup ($n = 35$) limits the statistical power of between-group comparisons and constrains the generalizability of the findings. The learning-curve analyses confirmed that current performance is data-limited, and further improvements are anticipated with larger cohorts. Furthermore, the study population does not fully capture the demographic heterogeneity of real-world OSA populations, including variations in ethnicity, comorbidities (e.g., obesity hypoventilation syndrome, COPD), continuous positive airway pressure (CPAP) use history, and upper airway anatomy, all of which could introduce systematic biases in TBS-based classification models. Compounding this, no independent external validation using a dataset from a separate clinical center has been conducted. Although the framework was evaluated on held-out blind test sets and through k-fold cross-validation, external validation on a demographically and geographically diverse cohort remains essential before clinical translation, as it would test the framework's ability to generalize beyond the acoustic and demographic characteristics of the training population. Additionally, inter-session variability in TBS features was not assessed; confirming that an individual's acoustic profile is sufficiently stable across recording sessions is a prerequisite for clinical deployment and should be addressed in future multi-site studies.

2. Cross-Sectional Design and Absence of Repeated-Measures Validation: The study employed a cross-sectional design, with OSA severity determined by a single night of PSG or home sleep testing. OSA severity is known to vary night-to-night and may be influenced by factors such as sleep position, alcohol consumption, sedation, and nasal congestion. The current framework does not account for this intra-individual variability. A repeated-measures design incorporating multiple nights of PSG assessments and corresponding TBS recordings within a short observation window (e.g., 3–7 consecutive nights) would better capture intra-individual night-to-night variability and would strengthen the validity of the reported diagnostic correlations.

3. Acoustic Feature Interpretability and Physical Grounding: While the SHAP-based framework and Structure–Function–Symptom analysis provide a meaningful layer of interpretability, a number of the most discriminative features, particularly bispectrum-derived image morphological descriptors (e.g., bounding-box texture energy, Euler number, connected components), remain phenomenologically abstract and lack a complete first-principles acoustic derivation. Future work should seek to formally bridge these statistical discriminators with physically grounded acoustic transmission models of the upper airway, such as those developed using computational fluid dynamics (CFD) and finite-element-based acoustic propagation frameworks.

4. Binary AHI-Based Ground Truth: The classification targets employed in this dissertation are derived from AHI thresholds defined by clinical consensus ($AHI < 5$: Non-OSA; $5 \leq AHI < 15$: Mild; $15 \leq AHI < 30$: Moderate; $AHI \geq 30$: Severe). However, AHI is itself an imperfect and reductionist metric that does not fully capture the physiological heterogeneity of OSA (e.g., oxygen desaturation index, arousal index, hypopnea phenotype). The assumption that TBS features map linearly onto AHI-derived severity strata may not hold across all physiological

subtypes of the disorder. Future frameworks should consider complementary severity metrics, including the oxygen desaturation index (ODI) and the cumulative time spent below SpO₂ thresholds.

5. Conformal Prediction Set Efficiency in High-Dimensional Multi-Class Settings: In the four-class conformal prediction framework, the average prediction set size reached up to 4 for some ambiguous test samples, effectively spanning the entire label space and offering limited diagnostic resolution. This behavior, while theoretically sound within the conformal guaranteed framework, reduces the clinical utility of the uncertainty estimates in cases where fine-grained severity discrimination is most needed. Developing more efficient conformal score functions, such as Regularized Adaptive Prediction Sets (RAPS) or Conformalized likelihood-ratio scores, may improve set efficiency without compromising coverage.

9.3 Future Work

Building on the findings and limitations identified in this dissertation, the following directions are proposed to advance clinical viability, methodological rigor, and physiological depth of the proposed TBS-based OSA screening framework.

1. Large-Scale Multi-Center Validation: The most critical next step is the collection and analysis of TBS data from a substantially larger, demographically diverse cohort across multiple clinical centers and geographic regions. A large-scale validation study should target 500–1000 participants, with balanced representation across OSA severity levels, sex, age, BMI ranges, ethnicity, and comorbidity profiles. Such a dataset would simultaneously address the data-limited regime identified by the learning-curve analyses, enable robust external validation, and support the development of severity-specific normative references for acoustic biomarkers.

2. Integration with Established Airway Phenotyping Tools: Future studies should systematically integrate TBS-based assessments with complementary upper airway measurement tools, including the pharyngometer, rhinometer, and rhinomanometer, which provide quantitative measurements of airway cross-sectional area, segment length, and airflow resistance. Combining these objective anatomical measurements with TBS-derived acoustic features would enable direct mapping between structural airway properties and the spectral characteristics of breathing sounds, thereby providing a mechanistic, physics-grounded validation of the acoustic biomarkers. This integration would also facilitate the development of a parametric acoustic model of the upper airway, linking geometric parameters (e.g., cross-sectional area at the velopharynx and oropharynx, airway length) to resonance frequencies and spectral envelope features in TBS.

3. Fusion with Medical Imaging Data: Combining TBS analysis with CT- or MRI-derived anatomical measurements would enable direct correlation of airway geometric features, such as minimum cross-sectional area at the retropalatal and retroglossal levels, airway volume, and pharyngeal length, with the discriminative acoustic features identified in this dissertation. This fusion approach would provide multimodal validation of the acoustic biomarkers and enable the development of anatomy-informed TBS classification models that account for inter-individual anatomical variability.

4. Development of a Portable, Automated Acquisition and Analysis System: To facilitate deployment outside specialized sleep laboratories, future work should develop an integrated hardware–software system that combines a miniaturized contact-microphone fixture for tracheal recording with an automated signal-processing and classification pipeline. The hardware design should ensure consistent geometric alignment between the recording sensor and the tracheal surface, minimizing inter-session and inter-operator variability. The accompanying software

should automate the entire workflow from raw signal acquisition and preprocessing through feature extraction, normalization, and severity classification with a user-friendly graphical interface accessible to non-specialist clinicians and, ultimately, to patients themselves in home-based settings.

5. Deep Learning and Foundation Model Approaches: Future investigations should systematically explore deep learning architectures, including convolutional neural networks (CNNs), recurrent networks (LSTMs, Gated recurrent unit (GRU)), transformer-based models, and hybrid CNN-LSTM architectures for end-to-end TBS-based OSA severity classification. Unlike the hand-crafted feature extraction pipeline employed in this dissertation, deep learning models can potentially learn latent acoustic representations directly from raw waveforms or spectrograms, circumventing the need for domain-specific feature engineering. The preliminary CNN-LSTM model developed in Chapter Three demonstrated the feasibility of this approach but was constrained by a limited dataset size. With a substantially larger corpus of TBS recordings, deep learning methods are expected to outperform hand-crafted feature frameworks. Furthermore, self-supervised pretraining on large unlabeled TBS repositories, followed by fine-tuning on labeled severity-stratified datasets, could address the annotation bottleneck and improve generalization.

6. Incorporation of Sleep-State Recordings and Multi-Modal Physiological Fusion: While the central contribution of this dissertation is the development of a wakefulness-based screening framework, future studies should investigate the complementary value of integrating TBS recorded during natural sleep or during CPAP titration trials with wakefulness-based features. The combination of awake and sleep-state acoustic signatures may capture complementary dimensions of upper airway vulnerability, anatomical predisposition during wakefulness, and dynamic collapsibility during sleep, yielding richer multimodal biomarker profiles for more

precise severity stratification. Integration of peripheral physiological signals, including photoplethysmography (PPG)-derived oxygen saturation (SpO₂), ECG-derived cardiorespiratory coupling indices, and actigraphy, could further enhance diagnostic accuracy and provide a more comprehensive characterization of OSA-related end-organ consequences.

7. Clinical Benchmarking Against Standard Screening Tools and Health Economic

Analysis: Future studies should rigorously benchmark the proposed TBS framework against currently validated OSA screening tools, including the STOP-Bang questionnaire, Berlin questionnaire, Epworth Sleepiness Scale, and home sleep testing (HST) devices, by assessing sensitivity, specificity, positive predictive value, and negative predictive value across population subgroups. A formal health economic analysis should also be conducted to quantify the cost-effectiveness of TBS-based pre-screening as a stratification tool to guide PSG referrals, taking into account device costs, acquisition time, personnel requirements, and healthcare system savings from avoided unnecessary PSGs.

8. Regulatory Pathway and Clinical Translation: Given the level of evidence accumulated in this dissertation and its companion publications, a structured pathway toward regulatory consideration should be initiated. This would require the development of a locked, validated classification algorithm with clearly defined performance specifications, training on a representative multi-center dataset, and evaluation using pre-specified primary endpoints (e.g., sensitivity $\geq 90\%$ for AHI ≥ 15 at a specificity $\geq 75\%$). The CUES composite evaluation metric, developed in Chapter Eight, integrating clinical utility, calibration, equity, and stability into a single multiplicative score, provides a regulatorily transparent framework for holistic model assessment and should serve as the primary evaluation instrument in future multi-center validation studies and regulatory submissions.

APPENDIX A

A.1 Anthropometric Missing Value Imputation

The imputation of missing anthropometric data was achieved within each severity group (Non-, Mild, Moderate and Severe OSA) using a k-nearest neighbors (k-NN) based approach [1]. The k-NN imputation estimates missing entries by leveraging the similarity among samples within the same OSA category, thus preserving the inherent structure and distribution of the data. This method enhances data completeness while minimizing the potential bias introduced by missingness [1].

By combining severity-specific grouping with localized imputation, the preprocessing approach ensures robust and consistent feature sets that improve the reliability of downstream modeling and statistical analyses [1,2]. Figure A1 illustrates the steps involved in filling in missing anthropometric data.

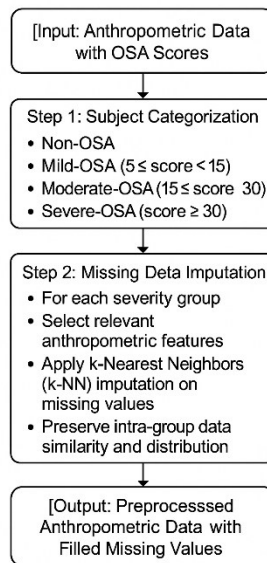


Fig. A-1 Flow chart of missing anthropometric data filling.

A.2 Automatic Feature Normalization

An automatic selection algorithm was implemented to choose the most suitable normalization method for a given dataset, thereby optimizing feature preprocessing. For each feature $X_i \in X \subset \mathbb{R}^{n \times d}$, four normalization techniques were evaluated: min–max scaling, z-score normalization, mean-range scaling, and robust scaling. Each feature was normalized individually and discretized into ten bins, and its dependency on the categorical labels y was quantified using mutual information [3–5]:

$$I(X; Y) = \sum_{i,j} p(x_i, y_j) \log \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right) \quad \text{A-1}$$

The normalization method that yielded the highest cumulative mutual information across all features was selected. This adaptive selection ensured that feature–label relationships were preserved and enhanced during preprocessing [3]. Then, the chosen normalization was performed using one of four methods:

Min–Max Scaling: Rescales data to the [0, 1] range [54] using

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad \text{A-2}$$

Z-Score Normalization: Standardizes features to a zero mean and unit variance [6]:

$$x' = \frac{x - \mu}{\sigma} \quad \text{A-3}$$

Mean-Range Scaling: Centers by the mean and scales by the range [6]:

$$x' = \frac{x - \mu}{\text{range}(x)} \quad \text{A-4}$$

Robust Scaling: Centers by the median and scales by the interquartile range (IQR) [6]:

$$x' = \frac{x - \text{median}(x)}{\text{IQR}(x)} \quad \text{A-5}$$

For each method, parameters (mean, standard deviation, minimum, maximum, median, IQR) were computed from the training data if not pre-specified, enabling their consistent application to the testing sets. The proposed automatic normalization selection ensures the preprocessing step is systematically adapted to the underlying data distribution. The method enhances feature relevance and model discriminability by maximizing the mutual information between rescaled features and categorical labels [3]. Furthermore, it increases resilience to outliers, non-Gaussian distributions, and varying feature ranges. This adaptive strategy enhances the robustness and generalization capabilities of subsequent learning models. Figure A2 shows a flow chart of the automatic feature normalization logic.

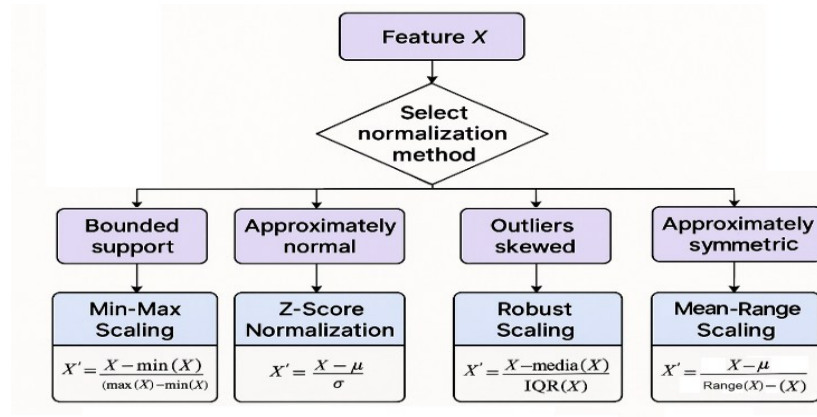


Fig. A-2 Flow chart of automatic feature normalization.

A.3 Selecting Best Features

We implemented a robust, three-stage feature selection pipeline to ensure that the learning algorithms, particularly ensemble models, operated on a compact and informative set of inputs. This section describes the methodology in detail and explains the rationale for its use in models' learning. The feature selection process combines filter, wrapper, and embedded methods to progressively refine the feature set. A pipeline flow chart is shown in Figure A3.

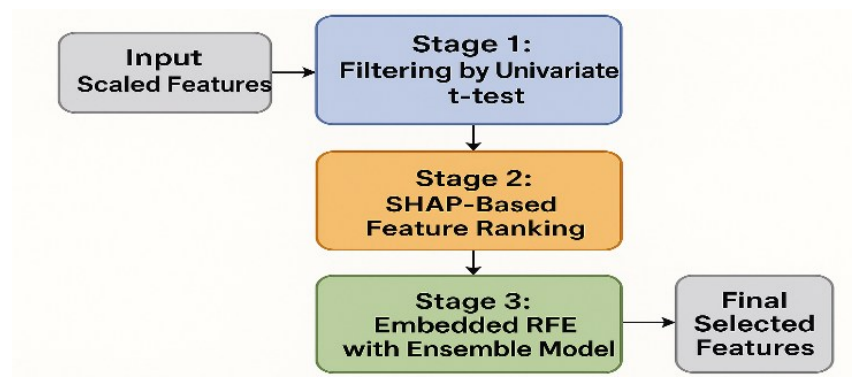


Fig. A-3 Flow chart of feature selection methodology.

The pipeline consisted of three main stages: univariate filtering using t-test, model-based feature ranking via SHAP values, and Recursive Feature Elimination (RFE) with RUSBoost ensemble. This hierarchical approach helped eliminate redundant, noisy, or non-informative features, thereby enhancing model generalizability and interpretability.

A.3.1 Stage 1: Filtering by Univariate t-Test

We applied a two-sample unpaired t-test to each feature in the first stage. The objective was to detect statistically significant differences in feature values between the two label groups. Features with a p-value ≤ 0.05 were retained for the next stage [7,8]. This method is computationally efficient and effective for removing features that are unlikely to contribute to class separation. However, it does not consider interactions between features or their impact on the learning algorithm's performance [7,8].

A.3.2 Stage 2: SHAP-Based Feature Ranking

Features retained from the t-test were further ranked using SHAP (SHapley Additive exPlanations) values. A classifier was trained using an Error-Correcting Output Codes (ECOC) ensemble model, and SHAP values were computed over multiple iterations [9,10]. In each iteration:

- The classifier was trained and evaluated.
- SHAP values were computed for each feature.
- The top $N/2$ features with the highest mean absolute SHAP value were stored.

Finally, we selected the features that appeared most frequently across iterations, ensuring robustness to sampling variance. This step introduces model explainability into the selection process, allowing for interpretability and greater confidence in the selected features [9,10]. Figure A4 shows the calculated shapely values for selected features when using the proposed method for base model 6.

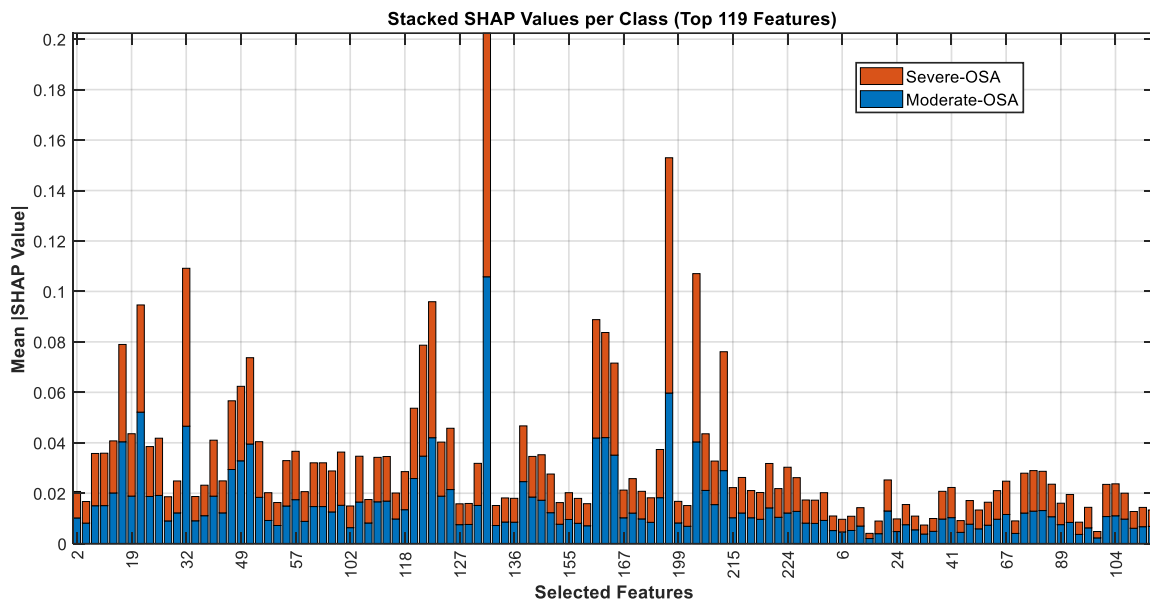


Fig. A-4 Shapley values calculated for selected features of base model 6.

A.3.3 Embedded Recursive Feature Elimination (RFE) with Ensemble Model

The final stage employed Recursive Feature Elimination (RFE) using a RUSBoost ensemble classifier [58,59]. This method was embedded within the model training process, recursively ranking and removing the least essential features until the desired number of features was retained. RFE was conducted within a cross-validation loop to enhance generalizability and

avoid overfitting a specific training partition. The most consistently retained features across folds were selected. This embedded method leverages the inherent feature importance estimates of the RUSBoost model and tailors feature selection to the classification task [11,12].

The proposed multi-stage feature selection pipeline offers several advantages, particularly within ensemble learning frameworks such as stacked generalization. The pipeline effectively balances speed, robustness, and task relevance by integrating filter methods such as t-tests, wrapper methods like RFE, and embedded techniques including SHAP-based ranking. Early-stage filtering removes non-discriminative features, reducing noise and enhancing the efficiency of subsequent modeling stages. SHAP values support model transparency and interpretability, which are essential in biomedical and clinical AI applications. Moreover, RFE aligns the selected features with the learning algorithm, ensuring they are optimized for predictive performance. As a result, the selected features are compact and stable and can be applied to the different experimental models used within the ensemble. In stacked ensemble systems, where individual experimental models are trained independently and a meta-learner combines their predictions, the quality of the features provided to each base model plays a critical role [13]. If experimental models are trained on noisy or inconsistent features, the ensemble's performance can suffer due to weak or conflicting individual predictions. The proposed pipeline addresses this by ensuring each base model receives a consistent and informative feature set. Consequently, the meta-learner benefits from high-quality base predictions, which improves the overall strength and stability of the ensemble. This also reduces the generalization gap across both training and unseen data distributions. Therefore, the proposed feature selection pipeline is crucial for enhancing the accuracy, interpretability, and robustness of ensemble learning systems [14]

Table A-1 Summary of extracted features for all models.

Power Spectrum Features	Bi-Spectrum Features	Wavelet	Wavelet	MFCC	TP	CDF
MeanPower	Hf2 2	Wavelet Approx Mean L1	Wavelet Approx Mean L4	MFCC Mean C1	TP Histogram 1	CDFMean
StandardDeviation	En F Bi D 3	Wavelet Approx Std L1	Wavelet Approx Std L4	MFCC Std C1	TP Histogram 2	CDFStd
Maximum	Mean Bi D F 3	Wavelet Approx Skewness L1	Wavelet Approx Skewness L4	MFCC Skewness C1	TP Histogram 3	Lyapunov
Slope	WCOB D Fx 3	Wavelet Approx Kurtosis L1	Wavelet Approx Kurtosis L4	MFCC Kurtosis C1	TP Entropy	Lyapunov Exponent Mean
SCBWCenter	WCOB D Fy 3	Wavelet Approx Entropy L1	Wavelet Approx Entropy L4	MFCC Median C1	TP Energy	Lyapunov Exponent Std
SCBWBandwidth	Hf1 3	Wavelet Approx Log Energy L1	Wavelet Approx Log Energy L4	MFCC Range C1	TP Skewness	Lyapunov Exponent Max
SpectralSkewness	Hf2 3	Wavelet Approx Max To Min Ratio L1	Wavelet Approx Max To Min Ratio L4	MFCC Entropy C1	TP Kurtosis	Lyapunov Exponent Min
SpectralKurtosis	Total Energy	Wavelet Approx Spectral Centroid L1	Wavelet Approx Spectral Centroid L4	MFCC Mean Abs Diff C1	TP Max Prob	Recurrence
PeakFrequency	Normalized Energy	Wavelet Approx Spectral Bandwidth L1	Wavelet Approx Spectral Bandwidth L4	Spectral Centroid C1	TP Ratio1	RecurrenceDeterminism
SpectralEnergy	Max Abs Bi	Wavelet Detail Mean L1	Wavelet Detail Mean L4	Spectral Bandwidth C1	TP Ratio2	Amplitude Modulation
SpectralEntropy	Mean Abs Bi	Wavelet Detail Std L1	Wavelet Detail Std L4	MFCC Mean C2	EP	AmplitudeModulationMean
ZeroCrossing Rate	Entropy Skewness	Wavelet Detail Skewness L1	Wavelet Detail Skewness L4	MFCC Std C2	EP Energy	AmplitudeModulationStd

RMS	Entropy Kurtosis	Wavelet Detail Kurtosis L1	Wavelet Detail Kurtosis L4	MFCC Skewness C2	EP Mean Energy	Amplitude ModulationMax
SpectralFlatness	Symmetry Metric	Wavelet Detail Entropy L1	Wavelet Detail Entropy L4	MFCC Kurtosis C2	EP Max Energy	Amplitude ModulationMin
FM2MFreq1	Asymmetry Ratio	Wavelet Detail Log Energy L1	Wavelet Detail Log Energy L4	MFCC Median C2	EP Min Energy	Amplitude ModulationMedian
FM2MFreq2	Sym Mean	Wavelet Detail Max To Min Ratio L1	Wavelet Detail Max To Min Ratio L4	MFCC Range C2	EP Std Energy	Miscellaneous
FreqSkewness1	Sym Max	WaveletApproxSkewnessL3	Wavelet Detail Spectral Centroid L4	MFCC Entropy C2	AVP	DFA_ScalingExponent
FreqSkewness2	Sym Std	WaveletApproxKurtosisL3	Wavelet Detail Spectral Bandwidth L4	MFCC Mean Abs Diff C2	AVP Histogram 1	
SpectralCrest	Sym Var	WaveletApproxEntropyL3	TQWT	Spectral Centroid C2	AVP Histogram 2	
BandPower	Mean Value	WaveletApproxLog EnergyL3	TQWT	Spectral Bandwidth C2	AVP Mean	
Bi-Spectrum Features	Std Value	Wavelet Approx Entropy L2	SpectralEntropy	MFCC Mean C3	AVP Std	
En T Bi	Skewness Value	Wavelet Approx Log Energy L2	BandPowerLow	MFCC Std C3	AVP Max	
En T Bi D	Kurtosis Value	Wavelet Approx Max To Min Ratio L2	BandPowerMid	MFCC Skewness C3	AVP Min	
En F Bi	Range Value	Wavelet Approx Spectral Centroid L2	BandPowerHigh	MFCC Kurtosis C3	AVP Entropy	
En F Bi D	Energy Value	Wavelet Approx Spectral Bandwidth L2	CQT	MFCC Median C3	AVP Energy	
Mean Bi T F	Median Value	Wavelet Detail Mean L2	CQT Mean Power	MFCC Range C3	DIR	
Mean Bi D F	IQR Value	Wavelet Detail Std L2	CQT Std Power	MFCC Entropy C3	DIR Histogram 1	

Mean Bi T F G	Coef Variation	Wavelet Detail Skewness L2	CQT Skewness Power	MFCC Mean Abs Diff C3	DIR Histogram 2	
Mean Bi D F G	Region Area	Wavelet Detail Kurtosis L2	CQT Kurtosis Power	Spectral Centroid C3	DIR Histogram 3	
Mean Bi T F H	Bounding Box Area	Wavelet Detail Entropy L2	CQT Total Energy	Spectral Bandwidth C3	MAG Mean	
Mean Bi D F H	Aspect Ratio	Wavelet Detail Log Energy L2	CQT Temporal Centroid	MFCC Mean C4	MAG Std	
WCOB Tx	Centroid X	Wavelet Detail Max To Min Ratio L2	CQT Temporal Spread	MFCC Std C4	MAG Max	
WCOB Ty	Centroid Y	Wavelet Detail Spectral Centroid L2	CQT Spectral Centroid	MFCC Skewness C4	MAG Min	
WCOB Dx	Perimeter	Wavelet Detail Spectral Bandwidth L2	CQT Spectral Bandwidth	MFCC Kurtosis C4	DIR Entropy	
WCOB Dy	Compactness	Wavelet Approx Mean L3	CQT Spectral Flatness	MFCC Median C4	DIR Energy	
WCOB T Fx	Bounding Box Diagonal	Wavelet Approx Std L3	CQT Time Entropy	MFCC Range C4	Noise Metrics	
WCOB T Fy	Peak Value	Wavelet Approx Skewness L3	CQT Freq Entropy	MFCC Entropy C4	NoiseToHar monicRatio	
WCOB D Fx	Frequency Centroid X	Wavelet Approx Kurtosis L3	CQT Gabor Energy Std	MFCC Mean Abs Diff C4	Shimmer	
WCOB D Fy	Frequency Centroid Y	Wavelet Approx Entropy L3	CQT Gabor Mean Mean	Spectral Centroid C4	Jitter	
H1	Frequency Bandwidth X	Wavelet Approx Log Energy L3	CQT Gabor Std Mean	Spectral Bandwidth C4	PBP	
H2	Frequency Bandwidth Y	Wavelet Approx Max To Min Ratio L3	CQT Gabor Skewness Mean	MFCC Mean C5	PBP Mean	
En F Bi D 1	Spectral Flux	Wavelet Approx Spectral Centroid L3	CQT Gabor Kurtosis Mean	MFCC Std C5	PBP Variance	
Mean Bi D F 1	Entropy Value	Wavelet Approx Spectral Bandwidth L3	CQT Freq Shifts Mean	MFCC Skewness C5	PBP Skewness	

WCOB D Fx 1	Texture Contrast	Wavelet Detail Mean L3	CQT Freq Shifts Std	MFCC Kurtosis C5	PBP Kurtosis	
WCOB D Fy 1	Texture Correlation	Wavelet Detail Std L3	CQT Freq Shifts Dynamic Range	MFCC Median C5	PBP Entropy	
Hf1 1	Texture Energy	Wavelet Detail Skewness L3	CQT Freq Intervals Mean	MFCC Range C5	LBP	
Hf2 1	Texture Homogeneity	Wavelet Detail Kurtosis L3	CQT Freq Intervals Std	MFCC Entropy C5	LBP Mean	
En F Bi D 2	Fractal Dimension	Wavelet Detail Entropy L3	CQT Bandwidth Mean	MFCC Mean Abs Diff C5	LBP Variance	
Mean Bi D F 2	Connected Components	Wavelet Detail Log Energy L3	CQT Bandwidth Std	Spectral Centroid C5	LBP Skewness	
WCOB D Fx 2	Euler Number	Wavelet Detail Max To Min Ratio L3	CQT Bandwidth Dynamic Range	Spectral Bandwidth C5	LBP Kurtosis	
WCOB D Fy 2	Num Holes	Wavelet Detail Spectral Centroid L3			LBP Entropy	

PSD: Power Spectral Density; Welch's method: Welch's method spectral estimation; CQT: Constant-Q Transform; MFCCs: Mel-Frequency Cepstral Coefficients; LBP: Local Binary Patterns; PBP: Probabilistic Binary Patterns; TP: Ternary Patterns; GBP: Gradient Binary Patterns; EP: Energy Patterns; AVP: Amplitude Variation Patterns; TQWT: Tunable Q-Factor Wavelet Transform; VMD: Variational Mode Decomposition; PSR: Phase Space Reconstruction; DFA: Detrended Fluctuation Analysis; NHR: Noise-to-Harmonics Ratio.

Table A-2 Summary of selected features for each model.

Feature Number	Non-OSA vs. Mild-OSA	Non-OSA vs. Moderate-OSA	Non-OSA vs. Severe-OSA	Mild-OSA vs. Moderate-OSA	Mild-OSA vs. Severe-OSA	Moderate-OSA vs. Severe-OSA
1	MouthExpiration_Gap_163-296_SpectralEntropy	Average_Gap_45-329_SpectralSkewness	NoseInspiration_Gap_816-885_SpectralKurtosis	NoseExpiration_Gap_299-707_MeanPower	MouthExpiration_Gap_367-515_SCBW_Bandwidth	NoseExpiration_Gap_384-567_SpectralCrest
2	MouthExpiration_Gap_163-296_SpectralCrest	NoseInspiration_Gap_321-577_SpectralSkewness	NoseInspiration_Gap_816-885_FreqSkewness1	MouthExpiration_BBox_229_712_22_26_iqrValue	Average_Average_BBboxes_EnFBiD_2	Average_BBboxes_584_362_21_35_numHoles

3	MouthExpiration_Gap_1375-1499_SpectralCrest	Average_BBox_548_370_24_29_entropyValue	NoseExpiration_Gap_986-1325_SCBW_Bandwidth	Mouth_Inspiration_FNMidFlow_MFCCMean_C1	NoseExpiration_BBx_787_244_15_15_rangeValue	Mouth_Inspiration_FNMidFlow_WaveletApproxEntropyL2
4	NoseInspiration_Gap_1066-1424_SCBW_Bandwidth	MouthInspiration_BBx_203_786_12_11_rangeValue	Average_BBox_527_294_11_18_textureEnergy	Nose_Expiration_FNMidFlow_WaveletApproxSkewnessL3	Average_Gap_73-307_SpectralCrest	Nose_Expiration_FNMidFlow_PeakCount
5	Mouth_Inspiration_FNMidFlow_MFCCMean_C1	Mouth_Expiration_FNMidFlow_WaveletApproxSpectralBandwidthL1	Average_BBox_526_336_15_18_centroidY	Mouth_Inspiration_FN_MFCCMean_C1	NoseInspiration_BBx_766_223_15_15_medianValue	Mouth_Inspiration_FN_MFCCMean_C2
6	Mouth_Inspiration_FNMidFlow_MFCCMedian_C5	Mouth_Expiration_FNMidFlow_WaveletDetailEntropyL3	NoseInspiration_BBx_0_0_22_23_aspectRatio	Mouth_Inspiration_FN_MFCCStd_C1	NoseInspiration_BBx_220_207_581_615_peakValue	MouthExpiration_Gap_282-460_SCBW_Bandwidth
7	Nose_Expiration_FNMidFlow_WaveletApproxMeanL3	MouthInspiration_BBx_229_756_15_17_iqrValue	MouthInspiration_BBx_0_934_73_89_eulerNumber	Mouth_Inspiration_FN_SpectralCentroid_C1	NoseInspiration_BBx_766_223_15_15_frequencyCentroidY	MouthInspiration_Gap_79-281_RMS
8	NoseExpiration_Gap_534-1003_BandPower	NoseInspiration_BBx_533_433_14_13_coefVariation	MouthInspiration_BBx_74_955_16_16_meanValue	Mouth_Inspiration_FN_SpectralBandwidth_C1	NoseInspiration_BBx_989_487_34_58_peakValue	NoseInspiration_Gap_327-532_MeanPower
9	NoseExpiration_Gap_1029-1329_SpectralSkewness	NoseExpiration_BBx_515_348_11_29_entropyValue	MouthInspiration_BBx_74_955_16_16_stdValue	Mouth_Inspiration_FN_MFCCKurtosis_C2	NoseInspiration_BBx_989_487_34_58_frequencyCentroidY	NoseInspiration_Gap_327-532_SCBW_Bandwidth
10	MouthInspiration_BBx_10_480_22_39_numHoles	Average_FNMidFlow_DIR_Histogram_2	MouthExpiration_Average_BBxes_totalEnergy	Mouth_Expiration_FN_WaveletDetailSpectralCentroidL2	NoseInspiration_BBx_989_487_34_58_spectralFlux	NoseInspiration_Gap_327-532_SpectralEnergy
11	MouthExpiration_BBx_30_0_11_13_frequencyCentroidX	Average_FNMidFlow_AVP_Histogram_2	MouthExpiration_Average_BBxes_normalizedEnergy	Mouth_Expiration_FN_CQTBandwidthDynamicRange	NoseInspiration_BBx_989_487_34_58_connectedComponents	NoseInspiration_Gap_327-532_BandPower
12	MouthExpiration_BBx_504_333_51_42_kurtosisValue	Mouth_Inspiration_FNMidFlow_ZeroCrossing1	MouthExpiration_Average_BBxes_stdAbsBi	Mouth_Expiration_FN_MFCCStd_C1	NoseInspiration_BBx_989_487_34_58_eulerNumber	MouthInspiration_BBx_290_464_45_80_regionArea

13	Mouth Expiration_BB ox_504_333_5 1_42_bounding BoxArea	Mouth_Inspiratio n_FNMidFlow_W aveletDetailSpectr alBandwidthL1	Mouth Expiration_Ave rage_BBboxes_s ymStd	Mouth_Expiration_FN_Spectr alCentroid_C1	Nose Inspiration_BB ox_226_683_94_ 101_iqrValue	Nose Inspiration_B Box_367_371_ 287_304_iqr Value
14	Mouth Expiration_BB ox_504_333_5 1_42_bounding BoxDiagonal	Mouth_Inspiratio n_FNMidFlow_W aveletDetailSpectr alBandwidthL4	Mouth Expiration_BB ox_588_259_1 69_356_perime ter	Mouth_Expiration_FN_MFC CSkewness_C2	Nose Inspiration_BB ox_226_683_94_ 101_frequencyC entroidX	Nose Inspiration_B Box_367_371_ 287_304_tex tureHomogen eity
15	Mouth Expiration_BB ox_977_478_4 6_82_numHole s	Mouth_Inspiratio n_FNMidFlow_M FCCMean_C2	Mouth Expiration_BB ox_266_279_4 35_530_centroi dY	Mouth_Expiration_FN_MFC CKurtosis_C2	Nose Inspiration_BB ox_241_721_42_ 42_boundingBo xArea	Nose Inspiration_B Box_367_371_ 287_304_nu mHoles
16	Mouth Expiration_BB ox_0_975_48_ 48_aspectRatio	Mouth_Inspiratio n_FNMidFlow_M FCCKurtosis_C5	Mouth Expiration_BB ox_774_364_2 3_29_perimete r	Mouth_Expiration_FN_MFC CKurtosis_C3	Nose Inspiration_BB ox_241_721_42_ 42_frequencyCe ntroidY	Nose Expiration_B Box_315_321_ 390_424_tex tureContrast
17	Mouth Expiration_BB ox_0_975_48_ 48_textureCont rast	Mouth_Inspiratio n_FNMidFlow_P BP_Skewness	Mouth Expiration_BB ox_300_553_1 9_16_centroid X	Nose_Inspirati on_FN_ZeroCr ossing2	Nose Inspiration_BB ox_243_745_14_ 14_textureEner gy	Nose Expiration_B Box_582_344_ 24_27_aspec tRatio
18	Nose Inspiration_BB ox_0_959_64_ 64_boundingB oxDiagonal	Mouth_Inspiratio n_FNMidFlow_T P_Histogram_1	Nose Inspiration_Av erage_BBboxes _EnFBiD_1	Nose_Inspirati on_FN_Wavele tApproxSkewn essL1	Nose Inspiration_BB ox_243_745_14_ 14_eulerNumber	Nose Expiration_B Box_582_344_ 24_27_perim eter
19	Nose Inspiration_BB ox_0_959_64_ 64_textureEner gy	Mouth_Inspiratio n_FNMidFlow_T P_MaxProb	Nose Inspiration_Av erage_BBboxes _MeanBiDF_1	Nose_Inspirati on_FN_Wavele tApproxSpectr alCentroidL2	Nose Inspiration_BB ox_223_765_15_ 16_stdValue	Average_FN MidFlow_Wa veletDetailMa xToMinRatio L3
20	Average_FNMid Flow_Wavelet ApproxMaxTo MinRatioL1	Mouth_Inspiratio n_FNMidFlow_E P_MaxEnergy	Nose Inspiration_Av erage_BBboxes _WCOBDFx_1	Nose_Inspirati on_FN_Wavele tDetailKurtosis L3	Nose Inspiration_BB ox_223_765_15_ 16_entropyValue	Average_FN MidFlow_Wa veletDetailKu rtosisL4
21	Nose_Inspirati on_FNMidFlo w_HurstExpon ent	Mouth_Expiration _FNMidFlow_Wa veletApproxKurto sisL1	Nose Inspiration_Av erage_BBboxes _WCOBDFy_1	Nose_Inspirati on_FN_CQTSt dPower	Nose Inspiration_BB ox_202_787_14_ 17_entropyValue	Mouth_Inspir ation_FNMid Flow_Lyapun ovExponentM ean
22	Nose_Expiration _FN_KatzFD	Mouth_Expiration _FNMidFlow_Wa veletApproxSkew nessL2	Nose Inspiration_Av erage_BBboxes _Hf1_1	Nose_Inspirati on_FN_CQTS kewnessPower	Nose Inspiration_BB ox_202_787_14_ 17_textureContr ast	Mouth_Inspir ation_FNMid Flow_BandPo werHigh

23	MouthInspiration_Gap_1217-1499_PeakFrequency	Mouth_Expiration_FNMidFlow_PB_P_Kurtosis	NoseInspiration_Average_BBoxes_Hf2_1	Nose_Inspiration_FN_CQTTemporalCentroid	NoseInspiration_BBox_202_787_14_17_textureEnergy	Nose_Inspiration_FNMidFlow_WaveletApproxEntropyL1
24	MouthExpiration_BBBox_504_333_51_42_centroidY	Mouth_Expiration_FNMidFlow_PB_P_Entropy	NoseInspiration_Average_BBoxes_reserved1	Nose_Inspiration_FN_CQTSpectralCentroid	NoseInspiration_BBox_0_989_39_34_stdValue	Nose_Inspiration_FNMidFlow_WaveletApproxEntropyL2
25	MouthExpiration_BBBox_504_333_51_42_peakValue	MouthInspiration_Gap_0-18_FM2MFreq1	NoseInspiration_Average_BBoxes_EnFBiD_2	Nose_Inspiration_FN_CQTBandwidthDynamicRange	NoseExpiration_BBBox_0_0_39_44_connectedComponents	Nose_Inspiration_FNMidFlow_WaveletDetailSpectralCentroidL3
26	MouthExpiration_BBBox_977_478_46_82_textureCorrelation	MouthInspiration_Gap_64-362_MeanPower	NoseInspiration_Average_BBoxes_MeanBiDF_2	Nose_Inspiration_FN_MFCCSkewness_C1	NoseExpiration_BBBox_465_0_66_46_medianValue	Nose_Expiration_FNMidFlow_WaveletApproxSkewnessL2
27	NoseInspiration_BBBox_0_0_38_38_rangeValue	MouthInspiration_Gap_64-362_RMS	NoseInspiration_Average_BBoxes_WCOBDFx_2	Nose_Inspiration_FN_MFCCKurtosis_C1	NoseExpiration_BBBox_465_0_66_46_iqrValue	Average_FN_AV_P_Mean
28	NoseInspiration_BBBox_477_0_55_39_stdValue	MouthInspiration_Gap_64-362_FM2MFreq1	NoseInspiration_Average_BBoxes_WCOBDFy_2	Nose_Inspiration_FN_SpectralBandwidth_C1	NoseExpiration_BBBox_465_0_66_46_textureCorrelation	Mouth_Inspiration_FN_KatzFD
29	NoseExpiration_BBBox_0_981_41_42_fractalDimension	MouthInspiration_Gap_64-362_BandPower	NoseInspiration_Average_BBoxes_Hf1_2	Nose_Inspiration_FN_MFCCKurtosis_C2	NoseExpiration_BBBox_465_0_66_46_textureEnergy	Mouth_Inspiration_FN_LyapunovExponentMax
30	NoseExpiration_BBBox_495_986_47_37_stdValue	MouthInspiration_Gap_378-565_StandardDeviation	NoseInspiration_Average_BBoxes_Hf2_2	Nose_Inspiration_FN_TP_Skewness	NoseExpiration_BBBox_465_0_66_46_fractalDimension	Mouth_Expiration_FN_MFCCMedian_C3
31	Mouth_Expiration_FNMidFlow_MFCCMedian_C1	MouthInspiration_Gap_378-565_FM2MFreq1	NoseInspiration_Average_BBoxes_reserved2	Nose_Expiration_FN_WaveletDetailSkewnessL2	NoseExpiration_BBBox_465_0_66_46_connectedComponents	Nose_Inspiration_FN_WaveletDetailEntropyL3
32	Mouth_Expiration_FNMidFlow_PBP_Kurtosis	MouthInspiration_Gap_584-810_FM2MFreq1	NoseInspiration_Average_BBoxes_bisEntropy	Nose_Expiration_FN_CQTMedianPower	NoseExpiration_BBBox_873_128_24_23_centroidY	Nose_Inspiration_FN_CQTGaborEnergyMean

33	Mouth_Expiration_FNMidFlow_TP_Histogram_1	MouthInspiration_Gap_837-1499_FM2MFreq_1	NoseInspiration_BBrox_140_140_733_744_spectralFlux	Nose_Expiration_FN_CQTSkewnessPower	NoseExpiration_BBrox_873_128_24_23_compactness	Nose_Inspiration_FN_MFCMedian_C3
34	Mouth_Expiration_FNMidFlow_TP_MaxProbability	MouthExpiration_Gap_0-353_StandardDeviation	NoseInspiration_BBrox_140_140_733_744_connectedComponents	Nose_Expiration_FN_CQTSpectralDynamicsStd	NoseExpiration_BBrox_841_162_21_21_energyValue	Average_Gap_306-496_Maximum
35	Mouth_Expiration_FNMidFlow_TP_Ratio1	MouthExpiration_Gap_815-1435_StandardDeviation	NoseInspiration_BBrox_782_151_70_73_skewnessValue	Mouth_Inspiration_FNMidFlow_ZeroCrossing1	NoseExpiration_BBrox_787_244_15_15_frequencyCentroidX	Average_Gap_1282-1359_SCBW_Bandwidth

References

- [1] Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* 2001, 17, 520–525. <https://doi.org/10.1093/bioinformatics/17.6.520>.
- [2] Batista, G.E.A.P.A.; Monard, M.C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Appl. Artif. Intell.* 2003, 17, 519–533. <https://doi.org/10.1080/713827181>.
- [3] Vergara, J.R.; Estévez, P.A. A Review of Feature Selection Methods Based on Mutual Information. *Neural Comput. Appl.* 2014, 24, 175–186. <https://doi.org/10.1007/s00521-013-1368-0>.
- [4] Zhao, Z.; Liu, H. Feature Selection Based on Mutual Information with Correlation Coefficient. *Appl. Intell.* 2022, 52, 1169–1180. <https://doi.org/10.1007/s10489-021-02524-x>.
- [5] Liu, S.; Motani, M. Improving Mutual Information Based Feature Selection by Boosting Unique Relevance. *arXiv* 2022, arXiv:2212.06143.

- [6] Singh, D.; Singh, B. Feature Wise Normalization: An Effective Way of Normalizing Data. *Pattern Recognit.* 2022, 122, 108307. <https://doi.org/10.1016/j.patcog.2021.108307>.
- [7] Haury, A.-C.; Gestraud, P.; Vert, J.-P. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE* 2011, 6, e28210. <https://doi.org/10.1371/journal.pone.0028210>.
- [8] Wang, D.; Zhang, H.; Liu, R.; Lv, W.; Wang, D. T-Test Feature Selection Approach Based on Term Frequency for Text Categorization. *Pattern Recognit. Lett.* 2014, 45, 1–10. <https://doi.org/10.1016/j.patrec.2014.02.013>.
- [9] Khoshgoftaar, T.M.; Wang, H.; Liang, Q.; Hancock, J.T. Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods. *J. Big Data* 2024, 11, 44. <https://doi.org/10.1186/s40537-024-00905-w>.
- [10] Gramegna, A.; Giudici, P. Shapley Feature Selection. *FinTech* 2022, 1, 72–80. <https://doi.org/10.3390/fintech1010006>.
- [11] Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Trans. Syst. Man Cybern. Part. A Syst. Hum.* 2010, 40, 185–197. <https://doi.org/10.1109/TSMCA.2009.2029559>.
- [12] Mounce, S.R.; Ellis, K.; Edwards, J.M.; Speight, V.L.; Jakomis, N.; Boxall, J.B. Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems. *Water Resour. Manag.* 2017, 31, 925–942. <https://doi.org/10.1007/s11269-017-1595-8>.
- [13] Wolpert, D.H. Stacked Generalization. *Neural Netw.* 1992, 5, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).

[14] Wang, M.; Qian, Y.; Yang, Y.; Chen, H.; Rao, W.-F. Improved Stacking Ensemble Learning Based on Feature Selection to Accurately Predict Warfarin Dose. *Front. Cardiovasc. Med.* 2024, 10, 1320938. <https://doi.org/10.3389/fcvm.2023.1320938>.

APPENDIX B

B.1 Top Selected Features for Each Model

Table B-1 Summary of Top Selected Features for Each Model.

Feature Number	Non-OSA vs. Mild-OSA	Non-OSA vs. Moderate-OSA	Non-OSA vs. Severe-OSA	Mild-OSA vs. Moderate-OSA	Mild-OSA vs. Severe-OSA	Moderate-OSA vs. Severe-OSA
1	MouthExpiration_Range_SpectralEntropy	Average_Range_SpectralSkewness	NoseInspiration_Range_SpectralKurtosis	NoseExpiration_Range_MeanPower	MouthExpiration_Range_SCBW_Bandwidth	NoseExpiration_Range_SpectralCrest
2	MouthExpiration_Range_SpectralCrest	NoseInspiration_Range_SpectralSkewness	NoseInspiration_Range_FreqSkewness	MouthExpiration_IQR	Average_Average_BBboxes_EnFBID	Average_BBboxes_NumHoles
3	MouthExpiration_Range_SpectralCrest	Average_BBboxes_Entropy	NoseExpiration_Range_SCBW_Bandwidth	MouthInspiration_MFCCMean	NoseExpiration_BBboxes_Range	MouthInspiration_WaveletApproxEntropy
4	NoseInspiration_Range_SCBW_Bandwidth	MouthInspiration_Range	Average_BBboxes_TextureEnergy	NoseExpiration_WaveletApproxSkewness	Average_Range_SpectralCrest	NoseExpiration_PeakCount
5	MouthInspiration_MFCCMean	MouthExpiration_WaveletApproxSpectralBandwidth	Average_BBboxes_CentroidY	MouthInspiration_MFCCMean	NoseInspiration_BBboxes_Median	MouthInspiration_MFCCMean
6	MouthInspiration_MFCCMedian	MouthExpiration_WaveletDetailEntropy	NoseInspiration_BBboxes_AspectRatio	MouthInspiration_MFCCStd	NoseInspiration_BBboxes_PeakValue	MouthExpiration_Range_SCBW_Bandwidth
7	NoseExpiration_WaveletApproxMean	MouthInspiration_IQR	MouthInspiration_BBboxes_EulerNumber	MouthInspiration_SpectralCentroid	NoseInspiration_BBboxes_FreqCentroid	MouthInspiration_Range_RMS
8	NoseExpiration_Range_BandPower	NoseInspiration_CoeffVariation	MouthInspiration_MeanValue	MouthInspiration_SpectralBandwidth	NoseInspiration_BBboxes_PeakValue	NoseInspiration_Range_MeanPower
9	NoseExpiration_Range_SpectralSkewness	NoseExpiration_BBboxes_Entropy	MouthInspiration_StdValue	MouthInspiration_MFCCKurtosis	NoseInspiration_BBboxes_FreqCentroid	NoseInspiration_Range_SCBW_Bandwidth
10	MouthInspiration_BBboxes_NumHoles	Average_DIR_Histogram	MouthExpiration_TotalEnergy	MouthExpiration_WaveletSpectralCentroid	NoseInspiration_BBboxes_SpectralFlux	NoseInspiration_Range_SpectralEnergy

11	MouthExpiration_BBBox_FreqCentroid	Average_AVP_Histogram	MouthExpiration_NormalizedEnergy	MouthExpiration_CQTBandwidthDynamicRange	NoseInspiration_BBBox_ConnectedComponents	NoseInspiration_Range_BandPower
12	MouthExpiration_BBBox_Kurtosis	MouthInspiration_ZeroCrossing	MouthExpiration_StdAbsBi	MouthExpiration_MFCCStd	NoseInspiration_BBBox_EulerNumber	MouthInspiration_BBBox_RegionArea
13	MouthExpiration_BBBox_Area	MouthInspiration_WaveletDetailSpectralBandwidth	MouthExpiration_SymStd	MouthExpiration_SpectralCentroid	NoseInspiration_BBBox_IQR	NoseInspiration_BBBox_IQR
14	MouthExpiration_BBBox_Diagonal	MouthInspiration_WaveletDetailSpectralBandwidth	MouthExpiration_Perimeter	MouthExpiration_MFCCSkewness	NoseInspiration_BBBox_FreqCentroid	NoseInspiration_BBBox_TextureHomogeneity
15	MouthExpiration_BBBox_NumHoles	MouthInspiration_MFCCMean	MouthExpiration_CentroidY	MouthExpiration_MFCCKurtosis	NoseInspiration_BBBox_BBBoxArea	NoseInspiration_BBBox_NumHoles
16	MouthExpiration_BBBox_AspectRatio	MouthInspiration_MFCCKurtosis	MouthExpiration_Perimeter	MouthExpiration_MFCCKurtosis	NoseInspiration_BBBox_FreqCentroid	NoseExpiration_BBBox_TextureContrast
17	MouthExpiration_BBBox_TextureContrast	MouthInspiration_PBP_Skewness	MouthExpiration_CentroidX	NoseInspiration_ZeroCrossing	NoseInspiration_BBBox_TextureEnergy	NoseExpiration_BBBox_AspectRatio
18	NoseInspiration_BBBox_Diagonal	MouthInspiration_TP_Histogram	NoseInspiration_Average_BBBox_EnFBiD	NoseInspiration_WaveletApproxBBox_Skewness	NoseInspiration_BBBox_EulerNumber	NoseExpiration_BBBox_Perimeter
19	NoseInspiration_BBBox_TextureEnergy	MouthInspiration_TP_MaxProb	NoseInspiration_Average_BBBox_MeanBiDF	NoseInspiration_WaveletApproxBBox_SpectralCentroid	NoseInspiration_BBBox_StdValue	Average_WaveletDetail_MaxToMinRatio
20	Average_WaveletApproxMaxToMinRatio	MouthInspiration_EP_MaxEnergy	NoseInspiration_Average_BBBox_WCOBDFx	NoseInspiration_WaveletDetailKurtosis	NoseInspiration_BBBox_Entropy	Average_WaveletDetail_Kurtosis
21	NoseInspiration_HurstExponent	MouthExpiration_WaveletApproxKurtosis	NoseInspiration_Average_BBBox_WCOBDFy	NoseInspiration_CQTStdPower	NoseInspiration_BBBox_Entropy	MouthInspiration_LyapunovExponentMean
22	NoseExpiration_KatzFD	MouthExpiration_WaveletApproxBBox_Skewness	NoseInspiration_Average_BBBox_Hf1	NoseInspiration_CQTSkewnessPower	NoseInspiration_BBBox_TextureContrast	MouthInspiration_BandPowerHigh
23	MouthInspiration_Range_PeakFrequency	MouthExpiration_PBP_Kurtosis	NoseInspiration_Average_BBBox_Hf2	NoseInspiration_CQTTemporalCentroid	NoseInspiration_BBBox_TextureEnergy	NoseInspiration_WaveletApproxEntropy

24	MouthExpiration_BBBox_CentroidY	MouthExpiration_PBP_Entropy	NoseInspiration_Average_BBBox_Reserved	NoseInspiration_CQTSpectralCentroid	NoseInspiration_BBBox_StdValue	NoseInspiration_WaveletApproxEntropy
25	MouthExpiration_BBBox_PeakValue	MouthInspiration_Range_FM2MFrequency	NoseInspiration_Average_BBBox_EnFBiD	NoseInspiration_CQTBandwidthDynamicRange	NoseExpiration_BBBox_ConnectedComponents	NoseInspiration_WaveletDetailSpectralCentroid
26	MouthExpiration_BBBox_TextureCorrelation	MouthInspiration_Range_MeanPower	NoseInspiration_Average_BBBox_MeanBiDF	NoseInspiration_MFCCSkewness	NoseExpiration_BBBox_Median	NoseExpiration_WaveletApproxSkewness
27	NoseInspiration_BBBox_Range	MouthInspiration_Range_RMS	NoseInspiration_Average_BBBox_WCOBDFx	NoseInspiration_MFCCKurtosis	NoseExpiration_BBBox_IQR	Average_AVP_Mean
28	NoseInspiration_BBBox_Std	MouthInspiration_Range_FM2MFrequency	NoseInspiration_Average_BBBox_WCOBDFy	NoseInspiration_SpectralBandwidth	NoseExpiration_BBBox_TextureCorrelation	MouthInspiration_KatzFD
29	NoseExpiration_BBBox_FractalDimension	MouthInspiration_Range_BandPower	NoseInspiration_Average_BBBox_Hf1	NoseInspiration_MFCCKurtosis	NoseExpiration_BBBox_TextureEnergy	MouthInspiration_LyapunovExponentMax
30	NoseExpiration_BBBox_Std	MouthInspiration_Range_Std	NoseInspiration_Average_BBBox_Hf2	NoseInspiration_TP_Skewness	NoseExpiration_BBBox_FractalDimension	MouthExpiration_MFCCMedian
31	MouthExpiration_MFCCMedian	MouthInspiration_Range_FM2MFrequency	NoseInspiration_Average_BBBox_Reserved	NoseExpiration_WaveletDetailSkewness	NoseExpiration_BBBox_ConnectedComponents	NoseInspiration_WaveletDetailEntropy
32	MouthExpiration_PBP_Kurtosis	MouthInspiration_Range_FM2MFrequency	NoseInspiration_Average_BBBox_BisEntropy	NoseExpiration_CQTMeanPower	NoseExpiration_BBBox_CentroidY	NoseInspiration_CQTGaborEnergyMean
33	MouthExpiration_TP_Histogram	MouthInspiration_Range_FM2MFrequency	NoseInspiration_BBBox_SpectralFlux	NoseExpiration_CQTSkewnessPower	NoseExpiration_BBBox_Compactness	NoseInspiration_MFCCMedian
34	MouthExpiration_TP_MaxProb	MouthExpiration_Range_Std	NoseInspiration_BBBox_ConnectedComponents	NoseExpiration_CQTSpectralDynamicsStd	NoseExpiration_BBBox_Energy	Average_Range_Maximum
35	MouthExpiration_TP_Ratio	MouthExpiration_Range_Std	NoseInspiration_BBBox_Skewness	MouthInspiration_ZeroCrossing	NoseExpiration_BBBox_FreqCentroid	Average_Range_SCBW_Bandwidth

B.2 Analysis of Clinical Interpretation of Top Features for All Models

B.2.1 Appendix B.1. Non-OSA vs. Mild-OSA

- MouthInspiration_Range_FreqSkewness:** Measures asymmetry of the power spectrum during mouth inspiration within the relevant frequency range. Positive values indicate dominance

of lower frequencies, negative values indicate dominance of higher frequencies. Reflects shifts in frequency energy distribution due to airflow changes, turbulence, or airway geometry.

- **Average_BBBox_TextureEnergy:** Indicates uniformity of bispectral texture. Higher energy suggests consistent patterns, while lower values indicate heterogeneous local coupling. In OSA, irregular airflow can produce less uniform bispectral coupling patterns, consistent with more complex or intermittent nonlinear flow-tissue interactions.

- **Average_BBBox_FrequencyCentroidX:** Centroid of bispectral energy along the f1-axis. Shifts in this centroid indicate changes in dominant coupling frequency, reflecting airflow-tissue interactions or turbulence associated with airway narrowing.

- **Average_BBBox_TextureCorrelation:** Measures gray-level co-occurrence texture correlation. Higher values indicate structured bispectral patterns, suggesting stable nonlinear coupling. Reduced correlation reflects heterogeneous coupling, consistent with variable airflow instability.

- **Average_BBBox_Perimeter:** Perimeter of high-intensity regions within the bispectral bounding box. Larger perimeter suggests fragmented coupling, possibly indicating variable airflow instability.

- **Average_BBBox_TextureHomogeneity:** Measures uniformity of gray-level values. Higher values indicate smoother patterns; lower values indicate irregular textures, reflecting turbulent airflow.

- **Average_BBBox_IQRValue:** Range between the first and third quartiles of bispectral intensity. Larger IQR indicates greater variability in the central portion of the sound signal, potentially associated with airflow turbulence.

- **Average_BBox_Compactness:** How compact high-intensity regions are. Less compact patterns suggest more diffuse sound energy, consistent with irregular airflow.
- **MouthInspiration_Range_SpectralSkewness:** Asymmetry of the power spectrum during mouth inspiration within a relevant frequency range. Changes may indicate airflow limitation or obstruction.

B.2.2 Non-OSA vs. Moderate-OSA

- **Average_Range_Maximum:** Highest amplitude within the relevant frequency range, representing peak breathing sound energy. Higher values may indicate more turbulent airflow due to airway narrowing.
- **Average_BBox_BoundingBoxDiagonal:** Length of the bounding box diagonal for a bispectral patch. Larger diagonal implies broader distribution of nonlinear interactions, consistent with more complex flow disturbances in moderate OSA.
- **Average_Range_MeanPower:** Average spectral energy across the relevant frequency range. Reflects overall intensity of breathing sounds in that band.
- **Average_BBox_ConnectedComponents:** Counts distinct high-intensity regions in the bispectral patch. More regions suggest fragmented coupling, potentially due to multiple flow structures or vibration sites.
- **Average_BBox_TextureContrast:** Local contrast differences in bispectral patches. Higher contrast indicates stronger **differences between regions of coupling, compatible with intermittent obstructions.**
- **Average_BBox_FrequencyCentroidY:** Centroid of bispectral energy along the f2-axis. Shifts indicate changes in dominant coupling frequencies, reflecting altered airflow-tissue interactions.

- **Average_Range_BandPower:** Total spectral power within a relevant mid-frequency range. Higher power may indicate stronger turbulent airflow.
- **Average_Range_MeanPower:** Average spectral energy, indicative of airflow consistency.
- **Average_Range_RMS:** Standard amplitude measure within the frequency range. Higher RMS indicates more variable or intense breathing sounds, associated with stronger flow disturbances.

B.2.3 Non-OSA vs. Severe-OSA

- **Average_BBox_MeanValue:** Average amplitude or intensity of the sound signal. In OSA, increased respiratory effort produces louder, higher-amplitude sounds.
- **Average_BBox_TextureEnergy:** Uniformity of bispectral texture. Lower energy indicates more heterogeneous local coupling, reflecting complex airflow-tissue interactions.
- **MouthInspiration_Range_FreqSkewness:** Asymmetry of the power spectrum during mouth inspiration within the relevant frequency range. Positive values indicate low-frequency dominance, negative values high-frequency dominance.
- **Average_BBox_FractalDimension:** Complexity or irregularity of local patterns. Higher values indicate more intricate structures, compatible with chaotic airflow in severe OSA.
- **Average_BBox_MedianValue:** Median local intensity; less sensitive to outliers than the mean. Changes reflect shifts in breathing sound intensity.
- **Average_BBox_EnergyValue:** Total energy within the bounding box. Higher energy indicates stronger, more intense breathing sounds, often associated with airway narrowing.

- **Average_BBox_EulerNumber:** Topological feature reflecting the number of connected components minus holes. Indicates complexity of sound events, altered by intermittent airflow or airway vibrations.
- **Average_BBox_AspectRatio:** Width-to-height ratio of detected sound events; changes may reflect altered spectral characteristics due to airway dynamics.
- **Average_BBox_ConnectedComponents:** Counts distinct high-intensity regions; more regions suggest fragmented coupling due to multiple airflow structures.
- **Average_BBox_KurtosisValue:** Measures tailedness of the sound signal distribution. Higher kurtosis indicates sudden, sharp sound events associated with airway collapse or reopening.

B.2.4 Mild-OSA vs. Moderate-OSA

- **MouthExpiration_BBox_AspectRatio:** Width-to-height ratio during mouth expiration; indicates alterations in spectral characteristics due to airway dynamics.
- **Average_BBox_MedianValue:** Median local intensity. Changes reflect general shifts in breathing sound intensity.
- **Average_BBox_TextureEnergy:** Uniformity of sound texture; lower values suggest heterogeneous patterns due to turbulent airflow.
- **MouthExpiration_Range_FM2MFreq:** Ratio of frequency-modulated to mean spectral energy in a relevant range; **reflects redistribution of low-frequency energy, indicating airflow variations.**
- **Average_BBox_FrequencyBandwidthY:** Effective bandwidth along the frequency axis; broader bandwidth suggests **wider participation of frequencies, consistent with turbulent airflow.**

- **Average_BBox_Perimeter:** Perimeter of high-intensity regions; larger perimeter indicates more fragmented structure, consistent with variable airflow instability.
- **Average_BBox_FrequencyCentroidY:** Centroid of bispectral energy along f2-axis; shifts indicate changes in dominant frequency regions.
- **MouthExpiration_BBox_TextureEnergy:** Uniformity of sound pattern; lower values indicate less consistent sound patterns, reflecting turbulence or intermittent obstruction.

B.2.5 Mild-OSA vs. Severe-OSA

- **MouthInspiration_BBox_FrequencyCentroidX:** Centroid along f1-axis; shifts indicate changes in dominant frequency region, reflecting airflow dynamics.
- **Average_Average_BBoxes_Entropy:** Overall entropy across bounding boxes; higher values indicate irregular airflow or turbulent sounds in severe OSA.
- **Average_Range_RMS:** Standard amplitude within the relevant frequency range; higher RMS indicates variable or stronger breathing sound intensity.
- **Average_BBox_EnergyValue:** Total energy in bounding box; higher energy reflects increased respiratory effort or airway narrowing.
- **MouthInspiration_Range_StandardDeviation:** Variability of sound amplitude; higher SD indicates greater fluctuation in airflow or turbulence.
- **Average_Range_FM2MFreq:** Low-frequency energy ratio; reflects subtle airflow variations or airway patency changes.
- **MouthInspiration_BBox_TextureContrast:** Local contrast; higher values indicate strong differences between adjacent sound regions.
- **MouthExpiration_BBox_FrequencyBandwidthY:** Bandwidth along frequency axis; broader bandwidth indicates more turbulent airflow.

- **Average_Range_StandardDeviation:** Variability of amplitude; higher SD reflects irregular airflow.

- **MouthExpiration_Range_SpectralEnergy:** Total spectral energy within relevant range; higher energy corresponds to stronger, more turbulent airflow.

B.2.6 Moderate-OSA vs. Severe-OSA

- **Average_BBox_MedianValue:** Median intensity; indicates overall breathing sound level.
- **Average_BBox_IQRValue:** Range between first and third quartiles; larger IQR suggests greater variability, linked to turbulence.

- **Average_BBox_EnergyValue:** Total energy; higher values reflect stronger, more intense breathing sounds due to airflow restriction.

- **Average_BBox_RangeValue:** Difference between maximum and minimum values; larger range indicates pronounced variations in sound intensity.

- **Average_BBox_KurtosisValue:** Measures tailedness; higher kurtosis indicates sudden peaks in sound, linked to airway collapse or reopening.

- **Average_BBox_Compactness:** How compact high-intensity regions are; less compact indicates more diffuse sound energy due to turbulent airflow.

- **Average_BBox_StdValue:** Dispersion of sound values; higher SD indicates greater variability, reflecting irregular airflow.

- **Average_BBox_IQRValue:** Central variability of sound; larger values indicate more turbulence.

- **Average_BBox_EulerNumber:** Complexity of sound events; reflects structure or continuity of airflow events.

- **Average_BBox_EntropyValue:** Randomness or unpredictability of sound; higher entropy

indicates more chaotic or turbulent airflow.

B.3 Analysis for Non-OSA vs. Mild-OSA

Table B-2 Top 10 ranked features for Non-OSA vs. Mild-OSA classification based on AUC Test and correlation performance.

Feature Name	AUC Test	Corr Test
MouthInspiration_Range_FreqSkewness	0.787037	0.498762
Average_BBox_TextureEnergy	0.666667	0.301309
Average_BBox_FrequencyCentroidX	0.657407	0.145422
Average_BBox_TextureCorrelation	0.657407	0.145286
Average_BBox_Perimeter	0.657407	0.0772519
Average_BBox_TextureEnergy	0.644444	0.0914112
Average_BBox_TextureHomogeneity	0.638889	0.0837961
Average_BBox_IQRValue	0.62963	0.211449
Average_BBox_Compactness	0.62963	0.211449
MouthInspiration_Range_SpectralSkewness	0.62963	0.237871

Table B-3 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Non-OSA vs. Mild-OSA classification.

Feature Name	Abs Delta AUC
MouthInspiration_Range_SpectralEntropy	0.0102916
Average_BBox_TextureEnergy	0.0241135
Average_BBox_FrequencyCentroidY	0.0441738
Average_BBox_SpectralFlux	0.0498553
Average_BBox_EntropyValue	0.0613248
Average_BBox_BoundingBoxDiagonal	0.0613248
Average_BBox_Compactness	0.0627066
Average_BBox_SpectralFlux	0.0728632
Average_BBox_FrequencyCentroidX	0.0728632
Average_BBox_FractalDimension	0.0860684

Table B-4 Top 10 correlations between selected features and anthropometric characteristics in Non-OSA vs. Mild-OSA participants.

Feature Name	Anthropometric Feature	Pearson Correlation
Average_BBox_MeanValue	Sex	0.998692
Average_BBox_NumHoles	MPS	0.997553
Average_BBox_FrequencyCentroidX	MPS	0.997288
Average_BBox_StdValue	Sex	0.997255
Average_BBox_CoefVariation	Sex	0.997255
Average_BBox_Perimeter	Sex	0.99691

Average_BBox_MeanValue	MPS	0.994605
Average_BBox_Compactness	NC	0.994605
Average_BBox_TextureContrast	NC	0.993768
Average_BBox_NumHoles	NC	0.993507

Table B-5 Top 10 features showing the highest Pearson correlations with AHI in Non-OSA vs. Mild-OSA comparison.

Feature Name	Pearson Correlation
Average_BBox_MeanValue	0.994605
Average_BBox_Compactness	0.994605
Average_BBox_AspectRatio	0.984433
Average_BBox_Compactness	0.984433
Average_BBox_MedianValue	0.978554
Average_BBox_FrequencyCentroidY	0.973494
Average_BBox_FrequencyBandwidthY	0.969688
MouthExpiration_Range_SCBW_Bandwidth	0.969226
Average_BBox_TextureContrast	0.960109
Average_BBox_TextureEnergy	0.951338

B.4 Analysis for Non-OSA vs. Moderate-OSA

Table B-6 Top 10 ranked features for Non-OSA vs. Moderate-OSA classification based on AUC Test and correlation performance.

Feature Name	AUC Test	Corr Test
MouthInspiration_Range_FreqSkewness	0.787037	0.498762
Average_Range_Maximum	0.849794	0.458931
Average_BBox_BoundingBoxDiagonal	0.81893	0.537497
Average_Range_MeanPower	0.816872	0.4697
Average_BBox_ConnectedComponents	0.792181	0.400149
Average_BBox_TextureContrast	0.788066	0.421496
Average_BBox_BoundingBoxDiagonal	0.781893	0.479976
Average_BBox_FrequencyCentroidY	0.76749	0.336108
Average_Range_BandPower	0.738683	0.437269
MouthInspiration_Range_FreqSkewness	0.738683	0.436397

Table B-7 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Non-OSA vs. Moderate-OSA classification.

Feature Name	AbsDeltaAUC
Average_BBox_FrequencyBandwidthX	0.00244068
Average_BBox_RangeValue	0.00810185
Average_BBox_AspectRatio	0.015558
Average_BBox_EnergyValue	0.0266176

Average_BBox_MeanValue	0.0277022
Average_BBox_EulerNumber	0.0319368
Average_BBox_KurtosisValue	0.0366769
Average_BBox_RangeValue	0.0470624
Average_BBox_BoundingBoxArea	0.0470624
Average_BBox_FrequencyBandwidthX	0.047544

Table B-8 Top 10 correlations between selected features and anthropometric characteristics in Non-OSA vs. Moderate-OSA participants.

Feature Name	Anthropometric Feature	Pearson Correlation
Average_BBox_TextureHomogeneity	Sex	0.999783
Average_BBox_RangeValue	BMI	0.997506
NoseExpiration_Range_MeanPower	Smoke History	0.996784
Average_BBox_FrequencyCentroidY	NC	0.995961
Average_BBox_BoundingBoxDiagonal	Smoke History	0.990975
Average_BBox_TextureHomogeneity	MPS	0.989529
Average_BBox_IQRValue	NC	0.987245
Average_BBox_TextureHomogeneity	Age	0.982519
Average_BBox_MeanValue	Sex	0.982193
Average_BBox_ConnectedComponents	Smoke History	0.980487

Table B-9 Top 10 features showing the highest Pearson correlations with AHI in Non-OSA vs. Moderate-OSA comparison.

Feature Name	Pearson Correlation
Average_BBox_RangeValue	0.962649
Average_BBox_TextureCorrelation	0.96016
Average_BBox_Perimeter	0.930377
Average_BBox_MeanValue	0.917263
Average_BBox_TextureEnergy	0.911276
Average_BBox_FrequencyCentroidX	0.89864
Average_BBox_SkewnessValue	0.882517
Average_BBox_IQRValue	0.877315
Average_BBox_StdValue	0.844572
Average_BBox_FractalDimension	0.735205

B.5 Analysis for Non-OSA vs. Severe-OSA

Table B-10 Top 10 ranked features for Non-OSA vs. Severe-OSA classification based on AUC Test and correlation performance.

Feature Name	AUC Test	Corr Test
Average_BBox_MeanValue	0.860248	0.505319
Average_BBox_TextureEnergy	0.775463	0.444318

MouthInspiration_Range_FreqSkewness	0.775	0.413793
Average_BBox_FractalDimension	0.73913	0.429386
Average_BBox_MedianValue	0.733333	0.242264
Average_BBox_EnergyValue	0.729167	0.216856
Average_BBox_EulerNumber	0.716667	0.246949
Average_BBox_AspectRatio	0.708333	0.200236
Average_BBox_ConnectedComponents	0.708333	0.215239
Average_BBox_KurtosisValue	0.708333	0.267374

Table B-11 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Non-OSA vs. Severe-OSA classification.

Feature Name	AbsDeltaAUC
Average_BBox_SkewnessValue	0.00123128
Average_BBox_FrequencyBandwidthX	0.0025107
Average_BBox_FrequencyBandwidthX	0.00366667
MouthInspiration_Range_FreqSkewness	0.00366667
Average_BBox_Compactness	0.00418763
Average_BBox_FrequencyBandwidthX	0.004408
Average_BBox_KurtosisValue	0.00454831
Average_BBox_SkewnessValue	0.00578492
Average_BBox_EulerNumber	0.00608255
Average_BBox_ConnectedComponents	0.00933082

Table B-12 Top 10 correlations between selected features and anthropometric characteristics in Non-OSA vs. Severe-OSA participants.

Feature Name	Anthropometric Feature	Pearson Correlation
Average_BBox_EnergyValue	NC	0.998475
Average_BBox_AspectRatio	MPS	0.99834
Average_BBox_StdValue	Age	0.995421
Average_BBox_EulerNumber	Sex	0.995285
Average_BBox_Perimeter	MPS	0.99373
Average_BBox_IQRValue	Sex	0.991446
Average_BBox_NumHoles	Smoke History	0.989047
Average_BBox_ConnectedComponents	Age	0.988745
Average_BBox_EnergyValue	Age	0.988332
Average_BBox_MeanValue	Age	0.987492

Table B-13 Top 10 features showing the highest Pearson correlations with AHI in Non-OSA vs. Severe-OSA comparison.

Feature Name	Pearson Correlation
Average_BBox_PeakValue	0.895716
Average_BBox_BoundingBoxDiagonal	0.881183
Average_BBox_BoundingBoxDiagonal	0.85574
Average_BBox_Perimeter	0.846486
Average_BBox_KurtosisValue	0.835585

Average_BBox_Perimeter	0.828251
Average_BBox_MeanValue	0.769224
Average_BBox_FrequencyCentroidX	0.759227
Average_BBox_EnergyValue	0.752154
Average_BBox_SkewnessValue	0.734339

B.6 Analysis for Mild-OSA vs. Moderate-OSA

Table B-14 Top 10 ranked features for Mild-OSA vs. Moderate-OSA classification based on AUC Test and correlation performance.

Feature Name	AUC Test	Corr Test
MouthExpiration_BBox_AspectRatio	0.883333	0.496958
Average_BBox_MedianValue	0.788889	0.451368
Average_BBox_TextureEnergy	0.783333	0.258151
MouthExpiration_Range_FM2MFreq	0.6875	0.187878
Average_BBox_FrequencyBandwidthY	0.671875	0.251735
Average_BBox_FrequencyBandwidthY	0.655556	0.0760239
Average_BBox_Perimeter	0.655556	0.0760239
Average_BBox_FrequencyCentroidY	0.644531	0.221861
MouthExpiration_BBox_TextureEnergy	0.644444	0.220908
Average_BBox_TextureEnergy	0.641667	0.294824

Table B-15 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Mild-OSA vs. Moderate-OSA classification.

Feature Name	AbsDeltaAUC
MouthExpiration_BBox_CoefVariation	0.00180556
Average_BBox_TextureEnergy	0.00291667
Average_BBox_FrequencyCentroidY	0.00443837
MouthExpiration_Range_FM2MFreq	0.00593891
Average_BBox_FrequencyBandwidthX	0.00599845
Average_BBox_CentroidY	0.00813298
Average_BBox_FrequencyBandwidthY	0.0124323
Average_BBox_CentroidY	0.0142974
Average_BBox_SkewnessValue	0.0192591
Average_BBox_SkewnessValue	0.0218056

Table B-16 Top 10 correlations between selected features and anthropometric characteristics in Mild-OSA vs. Moderate-OSA participants.

Feature Name	Anthropometric Feature	Pearson Correlation
Average_BBox_IQRValue	Sex	0.998547
MouthExpiration_FrequencyCentroidX	MPS	0.998212
Average_BBox_CentroidY	MPS	0.99816
MouthExpiration_StdValue	NC	0.997504
Average_BBox_MeanValue	MPS	0.997397

MouthExpiration_FrequencyCentroidX	NC	0.996663
Average_BBox_FrequencyCentroidY	Sex	0.996409
Average_BBox_Perimeter	Age	0.996071
Average_BBox_FrequencyCentroidX	Sex	0.995729
Average_BBox_Perimeter	MPS	0.994929

Table B-17 Top 10 features showing the highest Pearson correlations with AHI in Mild-OSA vs. Moderate-OSA comparison.

Feature Name	Pearson Correlation
Average_BBox_TextureHomogeneity	0.990363
Average_BBox_BoundingBoxDiagonal	0.984096
Average_BBox_EntropyValue	0.9703
Average_BBox_EntropyValue	0.964663
Average_BBox_FrequencyBandwidthX	0.962357
Average_BBox_CentroidY	0.962357
Average_BBox_FrequencyCentroidX	0.956978
Average_BBox_MedianValue	0.953038
Average_BBox_NumHoles	0.952526
Average_BBox_TextureEnergy	0.951101

B.7 Analysis for Mild-OSA vs. Severe-OSA

Table B-18 Top 10 ranked features for Mild-OSA vs. Severe-OSA classification based on AUC Test and correlation performance.

Feature Name	AUC Test	Corr Test
MouthInspiration_FrequencyCentroidX	0.844444	0.157994
Average_BBox_EnTBiD	0.777778	0.483544
Average_Range_RMS	0.744444	0.67878
Average_BBox_EnergyValue	0.733333	0.513402
MouthInspiration_Range_StdDev	0.722222	0.776042
Average_Range_FM2MFreq	0.714286	0.572164
MouthInspiration_TextureContrast	0.711111	0.380311
MouthExpiration_FrequencyBandwidthY	0.69375	0.214055
Average_Range_StdDev	0.688889	0.621574
MouthExpiration_Range_SpectralEnergy	0.688889	0.516899

Table B-19 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Mild-OSA vs. Severe-OSA classification.

Feature Name	AbsDeltaAUC
MouthExpiration_Perimeter	0.00291667
MouthExpiration_FrequencyBandwidthX	0.00291667
MouthInspiration_PeakValue	0.0041751
Average_BBox_NumHoles	0.0042735
MouthExpiration_Perimeter	0.00583333

MouthInspiration_ConnectedComponents	0.00600058
MouthExpiration_CentroidX	0.00666667
MouthExpiration_NumHoles	0.0075
MouthExpiration_PeakValue	0.0075
MouthInspiration_StdValue	0.00816946

Table B-20 Top 10 correlations between selected features and anthropometric characteristics in Mild-OSA vs. Severe-OSA participants.

Feature Name	Anthropometric Feature	Pearson Correlation
MouthExpiration_PeakValue	NC	0.99998
MouthExpiration_EntropyValue	BMI	0.99901
MouthInspiration_FrequencyCentroidX	BMI	0.998535
MouthInspiration_SpectralFlux	BMI	0.997205
MouthExpiration_PeakValue	NC	0.996188
MouthInspiration_CentroidY	MPS	0.996019
MouthExpiration_Range_RMS	NC	0.99593
MouthExpiration_Range_MeanPower	Sex	0.994836
MouthInspiration_EulerNumber	Sex	0.99472
MouthExpiration_Range_BandPower	Sex	0.994715

Table B-21 Top 10 features showing the highest Pearson correlations with AHI in Mild-OSA vs. Severe-OSA comparison.

Feature Name	Pearson Correlation
MouthInspiration_IQRValue	0.988287
MouthInspiration_EulerNumber	0.986967
MouthInspiration_AspectRatio	0.980666
MouthInspiration_CentroidX	0.965072
MouthInspiration_FrequencyBandwidthX	0.958011
MouthInspiration_RegionArea	0.942624
MouthExpiration_StdValue	0.902562
MouthInspiration_KurtosisValue	0.879297
MouthInspiration_IQRValue	0.875097
MouthInspiration_MedianValue	0.872257

B.8 Analysis for Moderate-OSA vs. Severe-OSA

Table B-22 Top 10 ranked features for Moderate-OSA vs. Severe-OSA classification based on AUC Test and correlation performance.

Feature Name	AUC Test	Corr Test
Average_BBox_MedianValue	0.71875	0.239624
Average_BBox_IQRValue	0.690972	0.354004
Average_BBox_EnergyValue	0.680556	0.361267
Average_BBox_RangeValue	0.642361	0.211027

Average_BBox_KurtosisValue	0.637153	0.397161
Average_BBox_Compactness	0.615625	0.423041
Average_BBox_StdValue	0.607143	0.121142
Average_BBox_IQRValue	0.607143	0.121142
Average_BBox_EulerNumber	0.59375	-0.0542083
Average_BBox_EntropyValue	0.59375	-0.0327421

Table B-23 Top 10 most stable features based on absolute change in AUC (AbsDeltaAUC) for Moderate-OSA vs. Severe-OSA classification.

Feature Name	AbsDeltaAUC
Average_BBox_MedianValue	0.00355392
Average_BBox_RegionArea	0.0053935
Average_BBox_FrequencyCentroidX	0.0101997
Average_BBox_IQRValue	0.0132378
MouthInspiration_Range_SpectralEnergy	0.0278361
Average_BBox_KurtosisValue	0.0325428
Average_BBox_PeakValue	0.0373264
Average_BBox_EnergyValue	0.0503472
Average_BBox_CoefVariation	0.0576721
Average_BBox_SpectralFlux	0.0649846

Table B-24 Top 10 correlations between selected features and anthropometric characteristics in Moderate-OSA vs. Severe-OSA participants.

Feature Name	Anthropometric Feature	Pearson Correlation
Average_BBox_BoundingBoxArea	NC	0.997757
Average_BBox_RangeValue	Sex	0.997111
Average_BBox_NumHoles	Smoke History	0.996614
Average_BBox_Perimeter	Smoke History	0.996614
Average_BBox_FrequencyBandwidthX	Smoke History	0.995328
Average_BBox_RegionArea	NC	0.992481
Average_BBox_EnergyValue	MPS	0.991225
NoseExpiration_Range_SpectralCrest	Age	0.990493
Average_BBox_PeakValue	Sex	0.988759
Average_BBox_Perimeter	Sex	0.987818

Table B-25 Top 10 features showing the highest Pearson correlations with AHI in Moderate-OSA vs. Severe-OSA comparison.

Feature Name	Pearson Correlation
Average_BBox_RegionArea	0.977981
Average_BBox_ConnectedComponents	0.966506
Average_BBox_TextureHomogeneity	0.937707
Average_BBox_MeanValue	0.937068
Average_BBox_RegionArea	0.925231
Average_BBox_BoundingBoxArea	0.915942
Average_BBox_EnergyValue	0.909118

Average_BBox_CentroidY	0.906971
Average_BBox_FrequencyCentroidX	0.903769
Average_Range_StdDev	0.890411

APPENDIX C

C.1 Proof Sketches for Theoretical Properties of the CUES Metric

This appendix provides detailed proof sketches supporting the theoretical properties of the proposed CUES (calibration–utility–equity–stability) metric introduced in Section 8.2. The results formalize the boundedness, monotonicity, sensitivity, consistency, and asymptotic behavior of CUES under the assumptions stated in the main text. Throughout, we adopt the notation and component definitions introduced in Sections 8.2.2–8.2.6.

C.1.1 Definition and Notation

Let $C \in [0,1], U \in [0,1], E \in [0,1], S \in [0,1]$ denote the normalized calibration, utility, equity, and stability components, respectively. Each component is constructed as a bounded functional of the data-generating distribution P , as described in Sections 8.2.2–8.2.6. The CUES metric is defined as $\text{CUES} = f(C, U, E, S) := (C \cdot U \cdot E \cdot S)^{1/4}$. Let $\hat{C}, \hat{U}, \hat{E}, \hat{S}$ denote the corresponding empirical estimators.

C.1.2 Boundedness the CUES Metric Satisfies $0 \leq \text{CUES} \leq 1$

Proof. By construction, each component is normalized to lie in the unit interval. Therefore,
$$0 \leq C \cdot U \cdot E \cdot S \leq 1 \tag{C-1}$$

The function $x \mapsto x^{1/4}$ is continuous and monotone increasing on $[0, 1]$, preserving bounds. Hence, $0 \leq (C \cdot U \cdot E \cdot S)^{1/4} \leq 1$. This ensures that CUES is interpretable on a standardized scale comparable across datasets and models.

C.1.3 Failure Dominance and Non-Compensability

If any component equals zero, then $\text{CUES} = 0$.

Proof. Suppose $C = 0$. Then $C \cdot U \cdot E \cdot S = 0$, and consequently, $\text{CUES} = 0^{1/4} = 0$.

The same argument applies if $U = 0$, $E = 0$, or $S = 0$. This property formalizes the non-compensatory design principle of CUES: severe failure in calibration, utility, equity, or stability cannot be offset by strong performance in other dimensions.

C.1.4 Monotonicity in Each Component

CUES is strictly increasing in each component on $(0, 1]^4$

Proof. Consider the partial derivative with respect to C :

$$\frac{\partial f}{\partial C} = \frac{1}{4}(\text{CUES})^{-3/4} \cdot UES. \quad \text{C-2}$$

Since $U, E, S > 0$, the derivative is strictly positive for all $C \in (0, 1]$. By symmetry, identical expressions hold for U, E, S . Therefore, improving any single component while holding the others fixed strictly increases CUES. This guarantees that CUES respects component-wise performance improvements without introducing paradoxical behavior.

C.1.5 Relative Sensitivity and Balanced Contribution to First Order, CUES Responds

Equally to Relative Changes in Each Component

Proof. Taking logarithms, $\log(\text{CUES}) = \frac{1}{4}(\log C + \log U + \log E + \log S)$.

Differentiating yields

$$\frac{d \text{CUES}}{\text{CUES}} = \frac{1}{4} \left(\frac{dC}{C} + \frac{dU}{U} + \frac{dE}{E} + \frac{dS}{S} \right). \quad \text{C-3}$$

Thus, a 1% relative increase in any single component produces an approximately 0.25% relative increase in CUES, assuming the other components remain fixed. This property ensures balanced weighting across dimensions without requiring explicit tuning parameters.

C.1.6 Continuity and Consistency of the CUES Estimator

If $\hat{C} \xrightarrow{p} C, \hat{U} \xrightarrow{p} U, \hat{E} \xrightarrow{p} E, \hat{S} \xrightarrow{p} S$, then $\widehat{CUES} \xrightarrow{p} CUES$

Proof. The mapping $f(c, u, e, s) = (cues)^{1/4}$ is continuous on $[0, 1]^4$. Since each component estimator is consistent under standard regularity conditions (e.g., empirical risk minimization for utility, law of large numbers for calibration error, subgroup-wise averaging for equity, and bootstrap resampling for stability), the joint convergence holds. By the Continuous Mapping Theorem, $f(\hat{C}, \hat{U}, \hat{E}, \hat{S}) \xrightarrow{p} f(C, U, E, S)$.

C.1.7 Asymptotic Normality via the Delta Method Assume

$$\sqrt{n} \left(\begin{bmatrix} \hat{C} \\ \hat{U} \\ \hat{E} \\ \hat{S} \end{bmatrix} - \begin{bmatrix} C \\ U \\ E \\ S \end{bmatrix} \right) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad \text{C-4}$$

for some covariance matrix Σ . Then,

$$\sqrt{n}(\widehat{CUES} - CUES) \xrightarrow{d} \mathcal{N}(0, \nabla f^\top \Sigma \nabla f).$$

Proof. The function f is continuously differentiable on $(0, 1]^4$. Applying the multivariate delta method yields asymptotic normality with gradient

$$\nabla f = \frac{1}{4} (CUES)^{-3/4} \begin{bmatrix} UES \\ CES \\ CUS \\ CUE \end{bmatrix}. \quad \text{C-5}$$

This result provides theoretical justification for inference on CUES when component estimators are asymptotically normal.

C.1.8 Bootstrap Validity for Finite-Sample Inference

While the asymptotic variance expression is analytically tractable, it depends on the unknown covariance matrix Σ . In practice, we therefore rely on bootstrap resampling, as described

in Section 8.2.6.7, to estimate the sampling distribution of CUES. Under standard regularity conditions, the bootstrap consistently approximates the distribution of smooth functionals such as CUES, making it well suited for uncertainty quantification in moderate sample sizes.