Early Detection of Diabetic Foot Ulceration from Thermograms and Rheumatoid Arthritis Joint and Damage prediction Using Artificial Intelligence

by

Saqib Al Islam

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Computer Science

University of Manitoba

Winnipeg, MB, Canada

August 2023

Copyright © 2023 by Saqib Al Islam

Abstract

Diabetes foot ulceration (DFU) and rheumatoid arthritis (RA) are two complex diseases that affect many people around the world. In this M.Sc. thesis, I present two approaches to address early detection and prediction challenges for these conditions. For DFU, previous trials have shown the potential of thermograms in detecting increased plantar temperature before DFU, but identifying it early has been challenging due to complex temperature distribution. Hence, I develop a multi-stream, multi-view convolutional neural network (CNN) to incorporate features from ipsilateral and contralateral foot images. This CNN model achieves an area under the receiver operating characteristic curve (ROC-AUC) of 0.94 on the test set, demonstrating potential for early DFU detection. For RA, I use deep learning to automate the detection of joints and prediction of Sharp-van-der-Heijde (SvdH) scores on hand radiographs. I develop a two-stage system, with a joint detection model trained on pediatric hand radiographs from healthy individuals and a vision-transformer model trained on adult hand radiographs predicting erosion and joint-space-narrowing (JSN) scores. The system addresses class imbalance. The joint detection model achieves an F1 score of 0.98 on internal testing and ~0.92 mean precision on an external set. At best, the joint damage prediction model achieves balanced accuracy values of ~0.93 for erosion and JSN scores, and ~0.90 for individual wrist joints. In summary, my AI-based approaches show potential for early detection in DFU, and automated joint detection and scoring in RA. They have potential to improve detection, prediction and treatment of the DFU and RA.

Acknowledgments

I am incredibly appreciative of everyone who has supported me and helped me along the way. I want to convey my sincere gratitude to my supervisor, Dr. Pingzhao Hu, for his tenacity in giving me encouragement and direction as I pursued my MSc at the University of Manitoba. He has offered me invaluable assistance and solutions to the problems I have run into while conducting my research for my MSc. He has provided unique suggestions for my projects. There are no adequate words to express my gratitude to Dr. Hu. I want to express my gratitude to Dr. Carson Leung, my co-supervisor, who has provided me with a wealth of advice, direction, and support.

I want to express my gratitude to Dr. Yiming Qian and Dr. Liam O'Neil for serving as my advisory committee members and for their insightful comments and recommendations.

Thank you to Drs. Carol Hitchon and Liam O'Neil for their ongoing comments on the rheumatoid arthritis project and their assistance in obtaining data on rheumatoid arthritis x-ray pictures for our study. The rheumatoid arthritis research has been made possible by their input and assistance. I had incredible support and assistance from my colleagues and friends in Hu Lab, Qian Liu, Leann Lac, Judah Zammit, Daryl Fung, Wasif Khan, Yan Sun, and Hamid Hadipour, during my MSc degree.

Finally, I would want to express my gratitude to my family for their unwavering love, support, and advice as well as for giving me the incredible chance to further my studies in Canada. Without their support and affection, I could not have reached where I am now. I also want to thank my friend Faria Leera who has been a constant support and made the transition to residing in Winnipeg easier.

Table of Contents

Abstract 2

Acknowledgments 3

List of Tables 6

List of Figures 7

List of Abbreviations 8

Chapter 1 Introduction and Background 9

1.1 Diabetic Foot Ulceration 9

1.1.1 Infrared thermogram Diabetic Foot Ulceration 9

1.2 Rheumatoid Arthritis (RA) 10

1.2.1 Sharp van der Heijde score 10

1.3 Machine learning 11

1.3.1 Deep learning 11

1.3.2 Multi-task Learning 12

1.4 Summary 13

Chapter 2 Motivation and Research Objectives 14

2.1 Motivations 14

2.1.1 Early detection of DFU 14

2.1.2 Damage prediction in RA 14

2.2 Research Objectives 16

Chapter 3 Early Detection of Diabetic Foot Ulceration from Thermal Images Using Machine Learning and Image Registration 17

3.1 Introduction 17

3.2 Materials and Methods 21

3.2.1 Dataset Description 21

3.2.2 Data Preparation 21

3.2.3 Image registration-based ROI extraction 22

3.2.4 Image registration and obtaining the coordinates of ROIs. 24

3.2.5 Feature extraction from ROI patches 29

3.2.6 Generation of feature table and feature selection 30

3.2.7 Classification of DM using machine learning approach 32

3.3 Multi-view and multi-scale feature representation DL model 34
3.3.1 Image enhancements and model inputs 37
3.3.2 Model interpretation and observations 37
3.4 Performance metrics 38
3.5 Results and Discussion 39
3.5.1 ROI feature-based detection 39
3.5.2 Deep learning and Multi-view CNN-based approach 43
3.6 Summary 47
Chapter 4 Deep learning-based joint detection and SvH score prediction in Rheumatoid arthritis hand radiographs. 49
4.1 Introduction 49
4.2 Methods 50
4.2.1 Study design & data sources 50
4.3 Deep learning pipeline 54
4.4 Model training and evaluation 59
4.4.1 Joint detection metrics 60
4.4.2 Score prediction metrics 62
4.5 Results 63
4.5.1 Joint detection and classification 63
4.5.2 Joint damage score prediction performance 65
4.6 Discussion 71
Chapter 5 Conclusion and Future Direction 73
5.1 Conclusion 73
5.2 Future Directions 75

List of Tables

Table 3.1 Number of images after dataset split and generation of augmented sets.. 22

Table 3.2 Description of the First and second order (FSStat) features extracted from the ROI.. 27

Table 3.3 Hyperparameter search space for the ML algorithms 33

Table 3.4 Test set results for the Machine learning ROI-based approach..... 40

Table 3.5 Best results acquired from using different areas for feature extraction 43

Table 3.6 Results using multi-view approach the subnetwork using both left and right images as input 44

Table 3.7 Best result acquired by different CNN models using different inputs 47

Table 4.1 Demographics and clinical features of CATCH subjects with SvdH scores..... 52

Table 4.2 CATCH data stratified split details. Number of samples in each score class 53

Table 4.3 Joint detection model metrics on internal test data (RSNA)..... 63

Table 4.4 Joint mean model metrics on external test data (Manitoba EAC patients)..... 64

Table 4.5 Performance on test set for multi-task models trained on wrist images (Erosion score)..... 64

Table 4.6 Performance on test set for multi-task model trained on wrist images (JSN scores)..... 65

Table 4.7 Performance on CATCH test set for multi-task model trained on different joint type images..... 65

Table 4.8 Performance on test set for ordered regression model and CNN models 69

List of Figures

Figure 3.1 Overview of obtaining landmarks from unsupervised image registration technique..... 24

Figure 3.2 Some results of obtaining landmarks from unsupervised image registration technique..... 29

Figure 3.3 Patches from region of interests..... 30

Figure 3.4 Machine learning-based DFU classification overview..... 32

Figure 3.5 Proposed multi-view CNN network for DFU classification..... 36

Figure 3.6 Results of the ML approach using image-only features extracted from ROIs..... 41

Figure 3.7 Multi-view CNN network visualization and interpretation..... 46

Figure 4.1 Overview of joint detection and joint damage prediction study design..... 54

Figure 4.2 Overview of joint detection and damage prediction model training..... 55

Figure 4.3 Architecture of YOLOv5..... 57

Figure 4.4 Multi-task joint score prediction..... 58

Figure 4.5 Example from the joint detection model & IOU calculation for joint prediction..... 61

Figure 4.6 Results of SvdH score prediction models..... 67

Figure 4.7 Positive and negative examples of full model..... 68

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Networks
CNN	Convolutional neural networks
DFU	Diabetic Foot Ulceration
DM	Diabetes Mellitus
FSStat	First and Second order statistical features
IRT	Infrared Thermography
KNN	K- Nearest Neighbours
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MTL	Multi-task learning
NCC	Negative Cross Correlation
RA	Rheumatoid Arthritis
ROI	Region of Interest
SvDH	Sharp van der Heijde
SVM	Support Vector Machines
US	Ultrasound

Chapter 1 Introduction and Background

1.1 Diabetic Foot Ulceration

Diabetes mellitus has serious consequences such as heart disease, stroke, renal failure, blindness, and diabetic foot ulceration (DFU) causing lower limb amputation. Diabetes-related foot ulcers are associated with increased medical costs, decreased quality of life, infection, amputation, and mortality. People who are at risk of DFU would benefit from timely diagnosis and treatment to avoid foot ulcers, amputation, and death. Self-care monitoring for evidence of DFU without professional medical assistance may allow for timely treatment to prevent skin breakdown and wound development. Visual inspection to identify DFU in its early stages has limitations since patients with obesity or vision impairment cannot see their ulceration location (Alavi et al., 2014) . Recent investigations, however, have shown that temperature monitoring can predict the development of DFU in 97 percent of patients (Frykberg et al., 2017; van Doremalen et al., 2020). Additionally, patients undergoing foot temperature monitoring in a remote home setting had a lower risk of DFU (Crisologo & Lavery, 2017).

1.1.1 Infrared thermogram Diabetic Foot Ulceration

Infrared thermogram (IRT) analysis has shown to be an effective utility in the diagnosis of many diseases (Magalhaes et al., 2021). Several studies have proposed thermogram-based techniques for identifying those at risk of DFU (Armstrong & Lavery, 1998; Filipe et al., 2021; Kumar et al., 2022; Nag et al., 2021a) by identifying a characteristic thermal distribution in the infrared images of the feet and other organs. Machine learning (ML) approaches, such as artificial neural networks (ANN), k-closest neighbors (KNN), and support vector machines (SVM), have been widely employed for automatic image categorization utilizing feature extraction, feature ranking, and various ML models such as ANN, KNN, and SVM (Prabha et al., 2021; Raghavendra et al., 2019; Xu et al., 2020).

1.2 Rheumatoid Arthritis (RA)

Rheumatoid arthritis (RA) is an autoimmune disease that primarily affects joints, causing inflammation, stiffness, and pain. Early detection of RA is essential to prevent joint damage and improve patient outcomes (Monti et al., 2015). Currently, the diagnosis of RA relies on a combination of clinical evaluation, laboratory tests, and imaging studies. Serial radiographs or ultrasound images are usually used to evaluate joint damage and scoring methods are used to quantify the progression of joint damage over time (Ravindran & Rachapalli, 2011; D. Van Der Heijde, 2000). However, the radiographic evaluation of joint damage requires expertise and detailed scoring of radiographs ("joint images") can be time-consuming and costly. Therefore, there is a growing interest in developing automated and accurate methods for detecting RA damage using deep learning (Maziarz et al., 2021).

1.2.1 Sharp van der Heijde score

The Sharp van der Heijde (SvdH) score is a widely used measurement tool in rheumatology that quantifies the structural damage to the joints caused by RA (D. Van Der Heijde, 2000). The SvdH score is computed based on radiographs of the hands and feet and is often used as an endpoint in clinical trials and longitudinal studies of RA. Manual scoring of the SvdH score is time-consuming and subject to inter- and intra-observer variability (Sharp et al., 2004). Therefore, there has been growing interest in using artificial intelligence (AI) techniques, such as deep learning, to automate the detection and scoring of the SvdH score.

1.3 Machine learning

Artificial intelligence refers to computer systems capable of performing tasks that typically require human intelligence. To advance this technology further, the concept of machine learning was introduced. Machine learning models can learn from data and make predictions without explicit programming (Bell, 2022). This is achieved by the models learning and recognizing general patterns in labeled input data. For example, an image of a dog with the corresponding label "dog" can be used to train a machine learning model to classify similar images as "dog". However, machine learning models can struggle when dealing with complex data distributions. There are different approaches to train machine learning models, two of which are supervised learning and unsupervised learning. Supervised learning involves using datasets that consist of input data along with corresponding labels. These labels are utilized to train the models to classify or predict outcomes based on the given inputs. On the other hand, unsupervised learning utilizes datasets that only contain input data without any labels. This approach helps machine learning models discover hidden patterns in the data by performing tasks such as clustering, association, and dimensionality reduction.

1.3.1 Deep learning

Deep learning outperforms machine learning when it comes to modelling complex distributions because deep learning models are significantly better at generalizing complex distributions. The term "deep" learning refers to the neural network building pieces that are stacked on top of one another to form deep layers. However, deep learning needs more samples to learn from in order to predict more complicated distributions and perform better than machine learning models. If not, deep learning models would overfit the dataset or memorize it rather than generalizing or comprehending the data itself. Deep learning has been applied extensively in both academia and industry in recent years. It has been applied to diverse fields including computer vision (Esteva et al., 2021; Islam et al., 2016; Jiao

et al., 2020; L. Li et al., 2019), natural language processing (Otter et al., 2021; Wu et al., 2020), speech recognition and generation (Bhangale & Kothandaraman, 2022; Y. J. Lu et al., 2022; Oruh et al., 2022) and robotics (Karoly et al., 2021; Soori et al., 2023).

Deep learning has also been used to analyze clinical, radiological, and administrative health data. These data might be highly dimensional and have a complex distribution, making deep learning suitable for them. Deep learning can be used to locate the associated disease through the health data by learning and extracting complicated patterns, much as distributions in the health data can be used to find associations with disease status, disease severity, and disease classification. They have been used extensively in identifying diseases from medical imaging such as endoscopy, ophthalmology, cancer, and tumor segmentation and classification (Dong et al., 2021; Esteva et al., 2021; Gao et al., 2022; Morid et al., 2021; Paderno et al., 2022). The use of transfer learning and augmentation has proven to be effective when applied to the medical domain (Chlap et al., 2021; Morid et al., 2021).

1.3.2 Multi-task Learning

Multi-task learning (MTL) (Caruana et al., 1997) is a learning paradigm developed with the inspiration of how humans learn activities by taking knowledge from previously learned activities. The goal of MTL is to leverage the shared knowledge between multiple related tasks to improve the generalization performance of all tasks (Vandenhende et al., 2020). When given the same information, pixel-level tasks in visual understanding frequently share characteristics that could be leveraged to improve MTL performance (Zhang et al., 2019). This is a useful characteristic when training using medical images which can contain necessary information from which to make multiple inferences (Dong et al., 2021).

1.4 Summary

Considering these recent advancements in AI assisted medical imaging diagnostic methods. There is a need to evaluate the effectiveness of using AI techniques for early disease detection of two diseases Rheumatoid Arthritis and Diabetic Foot Ulceration. The utility of such a tool would be beneficial for both healthcare providers and patients. AI assisted tools for detecting RA from radiographs and DFU from thermograms will provide a crucial tool for clinicians. Especially, those in underdeveloped countries or remote areas with limited healthcare resources. These tools can accurately and timely monitor RA and DFU. There have been recent advancements in AI tools used for medical applications. Some have outperformed physician's performance reinforcing the potential of developing such a tool (Amann et al., 2020; Tajbakhsh et al., 2016). Furthermore, the low-cost and accessibility will reduce the burden of limited resources in remote areas. They will make the work of healthcare clinician's easier and assist them in reducing their workload. Thus, they can effectively manage their time and prioritize patient's based on their conditions.

Chapter 2 Motivation and Research Objectives

2.1 Motivations

2.1.1 Early detection of DFU

Current diabetic foot ulceration diagnosis methods involve high-frequency ultrasonography, angiography, magnetic resonance imaging, and physical examination. All of which them are either invasive or expert resource dependent. Asymmetry between the left and the right organs is an indication for DFU. It would be beneficial to develop an algorithm that can incorporate features from multiple organs (D. Hernandez-Contreras et al., 2017). Thermal imaging is non-invasive and can be used in a home setting for monitoring disease progression. Current ML approaches require observations from carefully annotated Region of interests (ROIs) by experts. We present an unsupervised image registration technique to identify ROIs in anatomically similar images and utilize these to extract relevant features from thermograms to make a disease outcome prediction based on an interpretable tree-based classifier.

2.1.2 Damage prediction in RA

Rheumatoid arthritis (RA) is an autoimmune disease that primarily affects joints, causing inflammation, stiffness, and pain. Early detection of RA is essential to prevent joint damage and improve patient outcomes. Currently, the diagnosis of RA relies on a combination of clinical evaluation, laboratory tests, and imaging studies. Serial radiographs or ultrasound images are usually used to evaluate joint damage and scoring methods are used to quantify the progression of joint damage over time (Ory, 2003; D. M. F. M. Van Der Heijde, 2000). However, the detailed scoring of joint images can be time-consuming and expensive, and deep learning (DL) algorithms have been

shown to be a feasible and promising approach that may be used in research studies with the aim of moving into clinical practice (Sun et al., 2022). Therefore, there is a growing interest in developing automated and accurate methods for detecting RA damage using DL.

X-ray images are commonly used for detecting erosive disease and monitoring the progression of joint damage. The SvdH scoring method (van der Heijde et al. 1999), which grades joint damage based on joint space narrowing and cortical damage, is widely accepted and validated for quantifying joint damage. It is considered as a gold standard method for scoring radiographs in inflammatory arthritis, and it has been utilized in numerous clinical trials and cohort studies, including those focusing on early disease (Dias et al., 2008; Smolen et al., 2006). However, early radiographic damage can be subtle, and the SvdH scoring method does not account for juxta-articular osteopenia, which can occur before cortical damage (Ravindran & Rachapalli, 2011). Additionally, manual scoring tools for SvdH are time-consuming, require expertise, and are prone to subjectivity.

Recently, computer-assisted radiograph interpretation employing deep learning (DL) techniques has been applied to standard radiographs (Larson et al., 2017; M. T. Lu et al., 2019). Deep learning models, specifically convolutional neural networks (CNNs), have demonstrated comparable performance to medical professionals in various medical imaging tasks, including the evaluation of RA progression using radiographs (Hirano et al., 2019; Honda, Yano, & Tanaka, 2022).The application of automated radiograph interpretation has the potential to enhance the sensitivity in identifying important findings, discover clinically relevant features, reduce interpretation time, and assist in diagnostics, particularly in remote areas where specialists may not be readily available.

2.2 Research Objectives

This thesis covers two research objectives:

- **Aim 1**: To use machine learning and image registration to analyze features from feet thermal images to classify early detection of diabetic foot ulceration (DFU) (Chapter 3).
- **Aim 2:** To develop a tool for joint detection and damage scoring of RA, which is accessible and user-friendly for clinicians, particularly those in vulnerable or remote areas (Chapter 4).

Chapter 3 Early Detection of Diabetic Foot Ulceration from Thermal Images Using Machine Learning and Image Registration

3.1 Introduction

Diabetes Mellitus (DM) has severe consequences, including heart disease, stroke, renal failure, and blindness. Another complication it leads to is diabetic foot ulceration (DFU), which can result in lower limb amputation. DFUs are associated with increased medical costs, reduced quality of life, infections, amputations, and mortality. It is crucial for individuals at risk of DFU to receive timely diagnosis and treatment to prevent foot ulcers, amputations, and death. While self-care monitoring for evidence of DFU can help with early treatment, it has limitations, especially for patients with obesity or vision impairment who may not identify ulceration locations (Alavi et al., 2014).

Recent studies have found that temperature monitoring can predict DFU development in 97 percent of patients (Frykberg et al., 2017; van Doremalen et al., 2020). Moreover, conducting foot temperature monitoring in a remote home setting has been linked to a reduced risk of DFU (Crisologo & Lavery, 2017). Infrared thermogram (IRT) analysis has proven to be effective in diagnosing various diseases (Magalhaes et al., 2021) and has been explored in several studies as a means of identifying those at risk of DFU (Armstrong & Lavery, 1998; Filipe et al., 2021; Kumar et al., 2022; Nag et al., 2021a). These studies have identified characteristic thermal distributions in infrared images of the feet and other organs to assess the risk of DFU.

To facilitate automatic image categorization for this purpose, machine learning (ML) approaches such as artificial neural networks (ANN), k-nearest neighbors (KNN), and support vector machines (SVM) have been widely employed (Prabha et al., 2021; Raghavendra et al., 2019; Xu et al., 2020). These

techniques utilize feature extraction, feature ranking, and various ML models like ANN, KNN, and SVM to support the identification process (Prabha et al., 2021; Raghavendra et al., 2019; Xu et al., 2020).

Benefits of thermography in DM classification over other methods include ease of use and the possibility to be monitored in a home setting. Early detection of DFU can be an indicator for clinical risk assessment. A computer-aided expert-invariant diagnosis approach with low cost and low overhead to implement would be useful in less developed areas or home settings. Our objective is to conduct a comparative analysis between two approaches. Firstly, we will employ an unsupervised image registration technique to identify regions of interest (ROIs) and extract features from these regions. These features will be used to train ML classifiers for predicting disease outcomes. Secondly, we will utilize a multi-view deep-learning architecture that incorporates both ipsilateral and contralateral feet as input to the model. The model is trained to predict disease outcome.

Many studies have provided evidence to support the viability of using thermal imaging to capture the clinical evidence of the presence of diabetic foot ulceration and a correlation between the rise in plantar temperature with diabetic foot neuropathy has been established (Bagavathiappan et al., 2010; Van Netten et al., 2013). Asymmetry analysis of thermal imaging has also proved to be effective in identifying the presence of the disease based on the discrepancies between ipsilateral and contralateral foot (Saminathan et al., 2020). Computer-aided detection of DFU using thermal imaging has been conducted by a few studies. In the study (Saminathan et al., 2020) they chose 11 ROIs from individuals at risk to develop ulcerations. They manually annotated them in both left and right feet images. From these regions they extracted gray-level co-occurrence matrix (GLCM) texture features which consider the spatial relationship of pixels in an image and first-order statistical features namely mean, standard deviation, skewness, and kurtosis. They trained a SVM classifier using the difference calculated

between the features in ipsilateral and contralateral feet. They managed to achieve high accuracy of ~95% using the temperature difference and extracted texture features. The results were reported based on a small private dataset. Their algorithm also consisted of a time-consuming and labor-intensive manual annotation of ROIs. Hernandez, et.al. (D. A. Hernandez-Contreras et al., 2019) proposed a quantitative thermal change index (TCI) where the difference in mean temperature values of angiosomes found in diabetic feet is compared to a baseline control value. The TCI was used to evaluate the thermal change in the plantar region of diabetes patients to identify patients from controls. Hernandez-Contreras et al in their study (D. Hernandez-Contreras et al., 2017) used the TCI to classify subjects into five classes 1 to 5 based on the spatial temperature distribution and temperature range to act as a reference for experts to identify progress of the disease.

Advances in deep learning-based models for medical image analysis is a flourishing field of research. Deep neural network (DNN) learning has demonstrated advantages over traditional methods especially when it comes to tasks that require hand-crafted features by an expert in the field (Khan & Arora, 2018; Magalhaes et al., 2021; Vardasca et al., 2019). However, current literature in thermogram disease detection lacks in a few aspects, like automating ROI detection and using information from both left and right organs as well as feet and hands together in a single model. Multi-view convolution networks are employed in various domains including medical, where features are aggregated from multiple instances or views of the same object to a compressed representation using different fusion strategies to aid in classification (Seeland & Mäder, 2021; Setio et al., 2016). In the study by Jin et al (2021), the authors used a multi-stream Siamese network, fusing intermediate feature representations of a U-net model. The model was trained to predict the pathological complete response from pre-treatment and post-treatment longitudinal magnetic resonance imaging (MRI) scans of rectal cancer patients. They also performed segmentation of tumor in an end-to-end model which proved effective in multi-institution validation. Multi-scale feature representation where features from intermediate

representations are used to capture more information from representations learned by deep neural networks have also proved to be effective in improving predictability outcome prediction. Li et al. (2020) demonstrated an unsupervised approach to whole-slide imaging classification, which is a challenging domain that requires expert annotations of ROIs. Multi-view and multi-scale feature representation would be a suitable approach to incorporate the asymmetric characteristics in hands, hence increasing the predictability of the neural network model for the disease prediction task.

3.2 Materials and Methods

3.2.1 Dataset Description

In this study we used Plantar Thermogram Database for the Study of Diabetic Foot Complications (D. A. Hernandez-Contreras et al., 2019). The publicly available dataset comprises 334 individual foot thermograms from 122 diabetic subjects and 45 non-diabetic subjects. Thermograms captured are labeled as Control or Diabetic. Images were captured using thermal cameras (FLIR E60 and FLIR E6). An infrared radiation obstructive device was placed behind the feet to prevent background temperature from leaking into the final image. The captured thermograms were labeled as Control or Diabetic. Before capturing an image 15 mins time was given to reach an equilibrium temperature. Segmentation was performed on captured images to remove the background using the edge detection method.

3.2.2 Data Preparation

Limited sample size is a challenging constraint in machine learning, especially when involving health care, in which it is challenging to get access to large datasets. Thus, we employed several procedures to mitigate this challenge and produce a comprehensive result. We split the data into train, validation, and test sets. We also generated three more sets of data from these splits using augmentation on the training set and validation set, respectively. This produced three augmented sets in incrementing order of the total images in the set (~1k images, ~3k images, ~8k images). The detailed count and data-split strategy are outlined in **Table 3.1**. The test set was always held out and not augmented in any way. The idea behind creating three sets of augmented data was to identify the consistency of results and observe any subtleties associated with using larger sets of augmented data. We only used position augmentations (rotation, shear, zoom, and scale) to mimic the different anatomical variations of the

feet found in humans. Data augmentation has also been highly utilized when training machine learning or deep learning models to prevent overfitting (Chlap et al., 2021). The augmentations were carried out at the same time to the right and left feet images of a patient to ensure that each pair of images (left and right) correspond to the same patient and synchronicity is not broken.

Table 3.1 Number of images after dataset split and generation of augmented sets.

Dataset Contreras et al.(D. A. Hernandez-Contreras et al., 2019)	Class	Training Set	Validation Set	Test Set
Original Set	Diabetes	86	23	13
	Control	30	9	6
Augmented Set 1	Diabetes	1050	300	13
	Control	1047	300	6
Augmented Set 2	Diabetes	1750	499	13
	Control	1740	498	6
Augmented Set 3	Diabetes	3495	996	13
	Control	3471	998	6

3.2.3 Image registration-based ROI extraction

The goal of this step is to evaluate an approach to identify ROIs that are relevant and important when identifying the presence of the disease. Some spatial locations that are at higher risk for ulceration were defined as ROIs for DFU by Sibbald et al., (2012). In a clinical setting, the physical examiner prioritizes these locations in the plantar region of the foot to identify the risk of disease progression.

We performed an experimental analysis of our ROI extraction method using the clinically relevant ROI for diabetic foot ulcerations. Using a machine learning-based deformable image registration model VoxelMorph (Balakrishnan et al., 2018), given two images (which we call “moving” and “fixed”), the goal of image registration is to learn the deformation between two image. In learning-based methods, we use a neural network that takes in two images m ("moving") and f ("fixed") (e.g.,

MNIST digits of size 32x32) as input, and outputs a dense deformation Φ (e.g., size 32x32x2, because at each pixel we want a vector telling us where to go. Intuitively, this deformation Φ gives the correspondences between the images and tells us how to move the moving image to match up with the fixed image. We utilized this learned deformation required to map a moving image onto a fixed image. This deformation was used to transform ROIs marked on the template (fixed) image to align properly to other un-labeled images. The spatial locations of the ROIs were preserved in all the images. From our experiments we observed that minimizing a normalized cross-correlation loss (NCC) when training the registration model performs better compared to using a mean squared error loss (MSE). This enables us to use an unsupervised approach to find the relevant ROIs in unlabeled images. Limiting the need for time consuming and expensive annotation process on hundreds of images. The image registration-based ROI extraction is summarized in **Figure 3.1**. However, we needed to manually annotate the fixed image (1-5 images of the dataset) and then train the image registration model to learn the deformation field based on these fixed images. The corresponding annotations are then mapped using the inverse of the deformation field output from the trained registration model.

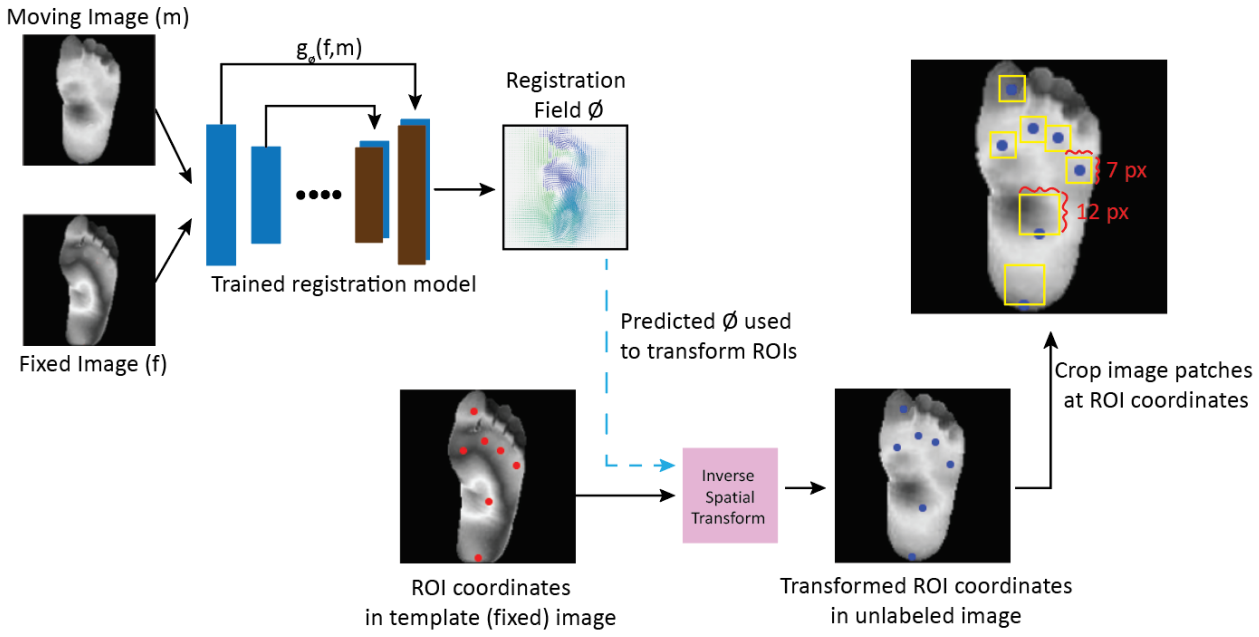

Figure 3.1 Overview of obtaining landmarks from unsupervised image registration technique. This shows the procedure using trained registration model to predict deformation field on unlabeled images and using the inverse of the field to transform ROI coordinates marked on the fixed image. The cropping scheme to generate patches from the image is as follows: **(1)** use larger base size 12x12 pixels for calcaneal heel and mid sole of feet ROIs, coordinate position used as the base of the crop box. **(2)** use smaller 7 x 7 pixels base size for the ROIs near the joints of the toes, coordinate position used as the center of the crop box.

3.2.4 Image registration and obtaining the coordinates of ROIs.

To set the template (reference) locations to use as ROIs, we first manually marked a "fixed image" for both the left and right feet according to the spatial locations identified by Sibbald et al. (2012). We manually marked two images (left and right) with the reference points by a clinician and noted down their coordinates denoted by $f(P_{x,y})$. In the next step, we trained the learning-based image registration model VoxelMorph using the template "fixed images" as the fixed image (f) to train the model and used the rest of the images available in the dataset as the moving images (m). The registration model learns a deformation field Φ which can map the moving image onto the fixed image $f = m\ o\ \Phi$. The overall loss function **Equation 1** was minimized for the model consists of a similarity loss denoted by $\mathcal{L}_{sim}(f, m \circ \phi)$ and a smoothing loss $\lambda\mathcal{L}_{smooth}(\phi)$ regularized by the hyperparameter λ .We trained

the model using a negative cross correlation similarity loss **Equation 2**. This was more robust to intensity variations found across scans and datasets (Avants et al., 2008). The major learning parameters were set at an initial learning rate of 0.005, λ=0.17 for around 200 epochs, steps_per_epoch = 150, until convergence. The learned weights of the model were then saved for inference.

The overall loss for the model is.

$$\mathcal{L}_{us}(f, m, \phi) = \mathcal{L}_{sim}(f, m \circ \phi) + \lambda \mathcal{L}_{smooth}(\phi) \quad (1)$$

Cross-correlation loss between the two images is defined as:

$$CC(f, m\ o\ \phi) = \sum_{\mathbf{p} \in \Omega} \frac{\left(\sum_{\mathbf{p}_i} \left(f(\mathbf{p}_i) - \hat{f}(\mathbf{p})\right)\left([m \circ \boldsymbol{\phi}](\mathbf{p}_i) - [\hat{m} \circ \phi](\mathbf{p})\right)\right)^2}{\left(\sum_{\mathbf{p}_i} \left(f(\mathbf{p}_i) - \hat{f}(\mathbf{p})\right)^2\right)\left(\sum_{\mathbf{p}_i} \left([m \circ \boldsymbol{\phi}](\mathbf{p}_i) - [\hat{m} \circ \boldsymbol{\phi}](\mathbf{p})\right)^2\right)} \quad (2)$$

A higher CC indicates a better alignment, so the similarity loss function yields to: $\mathcal{L}_{sim}(f, m \circ \phi) = -CC(f, m\ o\ \phi)$.

After the model training is completed, we use the trained model to predict the deformation field ϕ_i for every unlabeled image m_i in the dataset and transform the fixed annotation coordinates $f(P_{x,y})$ to get the new annotation for the unlabeled image $m_i(P_{x,y})$. Intuitively the deformation field gives us a 2D vector that tells where to move each pixel in the moving image to match the fixed image shape. However, we do not want to morph the image. Instead, we want to transform the coordinates in the fixed image to match the moving image. We do this by taking the reverse direction that the ϕ_i yields called the inverse spatial transform operation **Equation 3.**

$$m_i(P_{x,y}) = f(P_{x,y})\ o\ {\phi^{-1}}_i \quad (3)$$

Although this method does not yield perfect results since it is limited by the image registration accuracy, and noise in the deformation field due to varying contrast in the images. It still manages to

produce approximate coordinate prediction of the ROIs without the overhead of requiring any supervised labels. Some examples of images with their predicted ROIs are shown in **Figure 3.2**. Based on the transformed coordinates of the ROIs in each image, we cropped the image into square (N x N) patches at each coordinate location to extract features from each landmark location.

Table 3.2 Description of the First and second order features extracted from the ROI.

Local features extracted from ROI patches			
Method	Feature Name	Formulas. $P_{ij} = Element\ i,j\ of\ the\ normalized\ GLCM$ $N = Number\ of\ levels$ $\mu = GLCM\ Mean$ $\sigma^2 = Variance$	Description (Mohanaiah et al., 2013; Saminathan et al., 2020)
GLCM	Contrast	$\sum_{i,j=0}^{N-1} P_{ij}(i-j)^2$	Contrast represents the measure of the contrast intensity between a pixel and its neighboring pixels over the image.
GLCM	Correlation	$\sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\delta^2}$	Measure of how closely a pixel relates to its neighbors
GLCM	Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$	It is the measure of the closeness of the distribution of elements in the gray level co-occurrence matrix to its diagonal.
GLCM	Dissimilarity	$\sum_{i,j=0}^{N-1} P_{i,j} \mid i-j \mid$	Dissimilarity is a measure of distance between pairs of pixels in the patch of image (region of interest)
GLCM	Variance	$\sum_{i,j=0} j(i-\mu)^2 P(i,j)$	It is the square of the SD. A small variance indicates that the data points are close to the mean, whereas the large variance means that the data points are spread out from the mean
GLCM	Angular Second Moment (ASM)	$\sum_{i,j=0}^{N-1} (P_{ij})^2$	Shows the size of the shape irregularities of an image, defined as the sum of squared elements in gray-level co-occurrence matrix.

GLCM	Mean	$\frac{1}{n}\sum_{i,j} P(i,j)$	Shows the measure of average intensity of the pixels in the image of interest.
GLCM	Energy	$\sum_{i,j=0}^{N-1}\sqrt{(P_{ij})^2}$	When an image has excellent homogeneity or when the pixels are substantially similar, the energy is high.
GLCM	Entropy	$\sum_{i=0}^{N\,g-1}\sum_{j=0}^{N\,g-1} -P_{ij} * \log P_{ij}$	Entropy demonstrates how much information from the image is required for image compression. Entropy assesses both the picture information and the message or information loss in a transmitted signal.
First-order statistical feature	Skewness	$\sum_{i,j=0}^{N-1}\frac{\left(P_{ij}-\mu\right)^2 / N}{\delta^3}$	Skewness shows the relative tilt level of the histogram curve on an image. Skewness value can be positive or negative
First-order statistical feature	Kurtosis	$\sum_{i,j=0}\left[\frac{P_{ij}-\mu}{N}\right] / \sigma^4$	Kurtosis shows the relative tallness of the histogram curve of an image.
Global Features extracted from entire image			
First-order statistical feature	Mean	As defined above	Shows the measure of average intensity of an image.
First-order statistical feature	Variance	$\sum_{i,j=0} j(i-\mu)^2 P(i,j)$	It is the square of the SD. A small variance indicates that the data points are close to the mean, whereas the large variance means that the data points are spread out from the mean
First-order statistical feature	Kurtosis	As defined above	Kurtosis shows the relative tallness of the histogram curve of an image.
First-order statistical feature	Skewness	As defined above	Skewness shows the relative tilt level of the histogram curve on an image. Skewness value can be positive or negative

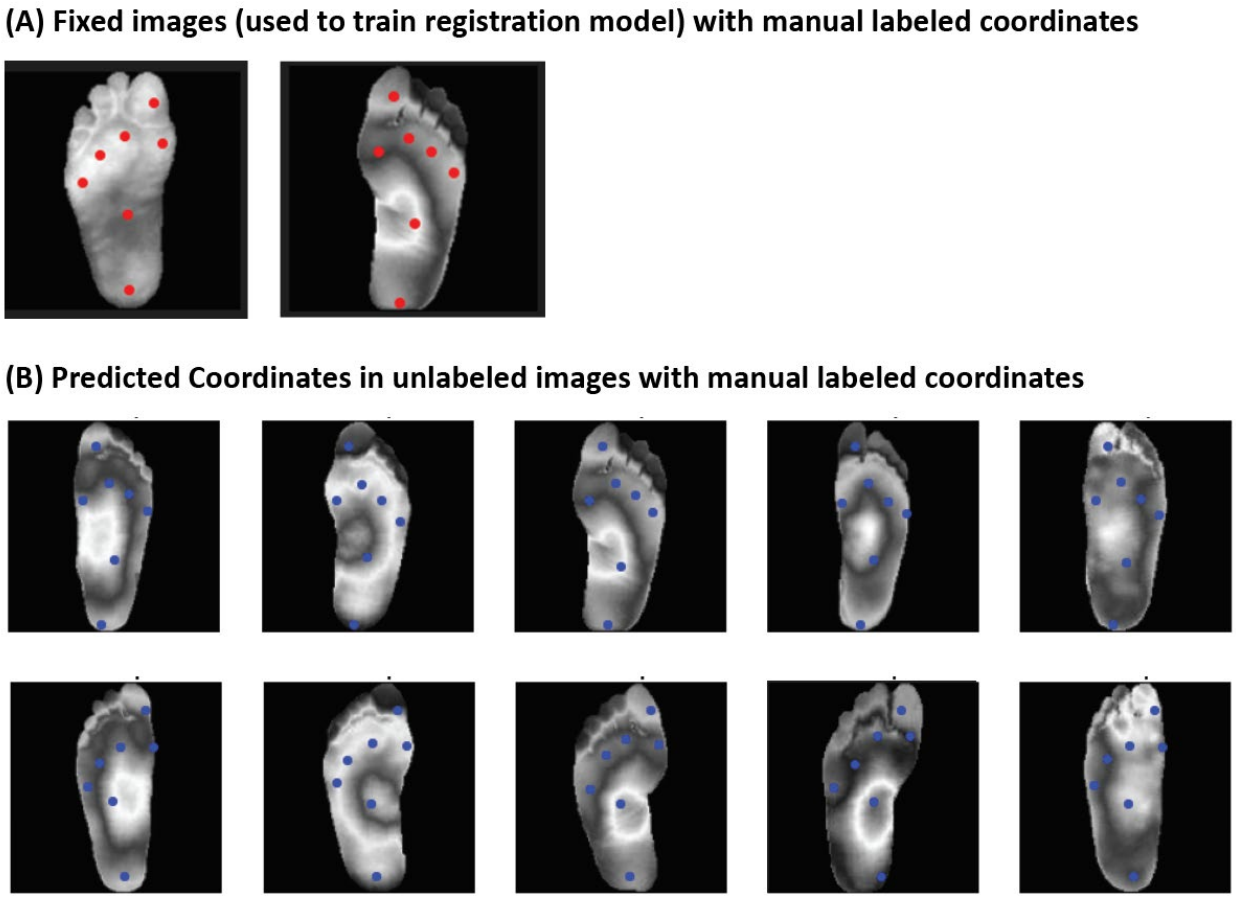

Figure 3.2 Some results of obtaining landmarks from unsupervised image registration technique. **(A)** Fixed images used to train the registration model, with the manually annotated coordinates in red by a clinician. **(B)** Examples of left and right foot thermograms using the transformed coordinates by the deformation field predicted by the registration model.

3.2.5 Feature extraction from ROI patches

The ROI patches in the foot were labelled as R1-R7 (**Figure 3.3**) The toe apices (great toe) (R1), 5 metatarsal heads behind the phalanges of the toe apices (R2-R6) and calcaneal heel (R7) are considered as primary ROIs 'at risk' for developing diabetic foot ulcer (Sibbald et al., 2012; Macdonald et al., 2016). Based on the ROI coordinates predicted by the image registration model, we identified patches of size (n x n) where n was increased at increments of (10%, 20%, 30%, and 40%)

with a base size of 7x7 pixels for R1 to R5 and a larger 12x12 pixels for R6 and R7, representative of the calcaneal heel and plantar region as these required more surface area to cover.

From these patches features were extracted to train the ML classifiers. Several feature extraction techniques have been employed by different studies when analyzing thermal images using ML approaches. Of them the most popular and effective approach is based on the first and second order statistical (FSStat) features extracted from the ROI patches using gray-level co-occurrence matrix (GLCM (Nag et al., 2021b; Rasyid et al., 2018). Details of the extracted features are summarized in **Table 3.2.** The GLCM (Mohanaiah et al., 2013) is a renowned statistical technique based on joint probability distribution of pair of pixels. For our study we extracted both the first and second order statistical features from the individual patches identified in each thermogram to use as relevant features to train the machine learning classifiers.

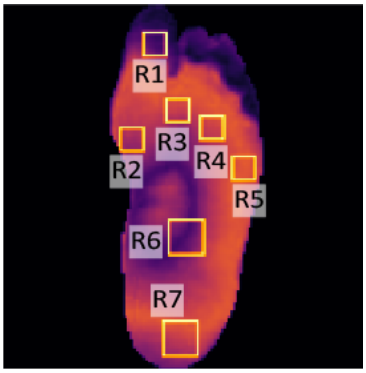

Figure 3.3 Patches from region of interests

3.2.6 Generation of feature table and feature selection

From each thermogram of the feet we identified 7 ROIs. We extracted 11 FSStat features from these ROIs as shown in **Table 3.2**. We arrange these in columns where each feature in the table is named using the extracted feature name as prefix and the ROI number as the suffix. So, for instance, the 'mean' from the 6th ROI patch is labelled as 'mean-6'. This gives us a table consisting of 77 feature

columns (7 x 11) and each row representing one patient's thermogram. Since this is a large feature set compared to the sample set size, we had to carry out feature selection and dimensionality reduction to remove multicollinearity as well as limit the degrees of freedom of the classifier to prevent overfitting. This will also result in a more easily interpretable depiction of the target concept, directing the output's attention to the most important variables.

As an initial pre-processing step, highly correlated features (threshold set to 85%) in the feature set were dropped, resulting in a feature set consisting of 55 columns. To further reduce the feature set after the multicollinearity optimization the top features were identified for binary classification. Three different sets of feature ranking were identified using the Multi-Tree Extreme Gradient Boost (Xgboost) (Chen & Guestrin, n.d.), random forest and lasso. To prevent overfitting default parameters were used to train the feature selection models. The best performing top-ranked features from the various feature ranking methodologies are used to discover the best feature combination utilizing a rigorous investigation to determine the best feature combination that provided the best performance. The final feature set was limited to 12 top features to train the classifiers.

3.2.7 Classification of DM using machine learning approach

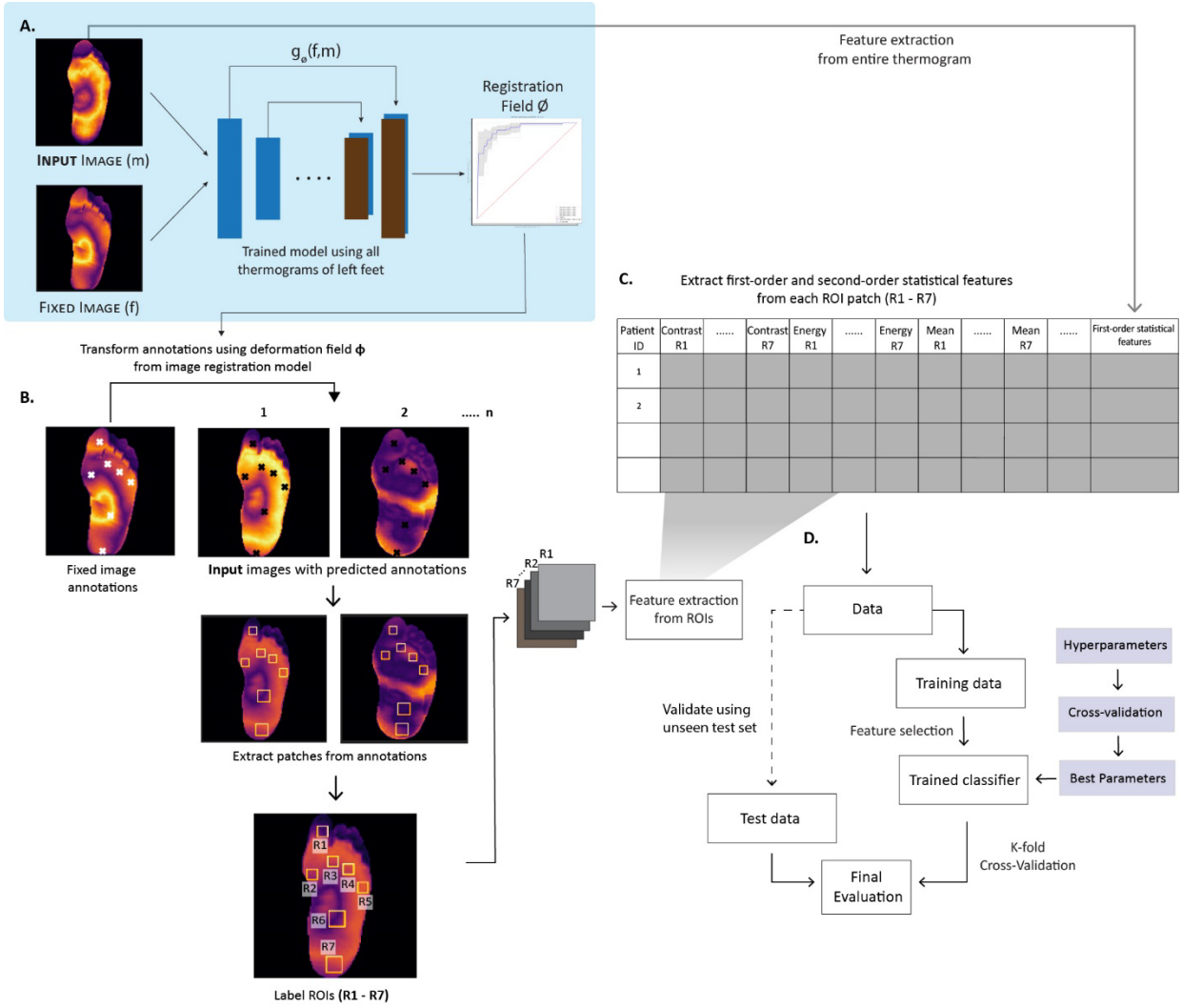

Figure 3.4 Machine learning-based DFU classification overview (A.) Find a deformation field to transform ROIs for unlabeled images and train a learning-based deformable image registration model (Balakrishnan et al., 2018) using all the images and one fixed image. (B.) Input to the registration model will output a deformation field Φ; the inverse of which is utilized to warp the ROIs on fixed images to get the corresponding ROIs in unlabeled images. (C.) Statistical and textural feature extraction was performed on the ROI patches and synthesized into a table; features extracted from the entire image were also constituted in this table. (D.) Data was split into train and test sets. Highly correlated features dropped, feature selection performed using Xgboost feature importance, and machine learning classifiers trained on these extracted features using K-fold cross-validation (K = 5) to classify image-level ground-truth (diabetic or control).

Both five-fold cross-validation and train test validation split was used in this study, where each fold was divided into an 80% training and 20% testing set. And for the train, validation and test split the data was divided into 70% training, 20% validation and 10% test set. Both the approaches were used to ensure consistency in results invariant of the data-split strategy used. The feature set was first optimized by removing any redundant features, i.e., features with correlation more than 85% were removed. The hyperparameter search space used to train and cross-validate the algorithms are summarized in **Table 3.3**.

Table 3.3 Hyperparameter search space for the ML algorithms

Method	Hyperparameter	Range of values
Machine learning algorithm		
XGBoost	max_depth	[2, 3, 4, 5]
	learning_rate	[0.01, 0.05, 0.1, 0.2]
	subsample	[0.1, 0.2, 0.3, 0.4, 0.5]
	alpha	[2.5,2.6,2.7,2.8]
	gamma	[4.5,4.6,4.7,4.8, …, 5.5]
	lambda	[0.6, 0.7, 0.8, 0.9, 1.0]
	colsample_bytree	[0.1,0.2,0.3,0.4,0.5,0.6]
	colsample_bylevel	[0.1, 0.2, 0.3,0.4,0.5]
	n_estimators	[20, 50, 80,120]
Random Forest	bootstrap	[True, False]
	max_depth	[10, 20, 30, None]
	max_features	['auto', 'sqrt']
	min_samples_leaf	[1, 2, 4]
	min_samples_split	[2, 5, 10]
	n_estimators	[200, 400, 600]

SVM	Regularization parameter, C	[0.1,1, 10, 100]
	gamma	[1,0.1,0.01,0.001]
	kernel	['rbf', 'poly', 'sigmoid']
Image Registration (VoxelMorph (Balakrishnan et al., 2018)**)**		
VoxelMorph	Lambda λ	[0.13, 0.15, 0.17, 0.19, 0.21]
	Similarity loss $\mathcal{L}_{sim}$	['mean square error', 'normalized cross-correlation']
	Initial learning rate	0.005
	Batch size	1
Multi-view CNN		
Resnet50 / Mobilenetv2	Initial learning rate	0.000025
	Learning rate schedule	Step decay
	FC layers	256-128-1
	Dropout in FC	0.5
	L2 Regularization in FC	0.1-0.4
	Conv weight initializer	glorot uniform (xavier uniform)
	Batch size	16
	FC Activation	Sigmoid

3.3 Multi-view and multi-scale feature representation DL model

We propose a multi-view and multi-stream CNN network using pairwise images and multi-scale feature representation integration for effective prediction. To effectively capture asymmetric information between the left and right images of feet. We propose a multi-stream learning framework with deep neural network architecture. The network consists of two main components: (1) an encoding network for feature extraction based on convolutional neural networks (CNN), and (2) a multi-stream Siamese subnetwork for diagnostic prediction trained on left and right images simultaneously. The overview of the architecture is outlined in **Figure 3.5**. The feature extraction subnetwork consists of two identical pre-trained CNN networks (pre-trained on ImageNet). We tested the model using

different subnetworks from popular CNN models (Resnet50 and MobilenetV2). The classification subnetwork combines features extracted from three distinct intermediate layers of the feature extraction subnetworks via depth-wise convolutions (Y. Guo et al., 2019). This ensures that both shallow and deep feature representations obtained by the Siamese feature extraction sub-networks can be fully integrated and applied to the disease classification task. Therefore, integrating multi-scale feature representations and pairwise comparison of images from different views (left, right, or hands, feet) of the patient. The classification subnetwork used a binary cross entropy loss or log loss **Equation 4**.

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i\,.\,log\,p(y_i) + (1-y_i).\,log\big(1-p(y_i)\big) \quad (4)$$

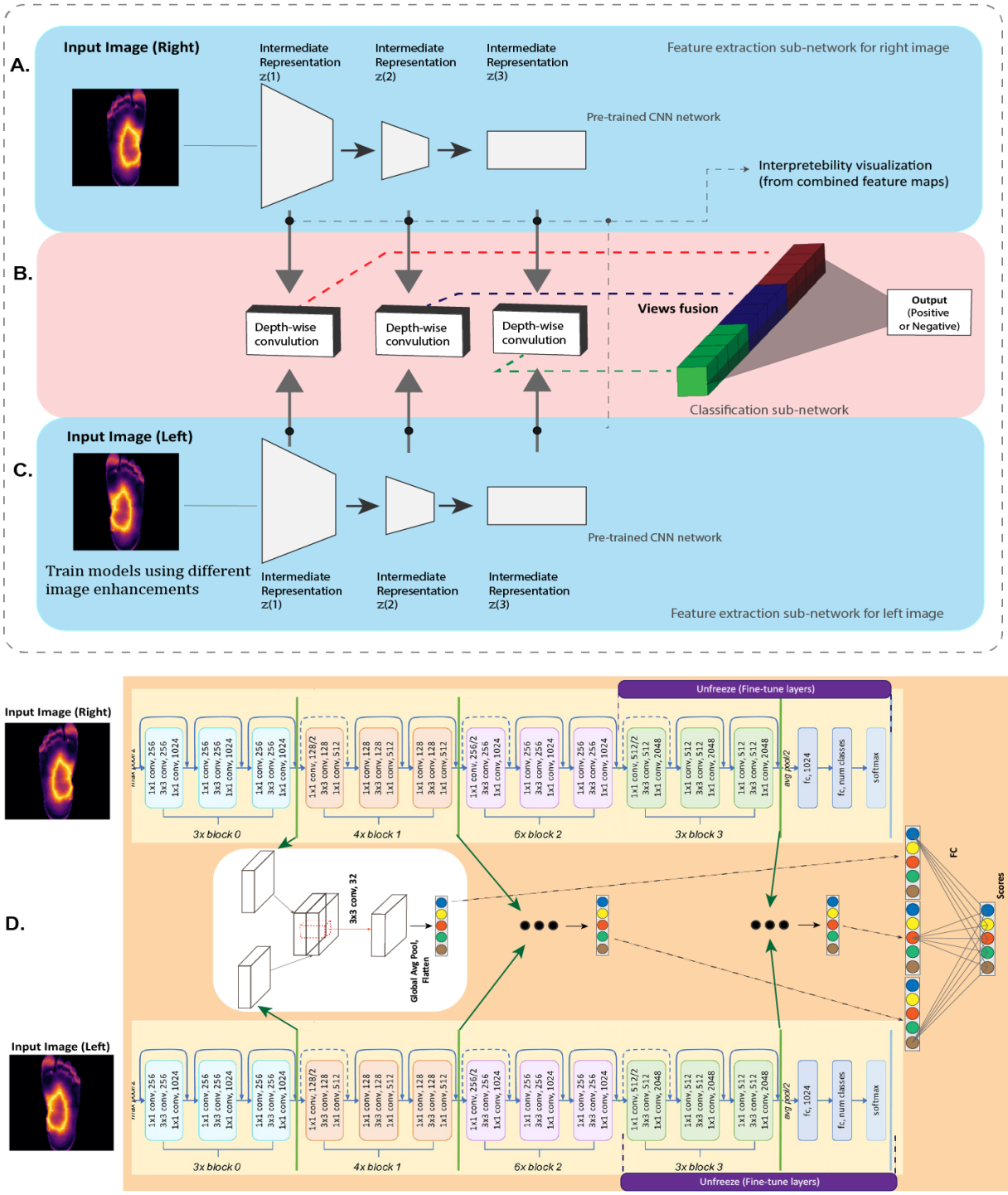

Figure 3.5 Proposed multi-view CNN network for DFU classification. The multi-stream deep learning network consists of three subnetworks (A, C & B): two for feature extraction, and one for classification. The two feature extraction subnetworks (A, C) each take an image as input, where the four models will be trained with different inputs which are single-channel enhanced images (No enhancement, CLAHE enhanced, and gamma enhanced) and a stacked 3-channel image (a). The first and second feature extraction subnetworks input left and right images, respectively. The classification network (B) uses depth-wise convolutions (b) to combine intermediate features from three distinct layers in the feature extraction subnetworks for multi-scale feature integration and prediction. (D) Visualizes the overall flowchart of the depth wise convolutions at different levels in the ResNet50 network.

3.3.1 Image enhancements and model inputs

Contrast-Limited Adaptive Histogram Equalization (CLAHE): Adaptive histogram equalization (AHE) is a computer image processing technique for improving image contrast. CLAHE is an AHE version that reduces noise amplification.

Gamma Adjustment: Gamma correction or gamma is a nonlinear operation used to encode and decode luminance A gamma value $g < 1$ is called an encoding gamma, and the process of encoding with this compressive power-law nonlinearity is called gamma compression, we used a value of ($g=0.6$)

All Enhancements (Stacked channel-wise): We tried to capture the information enhanced by all the enhancement techniques by combining three enhancements performed on single channel image and then these are stacked to form a 3-channel image (None, Gamma, and Clahe). Finally, as a baseline no enhancements are performed on the images.

3.3.2 Model interpretation and observations

Interpretability of deep learning models is an active field of research and has shown effective utility in ensuring the clinical validity of models when performing classification. Interpretability techniques help to visualize the importance the model assigns to spatial locations of the image during inference and better understand which areas of the image and what kind of features contributed to the network's output. We note that since feature representations from multiple layers in the feature extraction subnetworks are integrated to train the overall model. The shallow layers will capture low-level structural representations whereas deeper layers will capture high-level semantics and the depth wise convolution performed to integrate these representations from both the feet will contain information pertaining to the characteristics present in both the feet representative of identifying DFU inflammation. We used the approach GradCam (Selvaraju et al., 2020) to visualize the feature maps

representing the activations in the prediction network in the intermediate representations of both the feet.

3.4 Performance metrics

It is vital to evaluate the performance of the proposed algorithms with baselines and each other using performance metrics. Due to the limited sample size, in the machine learning based approach cross-fold validation will be used to train the model, and grid search used to tune the hyperparameters (number of trees, tree-depth, learning rate, number of layers, kernel size, etc.). For performance evaluation, Receiver Operating Characteristics Area under the curve (ROC-AUC) and F1 score for the classification task will be used to measure the performance of both the ML and CNN models. The ROC (receiver operating characteristics) curve is used to calculate the AUC score. ROC measures the performance of the models using different thresholds. Higher AUC reflects better performance. In addition to this we also divided the data into training, validation and test split and trained both the approaches using the training set and results reported on the hold-out (test set) using three sets of augmented data (1k samples, 3k samples and 8k samples) to better validate the generalizability of the model outputs.

$$Accuracy = \frac{TP+TN}{(TP+FN)+(FP+TN)} \quad (5)$$

$$F1\ Score = \frac{2*(Precision*Sensitivity)}{Precision+Sensitivity} \quad (6)$$

where, Precision (Positive predictive value) = $\frac{\text{TP}}{\text{TP + FP}}$ and Sensitivity (True Positive Rate) = $\frac{\text{TP}}{\text{TP + FN}}$

$$Specificity = \frac{TN}{TN+FP} \quad (7)$$

$$AUC = Area\ under\ the\ Receiver\ operating\ characteristic\ (ROC)\ curve, threshold = 200$$

3.5 Results and Discussion

3.5.1 ROI feature-based detection

To first evaluate the baseline performance without using any region of interest, we used whole images and extracted FSStat features from them, followed by feature selection and ranking using Xgboost, Random Forest and SVM. Default hyperparameters were used to train the models to identify most important features, without overfitting to the training data. The models were then trained and hyperparameters optimized using grid search using the top 12 features, the maximum accuracy achieved by the best model using whole image extracted features was ~85% **Table 3.4.** Following that we trained the same dataset instead of using the whole image to extract features from, we first used the output from the registration model to identify region of interests (ROIs) and extract features from nxn patches based on these spatial locations. Similar procedure of feature ranking and selection and randomized cross validation search to identify the best hyperparameters. Finally, the top 12 features selected according to the feature importance calculated by the default parameter models were noted. These were used to train the final model. The results on the hold-out (test) set are summarized in **Table 3.4.**

Table 3.4 Test set results for the Machine learning ROI-based approach. The ROI patch sizes are denoted by the Model name - X%.

Model (size – ROI size %)	Test Acc	Test AUC	Test F1-Score	Sensitivity	Specificity
Original					
Xgboost - 0%	0.888	0.811	0.927	0.883	0.845
Xgboost - 10%	0.905	0.832	0.939	0.893	0.877
Xgboost - 20%	0.883	0.856	0.906	0.854	0.824
Xgboost - 30%	0.950	0.943	0.926	0.900	0.890
Xgboost - 40%	**0.973**	**0.977**	**0.956**	**0.968**	**0.860**
Augemented2 (3k samples)					
Xgboost - 0%	0.928	0.929	0.930	0.917	0.940
Xgboost - 10%	0.905	0.906	0.907	0.884	0.929
Xgboost - 20%	0.963	0.962	0.964	0.961	0.964
Xgboost - 30%	0.943	0.943	0.934	0.928	0.958
Xgboost - 40%	0.964	0.963	0.962	0.950	0.976
Augmented1 (1k samples)					
Xgboost - 0%	0.876	0.875	0.867	0.853	0.902
Xgboost - 10%	0.871	0.870	0.864	0.860	0.889
Xgboost - 20%	0.926	0.851	0.842	0.83	0.872
Xgboost - 30%	0.945	0.876	0.885	0.897	0.902
Xgboost - 40%	0.966	0.903	0.904	0.907	0.881

Although an accurate comparison could not be made with the previous study working on a similar problem on the same dataset (Khandakar et al., 2021) due to different splits of data used and different augmentation strategy, however it is evident that our ML based approach using only image-specific features (excluding demographic information) yielded results like their reported results using demographic information during model training. Our best model achieved a test AUC of 0.977 and accuracy of 0.973 which outperforms their best reported result based on a similar tree-based random forest model which achieved an accuracy of 0.967. Saminathan et.al., (2020) reported results on a private dataset using an SVM classifier trained on extracted features from 11 manually annotated ROIs and used asymmetry analysis, their best model achieved an accuracy of 95.61%, sensitivity of 96.5% and specificity of 92.41%.

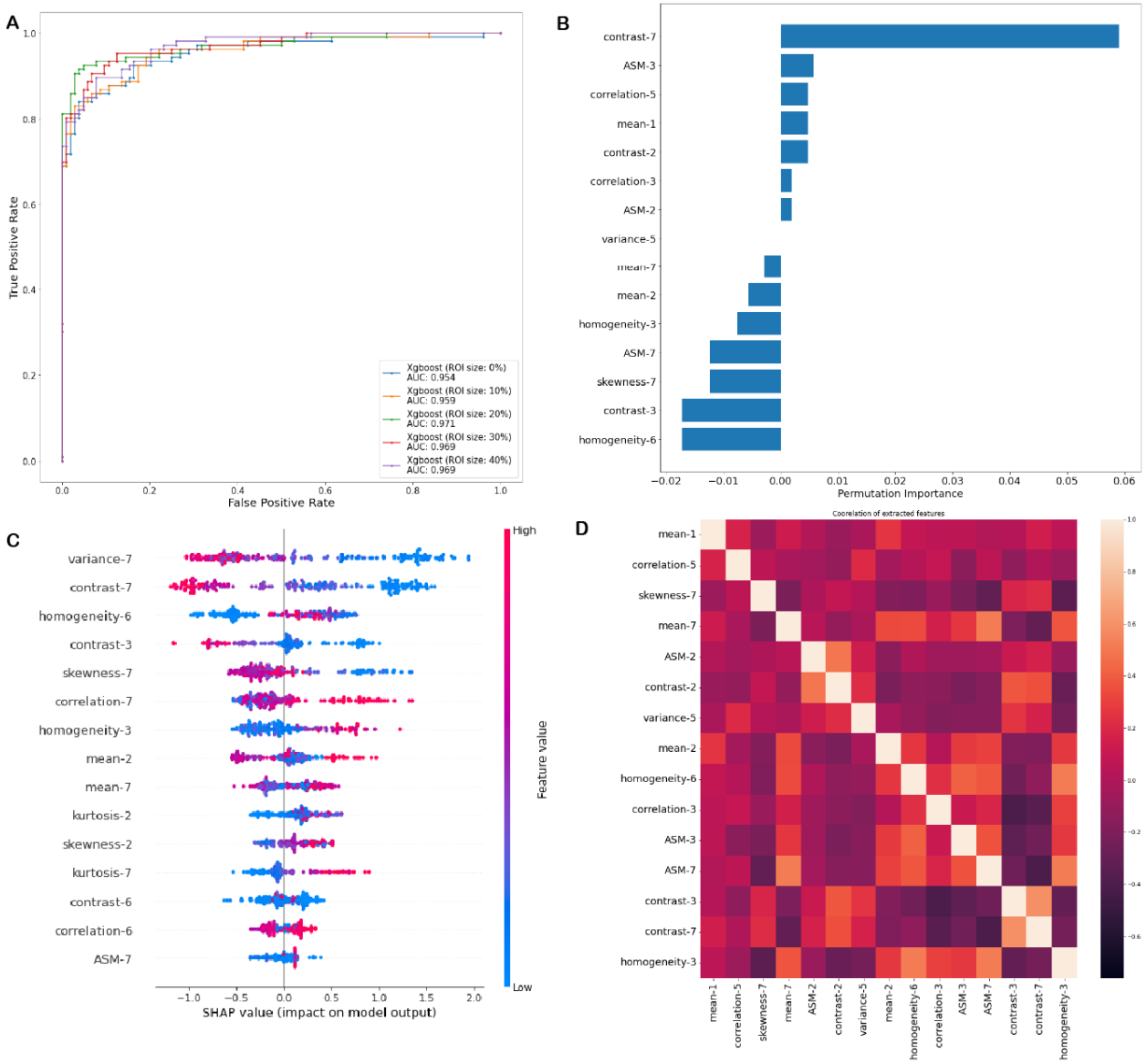

Figure 3.6 Results of the ML approach using image-only features extracted from ROIs (A) ROC for increasing ROI sizes for the best xgboost classifier (f=12). The optimal cutoff point was determined by maximizing the Youden's index on the ROC curve. (B) Average feature importance assigned by the gradient-boosted tree model by permuting the columns and assigning a value based on the magnitude of effect it has on the model's output. Top 15 features are shown. (C) Tree Explainer[33] shapley values from game theory to estimate how does each feature contribute to the prediction, where a negative SHAP value represents the classifier predicting a lower probability (0 or Control) whereas a higher SHAP value represents the model predicting a higher probability (1 or Diabetic). (D) Cross-correlation heatmap for the top 15 features selected by the default xgboost model.

We have investigated results using different feature extraction strategy from the thermal images. The ROI-based approach was carried out using increasing sizes of patch to extract FSStat features from. In our study the primary focus was to only use image level features to achieve the highest accuracy. Incorporating demographic information such as age, weight, ethnicity, etc. helps to improve model performance as reported by previous studies (Khandakar et al., 2021; Saminathan et al., 2020) however this has the drawback of overfitting to the data distribution used to train the model and might rely primarily on demographic features to predict disease outcome. From the results reported in **Table 3.4,** it was observed that increasing the ROI patch sizes used to extract features from improved the model performance consistently on both the original and augmented sets. Furthermore, from the feature importance calculated from the tree-based classifier it can be observed that the mean, contrast, and skewness of the pixel values (representative of temperature intensity of that region) of the ROI-7 which corresponds to the calcaneal heel of the foot have a higher importance score. Similarly, the ROI-1 and ROI-2 and ROI-6 features also have high importance values, which correspond to the toe apices (great toe), and the metatarsal head of the great toe region of the image. In previous studies, these areas have been established to show higher temperature changes during the progression of the disease (Gary Sibbald et al., 2012; Macdonald et al., 2016). The ROC for different investigations in **Figure 3.6** show a trend of higher AUC score with increasing ROI patch sizes used to extract features from, this indicates that increasing the neighboring pixels utilized when extracting features from the ROIs improves overall performance of the classifier, but this reaches a plateau after 30%, we considered a max 40% increase from the base size patch to ensure there is not a large overlap between the ROI patches..

From **Table 3.5** we can observe that extracting features from the whole image yields a poorer result when compared to extracting features from localized regions of the images. Thus, emphasizing the benefit of targeting specific spatial locations to reduce the noise and feature space the model must

learn from. Saminathan, et.al (2020) used asymmetry analysis in a small private dataset to find the difference in the extracted texture and temperature features between ipsilateral and contralateral foot regions (ROIs) to train an SVM classifier. Similarly, we also investigated this method using the extracted features from ROIs of both left and right feet, the results reported in **Table 3.5.** show that this did not yield the best results compared to using features from individual feet.

Table 3.5 Best results acquired from using different areas for feature extraction.

Feature extracted from	Accuracy	AUC	F1-Score	Sensitivity	Specificity
Whole image	0.778	0.789	0.712	0.857	0.723
Difference between Left and Right ROIs (asymmetry features)	0.848	0.821	0.898	0.917	0.667

3.5.2 Deep learning and Multi-view CNN-based approach

We investigated the multi-view network approach using two subnetwork architectures from popular CNN architectures used in literature, Resnet50 and MobilenetV2, as these are relatively shallower networks with fewer parameters to train and have been reported to perform better over other CNN architectures when trained using thermal images (Cruz-Vega et al., 2020; Khandakar et al., 2021). We have also analyzed the results obtained when using three image enhancement techniques when training the models namely (Gamma, Clahe, All) elaborated in Section 3.3.1.

Table 3.6 Results using multi-view CNN approach the subnetwork used using both left and right images as input. The image enhancements used for training are denoted by (None, Gamma, Clahe, All).

Model (image enhancement)	Test Acc	Test AUC	Test F1 Score	Sensitivity	Specificity
Augmented3 (8k)					
MobileNetV2 (None)	0.852	0.964	0.833	0.909	0.838
MobileNetV2 (Gamma)	**0.932**	**0.963**	**0.893**	**0.909**	**0.902**
MobileNetV2 (Clahe)	0.875	0.963	0.900	0.882	0.829
MobileNetV2 (All)	0.85	0.818	0.795	0.818	0.848
Augemented2 (3k)					
MobileNetV2 (None)	0.895	0.945	0.875	0.909	0.836
MobileNetV2 (Gamma)	**0.955**	**0.967**	**0.960**	**0.949**	**0.895**
MobileNetV2 (Clahe)	0.889	0.908	0.886	0.879	0.815
MobileNetV2 (All)	0.825	0.892	0.808	0.846	0.804
Augmented1 (1k)					
MobileNetV2 (None)	0.882	0.964	0.833	0.909	0.830
MobileNetV2 (Gamma)	**0.9375**	**0.963**	**0.944**	**0.927**	**0.889**
MobileNetV2 (Clahe)	0.875	0.927	0.900	0.887	0.780
MobileNetV2 (All)	0.900	0.964	0.908	0.961	0.802

A total of 12 different models were trained using the Mobilenetv2 architecture as the subnetworks of the Siamese network on all the combinations of the augmented training sets and image enhancements applied to the training set. Similarly, it was repeated using Resnet50 as the subnetwork architecture. From the results, it was evident that using Mobilenetv2 as the subnetworks of the multi-view network outperformed in all the evaluations. This was also more efficient in terms of computational load since it comprised a shallower network compared to other popular CNN architectures. The results of the best models on the combination of varied sizes of augmented training sets and image enhancements used on the training set are detailed in **Table 3.6.**

To explore other methods of incorporating ipsilateral and contralateral images in an end-to-end network, we trained pre-trained MobileNetV3 and Resnet CNN models using concatenated images as input. Two concatenation strategies were evaluated, firstly stacking the left and right images channel-

wise (or depth-wise) and secondly stacking them side by side (horizontal stacking). The best-performing model's test performance is summarized in **Table 3.7,** the models were trained using the same training set and evaluated using the same test set as the multi-view model.

Attention mechanisms in neural networks have been used extensively in the literature to provide an additional focus on a specific component of the input feature representation. We decided to train our full model by integrating two attention mechanisms Convolutional Block Attention Networks (CBAM) and Squeeze and Excitation networks (Hu et al., 2020; Woo et al., 2018). Both are modular attention mechanisms and are architecture agnostic for integration. CBAM takes an intermediate feature map and then infers attention maps along the channel and spatial dimensions, then the attention maps are multiplied together with the input feature map for adaptive feature refinement. Squeeze-Excitation blocks focus on channel-wise representations and adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels (Hu et al., 2020). We implemented our multi-view model introducing these attention blocks before each of the three intermediate representations. However, using the attention blocks did not yield a significant increase in the predictive accuracy of the model. Using the CBAM attention block performed better than the SE block approach.

Table 3.7 Best result acquired by different CNN models using different inputs.

Model	Accuracy	F1-Score	AUC	Sensitivity	Specificity
Resnet50 depth-wise stacking	0.802	0.812	0.799	0.816	0.785
MobilenetV2 depth-wise stacking	0.813	0.844	0.829	0.846	0.825
Resnet50 horizontal stacking	0.705	0.719	0.700	0.743	0.689
MobilenetV2 horizontal stacking	0.725	0.819	0.749	0.716	0.685

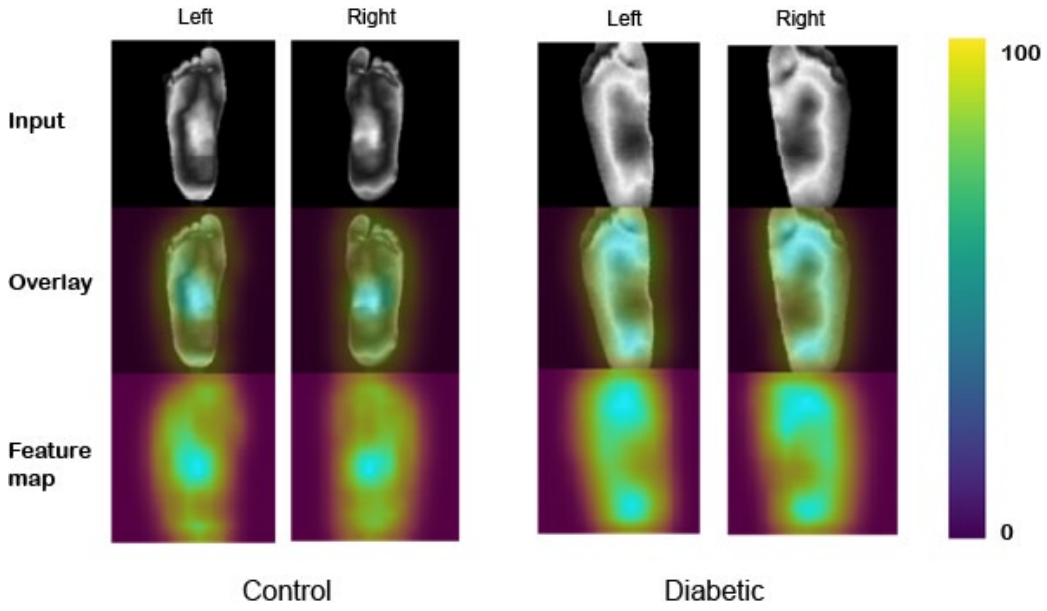

Figure 3.7 Multi-view CNN network visualization and interpretation. Left and right thermograms and the corresponding feature map visualization in key channels at the bottom of the multi-view network for two representative patients who were correctly predicted to be diabetic and control respectively.

Following that, we attempted to determine which portions of the image and what types of features contributed to the network's output and if there was a similar pattern in spatial locations of the feet which contributed more to the network's output. First, we note that the temperature values at different regions of the foot are useful for disease prediction, in addition to that the asymmetric features between the ipsilateral and contralateral feet also contribute to disease prediction. Different layers of the network typically capture two sorts of information. Shallow layers mostly reflect structural information of the images. Deep layer features, on the other hand, mostly represent high-level semantic properties from anatomical images. The classification network combines the features from the two subnetworks at different depths of the model essentially capturing both the spatial and asymmetric features (if any) between the left and right feet.

We selected two representative patients in the hold-out test set, one with diabetes and one control which the network successfully classified. The aggregated feature maps of the channels of the classification network were visualized as seen in **Figure 3.7.**

3.6 Summary

In our study, we investigated two computer-aided diagnostic approaches for early detection of DFU from thermograms of the feet. First, we used a tree-based machine learning classifier xgboost, which was trained using only image-level texture and temperature features (excluding demographic information) from 7 ROIs that are prone to get ulceration in foot regions selected from the thermal images. In addition, we also used an image registration-based model to identify the key regions of interest from the thermograms automatically. Second, we leveraged a multi-view CNN model to input multiple images more specifically ipsilateral and contralateral feet images, and trained end to end to predict the disease outcome using learned features from both feet simultaneously. The patients are classified into normal and diabetic using the ML classifier and multi-view CNN model with an overall AUC of 0.97 and 0.94 respectively. The performance achieved by the proposed methods was better than the other methods reported in the literature. In addition, both the models were able to achieve these results using only the temperature profile feature extracted from the thermal images and not incorporating demographic information such as age, gender, ethnicity, etc. This is advantageous since these features might not be available due to privacy concerns when collecting the data from patients. Also, in small datasets, it is challenging to capture a fair representative set that defines an entire population. In such cases, the trained model assigning a larger weight to a demographic feature such as age (Khandakar et al., 2021) might not perform well when validated using an external population. The study concludes that analysis of texture and temperature features extracted from targeted regions

of interests in foot thermograms were able to detect the DFU in patients with high accuracy and using an image registration-based approach provides an approximate result eliminating the overhead required for annotating the regions manually in the thermal images. Additionally, a multi-view CNN model can be used to incorporate ipsilateral or contralateral organs, and even images from different organs like hands and feet together in a single model to predict the disease outcome with reliable accuracy. In future studies, the region of interest identification process can be evaluated using a larger dataset and evaluate these methods to predict the risk of a disease which affects multiple organs together (hands and feet) like rheumatoid arthritis to extend the early findings obtained in this study.

3.6.1 Data Availability

Data used in the project is available open-source from https://ieee-dataport.org/open-access/plantar-thermogram-database-study-diabetic-foot-complications. This was published by D. Hernandez-Contreras et.al and published in 2019. It consists of left and right feet thermograms of 167 patients (122 diabetic and 45 control) taken from the plantar view. The feet were segmented from the background and the left and right feet images were separated.

3.6.2 Code Availability

The deep learning models were developed using standard libraries in open-source platforms including Keras and TensorFlow and Scikit-learn for the machine learning model. Custom codes for the deployment of the model are available https://github.com/sqbislam/Plantar_Thermogram_Diabetes_Project.

The paper and code for the deep learning-based image registration model can be found at https://ieeexplore.ieee.org/document/8633930/ and https://github.com/voxelmorph/voxelmorph, respectively

Chapter 4 Deep learning-based joint detection and SvH score prediction in Rheumatoid arthritis hand radiographs.

4.1 Introduction

Rheumatoid arthritis (RA) is an autoimmune disease affecting joints, requiring early detection to prevent damage and improve outcomes. Current diagnosis relies on clinical evaluation, lab tests, and imaging, but manual joint damage evaluation can be time-consuming. Deep learning shows promise in automating RA damage detection using methods like CNNs and vision transformers.

The widely used SvdH score quantifies RA-related joint damage based on radiographs. AI techniques, including deep learning, aim to automate SvdH scoring due to its labor-intensive and variable nature. Interpretability in AI models is essential for trust and effective decision-making in radiology practice, prompting the development of interpretability methods like visualization and semantics.

Studies in RA diagnosis have used SVMs and CNNs, while others focused on predicting SvdH scores for individual joints. Vision transformers have shown promise in natural imaging but not yet in predicting joint damage SvdH scores. Multi-task learning enhances performance by leveraging shared knowledge between related tasks in medical imaging.

In conclusion, deep learning methods offer non-invasive and accurate RA evaluation, but further research is needed for larger and diverse patient populations to ensure clinical usefulness and safe implementation in healthcare settings.

4.2 Methods

4.2.1 Study design & data sources

This study included three independent cohorts for model development and validation. The Radiological Society of North America (RSNA) 2017 Pediatric BoneAge Challenge dataset (Halabi et al., 2019) was used to develop the joint detection model. The dataset contains 14,236 hand skeletal images from children with mean skeletal age 129 months (range 0 to 216 months; 47% female). There was no description whether the children in the dataset were healthy or unhealthy as the dataset was developed to promote machine learning using X-ray hand images. The joints were annotated in https://github.com/razorx89/rsna-boneage-ossification-roi-detection from the RSNA Pediatric BoneAge Challenge dataset. The dataset contains 240 training images in total and 89 evaluation images annotated with regions of interest including: distal interphalangeal joint (DIP), proximal interphalangeal joint (PIP), and metacarpophalangeal joint (MCP) Wrist, Radius, and Ulna. For our purpose, since RA does not usually affect DIPs, we excluded these joints. We included only PIP, MCP, wrist, radius, and ulna thus each image contained 5 PIP joints, 5 MCP joints, 1 wrist, 1 radius, and 1 ulna joint.

COCO (Common Objects in Context) is a popular dataset in computer vision research and object identification applications (He et al., 2015). It is intended to provide a broad set of labelled images for training and evaluating object identification, segmentation, and captioning algorithms. More than 200,000 photos are labelled with object category labels, bounding box coordinates, and pixel-level segmentation masks for distinct objects in the COCO dataset. The collection contains 80 main kinds of things, including people, animals, cars, and ordinary items collected in a variety of scenarios and environments. The COCO dataset has become a standard for assessing the effectiveness of object

detection and segmentation algorithms, resulting in important advances in the field of computer vision. The joint detection model was pre-trained on the COCO dataset.

The joint detection model was validated using radiographic images from 4 adult RA patients followed in the Manitoba Early Arthritis Cohort (EAC). A total of 54 images that included both hands were obtained from 4 patients (2 females, 2 males; aged 46-70 years with < 12 months of symptoms at baseline, all seropositive for rheumatoid factor and/or anti-cyclical citrullinated protein antibodies, followed for 9-13 years with collection of clinical data including serial radiographs; 10-12 images per patient). Each image was manually annotated (CH) by an experienced RA clinician to identify the individual target joints. For the joint score prediction stage, we used data from the Canadian Early Arthritis Cohort (CATCH) (Bykerk et al., 2012). In brief, CATCH is a multicenter cohort of early arthritis patients with < 12 months symptoms, ≥2 swollen joints or 1 swollen MCP or PIP joint, plus at least one RA related disease factor at baseline. Participants are managed as part of routine care following standard treatment guidelines and detailed clinical data and radiographs are collected according to a study protocol. We used data from 530 patients with 2249 hand radiograph images of whom 381 had paired Svdh scores. As previously reported, sequential-order Svdh scores were assigned by one trained reader (intra-class correlation coefficient for intra-observer variability 0.74) (D. Van Der Heijde, 2000). Significant radiographic progression was defined as an annual increase above the smallest detectible change (SDC), which was 3.5 units/year in this cohort (Bruynesteyn et al., 2002). Radiographs performed 3–6 months prior to study entry were not repeated but were used as baseline films.

Table 4.1 summarizes the demographic and clinical features of the CATCH patients (Barnabe et al., 2015). For each patient, multiple hand radiographs were at different follow-up periods within in a fixed observation window of 0 to 36 months. Of these hand radiographs, there were 1298 images with

both hands present in the same image and 951 images with left and right hand captured on separate images. SvdH. The target hand joints were 5 PIPs/IP, 5 MCPs, 1 radius, 1 ulna, and 1 wrist (total 26 joints in both hands).

Table 4.1 Demographics and clinical features of CATCH subjects with SvdH scores

	All N=381	Only ERA N=360
Age (mean (SD))	54.2 (15.2)	54.4 (15.3)
Sex Female (n%)	259 (70.6%)	258 (71.7%)
Ethnic group white (n%)	344 (90.3%)	325 (90.3%)
Smoking status (n%)		
Never	162 (42.3)	150 (41.7%)
Current	63 (16.5%)	61 (16.9%)
Past	155 (40.7%)	147 (40.8%)
Education high school degree or above (n%)	167 (46.6%) Missing 23	158 (46.6) Missing 21
Income < 50 K/yr (n %)	116 (30.4%) Missing /not answered 191	114 (31.8%0 Missing /not answered 175
BMI (mean SD)	27/3 (5.8)	27.0 (5.7)
BMI category (n %)		
Obese I	52 (13.6%)	49 (13.6%)
Obese II	17 (4.5 %)	15 (4.2%)
Obese III	10 (2.6%)	9 (2.5%)
Overweight	87 (22.8%)	84 (23.3%)
Normal	107 (28.1%)	105 (29.4%)
Underweight	10 (2.6%)	10 (2.8%)
Missing	98 (25.7%)	88 (24.4%)
Osteoporosis (n%)	20 (5.2) missing 4	18 (5.1%) missing 4
(Rheumatic disease comorbidity index) median (range;25%, 75%)	1 (0-7; 0.2)	1 (0-7; 0,2)
Symptom duration at baseline mean (SD) days	162.8 (90.0)	164.0 (90.0)
RF Positive (n%)	180 (47.2%)	178 (49.4%)
ACPA Positive (n%)	205 (53.8%) missing 42	197 (61.4%) missing 39
Baseline DAS28 ESR (or CRP is missing ESR) median (IQR)	5.16 (1.57)	5.18 (1.57)
Physician global (1-10) (median IQR)	4.87 (2.63)	4.91 (2.62)
Patient global (1-10) (median IQR)	5.81 (2.95)	5.84 (2.93)
HAQ-DI baseline (0-3) (median IQR)	0.98 (0.68)	1.00(0.69)
Oral steroids in the first 3 months of follow-up	128 (33.6%)	119 (33.1%)
First RA treatment in the first 3 months (n%)		
Biologics,	7 (1.8%)	6 (1.7%)
MTX mono,	155 (40.7%)	147 (40.8%)
MTX triple,	36 (9.4%)	35 (9.7%)
MTX combo,	109 (28.6%)	105 (29.2%)
only non-biologics/non-MTX	50 (13.1%)	48 (13.3%)
none	24 (6.3%)	19 (5.3%)

Table 4.2 CATCH data stratified split details. Number of samples in each score class

Data Split	Images in each score class					
	0	**1`**	**2**	**3**	**4**	**5**
Train Set (60%)	21906	726	289	39	11	21
Validation Set (10%)	2435	81	32	4	1	2
Test Set (30%)	10433	346	138	19	5	9
Total images	34,774	1,153	459	62	17	32

4.3 Deep learning pipeline

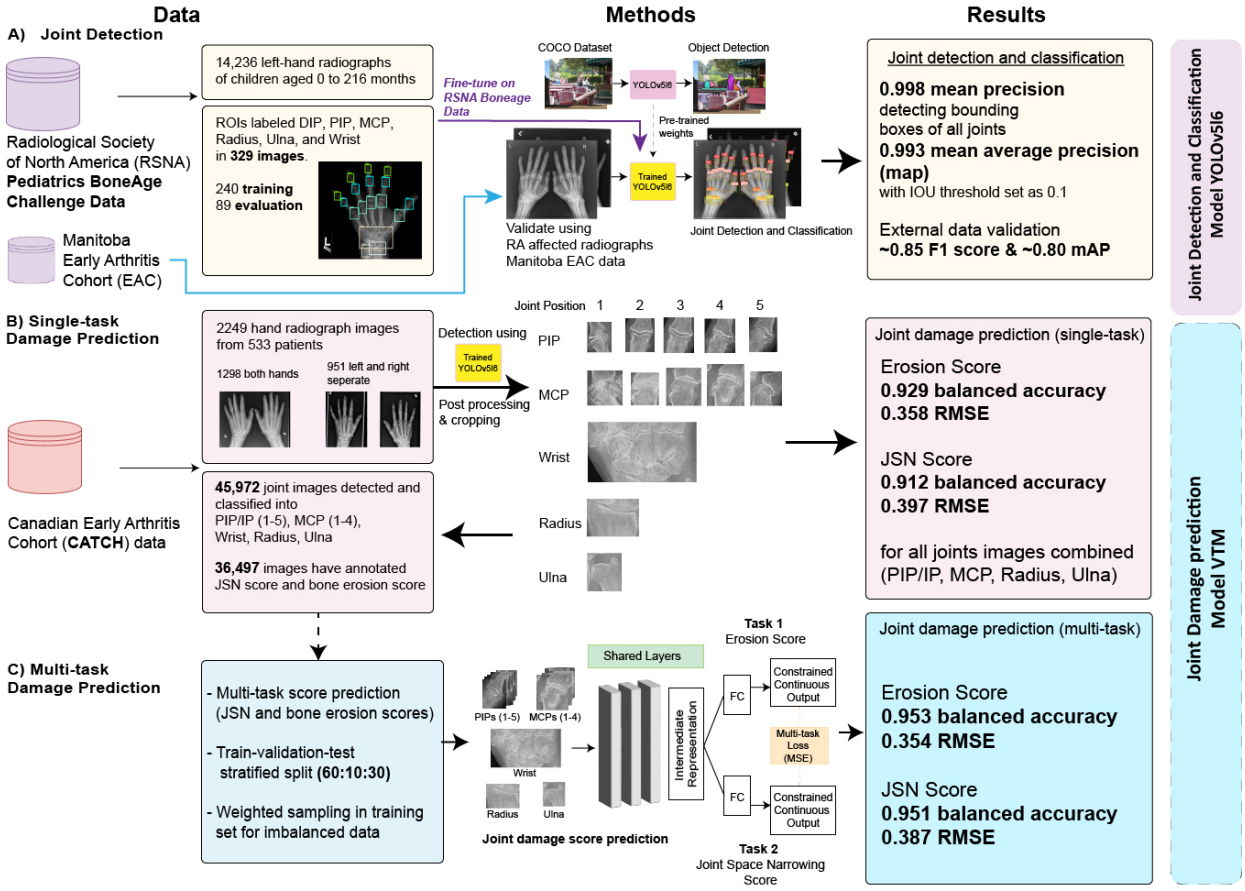

Figure 4.1 Overview of joint detection and joint damage prediction study design. We utilized two datasets, RSNA Boneage data and CATCH data. **A) Joint Detection** We used the Boneage data to train YOLOv5l6 object detection using transfer learning. The fine-tuned model achieved a high precision in detecting joints on the RSNA test data. The joint detection model was also externally validated on expert annotated Manitoba EAC data. **B) Single-task Damage Prediction** This trained model was then used to detect and classify all joints in CATCH data hand radiographs and DL models were trained to detect JSN score and Erosion scores. **C) Multi-task Damage Prediction** Finally, a multi-task approach was used to train a deep learning model to predict individual joint JSN and bone erosion scores together. Weighted sampling stratified split and data augmentations were used to highly mitigate class-imbalanced data.

We conducted our experiment in two stages. Firstly, the joint detection stage (**Figure 4.1A**), where we used YOLOv5l6 to detect PIP/IP, MCP, wrist, radius, and ulna labelled joints and removed the DIP labels as they are less relevant for rheumatoid arthritis. We started the training from a YOLOv5l6 model that was pretrained on the COCO dataset (**Figure 4.2A**). The architecture of YOLOv5l6 is illustrated in **(Figure 4.3)**. We fine-tuned the model on the RSNA Pediatric BoneAge challenge data.

We added several augmentations to the images before feeding them into YOLOv5l6 model including mixup (H. Zhang et al., 2017) (mixing up features of different images together into one), mosaic (Bochkovskiy et al., 2020) (combines 4 training images into 1 image), rotation, translation, scaling, and shearing augmentations. The final joint detection model was externally validated using 54 radiographs from 4 patient from the Manitoba EAC data. The Manitoba EAC data was manually annotated with bounding boxes for each joint in each radiograph by an expert.

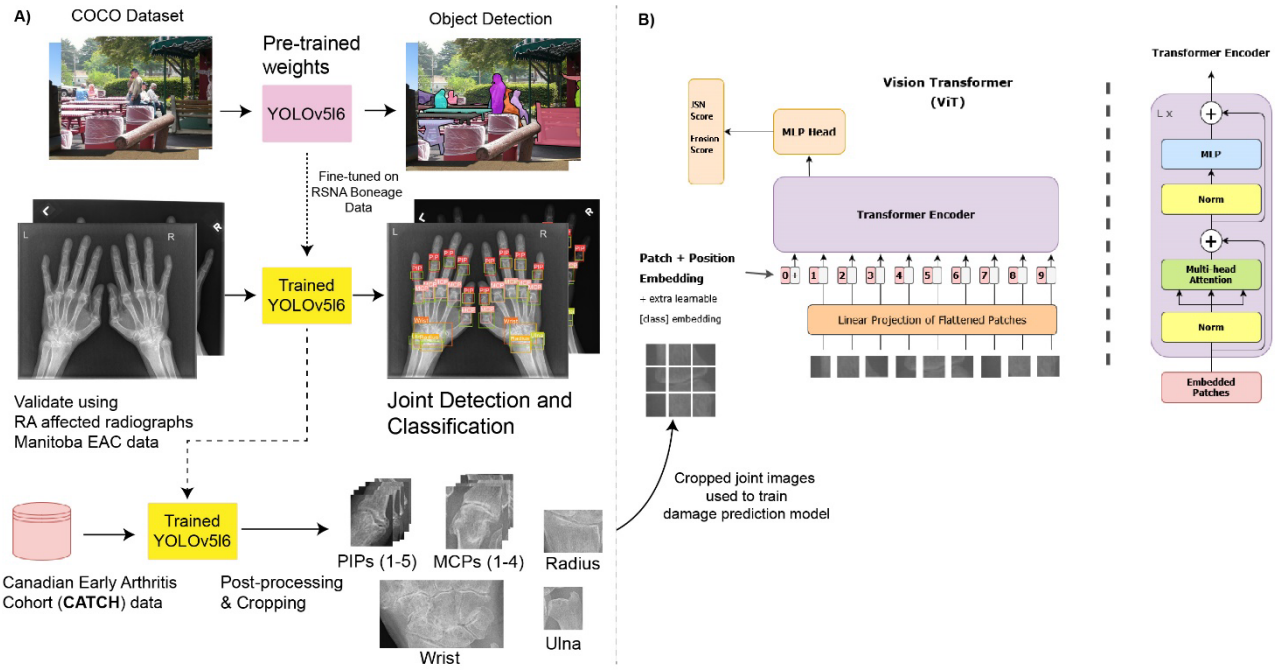

Figure 4.2 Overview of joint detection and damage prediction model training A) Transfer learning in joint detection and classification. The pretrained YOLOv5l6 model is further trained on RSNA BoneAge data to detect and classify joints. The model was then validated using Manitoba EAC data. This model was then used to detect and classify joints in all hand radiographs of the CATCH data. The model detects five types of joints and labels them (MCP 1-5, PIP/IP 1-5, Radius, Ulna, Wrist). **B) Architecture of vision transformer model** (Kolesnikov et al., 2021). Joint images (PIP/IP, MCP, Wrist, Radius, Ulna) were used to train the vision transformer. During training weighted sampling was used based on relative class weight, the class distribution was highly skewed therefore this was required to mitigate class imbalance. The MLP head consists of a fully connected multi-layer-perceptron (MLP) made of FC-relu-dropout-FC-relu-dropout-FC. Followed by a sigmoid transformation to output a constrained output. FC is fully connected layer; relu is rectified linear unit.

We then used this trained joint detection model to identify and classify joints of all patients' hand radiograph images in the CATCH data. The identified joints were the PIPs/IP, MCPs, wrist, radius and

ulna **Figure 4.5A**. We have two types of hand radiographs of patients in the CATCH data, one type is where both the left and right hand are captured in one image file. The other type is where the left hand and right-hand images of the patient exist as separate image files. The joint detection model can detect and classify joints in both the types of images without requiring further modifications. In the first case it detects 26 joints from one image, in the second case it identifies 13 joints from each image. The joint detection model also classifies each detected joint into either PIP/IP, MCP, Wrist, Radius or Ulna. The trained joint detection model successfully detected and classified 45,972 joint images from the 2249 hand radiographs available in the CATCH data. However, only 36,497 detected joints had an annotated SvdH score (both erosion and JSN) available in the dataset. The rest of the 9475 images had no score values, thus we did not use them in the score prediction stage. To label the joint positions of the PIP/IP and MCP joints we applied a co-ordinate based post-processing algorithm to label the positions (1-5) for the PIP/IP and MCP joints identified in each hand. For the detected bounding boxes in a hand radiograph, we sorted the boxes based on their relative coordinates in the image and assigned a label (1-5) based on their position relative to each other.

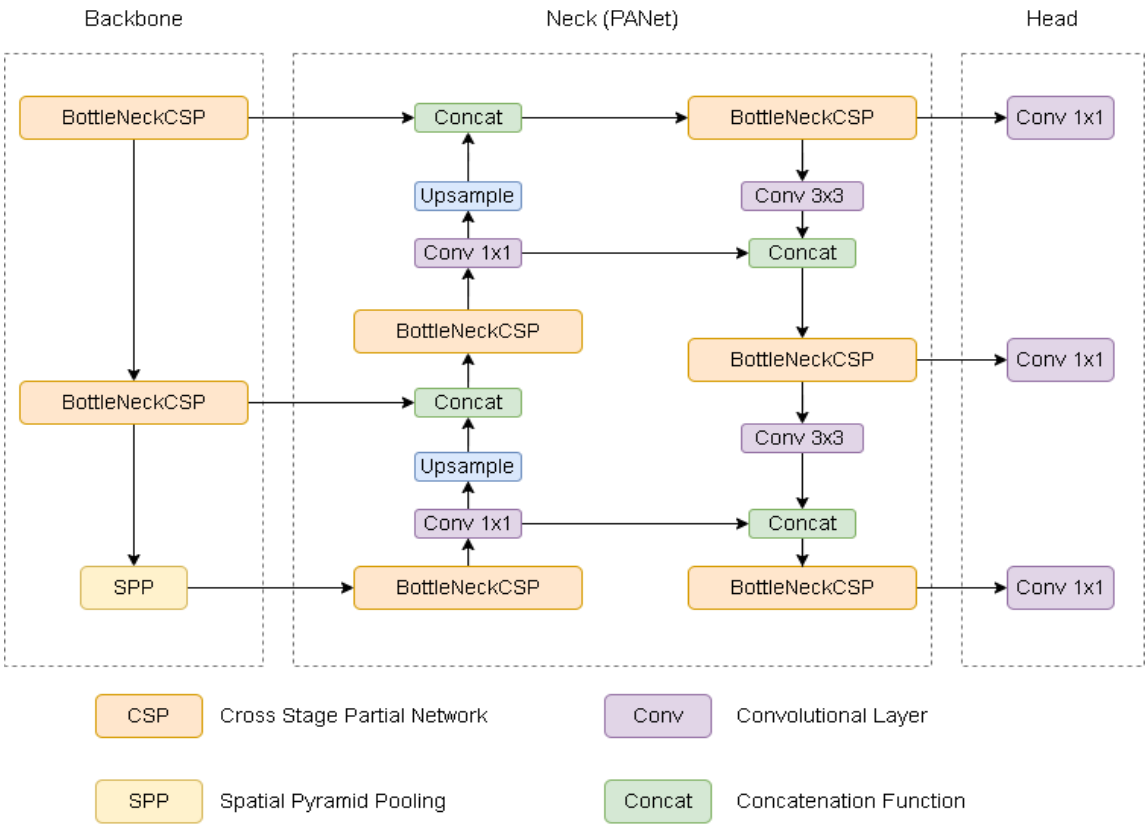

Figure 4.3 Architecture of YOLOv5. It consists of three parts: (1) Backbone: CSP, (2) Neck: PANet, and (3) Head: Yolo Layer. The data are first input to CSP for feature extraction, and then fed to PANet for feature fusion. Finally, YOLO Layer outputs detection results (class, score, location, size).

Each joint in the patients' hands had a joint space narrowing (JSN) score and the bone erosion score. Succinctly, the JSN score is defined as follows: 0, no JSN; 1, focal or doubtful; 2, generalized, >50% of the original joint space left; 3, generalized, <50% of the original joint space left or subluxation; and 4, bony ankylosis or complete luxation. The erosion score is defined as follows: 0, no erosion; 1, discrete; 2, larger, <50% of the joint surface; 3, extending over the middle of the bone; and 5, complete collapse (Hirano et al., 2019; D. Van Der Heijde, 2000).

We treated the score prediction as a multi-task constrained regression problem. The constraint was applied to ensure that the model outputs a continuous value but in a fixed range. In the case of the SvdH score prediction, the range was (0-5) which is the valid range for SvdH scores. In a study they identified that treating predicted scores as a regression problem improves accuracy when training on

small datasets (Honda et al 2022) but from a clinician's perspective, joint space narrowing (JSN) scores of three and four cannot be expressed as continuous values because they are defined by the presence or absence of JSN or dislocation of joints. To ensure the comparison and reliability of our model's output we also reported the classification metrics on the CATCH test set.

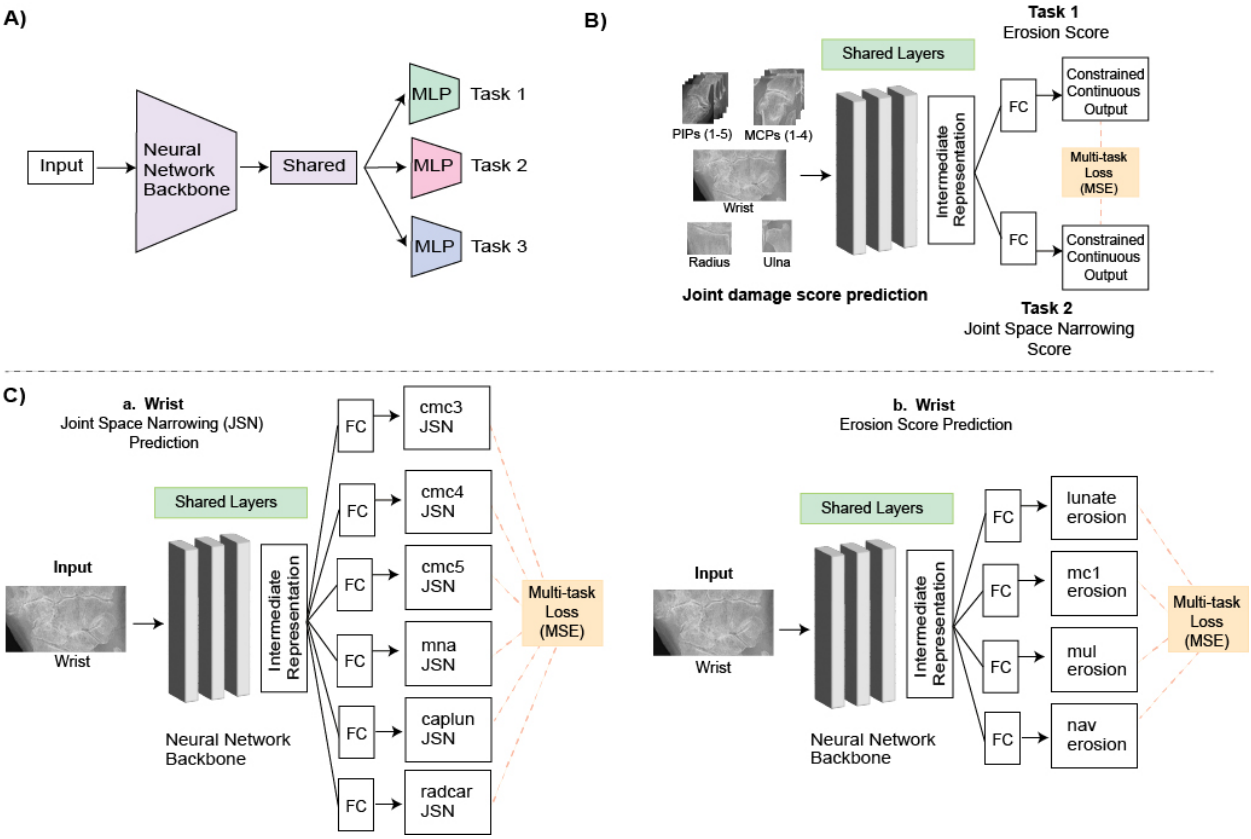

Figure 4.4 Multi-task joint score prediction. We used all detected joint images to train a multi-task model. Each joint had two annotated score values (erosion and JSN scores). **A)** We utilized a hard parameter sharing paradigm to train our multi-task models where the shared neural network learns on training images and the shared latent representation is connected to multiple outputs **B)** We trained the model using extracted joint images to predict JSN score and bone erosion scores in a supervised approach. The multi-task loss is a sum of MSE loss for both JSN and erosion scores. **C)** Wrist images were used to train a multi-task model. (a) Each wrist image has four associated joints in the image which have bone erosion scores namely (lunate, mc1, mul, nav) (b) Six joints with the labeled joint space narrowing scores for the joints namely (cmc3, cmc4, cmc5, mna, caplun, radcar). We trained separate models to predict the four erosion scores of the wrist and six JSN scores of wrist joints. MSE is Mean Squared Error; JSN Joint Space Narrowing.

We have corresponding annotated score values for the following joints (PIP/IP, MCP, Radius, Ulna) which we used as the labels when training. However, for the wrist joint, our joint detection model was trained to detect the whole wrist, which contains the specific individual wrist joints. We have corresponding scores in our dataset for the following joints in the wrist (mc1, mul, nav, lunate) for bone erosion and (cmc3, cmc4, cmc5, mna, caplun, radcar) for joint space narrowing. Since a direct correlation could not be made between a wrist image and the individual joint scores, we chose to train a separate multi-task model to detect wrist joint scores from whole wrist images **Figure 4.4C**.

To evaluate baseline models, we also trained an ordered logit regression model. An ordered logit regression model is a regression model for ordinal dependent variables (McCullagh, 1980); it can be thought of as an extension of logistic regression model. We used statistical features extracted from individual joint images to train the regression model to predict SvdH scores. The statistical features extracted were first and second order statistical features (FSStat) see **Table 3.2** for details.

4.4 Model training and evaluation

We used data augmentation, gradual unfreezing, and fine-tuning of the pretrained model for training the dataset. Data augmentations in training sets consisted of affine transformations only. Weighted sampling with replacement was used in generating training batches. Each sample in the training set was assigned a weight based on the relative class weight (determined by the number of samples available in that score class). During training, the mini batches were generated by sampling from the training set according to the assigned weight to ensure that the model sees samples from all the score classes a similar number of times. The training, validation and test set was also stratified split (see **Table 4.2**). This ensures that each split contains a consistent percentage of samples from each score class. We also used a 60:10:30 train-validation-test split at the patient-level for these experiments so that we had a larger test set to evaluate and ensuring that images from the same patients weren't

included in multiple splits. **Table 4.2** contains split details and number of samples in each score class. Using the same data-split, augmentation, and sampling strategies we trained recent CNN architectures EfficientNetV2 (Tan & Le, 2021) and MobilenetV3 (Howard et al., 2019) to compare their performance with the vision transformer model in joint score prediction.

We evaluated both our joint detection model and score prediction model separately using appropriate evaluation metrics. For the joint detection model, we evaluated the model's performance using mAP (mean absolute precision) which is a standard metric used in the literature for object detection models. To evaluate classification performance, we calculated the precision, recall and F1 score. For the score prediction model, we used balanced accuracy, Root Mean Squared Error (RMSE), and F1 Score to evaluate model performance.

4.4.1 Joint detection metrics

We used the following performance metrics to evaluate the joint detection model. Intersection over union (IOU) metric was used to evaluate the performance of the joint detection. IOU is a method that looks at the overlapping area of the predicted output and the ground truth mask. The equation for IOU is:

$$\text{IoU} = \frac{predicted\ output \cap ground\ truth\ mask}{predicted\ output \cup groud\ truth\ mask} \tag{8}$$

IOU measures the overlap between 2 bounding boxes. We used IOU to measure how much our predicted box overlaps with the ground truth (the real object box). We predefined an IOU threshold in classifying whether the prediction is a true positive (TP) or a false positive (FP) **Figure 4.5B**.

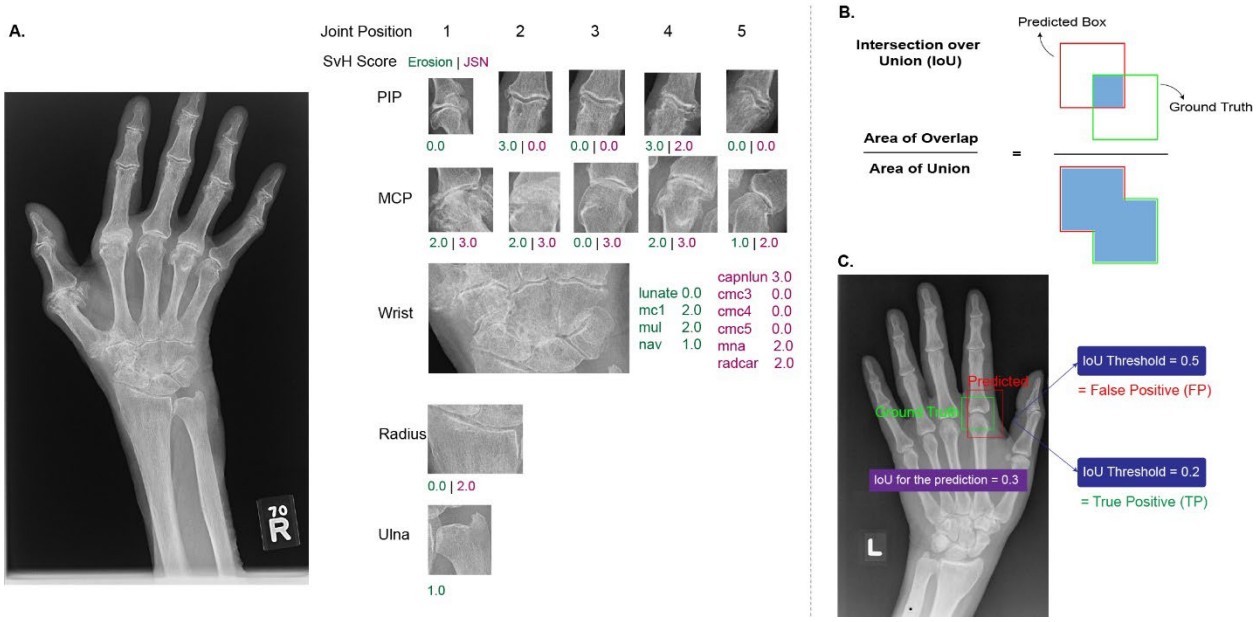

Figure 4.5 Example from the joint detection model & IOU calculation for joint prediction. (A) Joint position and type are classified by the model. Below each joint image is the associated Erosion score (in green) and Joint space narrowing score (in red). **(B)** IOU calculated using the area of overlap between two boxes divided by the area of union of the two boxes. **(C)** If the IOU for a prediction is 0.3 and IOU threshold is 0.2 then this prediction will be classified as a true positive (TP) however if the IOU threshold was 0.5 then this same prediction will be classified as false positive (FP). This means changing the IOU threshold can result in different binary outcomes for the same prediction.

Average Precision was calculated as the weighted mean of precisions at each threshold; the weight is the increase in recall from the prior threshold. Mean Average Precision (mAP) is the average of Average Precision (AP) of each class. The general definition for the AP is finding the area under the precision-recall curve. AP is the precision averaged over all classes. The mAP with IOU threshold set at 0.1 was reported as map0.1. AP was calculated for all IOU thresholds (0.1-0.55) at 0.05 intervals. The AP was averaged over all classes and 10 IOU thresholds reported as map.

4.4.2 Score prediction metrics

We used the following performance metrics to evaluate the multi-classification performance of the joint score prediction models. Both accuracy and macro-averaged F1-score were reported. Averaged F1-score is more robust to a class-imbalanced evaluation set when compared to accuracy.

Accuracy (global): $\frac{TP+TN}{TP+FP+TN+FN}$ (all score classes combined) (9)

F1 score (average) $2 . \frac{precision * recall}{precision+recall}$ (calculated for each class and averaged) (10)

Root mean squared error (RMSE) $\sqrt{\left(y_{true} - y_{pred}\right)^2}$ (all score classes combined) (11)

4.5 Results

The three data sources spanned a variety of patient populations and imaging protocols. The RSNA Boneage dataset was exclusively radiographs of children's hand skeleton with skeletal age between 0 to 216 months, it was not mentioned if the images were from healthy patients or not. The CATCH dataset included patients who were newly diagnosed with inflammatory rheumatoid arthritis and followed up for up to 36 months. With respect to demographic variables, patients in RSNA dataset had a mean age of about 127 months, slightly skewed towards male (53%). CATCH data represent an older population with a mean age above 50 years (SD ~15), predominantly female (~70%), and (where this information was available) majority White (90%). The external dataset EAC included serial radiographs (10-12 per participant, total 43 images with both hands) from 4 participants (2 females, 2 males; baseline age 46-70 years; all seropositive for rheumatoid factor and/or anti-cyclical citrullinated protein antibodies, followed for 9-13 years).

4.5.1 Joint detection and classification

The joint detection model attained a mean 0.998 mean precision for detecting joints and a mean average precision (map) of 0.993 with IOU threshold set as 0.1 for detecting bounding boxes of all joints on the internal test set (RSNA; n=89). Mean precision in detecting bounding boxes for joint detection refers to the average accuracy of a model in correctly localizing objects within an image using bounding boxes. It measures the precision or correctness of the predicted bounding boxes compared to the ground truth bounding boxes. A higher mean precision indicates that the model is more accurate in identifying and localizing objects within the image. Ablation studies showed that the use of transfer learning in the pipeline improved the mean precision by 0.050 compared to the use of COCO pre-trained weights **Table 4.3**. On the independent external test set (EAC; n=54), the model achieved a mean precision of 0.923 and map 0.871 when detecting bounding boxes of all joints **Table**

4.4.

Table 4.3 Joint detection model metrics on internal test data (RSNA). Results with and without pre-training.

Pre-trained on COCO						
	Images	Number targets	Mean precision	Mean recall	map0.1	map
All	89	1157	0.998	0.983	0.993	0.854
PIP	89	445	1	0.987	0.994	0.772
MCP	89	445	0.993	0.989	0.994	0.624
Wrist	89	89	1	0.986	0.994	0.958
Radius	89	89	1	0.976	0.994	0.983
Ulna	89	89	0.997	0.978	0.988	0.933
Not pre-trained						
All	89	1157	0.948	0.841	0.919	0.81
PIP	89	445	0.883	0.978	0.984	0.814
MCP	89	445	0.986	0.957	0.977	0.7
Wrist	89	89	0.87	0.978	0.987	0.932
Radius	89	89	1	0.955	0.977	0.957

Table 4.4 Joint mean model metrics on external test data (Manitoba EAC patients)

	RSNA radiographs pre-trained on COCO						
	Images	Number targets	F1 score	Mean precision	Mean recall	map0.1	map
All	89	1157	0.991	0.998	0.983	0.993	0.854
PIP	89	445	0.994	1	0.987	0.994	0.772
MCP	89	445	0.991	0.993	0.989	0.994	0.624
Wrist	89	89	0.993	1	0.986	0.994	0.958
Radius	89	89	0.998	1	0.976	0.994	0.983
Ulna	89	89	0.987	0.997	0.978	0.988	0.933
	Adult Rheumatoid Arthritis radiographs pre-trained on COCO						
All	54	1354	0.812	0.923	0.725	0.871	0.627
PIP	54	518	0.880	0.893	0.869	0.823	0.615
MCP	54	521	0.860	0.939	0.794	0.814	0.601
Wrist	54	102	0.883	0.965	0.814	0.886	0.876
Radius	54	109	0.821	0.909	0.750	0.828	0.661

4.5.2 Joint damage score prediction performance

The score prediction model was trained to predict both erosion and JSN scores from joint images. The vision transformer-based model attained a high accuracy of 0.91 and 0.93 for JSN and erosion scores respectively. The multi-task models trained to detect SvdH scores of individual joints in the wrist attained moderately high accuracy (0.6-0.91) for predicting erosion scores and JSN scores. However, for these models, the CNN-based EfficientNet model performed better than vision transformer-based models in detecting both erosion and JSN in wrist joints, with an average difference of 0.10 between the two models **Figure 4.6**.

Table 4.5 Performance on test set for multi-task models trained on wrist images. Predicting erosion score for 4 joints in wrist (lunate, mc1, mul, nav).

Model | Metrics	**EfficientNetV2***			ViT			MobileNetV3		
	Accuracy	RMSE	F1	Accuracy	RMSE	F1	Accuracy	RMSE	F1
lunate	0.8544	0.4255	0.9113	0.7533	0.9706	0.7592	0.6533	1.1404	0.7799
mc1	0.8556	0.4679	0.9071	0.7095	1.0911	0.7229	0.6095	1.0686	0.7361
mul	0.8178	0.5056	0.8831	0.7598	1.5606	0.6930	0.5598	1.5507	0.6986
nav	0.8604	0.4459	0.9049	0.7402	1.0736	0.7145	0.6402	1.0911	0.7596

Table 4.6 Performance on test set for multi-task model trained on wrist images. Predicting JSN score for 6 joints in wrist (cmc5, cmc4, cmc3 ,mna ,capnlun, radcar)

Model | Metrics	**EfficientNetV2***			ViT			MobileNetV3		
	Accuracy	RMSE	F1	Accuracy	RMSE	F1	Accuracy	RMSE	F1
cmc3	0.824	0.465	0.911	0.756	0.980	0.759	0.637	1.190	0.779
cmc4	0.835	0.647	0.927	0.719	1.281	0.722	0.619	1.258	0.736
cmc5	0.807	0.825	0.863	0.729	1.66	0.693	0.539	1.550	0.698
mna	0.760	0.445	0.874	0.69	1.073	0.714	0.590	1.091	0.759
capnlun	0.735	0.505	0.783	0.619	1.560	0.593	0.519	1.550	0.498
radcar	0.704	0.445	0.824	0.729	1.073	0.614	0.569	1.091	0.559

When compared to their non-pretrained counterparts, all the models performed better when transfer learning was employed. The models were started with pre-trained weights trained on ImageNet (Deng et al., 2009). They were later fine-tuned on the CATCH data and evaluated on the CATCH test set. The models were also trained without multi-task strategy, where separate models were trained to predict erosion and JSN scores. The results were comparable to the results attained by multi-task strategy training.

To understand if the model was better at predicting the extent of damage for certain joint types. We calculated the metrics for each joint type separately. For MCP and PIP joints the model consistently performed well with decreasing performance for radius, ulna, and wrist joints **Table 4.7**.

Table 4.7 Performance on CATCH test set for multi-task model trained on different joint type images.

Model	Joints trained on	Accuracy Erosion	F1 Erosion	RMSE Erosion	AUC Erosion	Accuracy JSN	F1 JSN	RMSE JSN	AUC JSN
ViT	All joints except wrist	0.953	0.956	0.354	0.893	0.951	0.931	0.387	0.876
ViT	MCP	0.935	0.922	0.421	0.910	0.936	0.921	0.624	0.905
ViT	PIP	0.944	0.971	0.376	0.892	0.940	0.969	0.444	0.912
ViT	Radius	0.871	0.911	0.584	0.852	0.848	0.901	0.751	0.812
ViT	Ulna	0.858	0.880	0.611	0.834	0.871	0.887	0.774	0.852

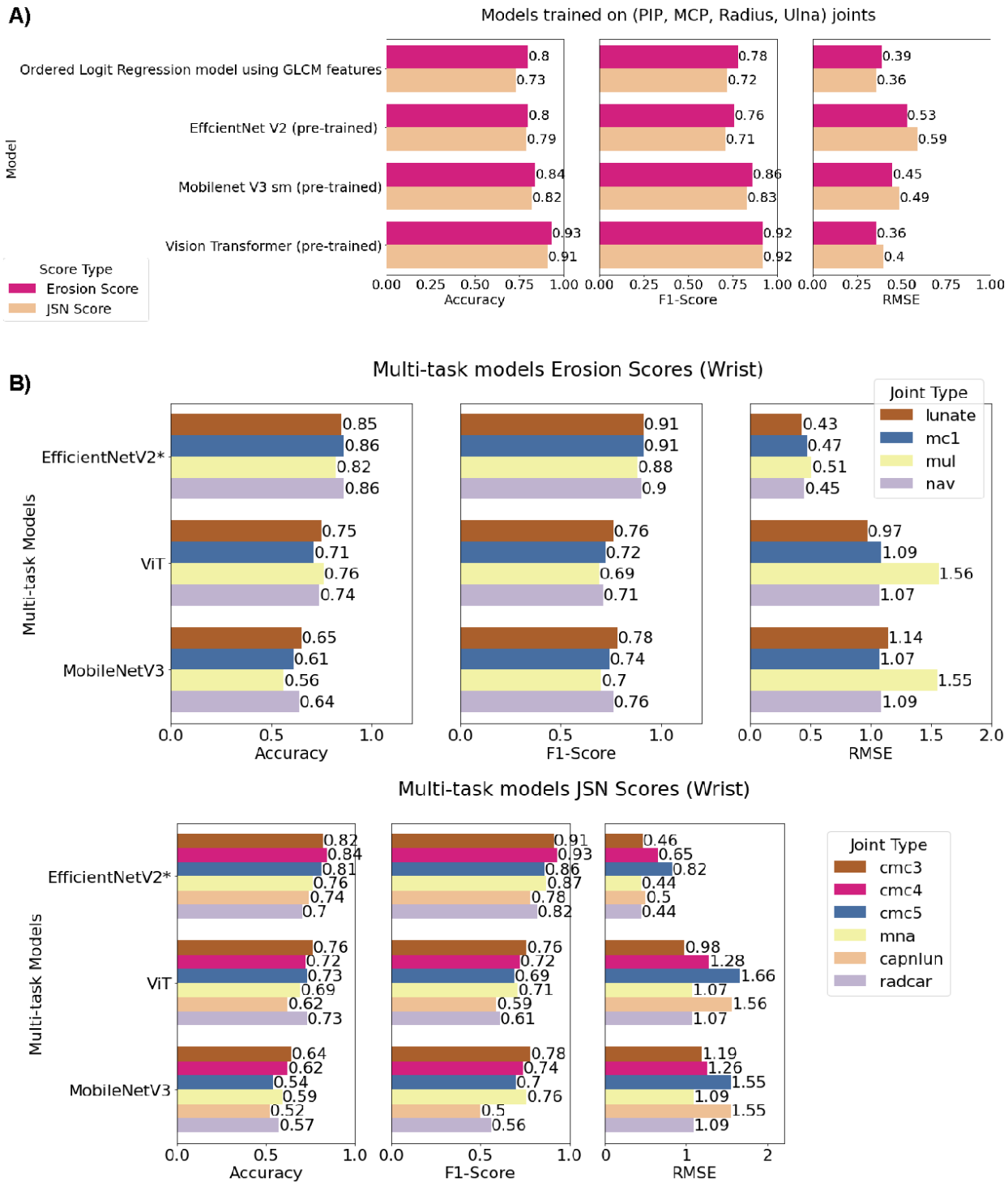

Figure 4.6 Results of SvdH score prediction models. A) Models were trained on all joints except wrist. The metrics reported were balanced accuracy, F1-Score and RMSE predicting the two scores (JSN and Erosion). **B)** Models were trained on wrist joints. Separate models trained to detect the erosion score and JSN scores.

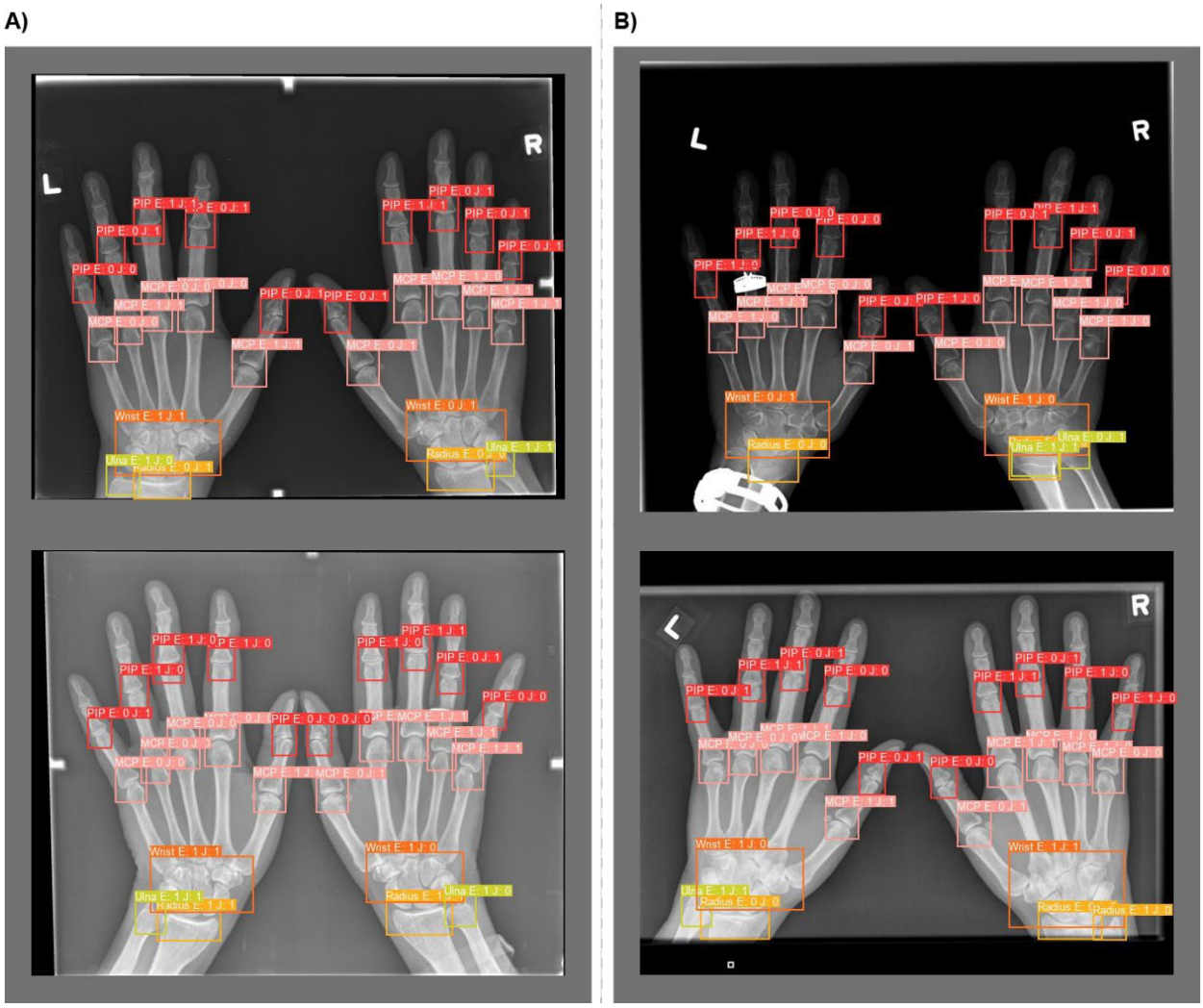

Figure 4.7 Positive and negative examples of full model A) All the joints were detected and classified correctly. The joint scores were also correctly predicted **B)** In the first image, the ulna joint was not detected for the left hand but detected twice for the right hand. In the second image, both radius and ulna of the right hand were classified as radius by the model.

Table 4.8 Performance on test set for ordered regression model and CNN models trained on all joint images.

Model	Trained on Joint type & Score type	Balanced Accuracy (global)	Test F1 (average)	Test RMSE (global)
Ordered Logit Regression model. (ML & GLCM features)	(MCP, PIP, Radius, Ulna) **Erosion score**	0.796	0.776	0.386
EffcientNet V2 (pre-trained) (DL)	(MCP, PIP, Radius, Ulna) **Erosion Score**	0.804	0.756	0.532
Mobilenet_v3 sm (pre-trained) (DL)	(MCP, PIP, Radius, Ulna) **Erosion Score**	0.836	0.856	0.453
Vision Transformer (Vit_Base_16) (pre-trained) (DL)	(MCP, PIP, Radius, Ulna) **Erosion Score**	0.929	0.924	0.358
Ordered Logit Regression model. (ML & GLCM features)	(MCP, PIP, Radius, Ulna) **JSN score**	0.731	0.725	0.363

EffcientNet V2 (pre-trained) (DL)	(MCP, PIP, Radius, Ulna) **JSN Score**	0.794	0.712	0.592
Mobilenet_v3 sm (pre-trained) (DL)	(MCP, PIP, Radius, Ulna) **JSN Score**	0.825	0.832	0.489
Vision Transformer (Vit_Base_16) (pre-trained) (DL)	(MCP, PIP, Radius, Ulna) **JSN Score**	0.912	0.922	0.397

4.6 Discussion

We describe a novel automated approach to assessing joint damage in hand radiographs from patients with rheumatoid arthritis that can also accurately determine SvdH erosion and joint space narrowing scores. We first showed that pre-trained object detection models can detect joints (PIP/IP, MCP, wrist, ulna, radius) of rheumatoid arthritis patients, even when the model was trained on joints of pediatric patients. Our model was able to achieve high performance on the test set with F1 score of more than 0.9 for all the joint types. We then trained a vision transformer-based model on extracted joint images to classify their JSN and erosion scores in a supervised approach. The data was highly class imbalanced, so we had to utilize weighted sampling during training and stratified train-test-split. The vision transformer-based model outperformed CNN based models in classifying erosion and JSN scores for joints (PIP/IP, MCP, Radius, Ulna). The CNN based EfficientNet model performed better when predicting scores of wrist joints.

Our study adds to literature in several areas. Firstly, we curate and develop a dataset of RA patient radiographs. The radiographs have been scored by the 'gold standard' expert clinicians. Secondly, we evaluate a vision transformer model on RA affected radiographs. This is the first paper in our knowledge exploring the performance of ViT based architectures evaluated on RA affected joints. From the results we can observe that ViT based model outperforms recent CNN based models when predicting RA damage. Although an exact comparison cannot be made between a recent publication as they used internal data to evaluate their models (Honda et al 2022). Our ViT outperform their CNN based damage prediction model when predicting damage of all joints in the hand. A strength of the study includes externally validating the joint detection model which was reported as a limitation in previous studies (Hirano et al., 2019; Honda, Yano, Tanaka, et al., 2022).

This study has some limitations. First, the models were trained on sampled patients from a broad range of disease durations. They were treated according to contemporary treatment strategies thus despite follow-up for up to 10 years, most and only mild or moderate joint damage. Evaluating the performance of joints with extensive damage would be useful. Secondly, the data was highly imbalanced where most joints did not have any damage. Although there were strategies employed to mitigate the imbalance in the dataset. Having a close to equal distribution of samples with extensive damage would improve the training of the deep learning models.

This opens new opportunities for the clinical adoption of such clinical decision support systems. One possible future work is that we could use the extracted joints, joint scoring model, and integrate the longitudinal information. This can enable us to determine if a rheumatoid arthritis patients' severity score will increase or decrease in the future. This application can help improve the efficiency of determining the usefulness of medicines for rheumatoid arthritis by predicting the severity of rheumatoid arthritis. This can aid in the early treatment and diagnosis of RA during its preliminary stages.

The development of an accessible and user-friendly online tool for joint detection and scoring of rheumatoid arthritis (RA) is driven by the need to address limited access to rheumatology specialists in vulnerable or remote areas, enable timely intervention and treatment, enhance diagnostic accuracy, and facilitate remote collaboration among healthcare professionals. By providing a standardized and automated approach to joint damage assessment, this tool can improve diagnostic accuracy, reduce inter-observer variability, and enable efficient remote consultations, ultimately improving patient outcomes in regions with limited resources.

Chapter 5 Conclusion and Future Direction

5.1 Conclusion

In conclusion, this thesis shows how deep learning can be applied to diabetic foot ulceration prediction and joint detection and joint damage prediction on rheumatoid arthritis data both healthy and RA-affected data. We could apply statistical models to the datasets; statistical models would be better than deep learning at identifying, elucidating, and establishing the significance of correlations between variables. However, because statistical models do not perform well with complex distributions, they would not be able to outperform deep learning algorithms. Deep learning models might be challenging to employ in the healthcare industry, despite their great performance. It is challenging to believe models with no explanation behind their predictions. Due of the numerous intricate causal relationships that exist in the healthcare industry, it might be challenging for deep learning to transfer information from one domain to another. However, research in improving explainability and interpretability of deep learning models can help overcome this gap. Where possible we try to visualize reasoning behind the deep learning model predictions to improve trust in the model and validate its clinical use.

In Chapter 3, we investigated two computer-aided diagnostic approaches for early detection of DFU from thermograms of the feet. First, based on a tree-based machine learning classifier xgboost trained using only image-level texture and temperature features (excluding demographic information) from 7 ROIs which are prone to get ulceration in foot regions selected from the thermal images, additionally use an image registration-based model to identify the key regions of interest from the thermograms automatically. Second, leverage a multi-view CNN model to input multiple images more specifically ipsilateral and contralateral feet images, and train end to end to predict the disease outcome using

learned features from both feet simultaneously. The patients are classified into normal and diabetic using the ML classifier and multi-view CNN model with an overall AUC of 0.97 and 0.94 respectively. The performance achieved by the proposed methods was better than the other methods reported in the literature.

In Chapter 4, we carried out joint detection on rheumatoid arthritis patients using YOLOv5l6. As there is a lack of dataset of rheumatoid arthritis patients, this poses a challenge for deep learning since deep learning requires a large amount of dataset to be able to achieve a high performance. Instead of using data from rheumatoid arthritis patients for joint detection, we used joints from oral patients to train YOLOv5l6 on, then used the trained YOLOv5l6 to detect joints on rheumatoid arthritis patients. Specifically, we predict PIP, MCP, wrist, radius, and ulna of the hands. We used RSNA 2017 Pediatric BoneAge Challenge for our training dataset of healthy joints and evaluated on the Manitoba dataset of joints with rheumatoid arthritis. We describe a novel automated approach to assessing joint damage in hand radiographs from patients with rheumatoid arthritis that can also accurately determine SvdH erosion and joint space narrowing scores. We trained a vision transformer-based model on extracted joint images to predict their JSN and erosion scores in a supervised approach. The data was highly class imbalanced, so we had to utilize weighted sampling during training and stratified train-test-split. The vision transformer-based model outperformed CNN based models in classifying erosion and JSN scores for joints (PIP/IP, MCP, Radius, Ulna). The CNN based EfficientNet model performed better when predicting scores of wrist joints.

5.2 Future Directions

As a future improvement we plan on implementing re-training/self-tuning capabilities in the tool used for joint detection and scoring of rheumatoid arthritis (RA) is to enable it to adapt to new data and continuously improve accuracy. RA is a complex disease with varying manifestations and progression patterns among patients (Q. Guo et al., 2018). As new data becomes available, such as additional labeled images or updated clinical guidelines, it is crucial to incorporate this information into the tool's algorithms to ensure accurate and up-to-date predictions.

Data drift refers to the phenomenon where the characteristics or distribution of the data used to train a deep learning model change over time. In other words, the data the model was originally trained on no longer accurately represents the data it encounters in real-world scenarios. Data drift can negatively impact the performance of machine learning and deep learning algorithms that were trained on historical data (Sahiner et al., 2023).

By implementing re-training/self-tuning capabilities, the tool can continuously learn and refine its algorithms based on new data. This adaptability allows it to capture evolving patterns and characteristics of RA joint damage, ultimately improving its accuracy in detection and scoring. As the tool encounters new cases and learns from them, it can update its models, fine-tune parameters, and optimize its performance to better reflect the nuances of RA progression and severity. Moreover, the implementation of re-training/self-tuning capabilities helps address the inherent variability and heterogeneity in RA. The disease presents differently among individuals, with variations in joint involvement, disease activity, and response to treatment (Roux-Lombard et al., 2001; Smolen et al., 2006). By allowing the tool to adapt and improve based on new data, it can better capture this heterogeneity and provide more accurate and personalized assessments. This enhances the tool's utility in clinical settings, where tailored treatment decisions and monitoring of RA progression require

precise and reliable joint detection and scoring. Ultimately, re-training/self-tuning capabilities empower the tool to continuously learn and improve and adjust to data drift, optimizing its accuracy and usefulness in supporting healthcare providers and improving patient care.

References

Alavi, A., Sibbald, R. G., Mayer, D., Goodman, L., Botros, M., Armstrong, D. G., Woo, K., Boeni, T., Ayello, E. A., & Kirsner, R. S. (2014). Diabetic foot ulcers: Part I. Pathophysiology and prevention. *Journal of the American Academy of Dermatology*, *70*(1), 1.e1-1.e18. https://doi.org/10.1016/J.JAAD.2013.06.055

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 1–9. https://doi.org/10.1186/S12911-020-01332-6/PEER-REVIEW

Armstrong, D. G., & Lavery, L. A. (1998). Diabetic foot ulcers: prevention, diagnosis and classification. *American Family Physician*, *57*(6), 1325-1332,1337-1338.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. *Medical Image Analysis*, *12*(1), 26. https://doi.org/10.1016/J.MEDIA.2007.06.004

Bagavathiappan, S., Philip, J., Jayakumar, T., Raj, B., Narayana, P., Rao, S., Varalakshmi, M., Diab, F., & Mohan, V. (2010). Correlation between Plantar Foot Temperature and Diabetic Neuropathy: A Case Study by Using an Infrared Thermal Imaging Technique. *Journal of Diabetes Science and Technology*, *4*(6). www.journalofdst.org

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2018). VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, *38*(8), 1788–1800. https://doi.org/10.1109/tmi.2019.2897538

Barnabe, C., Sun, Y., Boire, G., Hitchon, C. A., Haraoui, B., Thorne, J. C., Tin, D., Van Der Heijde, D., Curtis, J. R., Jamal, S., Pope, J. E., Keystone, E. C., Bartlett, S., Bykerk, V. P., Kraishi, M., Zummer, M., Baron, M., Bessette, L., Akhavan, P., … Lyddell, C. (2015). Heterogeneous Disease Trajectories Explain Variable Radiographic, Function and Quality of Life Outcomes in the Canadian Early Arthritis Cohort (CATCH). *PLOS ONE*, *10*(8), e0135327. https://doi.org/10.1371/JOURNAL.PONE.0135327

Bell, J. (2022). What is machine learning? *Machine Learning and the City: Applications in Architecture and Urban Design*, 209–216. https://doi.org/10.1007/978-3-319-18305-3_1/COVER

Bhangale, K. B., & Kothandaraman, M. (2022). Survey of Deep Learning Paradigms for Speech Processing. *Wireless Personal Communications 2022 125:2*, *125*(2), 1913–1949. https://doi.org/10.1007/S11277-022-09640-Y

Bruynesteyn, K., Van Der Heijde, D., Boers, M., Saudan, A., Peloso, P., Paulus, H., Houben, H., Griffiths, B., Edmonds, J., Bresnihan, B., Boonen, A., & Van Der Linden, S. (2002). Determination of the minimal clinically important difference in rheumatoid arthritis joint damage of the Sharp/van der Heijde and Larsen/Scott scoring methods by clinical experts and comparison with the smallest detectable difference. *Arthritis and Rheumatism*, *46*(4), 913–920. https://doi.org/10.1002/ART.10190

Bykerk, V. P., Jamal, S., Boire, G., Hitchon, C. A., Haraoui, B., Pope, J. E., Thorne, J. C., Sun, Y., & Keystone, E. C. (2012). The Canadian Early Arthritis Cohort (CATCH): patients with new-

onset synovitis meeting the 2010 ACR/EULAR classification criteria but not the 1987 ACR classification criteria present with less severe disease activity. *The Journal of Rheumatology*, *39*(11), 2071–2080. https://doi.org/10.3899/JRHEUM.120029

Caruana, R., Pratt, L., & Thrun, S. (1997). Multitask Learning. *Machine Learning 1997 28:1*, *28*(1), 41–75. https://doi.org/10.1023/A:1007379606734

Chen, T., & Guestrin, C. (n.d.). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563. https://doi.org/10.1111/1754-9485.13261

Crisologo, P. A., & Lavery, L. A. (2017). Remote home monitoring to identify and prevent diabetic foot ulceration. *Annals of Translational Medicine*, *5*(21), 973–980. https://doi.org/10.21037/ATM.2017.08.40

Cruz-Vega, I., Hernandez-Contreras, D., Peregrina-Barreto, H., Rangel-Magdaleno, J. de J., & Ramirez-Cortes, J. M. (2020). Deep learning classification for diabetic foot thermograms. *Sensors (Switzerland)*, *20*(6). https://doi.org/10.3390/s20061762

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dias, E. M., Lukas, C., Landewé, R., Fatenejad, S., & Van Der Heijde, D. (2008). Reliability and sensitivity to change of the Simple Erosion Narrowing Score compared with the Sharp–van der Heijde method for scoring radiographs in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, *67*(3), 375–379. https://doi.org/10.1136/ARD.2007.072785

Dong, N., Kampffmeyer, M., & Voiculescu, I. (2021). Self-supervised Multi-task Representation Learning for Sequential Medical Images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12977 LNAI*, 779–794. https://doi.org/10.1007/978-3-030-86523-8_47/TABLES/6

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. *Npj Digital Medicine 2021 4:1*, *4*(1), 1–9. https://doi.org/10.1038/s41746-020-00376-2

Filipe, V., Teixeira, P., & Teixeira, A. (2021). Measuring Plantar Temperature Changes in Thermal Images Using Basic Statistical Descriptors. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12953 LNCS*, 445–455. https://doi.org/10.1007/978-3-030-86976-2_30

Frykberg, R. G., Gordon, I. L., Reyzelman, A. M., Cazzell, S. M., Fitzgerald, R. H., Rothenberg, G. M., Bloom, J. D., Petersen, B. J., Linders, D. R., Nouvong, A., & Najafi, B. (2017). Feasibility and Efficacy of a Smart Mat Technology to Predict Development of Diabetic Plantar Ulcers. *Diabetes Care*, *40*(7), 973–980. https://doi.org/10.2337/DC16-2294

Gao, P., Shan, W., Guo, Y., Wang, Y., Sun, R., Cai, J., Li, H., Chan, W. S., Liu, P., Yi, L., Zhang,

S., Li, W., Jiang, T., He, K., & Wu, Z. (2022). Development and Validation of a Deep Learning Model for Brain Tumor Diagnosis and Classification Using Magnetic Resonance Imaging. *JAMA Network Open*, *5*(8), e2225608–e2225608. https://doi.org/10.1001/JAMANETWORKOPEN.2022.25608

Gary Sibbald, R., Ayello, E. A., Alavi, A., Ostrow, B., Lowe, J., Botros, M., Goodman, L., Woo, K., & Smart, H. (2012). Screening for the high-risk diabetic foot: A 60-second tool (2012) ©. *Advances in Skin and Wound Care*, *25*(10), 465–476. https://doi.org/10.1097/01.ASW.0000421460.21773.7B

Guo, Q., Wang, Y., Xu, D., Nossent, J., Pavlos, N. J., & Xu, J. (2018). Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. *Bone Research*, *6*(1). https://doi.org/10.1038/S41413-018-0016-9

Guo, Y., Li, Y., Wang, L., & Rosing, T. (2019). Depthwise convolution is all you need for learning multiple visual domains. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 8368–8375. https://doi.org/10.1609/AAAI.V33I01.33018368

Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M., Pan, I., Pereira, L. A., Sousa, R. T., Abdala, N., Kitamura, F. C., Thodberg, H. H., Chen, L., Shih, G., Andriole, K., Kohli, M. D., Erickson, B. J., & Flanders, A. E. (2019). The RSNA Pediatric Bone Age Machine LearningChallenge. *Radiology*, *290*(2), 498. https://doi.org/10.1148/RADIOL.2018180736

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 770–778. https://doi.org/10.48550/arxiv.1512.03385

Hernandez-Contreras, D. A., Peregrina-Barreto, H., Rangel-Magdaleno, J. D. J., & Renero-Carrillo, F. J. (2019). Plantar Thermogram Database for the Study of Diabetic Foot Complications. *IEEE Access*, *7*, 161296–161307. https://doi.org/10.1109/ACCESS.2019.2951356

Hernandez-Contreras, D., Peregrina-Barreto, H., Rangel-Magdaleno, J., Gonzalez-Bernal, J. A., & Altamirano-Robles, L. (2017). A quantitative index for classification of plantar thermal changes in the diabetic foot. *Infrared Physics & Technology*, *81*, 242–249. https://doi.org/10.1016/J.INFRARED.2017.01.010

Hirano, T., Nishide, M., Nonaka, N., Seita, J., Ebina, K., Sakurada, K., & Kumanogoh, A. (2019). Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. *Rheumatology Advances in Practice*, *3*(2). https://doi.org/10.1093/RAP/RKZ047

Honda, S., Yano, K., & Tanaka, E. (2022). *Development of a scoring model for the Sharp / van der Heijde score using convolutional neural networks and its clinical application in predicting radiographic progression using a graph convolutional network*.

Honda, S., Yano, K., Tanaka, E., Ikari, K., & Harigai, M. (2022). Development of a scoring model for the Sharp/van der Heijde score using convolutional neural networks and its clinical application. *Rheumatology*. https://doi.org/10.1093/RHEUMATOLOGY/KEAC586

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

Islam, S. M. S., Rahman, S., Rahman, M. M., Dey, E. K., & Shoyaib, M. (2016). Application of deep learning to computer vision: A comprehensive study. *2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016*, 592–597. https://doi.org/10.1109/ICIEV.2016.7760071

Jiao, S., Pan, Z., Hu, G., Shen, Q., Du, L., Chen, Y., & Wang, J. (2020). Multi-scale and multi-branch feature representation for person re-identification. *Neurocomputing*, *414*, 120–130. https://doi.org/10.1016/J.NEUCOM.2020.06.074

Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., Duan, X., Tang, J., Chang, D. T., Wu, X., Gao, F., & Li, R. (2021). Predicting treatment response from longitudinal images using multi-task deep learning. *Nature Communications 2021 12:1*, *12*(1), 1–11. https://doi.org/10.1038/S41467-021-22188-Y

Karoly, A. I., Galambos, P., Kuti, J., & Rudas, I. J. (2021). Deep Learning in Robotics: Survey on Model Structures and Training Strategies. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(1), 266–279. https://doi.org/10.1109/TSMC.2020.3018325

Khan, A. A., & Arora, A. S. (2018). Breast Cancer Detection Through Gabor Filter Based Texture Features Using Thermograms Images. *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications*, 412–417. https://doi.org/10.1109/ICSCCC.2018.8703342

Khandakar, A., Chowdhury, M. E. H., Ibne Reaz, M. Bin, Md Ali, S. H., Hasan, M. A., Kiranyaz, S., Rahman, T., Alfkey, R., Bakar, A. A. A., & Malik, R. A. (2021). A machine learning model for early detection of diabetic foot using thermogram images. *Computers in Biology and Medicine*, *137*, 104838. https://doi.org/10.1016/J.COMPBIOMED.2021.104838

Kumar, P., Gaurav, A., Rajnish, R. K., Sharma, S., Kumar, V., Aggarwal, S., & Patel, S. (2022). Applications of thermal imaging with infrared thermography in Orthopaedics. *Journal of Clinical Orthopaedics and Trauma*, *24*, 101722. https://doi.org/10.1016/J.JCOT.2021.101722

Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., & Langlotz, C. P. (2017). Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Https://Doi.Org/10.1148/Radiol.2017170236*, *287*(1), 313–322. https://doi.org/10.1148/RADIOL.2017170236

Li, B., Li, Y., & Eliceiri, K. W. (2020). Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 14313–14323. https://doi.org/10.48550/arxiv.2011.08939

Li, L., Xu, M., Wang, X., Jiang, L., & Liu, H. (2019). Attention Based Glaucoma Detection: A Large-Scale Database and CNN Model. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *2019-June*, 10563–10572. https://doi.org/10.1109/CVPR.2019.01082

Lu, M. T., Ivanov, A., Mayrhofer, T., Hosny, A., Aerts, H. J. W. L., & Hoffmann, U. (2019). Deep

Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA Network Open*, *2*(7), e197416–e197416. https://doi.org/10.1001/JAMANETWORKOPEN.2019.7416

Lu, Y. J., Chang, X., Li, C., Zhang, W., Cornell, S., Ni, Z., Masuyama, Y., Yan, B., Scheibler, R., Wang, Z. Q., Tsao, Y., Qian, Y., & Watanabe, S. (2022). ESPnet-SE++: Speech Enhancement for Robust Speech Recognition, Translation, and Understanding. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*, 5458–5462. https://doi.org/10.21437/Interspeech.2022-10727

Macdonald, A., Petrova, N., Ainarkar, S., Allen, J., Plassmann, P., Whittam, A., Bevans, J., Ring, F., Kluwe, B., Simpson, R., Rogers, L., Machin, G., & Edmonds, M. (2016). Thermal symmetry of healthy feet: a precursor to a thermal study of diabetic feet prior to skin breakdown. *Physiological Measurement*, *38*(1), 33. https://doi.org/10.1088/1361-6579/38/1/33

Magalhaes, C., Mendes, J., & Vardasca, R. (2021). Meta-analysis and systematic review of the application of machine learning classifiers in biomedical applications of infrared thermography. *Applied Sciences (Switzerland)*, *11*(2), 1–18. https://doi.org/10.3390/app11020842

Maziarz, K., Krason, A., & Wojna, Z. (2021). *Deep Learning for Rheumatoid Arthritis: Joint Detection and Damage Scoring in X-rays*. https://doi.org/10.48550/arxiv.2104.13915

Mohanaiah, P., Sathyanarayana, P., Gurukumar, L., & Professor, A. (2013). Image Texture Feature Extraction Using GLCM Approach. *International Journal of Scientific and Research Publications*, *3*(5). www.ijsrp.org

Monti, S., Montecucco, C., Bugatti, S., & Caporali, R. (2015). Rheumatoid arthritis treatment: the earlier the better to prevent joint damage. *RMD Open*, *1*(Suppl 1), e000057. https://doi.org/10.1136/RMDOPEN-2015-000057

Morid, M. A., Borjali, A., & Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine*, *128*, 104115. https://doi.org/10.1016/J.COMPBIOMED.2020.104115

Nag, U., Upadhayay, M., & Gupta, T. (2021a). *Detecting Diabetic Foot Complications using Infrared Thermography and Machine Learning; Detecting Diabetic Foot Complications using Infrared Thermography and Machine Learning*. https://doi.org/10.1145/3474906.3474919

Nag, U., Upadhayay, M., & Gupta, T. (2021b). Detecting diabetic foot complications using infrared thermography and machine learning. *ACM International Conference Proceeding Series*, 41–46. https://doi.org/10.1145/3474906.3474919

Oruh, J., Viriri, S., & Adegun, A. (2022). Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition. *IEEE Access*, *10*, 30069–30079. https://doi.org/10.1109/ACCESS.2022.3159339

Ory, P. A. (2003). Interpreting radiographic data in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, *62*(7), 597–604. https://doi.org/10.1136/ARD.62.7.597

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Paderno, A., Gennarini, F., Sordi, A., Montenegro, C., Lancini, D., Villani, F. P., Moccia, S., &

Piazza, C. (2022). Artificial intelligence in clinical endoscopy: Insights in the field of videomics. *Frontiers in Surgery*, *9*. https://doi.org/10.3389/FSURG.2022.933297

Prabha, A., Yadav, J., Rani, A., & Singh, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Computers in Biology and Medicine*, *136*, 104664. https://doi.org/10.1016/J.COMPBIOMED.2021.104664

Raghavendra, U., Gudigar, A., Rao, T. N., Ciaccio, E. J., Ng, E. Y. K., & Rajendra Acharya, U. (2019). Computer-aided diagnosis for the identification of breast cancer using thermogram images: A comprehensive review. *Infrared Physics & Technology*, *102*, 103041. https://doi.org/10.1016/J.INFRARED.2019.103041

Rasyid, M. B. Al, Yunidar, Arnia, F., & Munadi, K. (2018). Histogram statistics and GLCM features of breast thermograms for early cancer detection. *1st International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, ECTI-NCON 2018*, 120–124. https://doi.org/10.1109/ECTI-NCON.2018.8378294

Ravindran, V., & Rachapalli, S. (2011). An overview of commonly used radiographic scoring methods in rheumatoid arthritis clinical trials. *Clinical Rheumatology*, *30*(1), 1–6. https://doi.org/10.1007/S10067-010-1554-8/METRICS

Roux-Lombard, P., Eberhardt, K., Saxne, T., Dayer, J. M., & Wollheim, F. A. (2001). Cytokines, metalloproteinases, their inhibitors and cartilage oligomeric matrix protein: relationship to radiological progression and inflammation in early rheumatoid arthritis. A prospective 5-year study. *Rheumatology (Oxford, England)*, *40*(5), 544–551. https://doi.org/10.1093/RHEUMATOLOGY/40.5.544

Sahiner, B., Chen, W., Samala, R. K., & Petrick, N. (2023). Data drift in medical machine learning: implications and potential remedies. *Https://Doi.Org/10.1259/Bjr.20220878*. https://doi.org/10.1259/BJR.20220878

Saminathan, J., Sasikala, M., Narayanamurthy, V. B., Rajesh, K., & Arvind, R. (2020). Computer aided detection of diabetic foot ulcer using asymmetry analysis of texture and temperature features. *Infrared Physics & Technology*, *105*, 103219. https://doi.org/10.1016/J.INFRARED.2020.103219

Seeland, M., & Mäder, P. (2021). Multi-view classification with convolutional neural networks. *PLOS ONE*, *16*(1), e0245230. https://doi.org/10.1371/JOURNAL.PONE.0245230

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/S11263-019-01228-7/FIGURES/21

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., Wille, M. M. W., Naqibullah, M., Sanchez, C. I., & Van Ginneken, B. (2016). Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, *35*(5), 1160–1169. https://doi.org/10.1109/TMI.2016.2536809

Sharp, J. T., Wolfe, F., Lassere, M., Boers, M., Van Der Heijde, D., Larsen, A., Paulus, H., Rau, R., & Strand, V. (2004). Variability of precision in scoring radiographic abnormalities in

rheumatoid arthritis by experienced readers. *The Journal of Rheumatology*, *31*(6).

Smolen, J. S., Van Der Heijde, D. M. F. M., St.Clair, E. W., Emery, P., Bathon, J. M., Keystone, E., Maini, R. N., Kalden, J. R., Schiff, M., Baker, D., Han, C., Han, J., & Bala, M. (2006). Predictors of joint damage in patients with early rheumatoid arthritis treated with high-dose methotrexate with or without concomitant infliximab: Results from the ASPIRE trial. *Arthritis & Rheumatism*, *54*(3), 702–710. https://doi.org/10.1002/ART.21678

Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, *3*, 54–70. https://doi.org/10.1016/J.COGR.2023.04.001

Sun, D., Nguyen, T. M., Allaway, R. J., Wang, J., Chung, V., Yu, T. V., Mason, M., Dimitrovsky, I., Ericson, L., Li, H., Guan, Y., Israel, A., Olar, A., Pataki, B. A., Stolovitzky, G., Guinney, J., Gulko, P. S., Frazier, M. B., Chen, J. Y., … Bridges, S. L. (2022). A Crowdsourcing Approach to Develop Machine Learning Models to Quantify Radiographic Joint Damage in Rheumatoid Arthritis. *JAMA Network Open*, *5*(8), e2227423. https://doi.org/10.1001/JAMANETWORKOPEN.2022.27423

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Van Der Heijde, D. (2000). How to read radiographs according to the Sharp/van der Heijde method. *Journal of Rheumatology*, *27*(1), 261–263.

Van Der Heijde, D. M. F. M. (2000). Radiographic imaging: the “gold standard” for assessment of disease progression in rheumatoid arthritis. *Rheumatology (Oxford, England)*, *39 Suppl 1*(SUPPL. 1), 9–16. https://doi.org/10.1093/OXFORDJOURNALS.RHEUMATOLOGY.A031496

van Doremalen, R. F. M., van Netten, J. J., van Baal, J. G., Vollenbroek-Hutten, M. M. R., & van der Heijden, F. (2020). Infrared 3D Thermography for Inflammation Detection in Diabetic Foot Disease: A Proof of Concept. *Journal of Diabetes Science and Technology*, *14*(1), 46–54. https://doi.org/10.1177/1932296819854062

Van Netten, J. J., Van Baal, J. G., Liu, C., Van Der Heijden, F., & Bus, S. A. (2013). Diabetic Foot Disease: Emerging Technologies: Infrared Thermal Imaging for Automated Detection of Diabetic Foot Complications. *Journal of Diabetes Science and Technology*, *7*(5), 1122. https://doi.org/10.1177/193229681300700504

Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., & Van Gool, L. (2020). Multi-Task Learning for Dense Prediction Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3614–3633. https://doi.org/10.1109/TPAMI.2021.3054719

Vardasca, Magalhaes, & Mendes. (2019). Biomedical Applications of Infrared Thermal Imaging: Current State of Machine Learning Classification. *Proceedings*, *27*(1), 46. https://doi.org/10.3390/proceedings2019027046

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention

Module. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11211 LNCS*, 3–19. https://doi.org/10.48550/arxiv.1807.06521

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., & Xu, H. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, *27*(3), 457–470. https://doi.org/10.1093/JAMIA/OCZ200

Xu, L., Huang, J., Nitanda, A., Asaoka, R., Yamanishi, K., Xu, L., Huang, J., Nitanda, A., Asaoka, R., & Yamanishi, K. (2020). A Novel Global Spatial Attention Mechanism in Convolutional Neural Network for Medical Image Classification. *ArXiv*, arXiv:2007.15897. https://ui-adsabs-harvard-edu.uml.idm.oclc.org/abs/2020arXiv200715897X/abstract

Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., & Yang, J. (2019). Pattern-Affinitive Propagation across Depth, Surface Normal and Semantic Segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 4101–4110. https://doi.org/10.48550/arxiv.1906.03525