Deep Forest Classification-based Inference for Individual Level Infectious Disease Models

by

Winner Pathak

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

Master of Science

Department of Statistics
University of Manitoba
Winnipeg

Abstract

Infectious diseases spread rapidly and pose significant global health risks. Understanding the mechanisms of infectious disease transmission is essential for informing effective public health policy. To this end, individual-level models (ILMs) provide a flexible framework to incorporate covariate information when modelling. ILMs have been recently expanded to geographically dependent ILMs (GD-ILMs) to address spatially varying risk factors. Parameter estimation for these models is typically done using a Bayesian framework, which is computationally expensive. In this work, we propose to use a tree-based epidemic model classification method to bypass computationally intensive likelihood calculations. Specifically, we use a deep forest classifier and evaluate the predictive ability of the corresponding GD-ILMs to capture disease transmission dynamics. We validate this approach with simulated epidemic data and the UK 2001 foot-and-mouth disease outbreak, and compare its performance with that of the standard random forest classifier. Our findings suggest that the deep forest classifier generally outperforms the standard random forest classifier, as it uses an additional layer to extract more information from predictor variables.

Preface

This thesis is original work by Winner Pathak under the supervision of Dr. Gyanendra Pokharel and Dr. Elif Acar. Chapter 2 contains a manuscript that is to be submitted for publication.

Acknowledgment Page

I would like to express my deepest gratitude to my supervisors, Dr. Gyanendra Pokharel and Dr. Elif Acar, for their unwavering guidance, continuous encouragement, and invaluable mentorship throughout this journey. Their insights and support have been instrumental in the successful completion of this thesis. I would also like to extend my sincere thanks to my committee members, Dr. Ruppa Thulasiram and Dr. Shakhawat Hossain, for their thoughtful feedback, and constructive suggestions. I am also grateful to the Department of Statistics, University of Manitoba for providing essential resources that supported this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Infectious diseases have attracted significant research interest as they can spread quickly and pose a threat to global health security. Disease outbreaks such as COVID-19, H1N1 influenza, SARS, Ebola and foot-and-mouth disease (FMD) have profoundly impacted human populations and animals throughout history, highlighting the devastating effects of infectious diseases on global health and economy. For effective and timely control of disease outbreaks, it is important to understand the key factors such as host susceptibility and transmissibility that influence disease transmission between individuals in a population. It is also vital to understand how the disease spreads over time and what factors lead some individuals to become infected while others remain unaffected. To address these issues, several statistical models have been developed to understand the mechanisms of infectious disease transmission and inform public health policy decisions.

1.1 Infectious Disease Modelling

The modelling of infectious diseases and the identification of key risk factors are of significant interest to researchers, and public health professionals. Spatio-temporal modelling of infectious diseases has been widely used to investigate the risk factors associated with the disease transmission (Hughes et al., 1997; Keeling et al., 2001; Chis Ster and Ferguson, 2007; Deardon et al., 2010). Kermack and McKendrick (1927) proposed the deterministic susceptible-infected-removed (SIR) compartmental model, which laid the foundation for infectious disease models. This framework categorizes the population into three compartments according to individuals' infection status: susceptible (S) if they can contract the disease, infectious (I) if they have the disease and can infect others, and removed (R) if they are removed from the population, either because they have recovered or died and cannot be infected again or infect others. Individuals progress through the compartments sequentially: $S \rightarrow I \rightarrow R$. In recent years, studies have increasingly focused on the spatio-temporal analysis of infectious disease data across a range of applications, including the 2001 UK FMD outbreak (Keeling et al., 2001; Deardon et al., 2010; Peitsch et al., 2024), tomato spotted wilt virus (TSWV) (Hughes et al., 1997; Pokharel and Deardon, 2016), influenza (Mahsin et al., 2020), tuberculosis (Amiri et al., 2021), and COVID-19 (Amiri et al., 2023), among others. These spatio-temporal models help in understanding how diseases spread over space and time and what factors contribute to their transmission.

1.2 Individual Level Models (ILMs)

A notable class of spatio-temporal models is the discrete-time individual-level models (ILMs), developed by Deardon et al. (2010). ILMs are highly flexible and can incorporate individual-level covariates such as age, sex, and genetic information to provide insights into disease transmission. Disease progression is typically modeled within the SIR framework, and transmission is described at discrete time points. ILMs have been extensively applied over the years to model infectious disease dynamics. For example, Pokharel and Deardon (2014) applied ILMs to TSWV data with a uniformly gridded population and spatial stratification, using random forests for inference. They also fitted the same models using a full Bayesian approach with random walk Metropolis-Hastings Markov Chain Monte Carlo (MCMC) for comparison. Later, Pokharel and Deardon (2016) employed spatial ILM on the same TSWV dataset, incorporating Gaussian process emulators as a method of inference. Similarly, Malik et al. (2016) applied ILMs to the FMD outbreak and utilized sampling-based likelihood approximations for parameter estimation. More recently, Peitsch et al. (2024) applied ensemble learning methods to parameterize ILMs for a geographically clustered population, treating spatial clusters as strata in the context of the FMD outbreak data.

ILMs provide a common framework for capturing disease dynamics, but they primarily consider the spatial distance, and/or contact network between susceptible and infectious individuals but not their spatial location. To address this limitation, Mahsin et al. (2020) introduced geographically dependent ILMs (GD-ILMs), which incorporate spatial random effects to capture regional heterogeneity and applied it

to 2009 seasonal influenza outbreak data in Calgary, Canada. GD-ILMs have since been applied in analyzing diseases like tuberculosis (Amiri et al., 2021) and COVID-19 (Amiri et al., 2023).

1.3 Parameter Estimation

ILMs and GD-ILMs are typically fitted within a Bayesian framework. Closed-form solutions for the likelihood or posterior distributions in mechanistic models such as ILMs and GD-ILMs are often unavailable due to their complexity and the presence of latent variables. Therefore, MCMC methods are widely used to approximate these intractable posteriors, enabling flexible and practical Bayesian inference through iterative sampling.

In Bayesian framework, inference about unknown parameters $\boldsymbol{\theta}$ given observed data $\boldsymbol{D}$ (e.g., individual-level epidemic data) is achieved by constructing the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{D})$ using Bayes' theorem:

$$f(\boldsymbol{\theta} \mid \boldsymbol{D}) = \frac{f(\boldsymbol{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \tag{1.1}$$

where Θ is the parameter space; $f(\boldsymbol{D} \mid \boldsymbol{\theta})$ is the likelihood function; $p(\boldsymbol{\theta})$ is the prior distribution for the parameters; and $\int_{\Theta} f(\boldsymbol{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is a normalizing constant.

The posterior $f(\boldsymbol{\theta} \mid \boldsymbol{D})$ combines prior beliefs about the parameters $p(\boldsymbol{\theta})$ with the likelihood function $f(\boldsymbol{D} \mid \boldsymbol{\theta})$ of observed data. However, selecting appropriate priors

can be difficult, especially when strong prior information is unavailable. Additionally, Bayesian MCMC methods require repeated evaluations of the likelihood function, which can be computationally intensive. Various approximate methods within the Bayesian paradigm have been employed to mitigate these challenges, including Gaussian process approximations (Pokharel and Deardon, 2016), data sampling-based likelihood approximations (Malik et al., 2016), and approximate Bayesian computation (Almutiry and Deardon, 2020) for ILMs. Some studies have sought to avoid Bayesian methods altogether. For example, Amiri et al. (2021) and Amiri et al. (2023) used Expectation Conditional Maximization (ECM) for inference while fitting GD-ILMs. As an alternative direction, machine learning (ML) methods offer computational efficiency by avoiding explicit likelihood evaluation. These techniques have shown promise in parameterizing ILMs (Pokharel and Deardon, 2014; Augusta et al., 2019; Peitsch et al., 2024).

1.3.1 Machine Learning

The primary objective of this thesis is to investigate the use of ML classifiers for parameter estimation in ILMs and GD-ILMs. While ML techniques have been applied to ILMs, their use in parameterizing GD-ILMs remains unexplored. Additionally, deep forest has not been previously applied to infectious disease modelling in either ILMs or GD-ILMs. Hence, this thesis introduces deep forest classifier as a novel and computationally efficient alternative to traditional Bayesian inference methods. We assess the effectiveness of deep forest classifier, and compare its performance to

that of the random forest in predicting infectious disease transmission dynamics and estimating parameters. The evaluation uses both simulated epidemic data and FMD outbreak dataset.

The ML methods used in this work are briefly introduced below.

Random Forest Classification

Random forest is a tree-based method that aggregates the outputs of multiple decision trees to produce a single, more accurate result for predictive performance (Breiman, 2001). While building the tree, subset of the predictor variables are randomly sampled at each split, where the subset sample size is usually set to the square root of the total number of predictor variables (James et al., 2013). A new sample of predictor is considered at each split and the use of randomly selected predictor from a fresh set at each node to grow each tree minimizes the correlation between the trees and, therefore, reduces variability by combining multiple decision trees (James et al., 2013). The algorithm for random forest for a classification problem is stated below:

Algorithm 1 Random Forest Algorithm

Require: Training data $D = \{(x_i, y_i)\}_{i=1}^{n}$; number of trees B; number of predictors to consider at each split m

1: For $b = 1$ to B, repeat:

(a) Sample a bootstrap dataset D_b from D

(b) Train a decision tree T_b on D_b

(c) For each node in the tree:

- Randomly select m predictors from the total number of predictors
- Choose the best split based on Gini impurity or entropy
- Split the node and continue recursively

2: **Prediction:** For a new input x, return the majority class from $\{T_b(x)\}_{b=1}^{B}$

In this work, the random forest algorithm is applied to classify epidemic curves, where the number of infected individuals at each time point serves as the predictor and the underlying epidemic-generating model is the response.

Deep Forest Classification

Deep forest proposed by Zhou and Feng (2017) has two ensemble components: cascade forest and multi-grained scanning. The cascade forest structure consists of multiple layers where each layer represents an ensemble of decision tree forest. Multi-grained scanning is used to enhance the predictive ability of cascade forest by utilizing a sliding window structure to extract local predictor information by scanning raw input (Zhou and Feng, 2017). Figure 1.1 summarizes the overall procedure of gcForest.

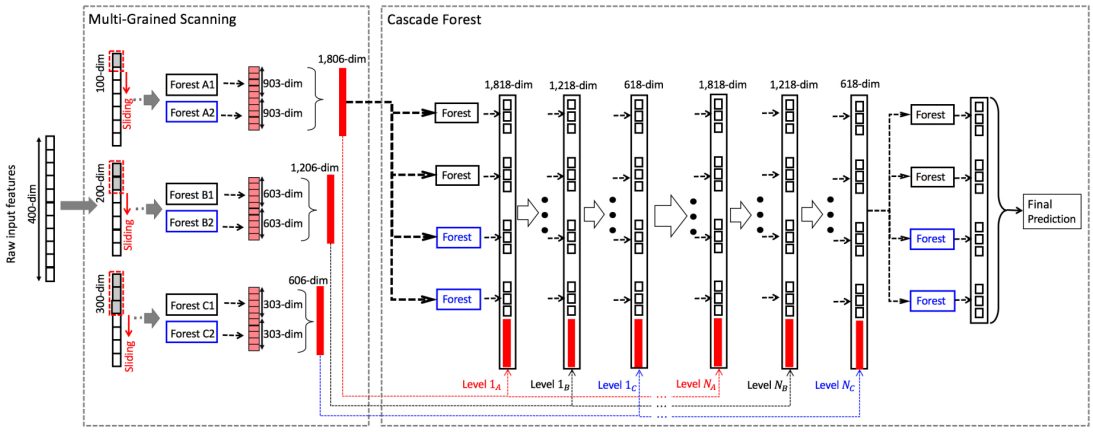


Figure 1.1: Overall procedure of gcForest. Suppose there are 3 classes to predict, 3 sizes of sliding windows and 400 dimensional raw predictors (Zhou and Feng, 2017).

Suppose the input consists of 400-dimensional raw predictor variables, the classification task involves 3 classes to predict, multi-grained scanning is performed using 3 different window sizes, and there are m training samples. A window of size 100 predictors will generate a data set of $301 \times m$ 100-dimensional training examples (number of raw predictors - size of the window +1). Two different forests are then trained using this data to obtain 1,806-dimensional predictor vector (3 classes to be predicted from each forest). This transformed predictor vector is used to train the first layer of cascade forest. The transformed predictor vectors, along with the class vectors produced by the preceding layer, are then used to train the second and third layers of the cascade forest, respectively. When a new layer is added, cross-validation or out-of-bag predictions can be performed to evaluate the overall performance of the cascade and the training process is automatically terminated once there is no significant performance gain. The test data (observed epidemic) is first processed through

the multi-grained scanning step to generate its transformed predictor representation. This representation is then passed through each level of the cascade structure. The final prediction is made by aggregating the class vectors produced at the final level (Zhou and Feng, 2017).

Based on these ML methods, this thesis explores their application to individual-level infectious disease models, with a focus on the novel use of deep forest classifiers.

1.4 Thesis Overview

The rest of the thesis is structured as follows. Chapter 2 contains a manuscript that is to be submitted for publication at a peer-reviewed journal. Section 2.2 provides relevant background information and the motivation for this research. Section 2.3 introduces the general ILMs, GD-ILMs and the two forest-based ensemble learning approaches applied in this work. Section 2.4 presents the simulation study demonstrating the application of ILM, and GD-ILM and the results from the simulation study. The 2001 FMD data description, application of ILMs and GD-ILMs to the FMD dataset, and the FMD data analysis results are presented in Section 2.5. Section 2.6 presents a discussion of the results and some concluding remarks. We then conclude this thesis with a brief summary and future research directions in Chapter 3.

Chapter 2

Deep forest classification-based inference for individual level infectious disease models[1]

2.1 Abstract

Infectious diseases spread rapidly and pose significant global health risks. Understanding the mechanisms of infectious disease transmission is essential for informing effective public health policy. To this end, individual-level models (ILMs) provide a flexible framework to incorporate covariate information when modelling. ILMs have been recently expanded to geographically dependent ILMs (GD-ILMs) to address spatially varying risk factors. Parameter estimation for these models is typically done using a Bayesian framework, which is computationally expensive. In this work, we propose to use a tree-based epidemic model classification method to bypass computa-

[1]This chapter contains the manuscript that is to be submitted for publication at a peer-reviewed journal.

tionally intensive likelihood calculations. Specifically, we use a deep forest classifier and evaluate the predictive ability of the corresponding GD-ILMs to capture disease transmission dynamics. We validate this approach with simulated epidemic data and the UK 2001 foot-and-mouth disease outbreak, and compare its performance with that of the standard random forest classifier. Our findings suggest that the deep forest classifier generally outperforms the standard random forest classifier, as it uses an additional layer to extract more information from predictor variables.

2.2 Introduction

Infectious diseases spread rapidly and pose a threat to global health security. Therefore, it is crucial to understand the underlying factors that influence the spread of the disease and, hence, predict how the disease progresses over time. The risk factors associated with disease transmission need to be identified, and effective containment strategies to mitigate the impact need to be devised quickly. Statistical models have been increasingly explored to identify risk factors and help guide public health policy decisions.

Built on this foundation, mechanistic models have been developed to model the transmission dynamics of infectious diseases in heterogeneous populations through time and space (Chis Ster and Ferguson, 2007). A class of discrete-time individual-level models (ILMs) developed by Deardon et al. (2010) offers highly flexible models that can incorporate individual-level covariates and provide insight into disease transmission dynamics. However, these ILMs only consider the spatial distance between susceptible

and infectious individuals and do not incorporate the effect of spatial heterogeneity due to their specific location. Extending the general framework of ILMs to spatially varying disease dynamics, Mahsin et al. (2020) introduced a new class called geographically dependent ILMs (GD-ILMs), and applied it to the 2009 Alberta seasonal influenza outbreak data. These models account for spatially varying risk factors as well as unobserved spatial structure in disease transmission using a mixed model with spatially defined random effects. ILMs and GD-ILMs are typically fitted within a Bayesian Markov Chain Monte Carlo (MCMC) framework. However, fitting ILMs and GD-ILMs to large datasets can be computationally demanding due to costly likelihood calculations.

Different methods have been developed in the literature to circumvent the computationally intensive likelihood calculation while fitting ILMs. In particular, machine learning (ML) methods have been used as an alternative to Bayesian MCMC methods to parameterize ILMs with an efficient model classification-based inference procedure. For example, Nsoesie et al. (2011) used supervised classification methods such as random forest, support vector machines, nearest neighbours, linear and flexible discriminant analysis to classify infectious disease models. Similarly, Pokharel and Deardon (2014) used a random forest classification to parameterize ILMs. Other machine learning approaches include neural network-based deep learning methods (Augusta et al., 2019) and ensemble learning classifiers for two-level ILMs (Liu et al., 2021). Most of these studies assume spatially homogeneous populations, implying no spatial variation in population characteristics such as density, growth, or interactions

between individuals, which may not be realistic for most human and animal populations. For example, Pokharel and Deardon (2014) used the random forest classifier to parameterize ILMs, but only considered a uniformly gridded population with spatial stratification. Peitsch et al. (2024) applied ensemble learning methods to parameterize ILMs for a geographically clustered population, considering spatial clusters as strata. While the latter approach is more realistic than the uniformly gridded population, it does not readily extend to capture spatial variation in GD-ILMs.

The machine learning approaches mentioned earlier could help address the time-sensitive challenges in controlling disease spread. However, a few deficiencies may limit their effectiveness in infectious disease modelling. For instance, a deep neural network (DNNs)-based deep learning approach requires a huge amount of training data (Guo et al., 2018). During an epidemic, we may not have immediate access to large amounts of data and might need to develop policies quickly using the available data. Moreover, the performance of the DNNs depends on many hyperparameters, and its effectiveness relies on the skill in tuning these parameters (Zhou and Feng, 2017). While ensemble learning methods such as random forest address some of the limitations of DNNs, they come with their own drawbacks. Random forests have a shallow structure because they consist of a single layer of trees working in parallel on the original inputs (Breiman, 2001), limiting their ability to capture a complex predictor representation. Therefore, it would be beneficial to employ an approach that is endowed with representation learning ability of a neural network but bypasses its aforementioned shortcomings. Deep forest, a decision tree ensemble

learning method proposed by Zhou and Feng (2017), is an alternative to DNNs that works well with small-scale data and is computationally efficient. It also has a fewer hyperparameters and is robust to hyperparameter settings compared to DNNs (Zhou and Feng, 2017). Similar to DNNs, deep forest has a multi-layer cascade structure. However, instead of neurons, each layer in the deep forest contains multiple random forests (Guo et al., 2018). In this work, we propose a deep forest-based classification approach for ILMs and GD-ILMs to bypass the computationally intensive likelihood calculations of the Bayesian MCMC framework while also addressing the limitations of random forest-based classification.

The deep forest classification is employed to specify complex infectious disease transmission dynamics, and its performance is subsequently compared with that of random forest classifier through a simulation study. We further illustrate these methods using data from the 2001 foot-and-mouth disease (FMD) outbreak in the UK, where we consider five neighboring counties: Cumbria, Northumberland, North Yorkshire, Dumfries and Galloway, and Durham, selected from the 48 counties in the UK. The data is modeled at the farm level, treating each farm as an individual unit. Our goal is to investigate whether there is evidence of spatial effects at the county level and whether such effects influence disease transmission between farms. We find that GD-ILMs effectively capture disease transmission dynamics while incorporating spatial heterogeneity into the model. Furthermore, both the standard random forest and deep forest methods perform well in characterizing the disease transmission for both simulated data and the 2001 FMD epidemic. Notably, deep forest outperforms random forest in most cases, likely due to its additional layer, which extracts more

complex information from the predictor variables.

The rest of the paper is structured as follows. In Section 2, the ILMs and GD-ILMs as well as the two forest-based ensemble learning methods used in this study are introduced. Section 3 presents the results of the simulation study evaluating the performance of the deep and random forest classifiers in ILMs and GD-ILMs. The 2001 FMD data description, application of ILMs and GD-ILMs to the FMD dataset, and FMD data analysis results are provided in Section 4. Section 5 presents a discussion of the results and some concluding remarks.

2.3 Methodology

2.3.1 Epidemic modelling

In this study, we model the spread of infectious disease using discrete-time ILMs and GD-ILMs. The general frameworks for the ILMs, and GD-ILMs are outlined below.

Individual Level Models (ILMs)

The discrete-time infectious disease ILM framework proposed by Deardon et al. (2010) is the main root of this research work. These ILMs describe the disease transmission between individuals at discrete time point $t = 1, ..., t_{\max}$, where $t_{\max}$ is the last time point observed. The disease progresses within a susceptible-infectious-removed (SIR) compartmental framework. In a finite population of n individuals, the individual i ($i = 1, ...n$) can be in any of the three compartments at time t. The sets of individuals

that are susceptible, infectious, and removed at time t are denoted as $S(t)$, $I(t)$, and $R(t)$, respectively. If the individual $i \in S(t)$, then they can contract the disease. If the individual $i \in I(t)$ then they have contracted the disease and can infect others. If the individual $i \in R(t)$, then they have been removed from the population, and cannot become infected again or infect others. An individual contracting the disease goes in order through all three compartments $S \to I \to R$. The time between I and R where an infectious individual is capable of transmitting the disease is called the *infectious period*, denoted by γ in this study.

Within the SIR framework, Deardon et al. (2010) defines the probability that the previously uninfected (susceptible) individual i becomes infected at time t as

$$P_{it} = 1 - \exp\left[-\left\{\Omega_{\mathcal{S}}(i) \sum_{j \in \mathcal{I}(t)} \Omega_{\mathcal{T}}(j)\mathcal{K}(i,j) + \epsilon(i,t)\right\}\right], \tag{2.1}$$

where $\Omega_{\mathcal{S}}(i) > 0$ is the susceptibility function for susceptible individual i describing the risk of individual i contracting the disease; $\Omega_{\mathcal{T}}(j) > 0$ is the transmissibility function for the infectious individual j representing the risk of individual j passing on the disease; $\mathcal{K}(i,j) > 0$ is the infection kernel that represents the shared risk factors between the susceptible individual i and infectious individual j based on their geographical distance or contact network; and $\epsilon(i,t) > 0$ is the spark function representing sources of infection from outside the population or that has not been described by the model framework.

Geographically Dependent Individual Level Models (GD-ILMs)

Mahsin et al. (2020) extended the ILMs into a new disease transmission framework called geographically dependent ILMs (GD-ILMs). These models allow to evaluate the effect of spatially varying risk factors on the transmission of infectious disease over the disjoint geographical regions. Similar to the ILMs, GD-ILMs can be defined within the SIR compartmental framework, and the probability of infection for the susceptible individual i at time t in geographical region k is given by

$$P_{ikt} = 1 - \exp\left[\left\{-\Omega_{\mathcal{S}}(i,k) \sum_{j \in \mathcal{I}(t,k,\xi(k))} \Omega_{\mathcal{T}}(j,k)\mathcal{K}(i,j)\right\} - \epsilon(i,k,t)\right], \tag{2.2}$$

where k represents the region index ranging from 1 to K; $\mathcal{I}(t,k,\xi(k))$ is the set of infectious individuals at time t in region k and its neighbouring regions; $\xi(k)$ is the set of neighbouring regions that are adjacent to region k; $\Omega_{\mathcal{S}}(i,k) > 0$ is the susceptibility function of potential risk factors for the susceptible individual i in region k; $\Omega_{\mathcal{T}}(j,k) > 0$ is the transmissibility function for the infectious individual j in region k; and $\epsilon(i,k,t) > 0$ is the spark function representing outside sources of infection.

The susceptibility and transmissibility functions can be used to incorporate individual-level covariates (e.g., age, gender) and region-level covariates (e.g., temperature, humidity) of interest. Let $\boldsymbol{X}(i,k)$ be the vector of covariates associated with susceptible individual i in region k, then the susceptibility function accounting

for the effects of covariates can be defined as $\Omega_{\mathcal{S}}(i,k) = \exp(\alpha + \boldsymbol{X}(i,k)'\boldsymbol{\alpha_1} + \phi_k)$, where $\alpha > 0$ is a constant infectivity parameter, $\boldsymbol{\alpha_1}$ is the vector of parameters for individual-level covariates. The spatial random effect, ϕ_k, represents the variation in disease spread between regions that cannot be explained by the observed variables. The inclusion of ϕ_k helps to better explain the spatial patterns of disease spread and improve the model's predictive accuracy by accounting for spatially structured heterogeneity (Mahsin et al., 2020). Following Mahsin et al. (2020), we consider a conditional autoregressive (CAR) model to capture the effects of spatially structured unobserved latent covariates or measurement error by spatial random effect ϕ_k (see Mahsin et al., 2020, for more details). The spatial random effects $\boldsymbol{\phi} = (\phi_1, \phi_2, ...\phi_K)$ are generated using the multivariate normal distribution $\phi \sim MVN(0, \Sigma)$, with $\Sigma = \sigma^2 \left[(\lambda D + (1-\lambda)I\right]^{-1}$. Here, I is the identity matrix of order K (number of regions), D is the spatial association or neighborhood matrix, and σ^2 controls the variance of random effects. The kl^{th} element of the neighborhood matrix D is defined as

$$D_{kl} = \begin{cases} f_k, & \text{if } l = k, \\ -1, & \text{if } l \sim k, \\ 0, & \text{otherwise,} \end{cases}$$

where f_k is the number of neighbors of region k, and $l \sim k$ indicates l and k are neighbors.

In this study, we use the power law spatial kernel; $\kappa(i,j) = d_{ij}^{-\beta}$, where d_{ij} is the Euclidean distance between individuals i and j, and β is a decay parameter which

determines the risk of infections that occur over varying distance. We also consider a constant transmissibility function, $\Omega_T(j,k) = 1$.

2.3.2 Classification Methods

Fitting ILMs and GD-ILMs using the Bayesian MCMC method is computationally expensive, particularly when applied to large scale real world datasets. To address this challenge, we use an alternative machine learning approach based on a deep forest classifier and compare its performance with that of the random forest classifier in identifying the underlying epidemic generating ILMs and GD-ILMs. For the classification-based inference, we first simulate epidemics over the parameter design matrix and form a library of epidemic curves, which contains the number of infections at each discrete time point. The epidemic curves are used as predictors and the epidemic generating models form a categorical response variable. The classifiers are then trained and used to predict the epidemic generating model for observed epidemic data. We briefly discuss random forest and deep forest classification methods below.

Random Forest

Random forest is an ensemble-based learning method that combines the outputs from multiple decision trees to reach a single result for better predictive performance (Breiman, 2001). Each tree in the random forest predicts a response, i.e., an epidemic generating model in our case. The overall prediction is determined by the majority vote across the decision trees (James et al., 2013). For the random forest classifier,

subset of the predictor variables are randomly sampled at each split while building decision trees. In our case, predictor variables are the number of infected individuals at each time point. A key disadvantage of random forests is their limited ability to capture complex predictor representations due to their relatively shallow ensemble structure, meaning that random forests consist of a single layer of decision trees that operate in parallel on the original input predictors without further transformation between trees. Alternatively, deep forest can be used to address this limitation, where a cascade structure is employed, consisting of multiple layers of forests (Zhou and Feng, 2017).

Deep Forest

The deep forest algorithm leverages the representation learning ability of deep neural networks, and incorporates ensemble learning techniques (Zhou and Feng, 2017). We will briefly discuss the two components of deep forest: cascade forest and multi-grained scanning.

The cascade forest structure consists of multiple layers, where each layer represents an ensemble of decision tree forest, essentially an ensemble of ensembles (Zhou and Feng, 2017). Two types of tree-based ensemble methods are employed in each layer to encourage diversity: extra-trees, which grow each tree using the entire training dataset and select split thresholds completely at random (Geurts et al., 2006), and random forests, which rely on bootstrapped samples and optimize split thresholds based on impurity measures (Breiman, 2001). In our case, for each epidemic curve, each forest

within a layer produces an estimate of class distribution (a vector of probabilities estimating the likelihood that the curve was generated by each possible epidemic model). Each layer in the cascade forest receives predictor information processed by its preceding layer, which then passes the processed results as an input to the next layer. After training, the model obtains out-of-bag (OOB) predictions from the current layer and estimates its generalization performance via OOB evaluation. If the performance surpasses that of all preceding layers, the cascade continues to grow. Otherwise, the training process is terminated, allowing the number of layers to be determined automatically (Zhou and Feng, 2017). The evaluation stage follows the same process as above.

Multi-grained scanning can be used to enhance the predictive ability of cascade forest by applying sliding windows over the predictors to generate a series of local low-dimensional predictor vectors. The low-dimensional predictor vectors are used to train a series of forests to obtain class distributions of input vectors, which are then used to train the first layer of cascade forest (Zhou and Feng, 2017). This transformed predictor vectors, combined with the class vector generated by the previous layer will then be used to train second and third layer of the cascade forests, respectively. During inference, an observed epidemic curve undergoes the same multi-grained scanning and cascade processing steps. The final prediction, the estimated epidemic generating model is obtained by aggregating the class vectors produced at the final cascade level (Zhou and Feng, 2017).

Hyperparameter Tuning

In this study, random forest-based classification was trained using the R package `randomForest` (Liaw and Wiener, 2002). The algorithm only requires two hyperparameters: predictor subset size considered at each split (`mtry`) and number of trees grown (`ntrees`). We used the default value of `mtry` $= \sqrt{p}$, where p is the number of predictors and `ntrees` $= 500$ in all the cases.

For deep forest, we used the Python package `deep-forest`. Of many hyperparameters used in this package, the most critical for tuning are the maximum depth of each tree, maximum number of cascade layers in the deep forest, number of bins used for non-missing values and number of trees in each forest (Zhou and Feng, 2017). Various combinations of these four hyperparameters were evaluated by measuring the corresponding error rates. However, the default hyperparameter settings yielded the optimal performance. This finding is consistent with previous research, which suggests that gcForest demonstrates strong robustness to hyperparameter selection and typically achieves excellent performance with its default configuration (Zhou and Feng, 2017).

2.4 Simulation Study

In this section, we evaluate the performance of random forest and deep forest classifiers in predicting disease transmission dynamics on data simulated using both ILMs and GD-ILMs.

2.4.1 ILM Epidemic Simulation

For the ILM epidemic simulation, we consider a population of 625 individuals on an equally-spaced uniform 25×25 grid. The susceptibility and transmissibility functions were kept constant assuming that all infected individuals had the same risk of contracting and transmitting the disease. Specifically, we set $\Omega_s(i) = \alpha$ and $\Omega_T(j) = 1$ where, α is the infectivity parameter that reflects the overall strength of the epidemic. We further assume $\epsilon(i,t) = \epsilon \geq 0$, which represents a purely random external infectious pressure and is not contingent on any other factors. We consider the power-law spatial kernel $\kappa(i,j) = d_{ij}^{-\beta}$, where β is the spatial decay parameter quantifying the risk of infection as a function of distance between susceptible and infected individuals, and d_{ij} is the Euclidean distance between susceptible individual i and infectious individual j. So, the probability of the susceptible individual i being infected at time t for the power-law ILM becomes

$$P_{it} = 1 - \exp\left[-\left\{\alpha \sum_{j \in \mathcal{I}(t)} d_{ij}^{-\beta} + \epsilon\right\}\right]. \tag{2.3}$$

A large number of parameter combinations were initially considered (Table 2.1) to create a 100×4 design matrix, resulting in 100 models for the power-law ILM. These parameter values were chosen arbitrarily to simulate informative epidemics, i.e. epidemics do not die out or infect the entire population too quickly.

Table 2.1: Parameter values used to simulate epidemics using power-law kernel.

Parameter	Values
Infectivity: α	0.100, 0.325, 0.550, 0.775, 1.000
Spatial decay: β	1.0, 1.5, 2.0, 2.5, 3.0
Infectious period: γ	2, 3
Spark: ϵ	0, 0.02

The first infectious individual was chosen from the center of the population at time point $t = 1$ as the initial seed to allow a reasonable progression, and the epidemic was allowed to spread using the probability model in Equation (2.3). The first 20 time points of each simulated epidemic were considered in the analysis, as most epidemics either died out or infected the entire population within this period.

We created an epidemic library of simulated epidemic curves, where each curve is represented by a vector of summary statistics,

$$\boldsymbol{\mathcal{X}}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_t^{(i)}\right),$$

where $\boldsymbol{\mathcal{X}}^{(i)}$ denotes the epidemic curve simulated using i-th parameter combination, and $x_t^{(i)}$ represents the number of infected individuals at time point t for $i = 1, 2, ..., n$. The simulation was replicated ten times for each parameter combination and an epidemic library was created.

The large number of parameter combinations in the design matrix leads to identifiability issues, i.e. , two sets of parameter produces similar epidemic resulting in a high misclassification rate. In previous studies, Pokharel and Deardon (2014) and

Peitsch et al. (2024) fixed the parameter α to avoid the model identifiability issue. In this study, we used k-means clustering with k=25 to cluster similar epidemics and narrow down the design matrix. This approach helps identify distinct models, ensuring that different parameter combinations in the trimmed design matrix produce unique epidemic curves reducing the misclassification rate. The ten replications of each epidemic that were generated using the parameter values in each row of the design matrix were first averaged. We then used k-means clustering on the averaged data so that similar epidemic curves were clustered together. Following the clustering process, we computed the sum of squared difference (SSD) between curves within each cluster. A threshold was then set arbitrarily to filter out the curves that were too similar. If the SSD exceeded the threshold, the curves were considered distinct and if the SSD was below the threshold, the curves were considered indistinguishable. Additionally, visual inspection was performed to manually select curves from each cluster such that the selected curves were sufficiently distinct from one another. By using clustering, we were able to systematically reduce the design matrix while maintaining diversity in the selected models.

We selected 31 distinct epidemic curves post clustering and recorded their corresponding epidemic generating models. The selected parameter combinations were then used to simulate epidemics using power-law ILMs described in Equation (2.3). For each parameter combination in the trimmed design matrix (see Appendix), we simulated 100 epidemic replications, resulting in an epidemic library of $31 \times 100 = 3100$ epidemic curves with 20 predictor variables corresponding to the number of infected individuals at the first 20 time points. The epidemic curves generated by

each parameter combination was assigned model names, $\mathbf{M} = \left(M^{(1)}, M^{(2)}, \ldots, M^{(n)}\right)^T$ where n is the total number of parameter combinations. We then trained both random forest and deep forest classifiers on this library and the trained classifiers were used to predict the generating model for another set of epidemic library simulated using same procedure but without the assigned model names (response). Confusion matrices were computed for both classifiers based on their predictions, and misclassification rates were calculated as the primary metric of comparison. To assess variability in classification error, we replicated the entire simulation process 10 times and recorded the classification error in each run.

2.4.2 ILM Simulation Results

The classification errors of random forest and deep forest across 10 replications under each model were averaged separately and tabulated in Table A1 in Appendix. The overall classification error was 20% for the random forest and 17% for the deep forest. We further analyzed the prediction errors of random forest and deep forest classifiers across a range of parameter values to understand the underlying transmission dynamics in the simulated epidemics. We found that the model with parameter set $(\epsilon, \gamma, \alpha, \beta)$ = (0.02, 2, 0.1, 3) had the lowest prediction error. The model prediction error for both random forest and deep forest classifiers is generally low for models with large spatial decay parameter β, especially when combined with low α. The parameter set $(\epsilon, \gamma, \alpha, \beta)$ = (0, 3, 0.325, 1) yielded the poorest model performance with the highest prediction error. This result is consistent with the literature (Pokharel and Deardon,

2014; Peitsch et al., 2024). When the spatial decay parameter is large, the epidemics are stronger, and the curves capture more information about the disease transmission dynamics (Peitsch et al., 2024). The above mentioned model with the worst model classification performance (labeled as M28 in Table A2, and A3) is being misclassified as model with parameter set $(\epsilon, \gamma, \alpha, \beta) = (0, 2, 0.775, 1.5)$ for both random forest and deep forest. This shows that despite using the clustering to narrow down the design matrix, the issue of model identifiability still persists. The spark term (ϵ) also has a strong marginal effect. Both random forest and deep forest classifiers exhibited lower prediction errors when $\epsilon = 0.02$, with average errors of 0.148 and 0.132, respectively. When no outside source of infection was introduced ($\epsilon = 0$), the prediction errors increased to 0.239 for random forest and 0.208 for deep forest. This increase suggests that the improved predictability could stem from existing identifiability issues, where introducing random background infections creates more distinct epidemic curves that the classifier can more easily differentiate and predict. The infectious period (γ) parameter does not appear to have a significant impact on prediction error in case of both classifiers. Deep forest generally maintained lower or comparable errors with narrower percentile intervals (Table A1 in Appendix) indicating its ability to capture uncertainty in prediction error rates compared to random forest.

We compare the performances of random forest and deep forest visually in Figure 2.1 using the class errors for the 31 models from each classifier. Each point in the plot represents the class error of an individual model where the x-axis and the y-axis represent the errors from the random forest, and deep forest, respectively. The

points are colored using a gradient scale corresponding to different β values, enabling assessment of classifier performance under varying spatial transmission patterns. A 45-degree reference line was included, where points above this line would indicate that the class error for the deep forest is higher than that for the random forest, and points below the line would suggest the opposite.

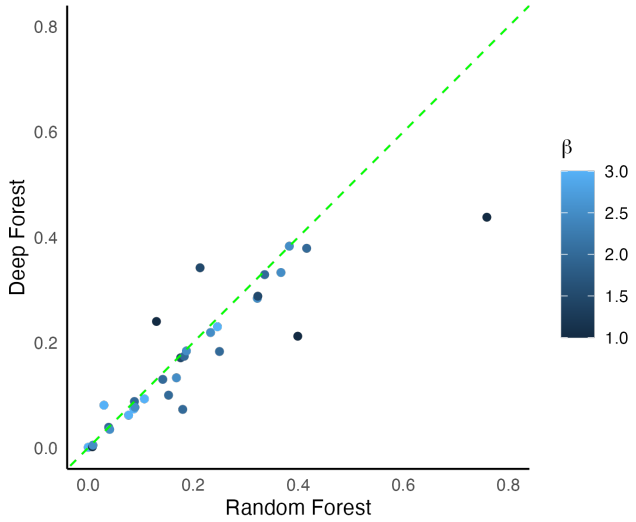


Figure 2.1: Scatterplot of model prediction error from the random forest and deep forest classifier for simulated data.

Among the 31 models considered, random forest outperformed deep forest only for 3 models (represented by points above the 45-degree line), while deep forest performed better than random forest for 16 models. The remaining 12 models exhibited comparable performance between the two classifiers. This indicates that deep forest generally resulted in a lower classification error than random forest. Based on these results, we can conclude that the deep forest classifier may perform better than the random forest classifier as it provides more information through the additional

layer during the classifier training process.

2.4.3 GD-ILM Epidemic Simulation

To simulate the epidemics using GD-ILM, we created a spatially clustered population of 1000 individuals over five spatial clusters using five different bivariate normal distributions following Peitsch et al. (2024) (the population plot is given in Appendix).

We consider two forms of the model: i) region-restricted GD-ILM and (ii) global GD-ILM. The probability of a susceptible individual being infected at time t in the geographical region k under these models is given by

$$P_{ikt} = 1 - \exp\left[-\left\{\exp(\alpha + \alpha_1 X_{i,k} + \phi_k) \sum_{j \in \mathcal{I}(t,k,\xi(k))} d_{ij}^{-\beta} + \epsilon\right\}\right], \tag{2.4}$$

and

$$P_{ikt} = 1 - \exp\left[-\left\{\exp(\alpha + \alpha_1 X_{i,k} + \phi_k) \sum_{j \in \mathcal{I}(t)} d_{ij}^{-\beta} + \epsilon\right\}\right], \tag{2.5}$$

respectively, where $X_{i,k}$ is a covariate associated with the susceptible individual i in the k^{th} cluster; $I(t)$ is the set of infectious individuals at time t in the global population considered; and $\mathcal{I}(t, k, \xi(k))$ is the set of infectious individuals at time t in the k^{th} region and its neighboring regions.

The region-restricted GD-ILM (2.4) allows the disease transmission to occur only within each cluster and their neighboring regions, $\xi(k)$ while in the global model (2.5), disease transmission occurs between individuals across the whole population. In this study, we used a uniform distribution from 0 to 1 to generate the individual-level covariate ($X_{i,k}$). In real data, these covariates can have different scales and distributions. Three different values for the spatial dependence hyperparameter λ were considered: i) weak ($\lambda = 0.3$), ii) moderate ($\lambda = 0.5$), and iii) strong ($\lambda = 0.8$). A total of nine different spatial random effects, ϕ were generated using the combinations of three spatial dependence and three variance hyperparameters (Table 2.2).

An individual located approximately at the center of each cluster was selected as the initial seed for both the region-restricted and global models. The epidemic was then allowed to spread according to the infection probabilities determined by Equations (2.4) and (2.5). Since the identifiability issue persisted even after clustering in the ILM simulation study, a fixed susceptibility parameter $\alpha = 0.5$ was considered in the GD-ILM study to avoid the model identifiability issue. We also assumed a minimal and constant infectious pressure from outside the population and set $\epsilon = 0.001$. Considering a constant infectious period $\gamma = 3$ time points for each individuals, we simulated the epidemics up to $t_{max} = 20$ time points. The rest of the parameter values considered for the simulation are given in Table 2.2. These parameter values are chosen arbitrarily to produce informative epidemics.

Table 2.2: Parameters used to simulate epidemics from GD-ILMs.

Parameter	Values
Covariate effect: α_1	2.5, 4.5, 6.5
Spatial decay: β	3.0, 3.5, 4.0, 4.5
Spatial dependence: λ	0.3, 0.5, 0.8
Spatial variance: σ^2	0.5, 0.7, 0.9

The combination of these parameter values yields 108 different models. Epidemics were simulated through these 108 models and replicated 200 times. We created an epidemic curve library, which consisted of $108 \times 200 = 21600$ rows and 20 columns representing 20 time points of the epidemic. The epidemic curves generated by each parameter combination were assigned model names and used to train random forest and deep forest classifiers. The trained classifiers were used to predict the generating model for another set of epidemic library simulated using same procedure but without the assigned model names. We then compare the performances of random forest and deep forest using the misclassification error rates.

2.4.4 GD-ILM Simulation Results

In this section, we evaluate and compare the classification performance of random forest and deep forest models on simulated data from GD-ILMs under various parameter combinations. To look at the spatial dependency in disease transmission, the classifiers were trained using simulated epidemic data over varying λ for a given variance, $\sigma^2 = 0.3, 0.7$, and 0.9. Classification error rates were recorded for each model (Table A4), and the performance comparison focused on both marginal and overall average errors

for both region-restricted and global models as shown in Figure 2.2.

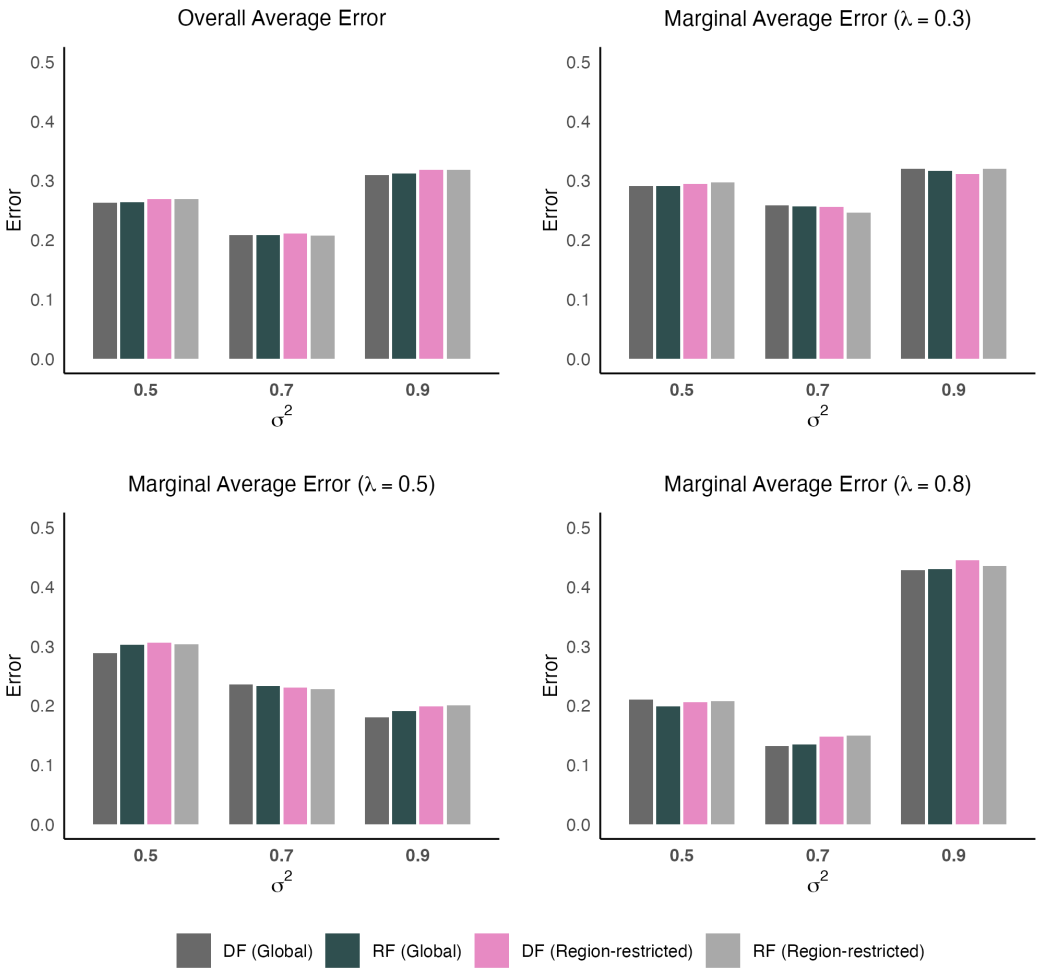


Figure 2.2: Misclassification errors for random forest and deep forest models under global and region-restricted settings across varying spatial variances (σ^2).

The results presented in Figure 2.2 show that both deep forest and random forest

classifiers are robust in classifying epidemic-generating models under both global and region-restricted settings. For both classifier models, performance differences emerge as the spatial dependence parameter (λ) varies, given a fixed spatial variance. Specifically, both deep forest and random forest achieved the highest classification accuracy under both global and region-restricted settings when the spatial variance was moderate ($\sigma^2 = 0.7$) and spatial dependence was high ($\lambda = 0.8$). This suggests that these classification methods may effectively capture the disease transmission dynamics when spatially structured heterogeneity exists and spatial variation and dependence influence the disease transmission mechanism between spatial regions. The GD-ILMs may generate more similar epidemic patterns when both spatial variance and dependence are high ($\sigma^2 = 0.9$ and $\lambda = 0.8$), potentially resulting in a poorer model fit due to the reduced model distinguishability. However, both deep forest and random forest classifiers perform similarly and reasonably well when the spatial variation of the disease transmission mechanism between the regions is low to moderate ($\sigma^2 = 0.5$, 0.7) and the disease transmission process is spatially strongly dependent ($\lambda = 0.8$).

When examining the individual model performances, we observed that model accuracy varies across spatial configurations, with deep forest generally offering improved classification accuracy over random forest, especially under complex spatial structures influenced by spatial dependence, spatial variance, and spatial distance (results are not shown for brevity but available in Table A4 in Appendix). As the spatial variance (σ^2) increases from 0.5 to 0.9, the classification error for both classifiers rises; however, deep forest consistently yields slightly lower or comparable

errors, particularly in the region-restricted setting. Deep forest shows an advantage when the spatial decay parameter is low ($\beta = 3.0$) and the covariate effect is high ($\alpha_1 = 6.5$), indicating its strength in handling localized and strong covariate-driven patterns. As the spatial correlation parameter (λ) increases, both classifiers tend to perform better overall in both global and region-restricted settings for more localized random effects ($\sigma^2 = 0.5, 0.7$), but deep forest continues to slightly outperform random forest in most configurations. The marginal averages across different parameter combinations consistently favor deep forest, especially in settings with high spatial variance and strong covariate influence. Overall, the results highlight deep forest's robustness and adaptability in modelling spatially structured disease transmission dynamics, particularly under more challenging conditions.

2.5 Real Data Analysis

To demonstrate the performance of deep forest classifier on real data, we consider the UK 2001 foot-and-mouth disease (FMD) dataset. FMD primarily affects the cloven-hoofed animals such as sheep and cattle. The dataset comprises information from 131,705 farms distributed across 48 counties. Due to the large size of this dataset, fitting ILMs and GD-ILMs within a Bayesian MCMC framework is computationally impractical. Also, the dataset can be partitioned by geographical regions, such as counties, allowing us to evaluate the effect of spatially varying risk factors. In this study, we consider five neighboring counties from the 48 counties in the UK: Cumbria, Northumberland, North Yorkshire, Dumfries and Galloway, and Durham. The five

counties were chosen because they were among the most affected during the actual 2001 FMD outbreak and are geographically adjacent, facilitating between-county transmission. The data is modeled at the farm level as the individual unit. We assume that the whole farm is infected if one of the animals in the farm is infected. Our aim is to investigate the epidemic transmission dynamics within counties and across the population considered.

2.5.1 FMD-ILMs

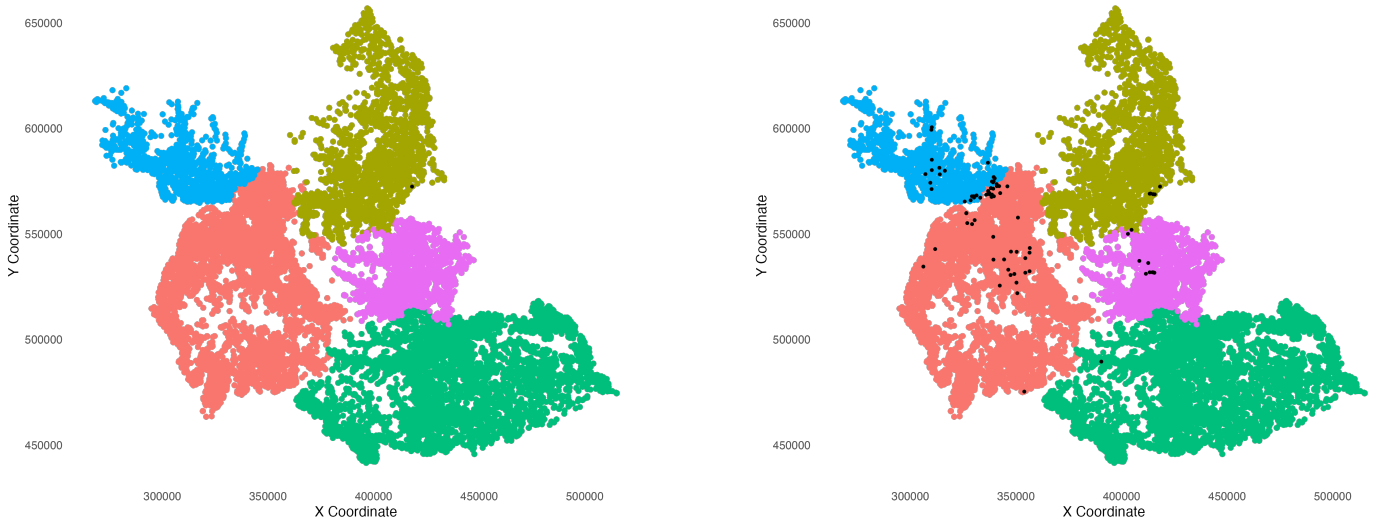
We modified ILM (2.3) to analyze the 2001 FMD data, taking into account both sheep and cattle on the farm, which may have different susceptibilities to the disease. Let N_i^s, and N_i^c represent the number of sheep and cattle on the susceptible farm i. Considering the number of sheep and cattle on susceptible and infected farms as risk factors for the FMD spread, the susceptibility function for i^{th} susceptible farm can be defined by $\Omega_s(i) = \alpha_s N_i^s + \alpha_c N_i^c$ and again we hold the transmissibility function constant, $\Omega_T(j) = 1$. Here α_s and α_c are the susceptiblity parameters for sheep and cattle, respectively. Now, considering the distance-based power-law kernel: $\kappa(i,j) = d_{ij}^{-\beta}$, the probability of susceptible farm i being infected at time t is then given by

$$P_{it} = 1 - \exp\left[-\left\{(\alpha_s N_i^s + \alpha_c N_i^c)\sum_{j\in\mathcal{I}(t)} d_{ij}^{-\beta} + \epsilon\right\}\right]. \tag{2.6}$$

We refer to model (2.6) as power-law FMD-ILM.

FMD Simulation and Model Fitting

To simulate FMD epidemics from FMD-ILMs, the design matrix was constructed using combinations of the parameter values: $\alpha_c \in \{0.01, 0.03, 0.04, 0.05\}$, $\beta \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$. The infectious period was fixed at $\gamma = 4$ days, consistent with previous studies (Deardon et al., 2010; Peitsch et al., 2024) and the susceptibility parameter was arbitrarily fixed at $\alpha_s = 0.0003$ to avoid identifiability issues. Again, a small but fixed infectious pressure, $\epsilon = 0.0001$ was considered. This resulted in $4 \times 5 = 20$ unique models. For each model, epidemics were generated, and replicated 25 times, giving a total of $20 \times 25 = 500$ simulated epidemics. These simulations were conducted over real farm-level geolocations across five counties in UK. The actual FMD infection spread for these five counties at 6 different time points is shown in Figure 2.3.



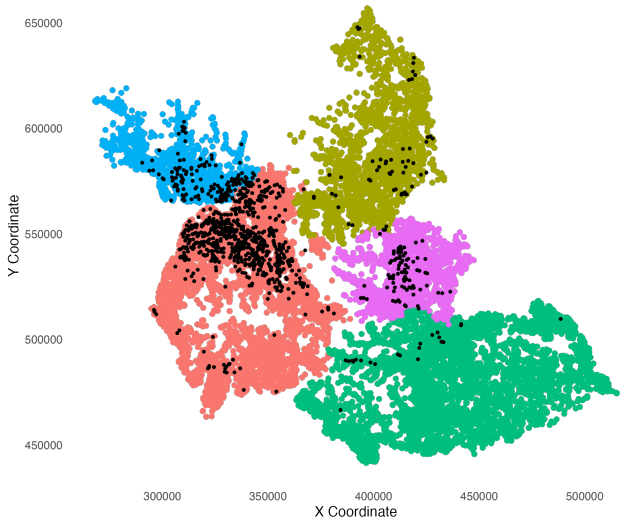
(a) Day 1

(b) Day 20

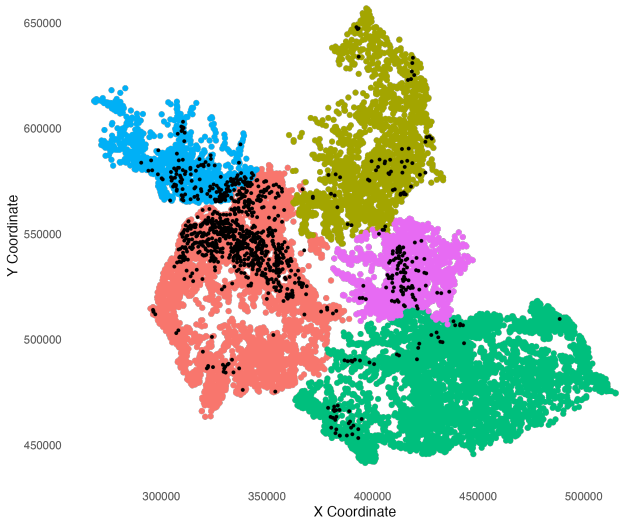
(c) Day 40

(d) Day 60

Figure 2.3: FMD infection spread across selected time points (Days 1–100). The black dots represent the farms infected during the true 2001 FMD epidemic.



(e) Day 80



(f) Day 100

(continued from previous page)

We extracted individual farm locations from the 2001 livestock dataset using the recorded geolocations. In total, there were 15,172 farms with 1,343 infected during the 2001 outbreak across five chosen counties. The first infected farm in each county, as identified from the actual 2001 FMD data, was used as the initial seed to simulate outbreaks. Each simulation ran for 100 days, reflecting the fact that the actual FMD outbreak largely subsided within 80 days in all counties. Each simulation resulted in a time series of infections across all counties, forming epidemic curves. The goal of simulating FMD epidemic was to create a training data for our classifier models, which then can be used to infer the spatio-temporal dynamics of observed FMD data in the 2001 outbreak. We constructed an epidemic curve library containing 500 epidemic

curves. Each curve was assigned a model name, corresponding to its underlying parameter configuration. These data were then used to train random forest and deep forest classifiers to estimate the FMD-ILM parameters.

FMD Data Analysis Results

The trained random forest and deep forest classifiers were used to predict the observed FMD epidemic. The model predictive probabilities and the corresponding models were recorded. Table 2.3 summarizes the model predictive probabilities for various combinations of the transmission parameters α_c and β using random forest and deep forest classifiers.

Table 2.3: Parameter estimates and model predictive probability from random forest and deep forest for the FMD epidemic data.

Parameter estimates		Predictive Probability	
α_c	β	Random forest	Deep forest
0.01	1.0	0.056	0.105
	1.1	0.006	0.020
Marginal predictive probability		0.062	0.125
0.03	1.0	0.012	0.005
	1.1	0.214	0.243
	1.2	0.012	0.018
Marginal predictive probability		0.238	0.266
0.04	1.1	0.124	0.063
	1.2	0.058	0.015
Marginal predictive probability		0.182	0.078
0.05	1.1	0.002	0.018
	1.2	0.506	0.493
Marginal predictive probability		0.508	0.511
Total predictive probability		0.99	0.98

The results show that 9 models out of 20 captured 99% of predictive probability mass under random forest and 98% under deep forest, indicating both classifiers perform similarly when the classifiers are trained using the epidemic simulated from standard FMD-ILM. The rest of the models have less than 1% predictive probability. Both classifiers identified the model with parameter values $\alpha_c = 0.05$, and $\beta = 1.2$ as the best performing model capturing close to 50% probability mass. Given a value of α_c, over varying spatial decay parameter, the deep forest classifier performs better

suggesting that the extra layer of deep forest can extract more spatial information of the disease transmission from the data. For instance, for $\alpha_c = 0.01$, deep forest yields a marginal predictive probability of 12.5% compared to 6.2% from random forest. Similarly, for $\alpha_c = 0.03$, deep forest assigns a higher marginal predictive probability (26.6%) than random forest (23.8%). Conversely, at $\alpha_c = 0.04$, random forest produces a higher marginal predictive probability (18.2%) compared to deep forest (7.8%). Overall, both models demonstrate comparable performances in identifying plausible FMD generating models, with deep forest exhibiting a better performance than random forest in majority of the models considered.

2.5.2 FMD GD-ILMs

We also fitted the GD-ILM to the 2001 FMD data from the UK by modifying equations (2.4) and (2.5). Let N_i^s, and N_i^c represent the number of sheep and cattle on the susceptible farm i, respectively. Considering the number of sheep and cattle on susceptible and infected farms as risk factors for the FMD spread, the susceptibility function for i^{th} susceptible farm is considered as $\Omega_s(i) = \alpha_s N_i^s + \alpha_c N_i^c$ and transmissibility function is set to a constant $\Omega_T(j) = 1$. Here α_s and α_c are the susceptiblity parameters for sheep and cattle, respectively. Now, considering distance-based power-law kernel $\kappa(i,j) = d_{ij}^{-\beta}$, the probability that the susceptible farm i is infected at time t in county k for the FMD transmission is given by

$$P_{ikt} = 1 - \exp\left[-\left\{\exp((\alpha_s N_i^s + \alpha_c N_i^c) + (\alpha_1 X_{i,k} + \phi_k)) \sum_{j \in \mathcal{I}(t,k,\xi(k))} d_{ij}^{-\beta} + \epsilon\right\}\right], \tag{2.7}$$

and

$$P_{ikt} = 1 - \exp\left[-\left\{\exp((\alpha_s N_i^s + \alpha_c N_i^c) + (\alpha_1 X_{i,k} + \phi_k)) \sum_{j \in \mathcal{I}(t)} d_{ij}^{-\beta} + \epsilon\right\}\right]. \tag{2.8}$$

As in the case of general GD-ILM, ϕ_k represents the spatial random effect assumed to follow a multivariate normal distribution, $\phi \sim MN(0, \Sigma)$, where $\Sigma = \sigma^2 \left[(\lambda D + (1-\lambda)I\right]^{-1}$, I is the identity matrix, D is a spatial association matrix, and $\sigma^2 \in \{0.5, 0.7, 0.9\}$ is the variance. α_1 is the parameter for individual-level covariate, and $X(i,k)$ is the covariate associated with susceptible farm i in area k, density of the farm in our case. Similar to the simulation study, we define two versions of the model: region-restricted FMD GD-ILM (2.7) and global FMD GD-ILM (2.8).

FMD Simulation and Model Fitting

We consider the actual coordinates of the farms in the counties to simulate the epidemics. Based on the real FMD data, we chose the first infected farm in each county as the initial seed of the epidemic. The true FMD epidemic almost died out

after the first 80 days in all five counties. Therefore, we set the epidemic length to 100 days and fixed α_s to 0.0001 to avoid identifiability issues following Peitsch et al. (2024). We also fixed the infectious period for FMD to 4 days which is known (Deardon et al., 2010), and the spark term (ϵ) to 0.001. The parameter values were chosen based on the previous literature (Mahsin et al., 2020; Peitsch et al., 2024). The rest of the parameter combination is tabulated in Table 2.4 below.

Table 2.4: Parameters used to simulate FMD using GD-ILMs.

Parameter	Values
Covariate effect: α_1	0.3, 0.6, 0.8
Susceptibility (cattle): α_c	0.01, 0.02, 0.03
Spatial decay: β	1.1, 1.2, 1.3, 1.4, 1.5
Spatial dependence: λ	0.3, 0.5, 0.8
Spatial variance: σ^2	0.5, 0.7, 0.9

A design matrix of size 405×5 was created using the combination of parameter values in Table 2.4. Each row of the design matrix was considered as a model and the epidemic was simulated using each row, recorded as epidemic curve and replicated 25 times. Epidemic library was created using the simulated epidemic curve, which consisted of $405 \times 25 = 10125$ rows and 100 columns. Both random forest and deep forest classifiers are trained using the simulated epidemic library. The trained classifiers are used to predict the observed FMD epidemic curve.

FMD Data Analysis Results

While training the classifiers, we considered epidemic curves as predictors and models as response. We recorded the model predictive probabilities for the 2001 FMD data for

both region-restricted and global models. The results for random forest and deep forest classifiers for FMD GD-ILM are presented in Tables 2.5(a) and 2.5(b), respectively.

The results in Table 2.5(a) show that 19 of 405 FMD GD-ILMs accounted for 92.2% and 91% of the model predictive probability mass under region-restricted and global models, respectively for the random forest classifier. The remaining 386 models have less than 1% predictive probability (results can be found in Appendix). The results in Table 2.5(b) show that 10 of 405 FMD GD-ILMs accounted for 94.8% and 95.2% of the model predictive probability mass under region-restricted and global models, respectively for the deep forest classifier. Examining the marginal predictive probabilities across λ reveals subtle differences in behavior. At $\lambda = 0.3$, deep forest performs slightly better than random forest with deep forest attaining region-restricted and global values of 38% and 40.7%, respectively and random forest reaching 38.6% and 34.6%. Analysis of probability mass concentration when $\lambda = 0.3$ reveals that both methods perform better at moderate values of the variance parameter σ^2 (specifically at 0.5 and 0.7), with deep forest showing particularly high predictive probabilities when $\sigma^2 = 0.5$. This means that classifiers perform better when the data contain a moderate level of random effect variability (σ^2). For example, for deep forest, a single parameter combination with $\lambda = 0.3$, $\sigma^2 = 0.5$, $\alpha_1 = 0.3$, and $\beta = 1.5$ yields a region-restricted predictive probability of 21% and a global probability of 16.5%. Similarly, combinations involving $\alpha_1 = 0.3$ and $\beta = 1.5$ appear to be consistently effective in both methods. The model with spatial parameter $\beta = 1.5$ captured 26.4% and 25% of the model predictive probability mass along with $\lambda = 0.3$, $\sigma^2 = 0.5$ and

0.7, and $\alpha_1 = 0.3$ for region-restricted and global models, respectively for random forest. The model with spatial parameter $\beta = 1.5$ captured 36% and 40% of the model predictive probability mass along with $\lambda = 0.3$, $\sigma^2 = 0.5$ and 0.7, and $\alpha_1 = 0.3$ for region-restricted and global models, respectively for deep forest.

Table 2.5: Parameter estimates and model predictive probability from random forest and deep forest for the FMD epidemic data. Both classifiers estimated $\alpha_c = 0.01$.

(a) Random Forest

Parameter estimates				Predictive Probability	
λ	σ^2	α_1	β	Region-restricted	Global
0.3	0.5	0.3	1.4	0.028	0.010
		0.3	1.5	0.152	0.110
		0.6	1.5	0.024	0.030
	0.7	0.3	1.4	0.028	0.030
		0.3	1.5	0.112	0.140
	0.9	0.6	1.5	0.020	0.020
		0.3	1.5	0.022	0.006
Marginal predictive probability				0.386	0.346
0.5	0.5	0.3	1.5	0.058	0.050
		0.6	1.5	0.012	0.004
	0.7	0.3	1.5	0.046	0.082
	0.9	0.3	1.4	0.012	0.016
		0.3	1.5	0.074	0.070
		0.6	1.5	0.014	0.014
Marginal predictive probability				0.216	0.236
0.8	0.5	0.3	1.5	0.038	0.042
	0.7	0.3	1.4	0.050	0.032
		0.3	1.5	0.176	0.160
		0.6	1.5	0.024	0.042
		0.8	1.5	0.006	0.012
	0.9	0.3	1.5	0.026	0.040
Marginal predictive probability				0.320	0.328
Total predictive probability				0.922	0.91

(b) Deep Forest

Parameter estimates				Predictive Probability	
λ	σ^2	α_1	β	Region-restricted	Global
0.3	0.5	0.3	1.5	0.210	0.165
	0.7	0.3	1.4	0.020	0.007
		0.3	1.5	0.150	0.235
Marginal predictive probability				0.380	0.407
0.5	0.5	0.3	1.5	0.045	0.040
	0.7	0.3	1.5	0.052	0.050
	0.9	0.3	1.5	0.128	0.080
Marginal predictive probability				0.225	0.170
0.8	0.5	0.3	1.5	0.025	0.015
	0.7	0.3	1.4	0.060	0.013
		0.3	1.5	0.250	0.323
	0.9	0.3	1.5	0.007	0.025
Marginal predictive probability				0.343	0.375
Total predictive probability				0.948	0.952

To assess the confidence on model predictions, we analyzed the cumulative predic-

tive probability distributions across classifiers (deep forest and random forest) and models (Global and Region-Restricted). The results are visualized using cumulative distribution plot (Figure 2.4), where x-axis represents ranked FMD generating models in Table 2.5 from highest to lowest probability and y-axis represents the cumulative sum of predicted probabilities, indicating how rapidly predictive certainty accumulates. A steeper curve that reaches high cumulative probability with fewer models suggests better concentration of predictive mass, meaning the classifier confidently assigns high probability to fewer models.

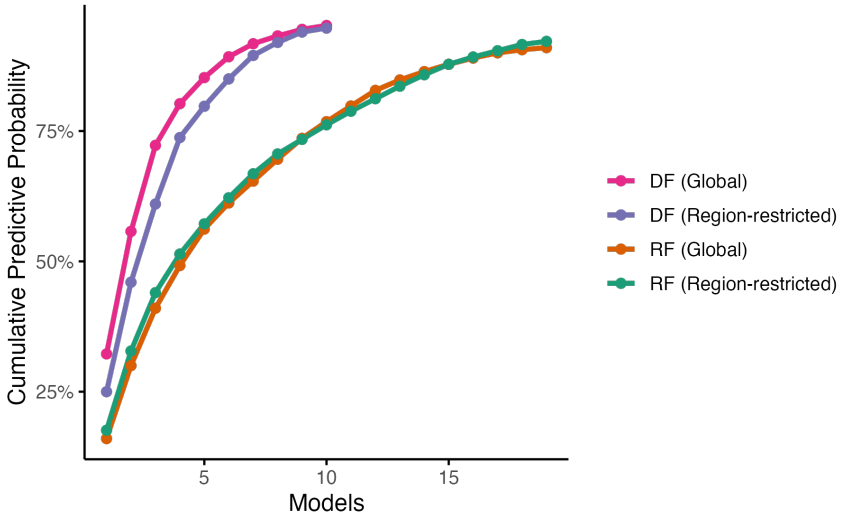


Figure 2.4: Cumulative predictive mass plot for region-restricted and global models.

Across both region-restricted and global models, deep forest demonstrates a notably steeper accumulation of predictive mass. In contrast, random forest curves rise more gradually, reflecting a broader distribution of belief and suggesting greater uncertainty among competing models. Deep forest assigns high probability to a smaller set

of models, an attribute desirable in scenarios such as epidemic modelling. Further supporting this, parameter analysis using Table 2.5 reveals that deep forest consistently selects specific parameter values with higher confidence. For instance, the parameter α_1 is fixed at 0.3 across all selected models in deep forest, whereas random forest selects 0.3, 0.6, and 0.8, suggesting less certainty. Similarly, the β parameter is predominantly 1.5 in both classifiers, though deep forest shows a stronger preference, occasionally selecting 1.4. Additionally, both deep forest and random forest reliably identify $\alpha_c = 0.01$.

Computation Time

We analyzed data from five counties of the UK affected by the FMD outbreak in 2001. The training times for our classifiers were considerably lower than those typically required for Bayesian MCMC approaches. Specifically, training the random forest classifier using the `randomForest` package (Liaw and Wiener, 2002) took approximately 3 minutes, while the deep forest classifier implemented via the `deep-forest` package (Zhou and Feng, 2017) required about 9 minutes (on an Apple MacBook Air with a 1.6 GHz Dual-Core Intel Core i5 processor and 8 GB RAM). In contrast, Bayesian MCMC methods have been reported to require significantly higher computational resources. For example, over 249 hours were needed to generate 20,000 MCMC iterations for just Cumbria county (Malik et al., 2016) while using a stronger machine, an Apple Mac Pro with 2 6-core Intel Xeon 2.93 GHz processors and 12 GB RAM. Although the deep forest classifier required more training time than the random forest classifier,

it achieved higher predictive accuracy, illustrating a trade-off between computation time and classifier performance. Its ability to capture a larger mass of predictive probability with fewer models, along with consistent parameter selection highlights its applicability in identifying plausible epidemic model configurations.

2.6 Discussion and Conclusion

In this study, we investigated the effectiveness of ensemble learning techniques, specifically the random forest and deep forest classifiers, in predicting infectious disease transmission dynamics and estimating parameters using both simulated and FMD outbreak data. The application of these methods to the 2001 UK FMD outbreak demonstrated their potential utility in real-world scenarios, allowing for the identification of the most plausible transmission parameters and contributing to a better understanding of outbreak dynamics. We conclude that the deep forest classifier provides more accurate predictions of FMD epidemic generating models and estimates of model parameters. The simulation study showed similar results, with deep forest performing better in most scenarios.

Despite these promising results, several limitations must be considered. First, the training data were largely based on simulations generated from assumed mechanistic models, which may not fully capture the complexity and stochasticity of real epidemics, including heterogeneities in host behavior, spatial structure, and reporting delays. Second, the clustering approach used for the parameter space reduction in the ILM simulation study, while effective in improving classification accuracy, may lead to loss

of granularity, potentially obscuring subtle but epidemiologically important differences between parameter sets. Third, the inference methods rely on the assumption that the underlying models are correctly specified; model misspecification could lead to biased or misleading estimates. Lastly, the challenge of parameter identifiability remains an important issue, as multiple parameter combinations can produce similar epidemic curves, which complicates precise estimation and requires further methodological development.

Future research should aim to address these limitations by incorporating more realistic and diverse data sources, such as temporal changes in contact patterns, and spatial heterogeneity to better capture the complexities of infectious disease transmission. Enhancing ensemble learning frameworks to provide uncertainty quantification, such as confidence intervals or posterior distributions for parameter estimates, would improve their applicability to decision making. This can be done by using the emulation technique proposed by Pokharel and Deardon (2016). In addition, expanding the analysis to include multiple pathogens, different transmission routes, and various geographic contexts would help evaluate the generalizability and robustness of these methods.

In this work, both deep forest and random forest classifiers are fitted to the global epidemic curve data. As concluded by Pokharel and Deardon (2014) and Peitsch et al. (2024), classifiers may be better informed about disease transmission patterns when trained on spatially stratified epidemic curve data. Therefore, fitting the deep forest classifier to spatially stratified data is of obvious interest. Further to this, we have fitted both classifiers to the complete epidemic curve data. However, in real-time

outbreak scenarios, the complete data may not be available, and there may be a need to inform policy decisions during the early stage of the epidemic. Pokharel and Deardon (2014) suggested that the random forest classifier can be used to predict the progression of an ongoing epidemic, but the data up to the epidemic peak are typically needed to achieve reliable and accurate predictions. This raises important questions: how much information is required for deep forest to provide reasonable and trustworthy predictions? Can deep forest provide accurate predictions with limited data or from the early stages of an epidemic? Investigating these questions could enhance the utility of deep forest classifiers in a public health context and represents a promising direction for future work.

Simulated and real data analyses were performed under the assumption that the infection times and infectious periods for each individual were known. However, this assumption is not realistic in practice, as the true infection times and infectious periods are typically difficult or even impossible to accurately determine. Consequently, further investigation is required to assess the performance and utility of the proposed method under uncertainty about the timing and duration of the infection.

Standard random forest and deep forest classifiers randomly sample covariates for splitting, disregarding the temporal dependency inherent in epidemic curves. However, the number of infections at later time points is influenced by the number of infections at earlier stages, at least until the epidemic reaches its peak. Ignoring these temporal dependency may lead to poor model fit and can represent the true predictive performance of classifiers. Therefore, further work is needed to evaluate model accuracy when temporal dependencies in the epidemic trajectory are explicitly

incorporated.

For simplicity, a uniformly gridded population was used in the ILM simulation study. However, this representation may not accurately reflect real-world population structures. In contrast, the GD-ILM simulation employed a clustered population. Therefore, when comparing results between the ILM and GD-ILM simulations, it is important to account for differences in the underlying population structures. To support a more consistent and meaningful comparison, the ILM simulations can also be conducted using a clustered population.

In summary, ensemble learning methods, especially deep forest classifiers show considerable promise for epidemic model classification and parameter estimation. With further refinement and validation, these approaches could become valuable components of infectious disease surveillance and control strategies.

Chapter 3

Conclusion

Infectious disease modelling frameworks such as individual-level models (ILMs) are well-established and allow for the incorporation of individual-level covariates. However, spatial heterogeneity also plays a critical role in disease dynamics, a factor that geographically-dependent ILMs (GD-ILMs) can account for, but standard ILMs cannot fully capture. This thesis aimed to evaluate the effectiveness of random forest and deep forest classification-based inference for ILMs and GD-ILMs to bypass the computationally intensive likelihood calculations of the Bayesian framework.

In Chapter 2, we used both simulated data and the 2001 UK FMD outbreak data to demonstrate the performance of the random forest and deep forest classifiers. For data analysis, we considered five neighboring counties: Cumbria, Northumberland, North Yorkshire, Dumfries and Galloway, and Durham, selected from the 48 counties in the UK. We found that the deep forest classifier provides more accurate predictions of FMD epidemic generating models and estimates of model parameters. The simulation study showed the similar results with deep forest performing better in most scenarios.

3.1 Future Work

In this work, we only fitted the classifiers to global epidemic curve data. Previous literature showed that the ensemble learning approaches such as random forest (Pokharel and Deardon, 2014; Peitsch et al., 2024) and gradient boosting (Peitsch et al., 2024) perform better when trained using spatially stratified epidemic curves. Therefore, we could further explore the performance of deep forest using stratified epidemic curves, and compare it with other ensemble learning approaches. The deep forest classifier typically utilizes random forest and extra trees as its default base classifiers. Its performance can be enhanced by appending an additional predictor to the end of the deep forest architecture (Zhou and Feng, 2017). Another potential avenue for future research could involve enhancing ensemble learning frameworks to provide uncertainty quantification, such as confidence intervals or posterior distributions for parameter estimates, which would improve their applicability to decision making. This can be done by using the emulation technique proposed by Pokharel and Deardon (2016), where the emulator can be constructed using the vector model predictive probabilities from the classifiers.

In addition, expanding the analysis to include multiple pathogens, different transmission routes, and various geographic contexts could help evaluate the generalizability and robustness of these methods. For instance, the GD-ILM models can be expanded to study SEIRS (susceptible-exposed-infected-removed-susceptible) framework, which could be common for some infectious diseases where reinfection is possible. We could further apply these models to diseases that primarily affect humans, such as influenza

and tuberculosis, and compare their transmission dynamics. Different kernels other than the power-law kernel can also be used to evaluate the ability of classifiers in estimating the model parameters, as well as the transmission kernel.

Another avenue of future research is to explore other ILMs like behavioural change ILM proposed by Ward et al. (2023), where Bayesian methods were used in model fitting. In such cases, applying the machine learning approaches used in this work could improve model fitting efficiency.

Bibliography

Almutiry, W. and R. Deardon (2020). Incorporating contact network uncertainty in individual level models of infectious disease using approximate bayesian computation. *The International Journal of Biostatistics 16*(1), 20170092. (Cited on page 5.)

Amiri, L., M. Torabi, and R. Deardon (2023). Analyzing COVID-19 data in the Canadian province of Manitoba: A new approach. *Spatial Statistics 55*, 100729. (Cited on pages 2, 4 and 5.)

Amiri, L., M. Torabi, R. Deardon, and M. Pickles (2021). Spatial modeling of individual-level infectious disease transmission: tuberculosis data in Manitoba, Canada. *Stat Med 40*(7), 1678–1704. (Cited on pages 2, 4 and 5.)

Augusta, C., R. Deardon, and G. Taylor (2019). Deep learning for supervised classification of spatial epidemics. *Spatial and Spatio-temporal Epidemiology 29*, 187–198. (Cited on pages 5 and 13.)

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32. (Cited on pages 6, 14, 20 and 21.)

Chis Ster, I. and N. Ferguson (2007). Transmission Parameters of the 2001 Foot and Mouth Epidemic in Great Britain. *PLoS ONE 2*(6), e502. (Cited on pages 2 and 12.)

Deardon, R., S. Brooks, T. Grenfell, M. Keeling, M. Tildesley, N. Savill, D. Shaw, and M. Woolhouse (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica 20*, 239–261. (Cited on pages 2, 3, 12, 16, 17, 37 and 44.)

Geurts, P., D. Ernst, and L. Wehenkel (2006). Extremely randomized trees. *Machine Learning 63*(1), 3–42. (Cited on page 21.)

Guo, Y., S. Liu, Z. Li, and X. Shang (2018). Bcdforest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC Bioinformatics 19*, 118. (Cited on pages 14 and 15.)

Hughes, G., N. McRobert, L. Madden, and S. Nelson (1997). Validating mathematical models of plant-disease progress in space in time. *IMA Journal of Mathematics Applied in Medicine and Biology 14*, 85–112. (Cited on page 2.)

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. (Cited on pages 6 and 20.)

Keeling, M., M. Woolhouse, D. Shaw, L. Matthews, M. Chase-Topping, D. Haydon, S. Cornell, J. Kappey, J. Wilesmith, and B. Grenfell (2001). Dynamics of the 2001

UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science 294*, 813–817. (Cited on page 2.)

Kermack, W. and A. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a Mathematical and Physical character 115*(772), 700–721. (Cited on page 2.)

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22. (Cited on pages 23 and 48.)

Liu, Z., R. Deardon, Y. Fu, T. Ferdous, T. Ware, and Q. Cheng (2021). Estimating parameters of two-level individual-level models of the COVID-19 epidemic using ensemble learning classifiers. *Frontiers in Physics 8*(11), Article 602722. doi: 10.3389/fphy.2020.602722. (Cited on page 13.)

Mahsin, M., R. Deardon, and P. Brown (2020). Geographically dependent individual-level models for infectious diseases transmission. *Biostatistics 23*(1), 1–17. (Cited on pages 2, 3, 13, 18, 19 and 44.)

Malik, R., R. Deardon, and G. P. S. Kwong (2016). Parameterizing spatial models of infectious disease transmission that incorporate infection time uncertainty using sampling-based likelihood approximations. *PLoS ONE 11*(1). (Cited on pages 3, 5 and 48.)

Nsoesie, E., R. Beckman, M. Marathe, and B. Lewis (2011). Prediction of an epidemic

curve: a supervised classification approach. *Statistical Communication in Infectious Diseases 3*(1). (Cited on page 13.)

Peitsch, J., G. Pokharel, and S. Hossain (2024). Ensemble learning methods of inference for spatially stratified infectious disease systems. *International Journal of Biostatistics 20*(2), 507–529. (Cited on pages 2, 3, 5, 14, 26, 28, 30, 37, 44, 50 and 54.)

Pokharel, G. and R. Deardon (2014). Supervised learning and prediction of spatial epidemics. *Spatial and Spatio-temporal Epidemiology 11*, 59–77. (Cited on pages 3, 5, 13, 14, 25, 27, 50, 51 and 54.)

Pokharel, G. and R. Deardon (2016). Gaussian process emulators for spatial individual-level infectious disease. *Canadian Journal of Statistics 44*(4), 480–501. (Cited on pages 2, 3, 5, 50 and 54.)

Ward, C., R. Deardon, and A. M. Schmidt (2023). Bayesian modeling of dynamic behavioral change during an epidemic. *Infectious Disease Modelling 8*(4), 947–963. (Cited on page 55.)

Zhou, Z. H. and J. Feng (2017). Deep forest: Towards an alternative to deep neural networks. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence IJCAI'17*, 3553–3559. (Cited on pages 7, 8, 9, 14, 15, 21, 22, 23, 48 and 54.)

Appendix A

Appendix

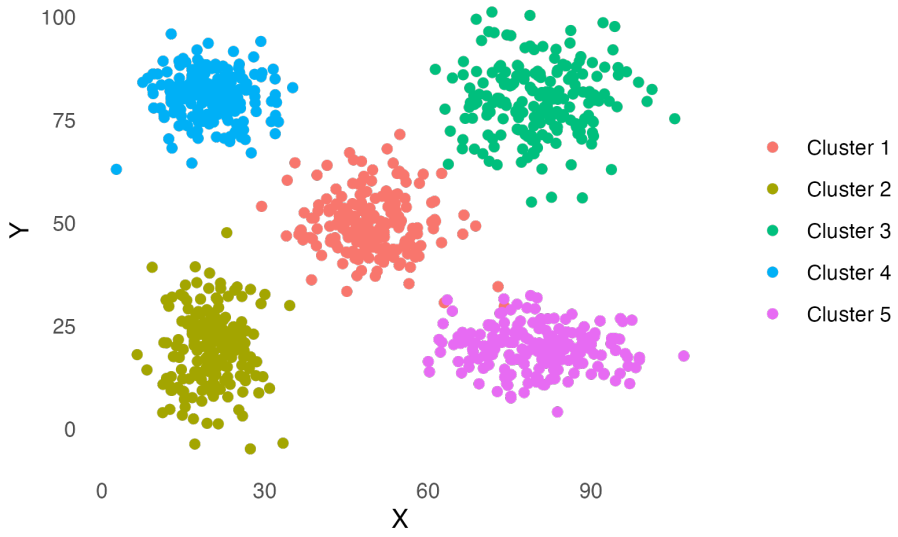


Figure A1: Simulated population for GD-ILM simulation study. Each circle represents an individual.

Table A1: Parameter values, average and 95% percentile interval of model prediction error from random forest and deep forest classifiers for simulated data from power-law kernel ILM.

Model	ϵ	γ	α	β	Random Forest Average (95 % PI)	Deep Forest Average (95 % PI)
M1	0.02	2	0.55	1.00	0.13 (0.05,0.21)	0.24 (0.16,0.32)
M10	0.02	3	0.33	2.00	0.416 (0.25,0.55)	0.379 (0.25,0.49)
M11	0.02	2	0.55	2.00	0.153 (0.06,0.2)	0.1 (0.05,0.15)
M12	0.02	3	0.78	2.50	0.233 (0.16,0.34)	0.219 (0.17,0.31)
M13	0.00	2	0.33	2.00	0.336 (0.25,0.46)	0.329 (0.22,0.48)
M14	0.02	3	0.55	2.50	0.367 (0.22,0.5)	0.333 (0.25,0.4)
M15	0.02	2	0.33	3.00	0.077 (0.04,0.15)	0.062 (0.03,0.1)
M16	0.02	2	0.10	2.00	0.039 (0.03,0.06)	0.039 (0.02,0.07)
M17	0.02	2	1.00	1.50	0.176 (0.09,0.3)	0.171 (0.09,0.26)
M18	0.00	3	0.55	1.00	0.399 (0.23,0.56)	0.212 (0.14,0.31)
M19	0.00	3	1.00	2.50	0.187 (0.12,0.28)	0.184 (0.11,0.25)
M2	0.00	2	0.78	2.50	0.383 (0.3,0.48)	0.383 (0.29,0.55)
M20	0.00	3	1.00	1.50	0.323 (0.15,0.55)	0.288 (0.22,0.38)
M21	0.00	2	0.78	3.00	0.107 (0.06,0.17)	0.093 (0.05,0.14)
M22	0.00	3	0.10	2.50	0.089 (0.05,0.15)	0.077 (0.04,0.14)
M23	0.00	2	0.10	2.50	0.322 (0.26,0.41)	0.284 (0.18,0.38)
M24	0.00	2	0.78	2.00	0.088 (0.06,0.12)	0.088 (0.03,0.13)
M25	0.00	3	0.55	2.50	0.168 (0.1,0.22)	0.133 (0.06,0.17)
M26	0.02	3	1.00	1.00	0.008 (0,0.02)	0.002 (0,0.01)
M27	0.00	2	0.78	1.50	0.213 (0.09,0.37)	0.342 (0.24,0.46)
M28	0.00	3	0.33	1.00	0.759 (0.48,1)	0.438 (0.41,0.49)
M29	0.02	2	0.78	2.00	0.25 (0.15,0.37)	0.183 (0.14,0.23)
M3	0.00	2	0.55	3.00	0.087 (0.04,0.13)	0.074 (0.03,0.13)
M30	0.00	3	0.10	2.00	0.142 (0.08,0.22)	0.13 (0.07,0.24)
M31	0.00	2	0.10	2.00	0.183 (0.13,0.23)	0.174 (0.13,0.23)
M4	0.02	2	1.00	2.00	0.18 (0.13,0.24)	0.073 (0.02,0.13)
M5	0.02	3	0.10	2.50	0.009 (0,0.03)	0.005 (0,0.03)
M6	0.02	2	0.10	3.00	0 (0,0)	0.001 (0,0.01)
M7	0.00	3	0.10	3.00	0.246 (0.2,0.32)	0.23 (0.17,0.29)
M8	0.00	2	0.10	3.00	0.03 (0.01,0.05)	0.081 (0.01,0.24)
M9	0.02	2	0.33	2.50	0.041 (0.02,0.06)	0.035 (0.02,0.06)
Overall average					0.198 (0.12, 0.28)	0.174 (0.12, 0.24)

Table A2: Confusion matrix for simulated ILM using random forest.

	M1	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M2	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M3	M30	M31	M4	M5	M6	M7	M8	M9
M1	89	0	0	0	0	0	0	0	2	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M10	0	63	0	18	0	40	0	0	0	0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M11	0	0	80	5	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	2	0	0	0	0	0	0	0	0	0
M12	0	9	3	71	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M13	0	0	0	0	73	0	0	0	0	0	0	15	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
M14	0	19	0	0	1	55	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M15	0	0	0	0	0	0	93	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M16	0	0	0	0	0	0	6	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M17	0	0	0	0	0	0	0	0	78	1	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M18	11	0	0	0	0	0	0	0	0	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M19	0	9	0	6	0	0	0	0	0	0	85	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M2	0	0	0	0	22	5	0	0	0	0	6	65	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
M20	0	0	0	0	0	0	0	0	20	9	0	0	77	0	0	0	0	0	0	7	8	0	0	0	0	0	0	0	0	0	0
M21	0	0	0	0	0	0	0	0	0	0	0	0	0	92	0	0	0	6	0	0	0	0	7	0	0	0	0	0	0	0	4
M22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	1	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0
M23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	59	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0
M24	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	89	0	0	0	0	6	0	0	0	0	0	0	0	0	0
M25	0	0	0	0	4	0	0	0	0	0	0	1	0	3	0	0	0	81	0	0	0	0	0	0	0	0	0	0	0	0	0
M26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
M27	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	76	49	0	0	0	0	3	0	0	0	0	0
M28	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	15	43	0	0	0	0	0	0	0	0	0	0
M29	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	85	0	0	0	14	0	0	0	0	0
M3	0	0	0	0	0	0	1	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0
M30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87	10	0	0	0	0	0	0
M31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	82	0	0	0	0	0	0
M4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	7	0	0	0	83	0	0	0	0	0
M5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
M6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
M7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	77	4	0
M8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	17	0	0	0	0	0	0	0	2	2	0	0	0	16	96	0
M9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	96

Table A3: Confusion matrix for simulated ILM using deep forest.

	M1	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M2	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M3	M30	M31	M4	M5	M6	M7	M8	M9
M1	76	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M10	0	63	0	3	0	28	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M11	0	0	89	4	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M12	0	15	4	74	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M13	0	0	0	0	65	1	0	0	0	0	0	27	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
M14	0	27	0	0	0	70	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M15	0	0	0	0	0	0	94	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M16	0	0	0	0	0	0	4	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M17	0	0	0	0	0	0	0	0	79	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M18	24	0	0	0	0	0	0	0	0	75	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M19	0	3	0	5	0	0	0	0	0	0	86	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M2	0	1	0	0	11	7	0	0	0	0	10	66	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
M20	0	0	0	0	0	0	0	0	12	0	0	0	75	0	0	0	0	0	0	4	9	0	0	0	0	0	0	0	0	0	0
M21	0	0	0	0	0	0	0	0	0	0	0	0	0	91	0	0	0	3	0	0	0	0	6	0	0	0	0	0	0	0	0
M22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
M23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	64	0	0	0	0	0	0	0	0	0	0	0	0	19	17	0
M24	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	88	0	0	0	0	4	0	0	0	0	0	0	0	0	0
M25	0	0	0	0	13	0	0	0	0	0	0	2	0	2	0	0	0	83	0	0	0	0	0	0	0	0	0	0	0	0	0
M26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
M27	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	74	23	0	0	0	0	2	0	0	0	0	0
M28	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	38	57	0	0	0	0	0	0	0	0	0	0
M29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	80	0	0	0	14	0	0	0	0	0
M3	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	95	1	0	0	0	0	0	0	0
M30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	9	0	0	0	0	2	0
M31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	11	82	0	0	0	0	2	0
M4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	96	0	0	0	0	0
M5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
M6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
M7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	78	16	0
M8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	98	0
M9	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96

Table A4: Parameter values and model classification error from random forest and deep forest for simulated data from GD-ILMs for given values of σ^2.

Parameter values			$\sigma^2 = 0.5$				$\sigma^2 = 0.7$				$\sigma^2 = 0.9$			
			Global		Region-restricted		Global		Region-restricted		Global		Region-restricted	
λ	α_1	β	RF	DF	RF	DF	RF	DF	RF	DF	RF	DF	RF	DF
0.3	2.5	3.0	0.150	0.170	0.135	0.120	0.140	0.145	0.185	0.195	0.295	0.325	0.350	0.335
		3.5	0.265	0.240	0.265	0.260	0.265	0.265	0.260	0.250	0.345	0.350	0.310	0.305
		4.0	0.420	0.490	0.495	0.480	0.300	0.315	0.235	0.270	0.445	0.465	0.375	0.375
		4.5	0.365	0.385	0.415	0.405	0.220	0.220	0.245	0.275	0.400	0.375	0.375	0.325
	4.5	3.0	0.120	0.125	0.090	0.075	0.110	0.115	0.175	0.185	0.215	0.255	0.210	0.235
		3.5	0.255	0.250	0.265	0.270	0.305	0.295	0.285	0.300	0.285	0.300	0.345	0.380
		4.0	0.400	0.340	0.355	0.390	0.310	0.285	0.290	0.290	0.360	0.340	0.345	0.335
		4.5	0.420	0.395	0.420	0.420	0.340	0.350	0.315	0.340	0.330	0.325	0.375	0.350
	6.5	3.0	0.150	0.155	0.170	0.180	0.215	0.205	0.160	0.190	0.225	0.225	0.210	0.160
		3.5	0.225	0.265	0.215	0.200	0.245	0.255	0.215	0.205	0.285	0.310	0.265	0.285
		4.0	0.355	0.330	0.330	0.330	0.315	0.335	0.300	0.280	0.275	0.265	0.300	0.270
		4.5	0.360	0.345	0.410	0.400	0.310	0.315	0.285	0.290	0.330	0.305	0.380	0.380
Marginal average			0.290	0.291	0.297	0.294	0.256	0.258	0.246	0.256	0.316	0.320	0.320	0.311
0.5	2.5	3.0	0.315	0.300	0.280	0.305	0.280	0.250	0.210	0.210	0.145	0.095	0.175	0.185
		3.5	0.510	0.430	0.530	0.530	0.355	0.385	0.400	0.410	0.285	0.305	0.270	0.270
		4.0	0.365	0.370	0.430	0.430	0.170	0.165	0.180	0.170	0.205	0.185	0.225	0.270
		4.5	0.355	0.370	0.255	0.260	0.235	0.215	0.120	0.115	0.160	0.185	0.245	0.265
	4.5	3.0	0.145	0.145	0.215	0.195	0.210	0.190	0.240	0.210	0.090	0.080	0.100	0.105
		3.5	0.445	0.390	0.330	0.355	0.275	0.285	0.330	0.350	0.230	0.190	0.240	0.170
		4.0	0.400	0.365	0.410	0.405	0.195	0.190	0.245	0.220	0.275	0.260	0.215	0.220
		4.5	0.255	0.260	0.285	0.280	0.215	0.210	0.160	0.165	0.255	0.270	0.300	0.285
	6.5	3.0	0.160	0.150	0.170	0.170	0.200	0.235	0.265	0.290	0.070	0.075	0.040	0.055
		3.5	0.190	0.170	0.190	0.195	0.150	0.160	0.135	0.165	0.110	0.085	0.160	0.155
		4.0	0.225	0.240	0.280	0.270	0.245	0.270	0.250	0.245	0.225	0.165	0.200	0.155
		4.5	0.260	0.265	0.260	0.270	0.260	0.265	0.190	0.210	0.240	0.260	0.230	0.250
Marginal average			0.302	0.288	0.303	0.305	0.233	0.235	0.227	0.230	0.191	0.180	0.200	0.199
0.8	2.5	3.0	0.020	0.045	0.065	0.055	0.050	0.045	0.050	0.045	0.380	0.425	0.340	0.320
		3.5	0.190	0.175	0.145	0.145	0.110	0.100	0.160	0.135	0.410	0.395	0.470	0.485
		4.0	0.235	0.225	0.300	0.275	0.160	0.180	0.220	0.205	0.515	0.505	0.490	0.485
		4.5	0.020	0.030	0.010	0.005	0.250	0.245	0.180	0.190	0.565	0.550	0.550	0.630
	4.5	3.0	0.010	0.005	0.045	0.045	0.030	0.025	0.005	0.005	0.310	0.250	0.360	0.335
		3.5	0.205	0.190	0.240	0.250	0.080	0.085	0.090	0.115	0.390	0.400	0.365	0.330
		4.0	0.450	0.495	0.400	0.410	0.225	0.245	0.250	0.255	0.425	0.460	0.430	0.465
		4.5	0.475	0.445	0.515	0.545	0.285	0.285	0.315	0.330	0.545	0.555	0.510	0.570
	6.5	3.0	0.015	0.025	0.025	0.025	0.055	0.040	0.055	0.035	0.320	0.295	0.355	0.415
		3.5	0.095	0.115	0.085	0.090	0.030	0.025	0.085	0.065	0.445	0.420	0.360	0.325
		4.0	0.305	0.340	0.265	0.250	0.080	0.075	0.090	0.080	0.420	0.410	0.480	0.485
		4.5	0.365	0.430	0.395	0.375	0.260	0.230	0.295	0.305	0.435	0.475	0.505	0.495
Marginal average			0.199	0.210	0.207	0.206	0.135	0.132	0.150	0.147	0.430	0.428	0.435	0.445
Overall average			0.264	0.263	0.269	0.269	0.208	0.208	0.208	0.211	0.312	0.309	0.318	0.318