A Machine-Learning Approach to Earthquake Declustering: Application to New Zealand Earthquake Catalogue

Submitted By:

Md Ashraf

Admission Number: 23MC0049

M.Sc (Tech) Applied Geophysics

Under the Guidance of:

Dr. Niptika Jana

Indian Institute of Technology (ISM) Dhanbad

14 November 2025

ACKNOWLEDGEMENTS

I would like to sincerely thank Dr. Niptika Jana, Department of Applied Geophysics, Indian Institute of Technology (Indian School of Mines) Dhanbad, for her patient guidance, valuable advice, and constant encouragement throughout the completion of this work.

I am also deeply grateful to Prof. Soumen Maiti, Head of the Department of Applied Geophysics at the Indian Institute of Technology (Indian School of Mines) Dhanbad, for providing the necessary facilities and supporting me in completing this part of my dissertation.

A special note of gratitude goes to Aditi Seal, whose constant help, technical guidance, and thoughtful suggestions were instrumental at many crucial stages. Her willingness to assist, clarify concepts, and review parts of my work added immense value to this thesis. I am sincerely grateful for her time and support.

I would also like to acknowledge the faculty and staff members of the department for their cooperation, academic inputs, and for maintaining a smooth research environment.

Finally, I would like to express my deepest gratitude to my parents, whose unconditional love, motivation, and unwavering belief in my abilities have been the driving force behind my academic journey. Their support has given me the strength to overcome moments of doubt and inspired me to keep moving forward.

Dr. Niptika Jana
Assistant Professor
Department of Applied Geophysics

Md Ashraf
Master of Science and Technology
Department of Applied Geophysics
Admission No.: 23MC0049

Date: 14/11/2025
Place: Dhanbad, Jharkhand

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Earthquake catalogues record both independent background events and dependent events such as aftershocks. For applications in seismic hazard studies, statistical seismology, and earthquake forecasting, it is necessary to distinguish between these types through a process called declustering. Methods such as those developed by Gardner and Knopoff, Reasenberg, and the nearest-neighbour distance (NND) approach have been quite useful in identifying aftershock sequences. However, these techniques still face a few issues. They depend on fixed time and distance windows or specific threshold values, and this often causes problems when earthquake clusters overlap with independent events, leading to misclassification in complex seismic regions.

This problem is particularly challenging in New Zealand, where seismic activity is frequent and complex. This study uses machine learning techniques to decluster the New Zealand earthquake catalogue covering 1980–2024. To train our supervised machine learning models, we require labelled earthquake events; therefore, we generate synthetic catalogues using the Epidemic-Type Aftershock Sequence (ETAS) model. From both synthetic and real catalogues, we compute NND-based metrics and derive features including rescaled time and distance, magnitude differences and counts of siblings and offspring. Using these features, we train and test different classifiers, including Random Forests, support vector machines, and gradient boosting models. We then apply the best-performing models to the New Zealand catalogue. By combining ETAS-based synthetic training data with over four decades of New Zealand seismic observations, this study provides a refined view of earthquake clustering patterns. The decluttered catalogue we obtain offers a stronger foundation for seismicity analysis and contributes to more robust hazard assessment in this tectonically active region.

Keywords: Declustering; Earthquake Catalog; ETAS Model; Nearest-Neighbor Method; Machine Learning

CHAPTER 1: INTRODUCTION

Earthquakes are considered among the most damaging natural events, with the potential to cause significant damage to infrastructure, loss of life, and social and economic problems. Understanding the fundamental mechanisms of seismicity and precisely predicting future earthquake events are essential goals in geophysical research. An important aspect of this Work is the capacity to differentiate between independent background seismicity and dependent seismic events, such as aftershocks induced by mainshocks and foreshocks that precede significant earthquakes.

The process of separating independent events, that is, background, from the dependent events (aftershocks) is known as declustering. This is important for various applications like seismic hazard assessment, tectonic stress pattern characterisation, statistical seismology and earthquake forecasting.

Generally, it is assumed that background seismicity follows the Poisson process in which events occur randomly and independently in time, while cluster events occur with temporal and spatial dependencies.

Traditional declustering methods rely on deterministic space-time windows or statistical nearest-neighbour approaches. While these techniques have been instrumental in advancing seismological studies, they possess significant shortcomings. Window-based methods, such as those proposed by Gardner and Knopoff (1974) and Reasenberg (1985), require subjective parameter tuning and often fail to capture the complexity of real seismic sequences, particularly in regions with swarm-like activity or overlapping aftershock sequences.

1.1 Problem Statement and Objectives

Despite the fact that several declustering techniques have already been established, most of them still encounter difficulties when applied to regions with complex and overlapping seismic activity. The primary issue is that traditional space–time window or statistical approaches employ fixed parameters, which may not be equally effective for all types of seismic environments. For example, in a tectonically active region like **New Zealand**, earthquakes occur due to different processes — subduction, strike-slip, and crustal faulting — all happening together. Because of this, aftershock sequences often overlap and make it hard to correctly separate background events from triggered ones.

To overcome these problems, this study applies a **machine-learning-based approach** that learns patterns from both real and synthetic earthquake data. Synthetic labelled catalogues are created using the **Epidemic-Type Aftershock Sequence (ETAS)** model, which helps in training the algorithms to identify the difference between background and aftershock events. Important features, such as inter-event time, distance, magnitude difference, and neighbour relationships, are extracted and used as inputs for various machine learning models.

The main objectives of the study are:

1. To generate synthetic earthquake catalogues using the ETAS model with clear background and triggered event labels.

2. To extract statistical and physical features from both synthetic and real earthquake catalogs.

3. To train and test different supervised machine learning models such as Random Forest, SVM, and Gradient Boosting.

4. To apply the best-performing model to the **New Zealand earthquake catalog (1980–2024)**.

5. To compare machine-learning-based declustering results with traditional approaches and study their effect on the regional seismic pattern.

1.2 Scope of the Study

The primary goal of this work is to create and evaluate a machine learning-based framework for earthquake declustering.

The study uses the **New Zealand earthquake data from 1980 to 2024**, obtained from the GeoNet catalogue.

The research focuses on data analysis, feature extraction, and classification using statistical and machine learning methods.

The study includes:

- Collecting and cleaning the earthquake catalogue data.
- Generating synthetic labelled data using ETAS simulation.
- Computing nearest-neighbour-based and time-magnitude-based features.
- Building and testing machine learning models for declustering.
- Comparing ML results with standard declustering techniques.

After generating the synthetic data through ETAS and combining them with observed real earthquake records,

this work aims to develop a more reliable method for distinguishing between background and triggered events. The outcome will help improve seismic hazard studies and provide a clearer understanding of earthquake clustering in tectonically active regions, such as New Zealand.

CHAPTER 2: LITERATURE SURVEY

2.1 Traditional Declustering Methods

The earliest declustering algorithms were based on **spatial and temporal window techniques**.
The **Gardner and Knopoff (1974)** method is one of the most commonly used empirical approaches for earthquake declustering. It sets fixed time and distance windows around each mainshock, and any smaller event that occurs within these limits is treated as an aftershock. Although the method is simple and works efficiently, it relies on region-specific constants and performs poorly in areas where seismic sequences are complex or overlapping.

The **Reasenberg (1985)** algorithm improved upon this idea by introducing a probabilistic framework where clusters are formed if the temporal and spatial distances between events satisfy predefined conditions. This method dynamically links events into clusters but still depends heavily on parameter tuning. As a result, the output can vary considerably across regions, and reproducibility is often limited.

Later studies, such as that by **Uhrhammer (1986)**, tried to improve these methods by introducing hierarchical clustering and magnitude-dependent scaling. However, even with these refinements, window-based techniques still depend heavily on empirical parameters and often find it difficult to clearly separate independent earthquakes from nearby

Aftershocks.

2.2 Statistical and Model-Based Approaches

A major advancement in earthquake declustering occurred with the introduction of the Epidemic-Type Aftershock Sequence (ETAS) model by **Ogata (1988)**.
The ETAS model treats earthquakes as a **branching process**, where each event can trigger its own aftershocks according to parameters that describe productivity, spatial decay, and temporal decay (often following the Omori–Utsu law). The model naturally separates background events from triggered ones by estimating the probability that an event was generated independently.

ETAS has become a fundamental tool in statistical seismology because of its physical and probabilistic foundation. However, later studies have pointed out that parameter estimation in the ETAS model can often be biased, particularly when the earthquake catalog is incomplete or when the stationarity assumption does not hold true.

D.S. Harte, in his study "Bias in Fitting the ETAS Model," shows and highlights that factors such as the chosen magnitude cut-off, the length of the catalogue, and inaccuracies in the background seismicity rate significantly affect our estimated productivity and temporal parameters. In simpler terms, while the ETAS model is powerful and widely used today for generating synthetic earthquake catalogues, it still requires careful calibration. Without proper tuning of parameters and validation, the results can lead to biased interpretations of seismic clustering.

Another important development was the **Nearest-Neighbour Distance (NND)** method proposed by **Zaliapin and Ben-Zion (2013)**. This method changes earthquake occurrences into a rescaled space–time–magnitude form and calculates a nearest-neighbour value to study how events are clustered. The NND values usually form a **bimodal pattern**, which helps to separate background earthquakes from aftershock clusters.

2.3 Limitations of Classical Approaches

Despite their contribution, classical and model-based approaches share several limitations:

- **Parameter Sensitivity:** Results are strongly influenced by chosen thresholds or estimated parameters.
- **Overlapping Clusters:** Many models cannot clearly distinguish overlapping aftershock sequences, especially in the case of NND.
- **Assumption of Stationarity:** Real earthquake processes often violate the assumption of a constant background rate.

These challenges have motivated the exploration of **data-driven machine-learning (ML) approaches** that can automatically learn patterns and relationships from seismic data without relying solely on fixed rules.

2.4 Machine Learning Approaches to Declustering

Recent studies have introduced machine-learning methods as a promising alternative to traditional techniques.
Since real earthquake catalogues do not contain labels identifying background and aftershock events, researchers use synthetic catalogues **generated by ETAS models** to produce labelled data for supervised learning.

Aden-Antoniów et al. (2022) introduced a flexible RF model that was trained on synthetic earthquake catalogues generated using the ETAS model. The RF classifier used several input features, including magnitude differences, time and distance between events, and nearest-neighbour parameters. Their results showed that the Random Forest model performed better than traditional declustering methods, mainly because it could capture nonlinear relationships between features and adapt well to different seismic regions.

Similarly, **Kothari et al. (2023)** applied a supervised RF framework to decluster regional earthquake catalogues. They emphasised the flexibility of tree-based models and the ability to evaluate feature importance, helping to interpret which parameters contribute most to event classification. Their results indicated that machine learning–based declustering provides more consistent and objective identification of background events compared to empirical or ETAS-based approaches.

The advantage of ML techniques lies in their **adaptability**: once trained on realistic synthetic data, they can be applied to real catalogues from any region, such as New Zealand, without requiring region-specific tuning.

CHAPTER 3: RESEARCH OBJECTIVES

3.1 Introduction

Earthquake declustering is essential to understanding the statistical nature of seismicity, estimating seismic hazard, and identifying long-term tectonic patterns. Despite several decades of research, declustering remains a challenging task because earthquakes are not distributed randomly in space and time. Instead, they occur in clusters due to stress transfer, fault interactions, and the triggering of aftershocks.
The literature review in the previous chapter highlights the evolution of declustering methods from empirical rule-based approaches to probabilistic and, more recently, machine-learning-based frameworks. Each of these techniques has contributed valuable insights, yet none of them completely resolves the complexities of real earthquake catalogs, especially in regions with overlapping clusters such as New Zealand.

3.2 Limitations of Existing Approaches

Traditional methods, such as the Gardner–Knopoff (1974) and Reasenberg (1985) algorithms, rely on predefined spatial and temporal windows to identify aftershocks. Although computationally simple, these methods are **region-dependent** and **sensitive to parameter selection**. A small change in window parameters can produce widely different results. They also fail to capture the true physical relationships between seismic events when multiple sequences overlap in both time and space.

The **ETAS (Epidemic-Type Aftershock Sequence)** model introduced a more physical and probabilistic view of earthquake occurrence. It can, in principle, separate background and triggered events based on modeled triggering probabilities. However, ETAS estimation is **computationally demanding**, requires **careful parameter fitting**, and can lead to **biased results** if the catalog is incomplete or non-stationary. Moreover, ETAS assumes that parameters such as productivity and background rate remain constant throughout the study period — an assumption that does not hold true for regions experiencing complex stress changes.

The **Nearest-Neighbour Distance (NND)** approach and other statistical methods have further improved the identification of clusters using distance-based metrics. Yet, these approaches still depend on empirical scaling factors and often oversimplify the nonlinear relationships between time, magnitude, and spatial proximity.

3.3 Scope for Machine-Learning Approaches

Recent advances in data-driven modelling have opened new possibilities for analysing seismic patterns. **Machine learning (ML)** techniques can automatically learn complex, nonlinear relationships between features without requiring strict parametric assumptions. When trained with **synthetic labelled data (for example, using ETAS-generated catalogues), ML algorithms can generalise** these learned relationships and apply them to real earthquake data to distinguish background events from clusters more objectively.

Studies such as **Aden-Antoniów et al. (2022)** and **Kothari et al. (2023)** have demonstrated the potential of **Random Forest (RF)** and other supervised classifiers for earthquake declustering. However, the majority of existing ML-based studies have been applied to limited regions such as California or Japan. The **New Zealand seismic environment**, characterised by its variable faulting styles and complex subduction processes, remains relatively unexplored using machine-learning-based declustering frameworks.

3.4 Identified Research Gap

1. **Deficiency in region-specific machine-learning studies:**
   Maximum studies which used ML declustering methods have been applied to limited datasets outside New Zealand. A comprehensive application using the long-term GeoNet catalogue is still missing.
2. **Dependence on fixed empirical parameters:**
   Traditional algorithms use rigid windowing or parameterised models that do not adapt to local variations in seismicity.
3. **Incomplete use of multi-dimensional features:**
   Many previous works use only temporal or spatial metrics, whereas the combined effect of time, magnitude, and distance — particularly in a multidimensional feature space — remains underexplored.
4. **Need for hybrid approaches combining ETAS and ML:**
   Although ETAS provides a physically grounded framework, it lacks flexibility. Machine learning, on the other hand, is flexible but requires meaningful labelled data. A hybrid approach can bridge this gap by generating ETAS-based synthetic data for training ML classifiers.
5. **Limited interpretability of results:**
   While ML models can classify events effectively, few studies attempt to interpret the physical significance of feature importance and model outcomes in the context of tectonic processes.

3.5 Problem Statement

Considering these research gaps, the main problem addressed in this study can be stated as follows:

> *In regions with complex tectonic settings, such as New Zealand, traditional declustering methods often fail to clearly distinguish between background earthquakes and clustered events. Statistical models, such as ETAS, are grounded in physical theory; however, they tend to be computationally intensive and sensitive to the choice of parameters. Empirical methods, on the other hand, are simpler but usually lack the flexibility needed to handle diverse seismic patterns. This study aims to overcome these issues by creating a* ***machine learning-based declustering*** *system that combines the flexibility of data-driven categorisation with the physical principles of ETAS modelling. This integrated approach aims to produce an objective, scalable, and more interpretable declustering of the New Zealand earthquake catalogue.*

The proposed framework leverages **nearest-neighbour-based metrics**, **magnitude differences**, and **spatio-temporal features** to train and test supervised ML-based algorithms such as Random Forest, Support Vector Machine, and Gradient Boosting.
By doing so, the research seeks to produce a refined declustered catalogue that more accurately represents the background seismicity of New Zealand and enhances the reliability of future seismic hazard assessments.

Chapter 4: Data Source and Study Area

4.1 Introduction

This chapter describes the data and study area used in the present research. The study aims to apply machine learning–based techniques for earthquake declustering and therefore requires a reliable earthquake catalogue representing a wide range of magnitudes, depths, and temporal intervals. New Zealand has been selected as the study region because of its active tectonic setting and complex seismic behaviour. The area provides an ideal natural laboratory for testing different declustering algorithms due to the frequent occurrence of clustered events, overlapping sequences, and variable background seismicity.

4.2 Study Area:

New Zealand is located along one of the most seismically active plate boundaries in the world, where the **Pacific Plate** converges with the **Australian Plate**. The interaction between these plates occurs through a combination of subduction, strike-slip, and oblique-slip faulting. The **Hikurangi Subduction Zone** in the North Island marks the region where the Pacific Plate is descending beneath the Australian Plate, while in the South Island, most of the relative motion is accommodated by the **Alpine Fault**, a major strike-slip fault with a significant reverse component

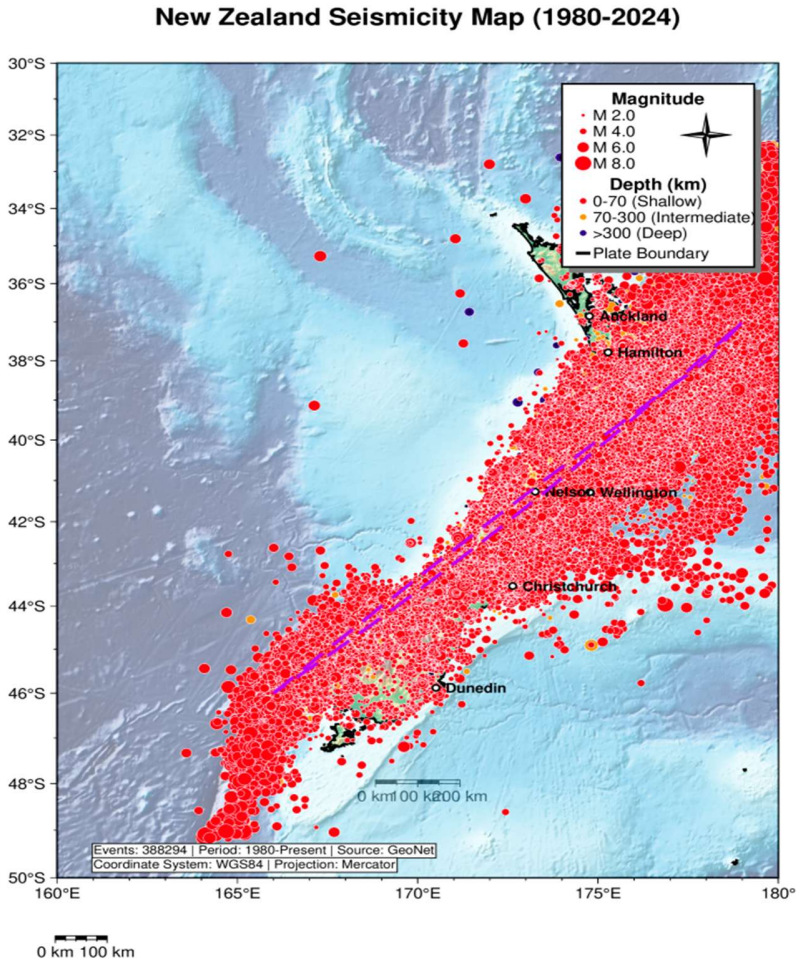
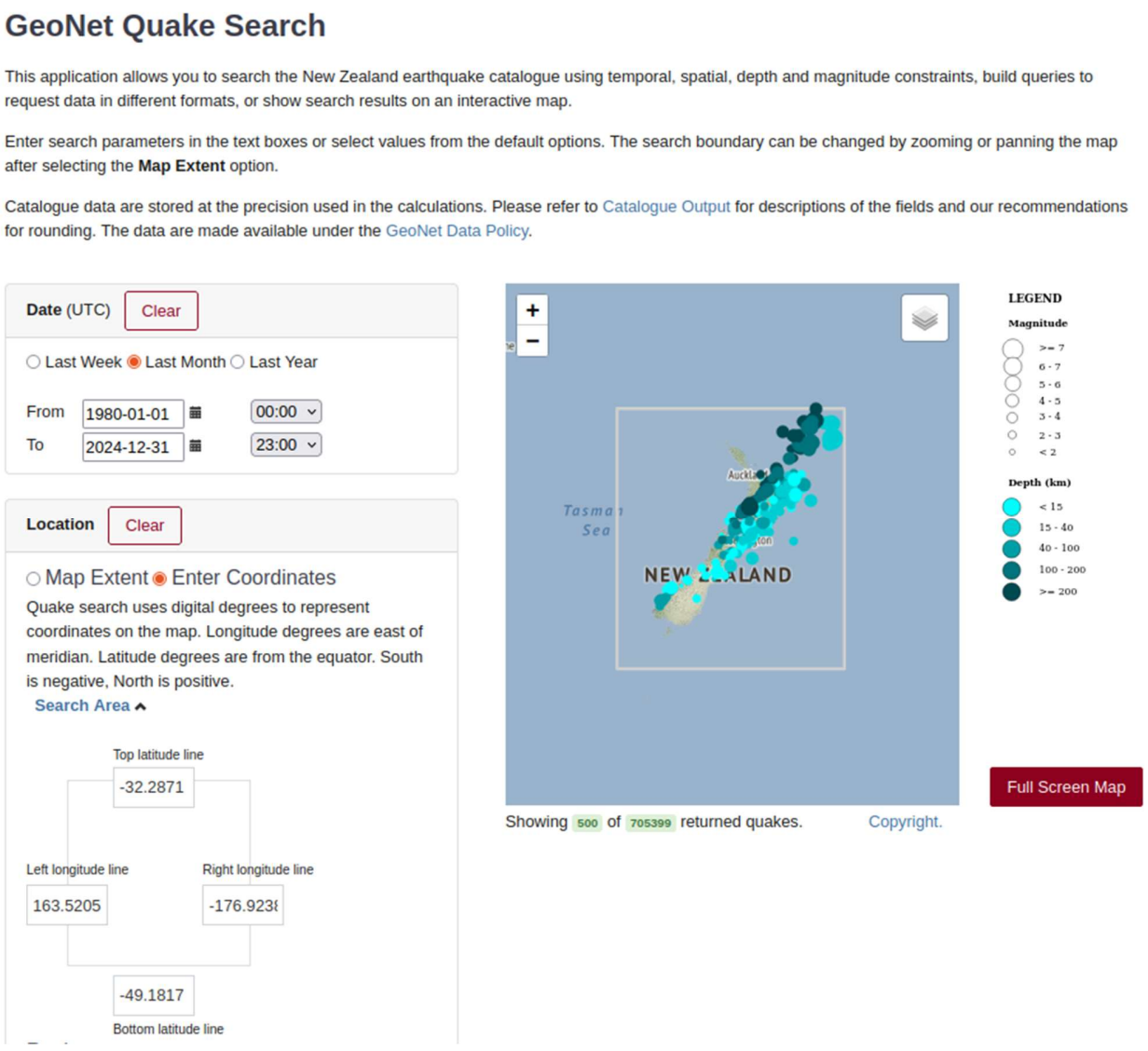


Figure 1. *Seismicity map of New Zealand (1980–2024) showing earthquake epicentres from the GeoNet catalogue. Circle size indicates magnitude, colour represents depth.*

The above figure shows the distribution of seismicity in and around New Zealand from 1980 to 2024. A dense concentration of shallow to intermediate-depth earthquakes is observed along the plate boundary, extending from Fiordland in the southwest to the Hikurangi margin in the northeast. The Alpine Fault dominates seismic activity in the South Island, while the North Island is characterised by subduction-related events. This spatial pattern reflects both crustal and subduction-driven processes, making New Zealand a suitable test region for machine-learning-based declustering.

The earthquake data used in this study were obtained from the **GeoNet Earthquake Catalogue**, operated by **GNS Science** in collaboration with New Zealand. GeoNet maintains a dense seismic network that continuously records seismic activity nationwide, ensuring high spatial and temporal coverage.

For this work, events spanning from **January 1, 1980, to December 31, 2024, were considered**. The catalogue includes origin time, latitude, longitude, focal depth, magnitude, and location uncertainty. All these parameters are essential for computing features related to space, time, and magnitude, which are later used in the machine learning framework.



CHAPTER 5: METHODOLOGY

5.1 Introduction

The entire process is explained in this part, which I followed in this research work for the declustering of the New Zealand earthquake catalogue using machine learning. The process combines both traditional seismological methods and modern data-driven techniques. Initially, I used the **Nearest-Neighbour Distance (NND)** approach to understand the clustering behaviour of earthquakes. After that, I generated a **synthetic earthquake catalogue** using the **ETAS** model with the help of an R module. This synthetic dataset provided labelled examples of background and triggered events, which are essential for supervised learning.
Finally, I trained different **machine learning algorithms** using features derived from the NND and ETAS-based data, and the trained model was applied to the real GeoNet earthquake catalogue of New Zealand.

5.2 Data Acquisition and Preprocessing

The earthquake data used in this study were obtained from the **GeoNet catalogue**, which provides comprehensive seismic records for the New Zealand region. The dataset includes information on origin time, latitude, longitude, depth, magnitude, and associated uncertainties for each recorded event from **1980 to 2024**.

To ensure data consistency and reliability, the following steps were performed:

- Duplicate and artificial events (such as quarry blasts) were removed.
- Events with a magnitude less than the magnitude of completeness are removed.

5.3 Gaussian Mixture Model (GMM)–Based Threshold Estimation

Before performing classification, the **Nearest Neighbour Distance (NND) parameter (η) was analysed** using a **Gaussian Mixture Model (GMM)** to statistically separate clustered (triggered) and background events.
The logarithm of the rescaled proximity distance, $\log10(\eta)$, was modelled as a bimodal distribution using two Gaussian components. The first Gaussian represents the **clustered (aftershock)** population, while the second corresponds to the **background seismicity**.

The GMM fitting revealed a clear **bimodal structure**, confirming the coexistence of two distinct seismic populations. The intersection point between the two Gaussian curves was adopted as the **threshold** for distinguishing background and triggered events.
In this study, the optimal threshold was determined to be approximately $\eta = 2.03$, which effectively separates the two modes of the NND distribution.

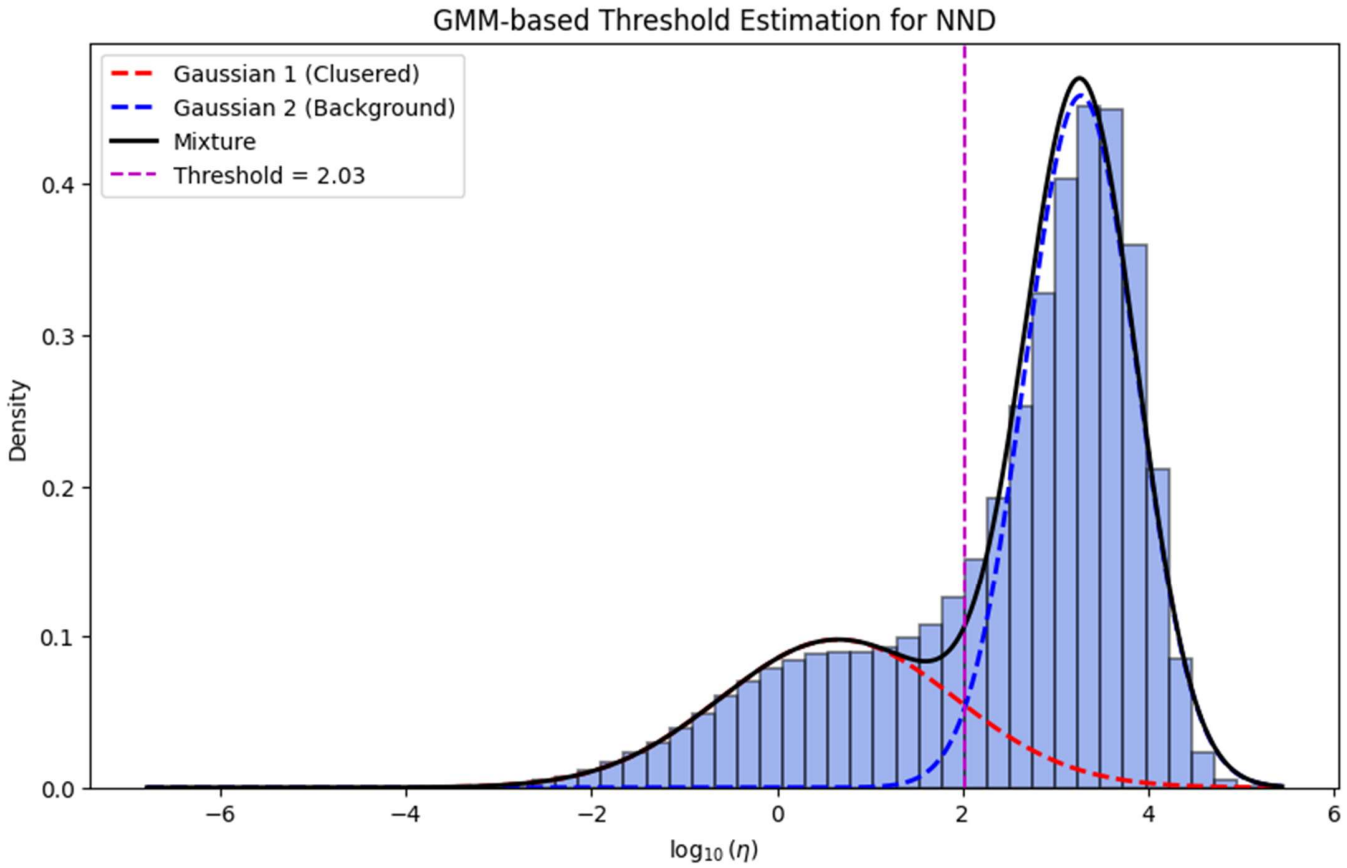


Figure: *Bimodal distribution of the rescaled proximity distance (log10(η)) obtained using a Gaussian Mixture Model. The red dashed curve represents clustered events, the blue dashed curve indicates background events, and the solid black line shows their mixture distribution. The magenta vertical line marks the estimated threshold (η=2.03) separating the two seismic populations.*

This analysis confirms that the NND metric captures the dual nature of seismicity — where lower η values correspond to temporally and spatially correlated aftershocks, and higher η values correspond to independent background earthquakes. The identified threshold was later used to label events in the **synthetic catalogue**, forming the training dataset for supervised machine-learning models.

5.4 Nearest-Neighbour Distance (NND) Analysis

The Nearest-Neighbour Distance (NND) method is one of the most powerful approaches for earthquake declustering. In this method, we study the distribution of inter-event distances between earthquake events and analyse the clustering properties of the earthquakes.

The concept was originally introduced by **Baiesi and Paczuski (2004)** and later refined by **Zaliapin and Ben-Zion (2013a, 2013 b, 2016, 2020)**. Their studies show that, for several regional earthquake catalogues, the frequency distribution of the key parameters, the rescaled proximity distance (η), often displays two distinct peaks: one linked to background seismicity and the other to clustered earthquake events.

For each earthquake i and a potential parent event j, the nnd is given as:

$$\eta_{ij} = T_{ij}R_{ij}, \quad \text{where} \quad T_{ij} = t_{ij}10^{\left(\frac{-bm_i}{2}\right)}, \quad R_{ij} = \left(r_{ij}\right)^{d_f}10^{\left(\frac{-bm_i}{2}\right)}$$

Where,

- $t_{ij} = t_j - t_i$ is the time interval between events.
- r_{ij} is how far apart the two events are in space,
- m_i is previous event magnitude,
- d_f is the fractal dimension.

To compute Δr_{ij} The spherical distance between the two epicentres was calculated as:

$$\Delta r_{ij} = R_E \cos^{-1}\left[\sin\phi_i \sin\phi_j + \cos\phi_i \cos\phi_j \cos(\lambda_i - \lambda_j)\right]$$

The **rescaled variables** for time and distance were obtained as:

$$T_{ij} = t_{ij} 10^{\left(\frac{-bm_i}{2}\right)}, \quad R_{ij} = r_{ij} 10^{\left(\frac{-bm_i}{2}\right)}$$

The probability density function (PDF) of log(η) generally exhibits a **bimodal distribution**, with one peak corresponding to background (independent) events and another to clustered (dependent) events. The threshold η* separating these two modes was determined using visual inspection and kernel density estimation.

While the NND approach provides valuable insight into the clustering pattern, it suffers from overlap between populations, which motivates the use of machine learning for more objective separation.

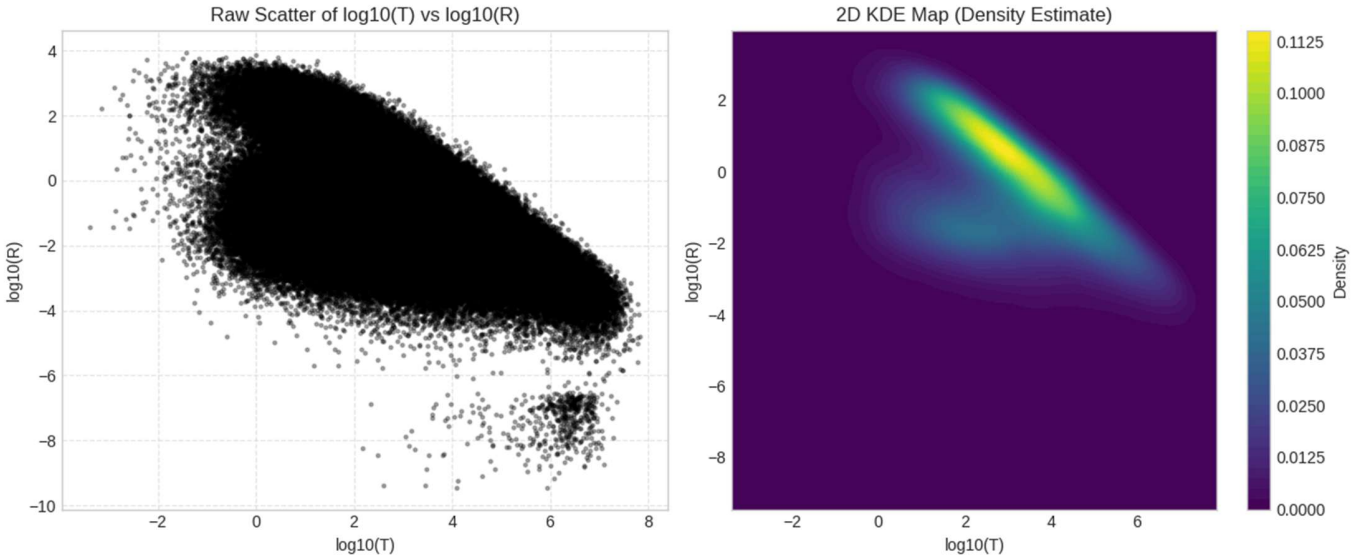


Fig. Density plot of η_{ij}

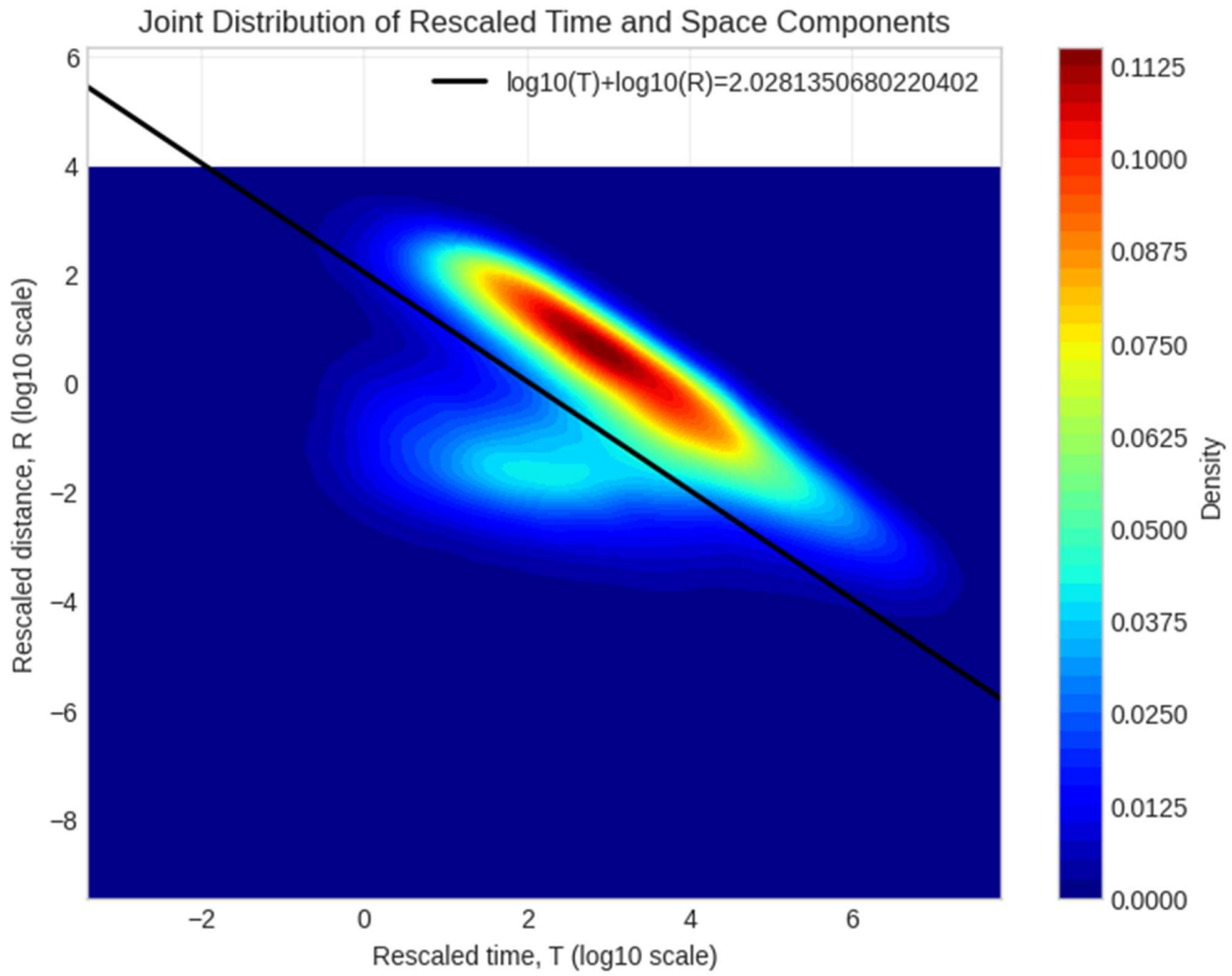


Fig: Rescaled Time and Distance Joint distribution

5.5 Synthetic Catalogue Generation Using the ETAS Model

The **ETAS model** represents a *self-exciting point process*, meaning that each earthquake increases the likelihood of future events for a certain period and within a specific spatial range. Mathematically, the conditional intensity function λ(t,x,y|Ht) — which defines the expected occurrence rate of earthquakes at a given time and location — can be expressed as the sum of a **background term** and a **triggering term** that accounts for previous earthquakes. This is written as:

$$\lambda(t, x, y \mid H_t) = \mu \cdot u(x, y) + \sum_{i:t_i<t} \xi(t - t_i, x - x_i, y - y_i; m_i)$$

Here,μu(x, y) represents the *spatially varying background rate*, while ξ(t, x, y; m)defines the contribution of a past earthquake of magnitude mi to the overall seismic rate at time t and position (x,y).

The triggering component $\xi(t, x, y; m)$ can be separated into three independent functions as:

$$\xi(t, x, y; m) = \kappa(m) g(t) f(x, y, m)$$

where

- $\kappa(m) = Ae^{\alpha(m-m0)k(m)}$ represents the **productivity** of an earthquake (i.e., how many aftershocks it is expected to generate above a reference magnitude m0),
- g(t) is the **temporal kernel** describing how the triggering rate decreases with time, following the **Omori–Utsu decay law**, and
- $f(x, y, m)$ is the **spatial kernel** controlling how the influence of an earthquake decreases with distance from its epicentre.

The temporal kernel is defined as:

$$g(t) = \frac{p-1}{c}\left(1 + \frac{t}{c}\right)^{-p}; t > 0$$

and the spatial kernel (Ogata and Zhuang, 2006) as:

$$f(x, y, m) = \frac{q-1}{\pi D^2 e\gamma^{(m-m_0)}}\left[1 + \frac{x2 + y2}{D^2 e\gamma^{(m-m_o)}}\right]$$

Here, the parameter γ determines how the spatial influence of an event scales with its magnitude.
The complete parameter set is given as

$$\theta = [\mu, A, \alpha, c, p, D, \gamma, q]$$

and is estimated by **maximising the log-likelihood function** over the target space–time region [Ts,Te]×S:

$$\log L(\boldsymbol{\theta}) = \sum_{\{i:\ (t_i, x_i, y_i) \in [T_s, T_e] \times S\}} \log \lambda_\theta\left(t_i, x_i, y_i \mid \mathcal{H}_{\{t_i\}}\right) - \int_{\{T_s\}}^{\{T_e\}} \iint_S \log \lambda_\theta(t, x, y \mid \mathcal{H}_t)\, dx\, dy\, dt$$

5.6 Stochastic Declustering (SD) Method

After fitting the ETAS model, the **Stochastic Declustering (SD)** method can be applied to distinguish between background and triggered events.
This approach computes the probability that a given earthquake j was **triggered** by a prior event i, based on the relative contribution of i to the total intensity at the time and location of j:

$$\rho ij = \frac{\kappa(mi) g(t_j - t_i) f(x_j - x_i, y_j - y_i; m_i)}{\lambda(t_j, x_j, y_j \mid Ht_j)}$$

Similarly, the probability that an event j occurred **independently** (as a background event) is:

$$\phi_j = \frac{\mu u(xj, yj)}{\lambda(tj, xj, yj \mid Htj)}$$

From this, the probability that an event j was triggered can be expressed as:

$$\rho_j = 1 - \phi_j = \sum \rho ij$$

Using these probabilities ($\phi_j \; and \; \rho_j$), earthquakes in the catalogue can be probabilistically assigned as background or triggered through a **random thinning process**.

For each event, a random number U_j is generated from a uniform distribution on [0, 1]
If $U_j < \phi_j$ The event is labelled as background; otherwise, it is labelled as **triggered**.

In this study, the **stochastic declustering approach** is used as a **benchmark method** to validate and compare the classification performance of the **supervised machine-learning** and **Nearest Neighbour Distance (NND)** models.
Unlike the NND-based approach, the SD method relies purely on the ETAS conditional intensity and does not depend on spatial or temporal thresholds, providing a statistically consistent reference for declustering analysis

5.7 Feature Engineering

To effectively distinguish **background earthquakes** from **aftershocks**, it is essential to select features that capture their **relative characteristics** rather than their absolute position or origin time. This ensures that the classification approach remains general and applicable across different regions and time periods.

In both **supervised** and **unsupervised** models, the **Nearest Neighbour Distance (NND)** framework provides a set of **relative metrics** that can be used as features. These features are derived from the spatial–temporal relationships between earthquakes and are designed to describe the clustering behaviour in seismicity.

To predict the label y_j for the jth event, up to **five features** are considered:

1. **Rescaled Time (Tj)** represents the nearest-neighbour rescaled time, defined as in Equation (1).
2. **Rescaled Distance (Rj)** captures the normalised spatial separation between the event and its nearest neighbour.
3. **Magnitude Difference** ($\boldsymbol{dm_j = m_i - m_j}$) — Here, i refers to the index of the nearest neighbour. A larger ΔM suggests that the event j is more likely an aftershock of a higher-magnitude parent event.
4. **Number of Parent events** (n_{parent}) — This is the count of events that share the same nearest neighbour as the event j. Higher values typically indicate swarm-like behaviour or multiple candidates acting as parent events
5. **Number of Triggered Offspring** (n_{child}) — This measure how many events are triggered by the event j, reflecting its role as a parent in aftershock generation.

These features can collectively be expressed as a vector:

$$x_j = \left(T_j^*, R_j^*, dm_j, n_{parent}, n_{child}\right)^T$$

where T denotes the transposed vector form.

To compute T_j^* and R_j^*, it is first necessary to evaluate the η and identify the **NND index** i for each event j. The computation of η requires an estimate of the **fractal dimension** d_f, which characterises the spatial clustering of earthquake epicentres.

The fractal dimension d_f is estimated using the **Minkowski–Bouligand (box-counting)** method.
In this approach, the study area is divided into grids (or boxes) of different sizes, and the number of boxes containing at least one event is counted. The relationship between the box size and the count of non-empty boxes follows a **power law**, whose exponent corresponds to the fractal dimension d_f.

For the synthetic catalogues generated in this study, d_f was found to remain relatively stable around **1.7**, showing only minor fluctuations. This consistency indicates that the simulated seismicity exhibits realistic spatial clustering behaviour comparable to natural earthquake distributions.

Once both the real and synthetic catalogues were available, I computed various **features** that describe the spatio-temporal relationships between events. Each earthquake event was represented as a vector of numerical attributes, such as:

Feature	Description
η_{ij}	Nearest-neighbour distance combining time, space, and magnitude
$T_{ij} +$	Rescaled Time
$R_{ij} +$	Rescaled Distance
ΔM	Magnitude difference between the previous event and its parent event
n_{parent}	Number of potential parent events
n_{child}	Number of triggered offspring events

CHAPTER 6: Model Training and Evaluation

Once the feature set was derived from the ETAS-based synthetic catalogue, it was used to train and validate a series of machine-learning classifiers for distinguishing between background and triggered seismic events.

This chapter explains the overall training workflow, data preparation, model configurations, hyper-parameter tuning, and evaluation protocol that ensured the robustness of the proposed classification framework.

6.1 Data Preparation

The labelled synthetic catalogue is generated through the Epidemic-Type Aftershock Sequence (ETAS) simulation, which contains both **background** and **triggered** events.
Each record included six attributes obtained from the Nearest-Neighbour Distance (NND) framework: rescaled distance (R^+), rescaled time (T^+), magnitude difference (dm^+), number of siblings (N^+), number of parent events (n_{parent}), and the number of child events (n_{child}).
All numerical features were **standardised (z-score normalisation)** to remove scale effects among variables and accelerate model convergence.
No missing data were present, and class balance was verified to prevent bias during model fitting.

6.2 Supervised Learning Framework

In this study, there were four supervised classifiers used, which are:

1. Random Forest (RF)
2. Support Vector Machine (SVM).
3. Gradient Boosting (GB) is a sequential ensemble that uses additive learning to reduce residual errors.
4. Extreme Gradient Boosting (XGBoost)

All models were created with the scikit-learn and XGBoost modules in Python 3.11.

6.3 Cross-Validation and Hyper-Parameter Optimisation

To reduce over-fitting, a 5-fold cross-validation process was used.In each cycle, the dataset was split into five equal folds, four of which were used for training and one for validation. The best configurations were chosen based on average performance across folds. Using a grid-search approach, hyperparameters were adjusted by methodically examining parameter combinations until the optimal F1-score was obtained. The tuning ensured that each model achieved a balanced bias–variance trade-off and stable convergence.

6.4 Workflow of Model Training and Evaluation

The complete process adopted in this study is summarised schematically in **Figure 6.1**.
It illustrates the logical sequence from feature extraction to final evaluation.

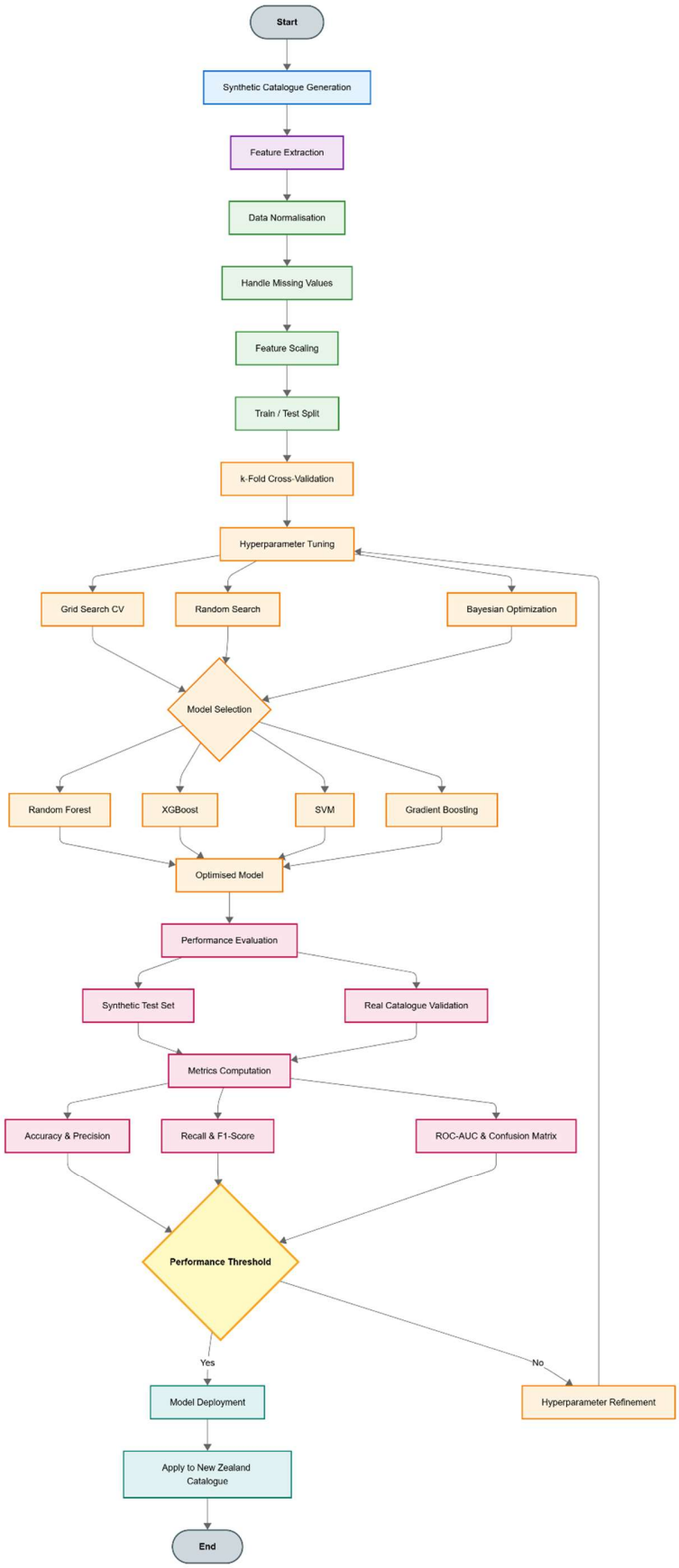


Figure 6.1: workflow of the ML pipeline from synthetic data generation to model deployment on the New Zealand catalogue

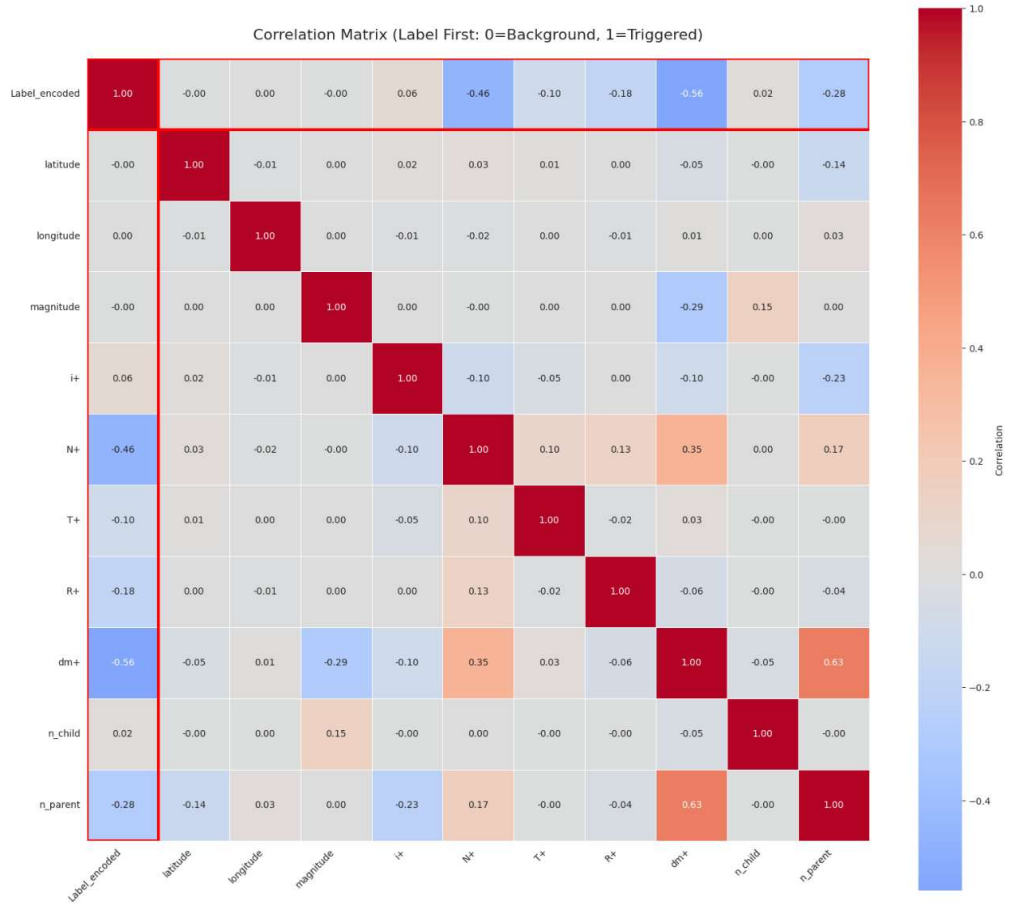
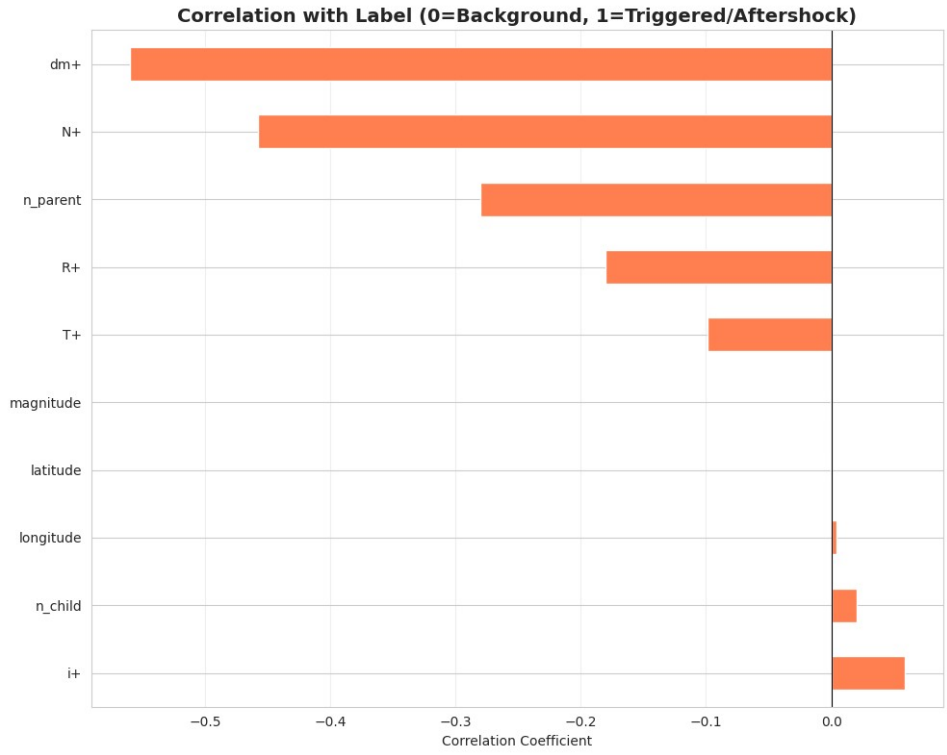


Figure 6. Heatmap showing pairwise correlations among seismic features, highlighting strong spatial–temporal dependencies



CHAPTER 7: Results and Discussion

The ML Algorithms were trained and tested using the labelled synthetic catalogue generated through the ETAS simulation framework. The results were analysed in two stages: (i) model performance on synthetic data, where true event labels were known, and (ii) validation on real earthquake catalogues to assess the generalisation ability of the trained models.

7.1. Model Performance on Synthetic Data

The classification models, including **Random Forest (RF)**, **Support Vector Machine (SVM)**, **XGBoost** and **Gradient Boosting (GB)**, were first applied to the synthetic dataset containing known background and triggered events. Among these, the **XGBoost model** achieved the highest accuracy of **97.44%**, followed closely by the Gradient Boosting model, while the SVM showed slightly lower but consistent performance.

The **confusion matrix** revealed that the models were able to correctly identify the majority of triggered events with high recall values, indicating their effectiveness in capturing clustered seismicity patterns. The precision values were also high, confirming that the models maintained low false-positive rates.

Feature importance analysis showed that $N+$, **rescaled distance** ($R+$) and **rescaled time** ($T+$) were the most influential predictors, followed by **magnitude difference** ($dm+$) and **number of parents** (n_{parent}). The **number of offspring** (n_{child}) contributed relatively less but still added valuable information about aftershock productivity.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9672	0.9622	0.9515	0.9791
SVM	0.9436	0.9448	0.9436	0.9440
Gradient Boosting	0.9711	0.9706	0.9889	0.9797
XGBoost	**0.9744**	**0.9766**	**0.9874**	**0.9820**

Table 7.1: Performance comparison of machine-learning models on the synthetic ETAS catalogue.

These findings are consistent with physical expectations — aftershocks tend to occur close in time and space to their parent events, making temporal and spatial proximity strong indicators of triggering.

Table 7.1 summarises the quantitative performance of all models. Ensemble-based approaches (RF, GB, XGB) performed exceptionally well, achieving an accuracy of around 97%, while SVM provided consistent results.

7.1.1 Random Forest Model

The Random Forest classifier produced reliable findings, with an F1-score of 0.979 and an accuracy of 96.7%.
The majority of triggered and background events were accurately detected, with little overlap close to the dec ision threshold, according to the confusion matrix (Fig. 7.1a).

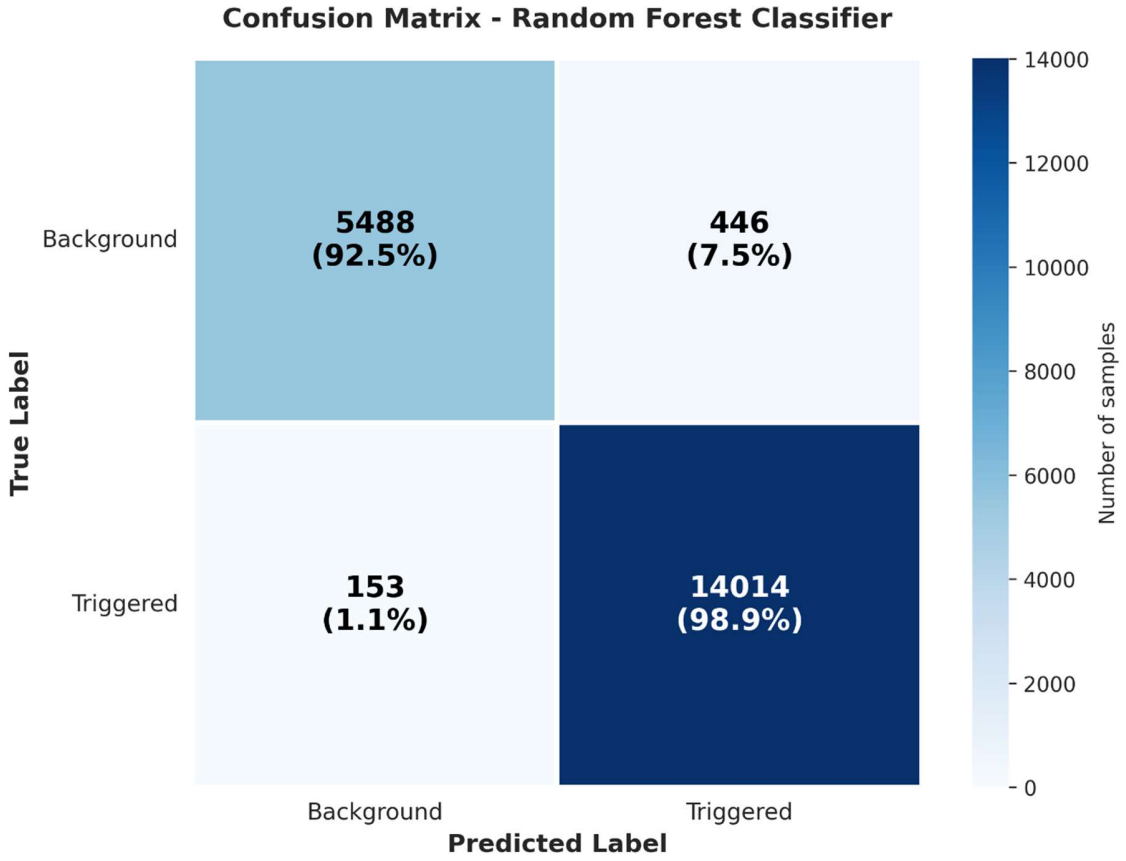


Figure 7.1a: *Confusion matrix illustrating classification performance of the Random Forest model on synthetic data.*

Fig. 7.1b shows a multi-dimensional evaluation graphic which provides an in-depth analysis of the Random Forest predictions. Aftershocks (yellow) cluster strongly below the diagonal line in the rescaled space-time map, pointing out their spatiotemporal proximity to mainshocks. The two populations are clearly separated by a bimodal structure in the GMM-based NND distribution. The classification ratio, which consists of 68.8% background and 31.2% aftershocks, can be seen in the pie chart. The temporal distribution corresponds with New Zealand's historical earthquake activity, with large spikes during major earthquakes like Kaikōura (2016) and Canterbury (2010).

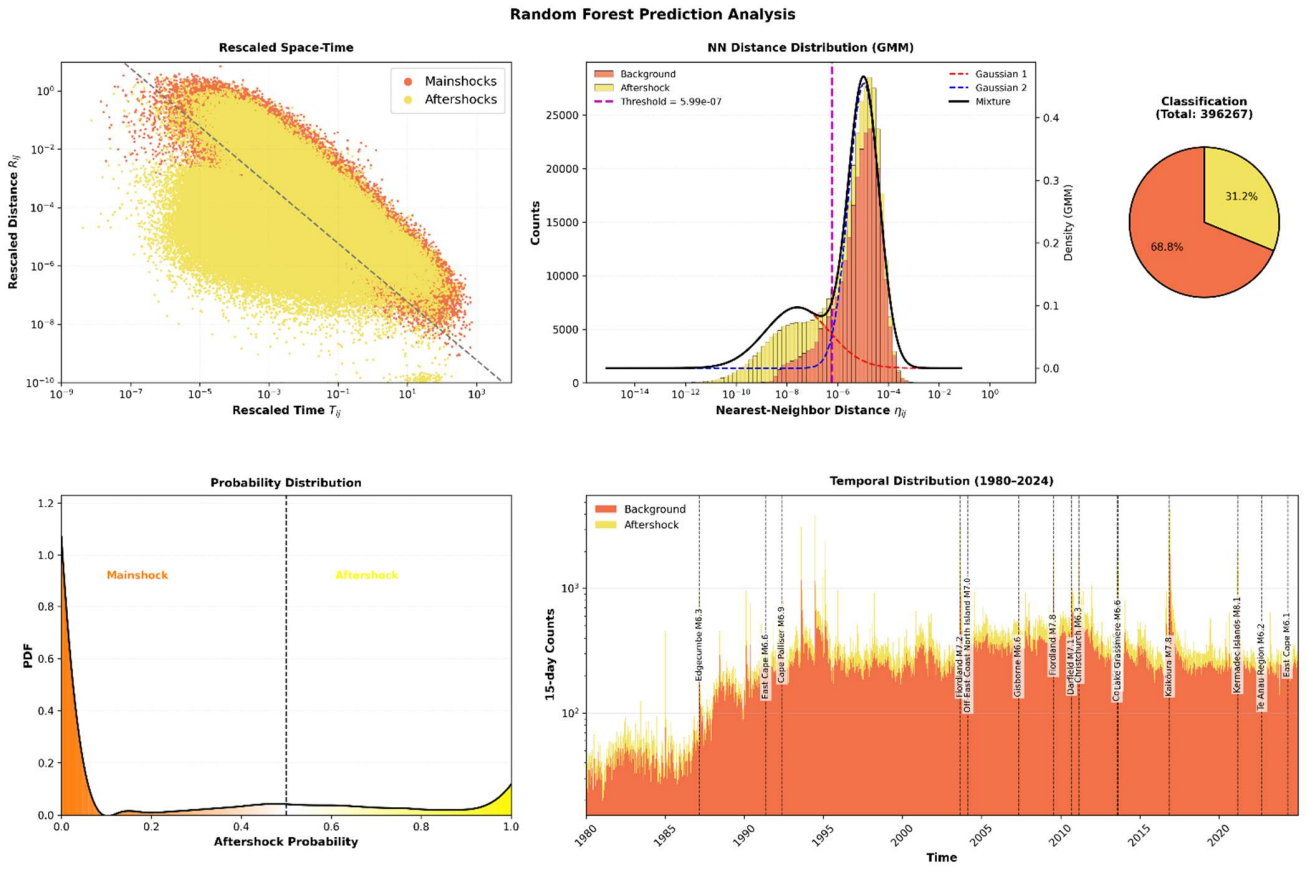


Figure 7.1b: *Random Forest–based classification of the New Zealand earthquake catalogue showing (top-left) rescaled space–time distribution, (top-right) nearest-neighbour distance histogram with Gaussian mixture fit, (bottom-left) aftershock probability density, and (bottom-right) temporal evolution of background and aftershock events (1980–2024).*

The number of siblings (N^+) and rescaled distance (R^+) are the most significant predictors, followed by rescal ed time (T^+) and the number of child events (n_child), according to feature-importance analysis (Fig. 7.1c). This illustrates how aftershock productivity, closeness, and recurrence are physically related.

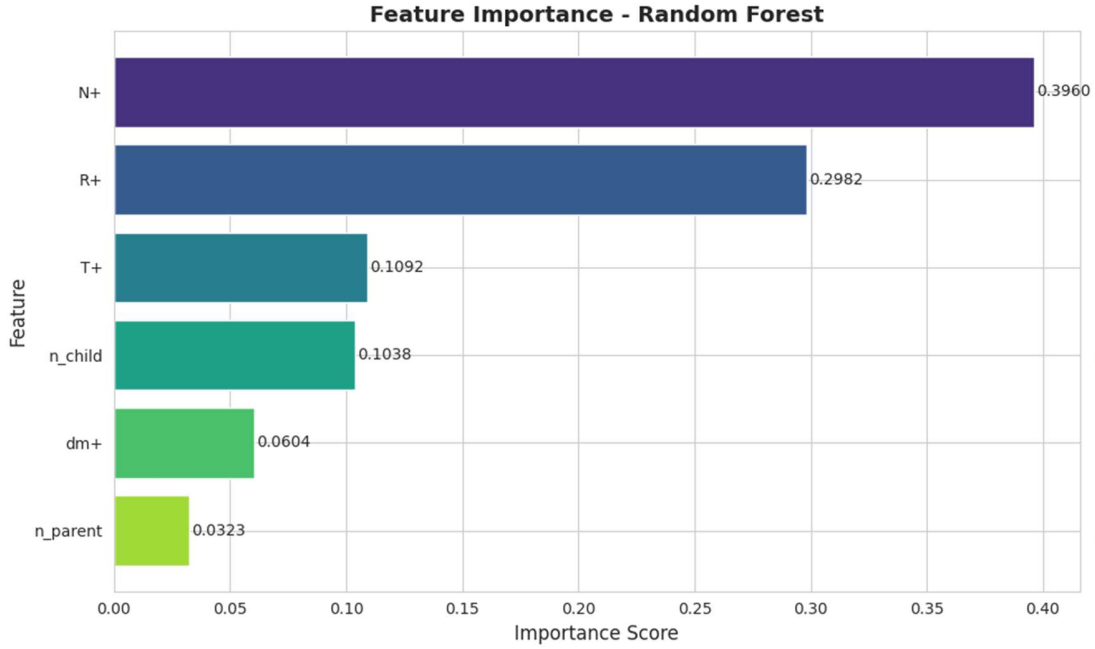


Figure 7.1c: *Feature-importance ranking for the Random Forest model showing the dominant contribution of spatio-temporal parameters.*

7.1.2 Gradient Boosting Model

Random Forest was enhanced by the Gradient Boosting method, which used additive learning to iteratively correct misclassified data. It had the highest recall value (0.9889) and 97.1% accuracy, showing a high sensitivity to triggered events. As shown by the confusion matrix (Fig. 7.2a), it is effective for minimising false negatives.

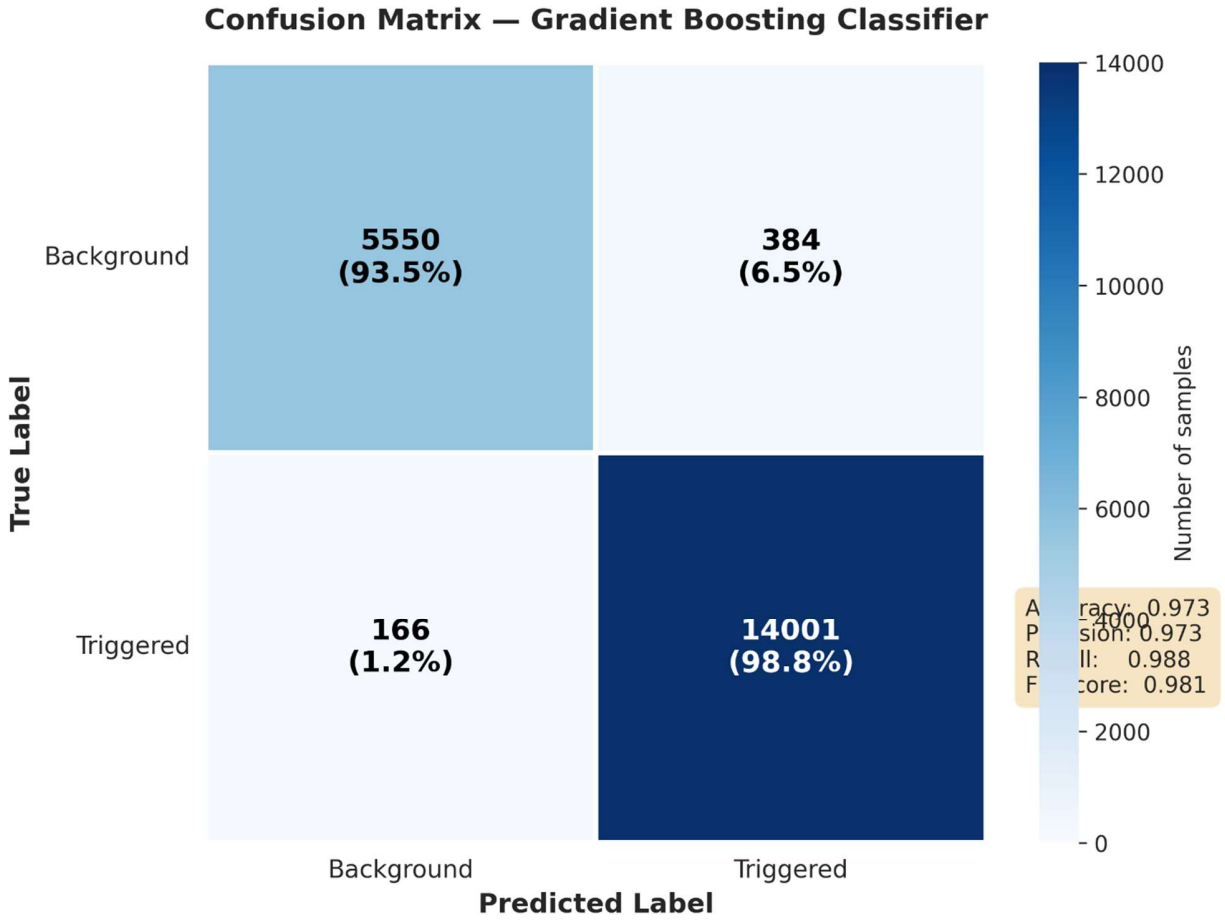


Figure 7.2a: *Confusion matrix for the Gradient Boosting classifier trained on the synthetic dataset.*

Feature-importance results (Fig. 7.2b) show that *temporal and spatial NND parameters (T, R*)** strongly influence model predictions, confirming that the model effectively captures seismic clustering patterns.

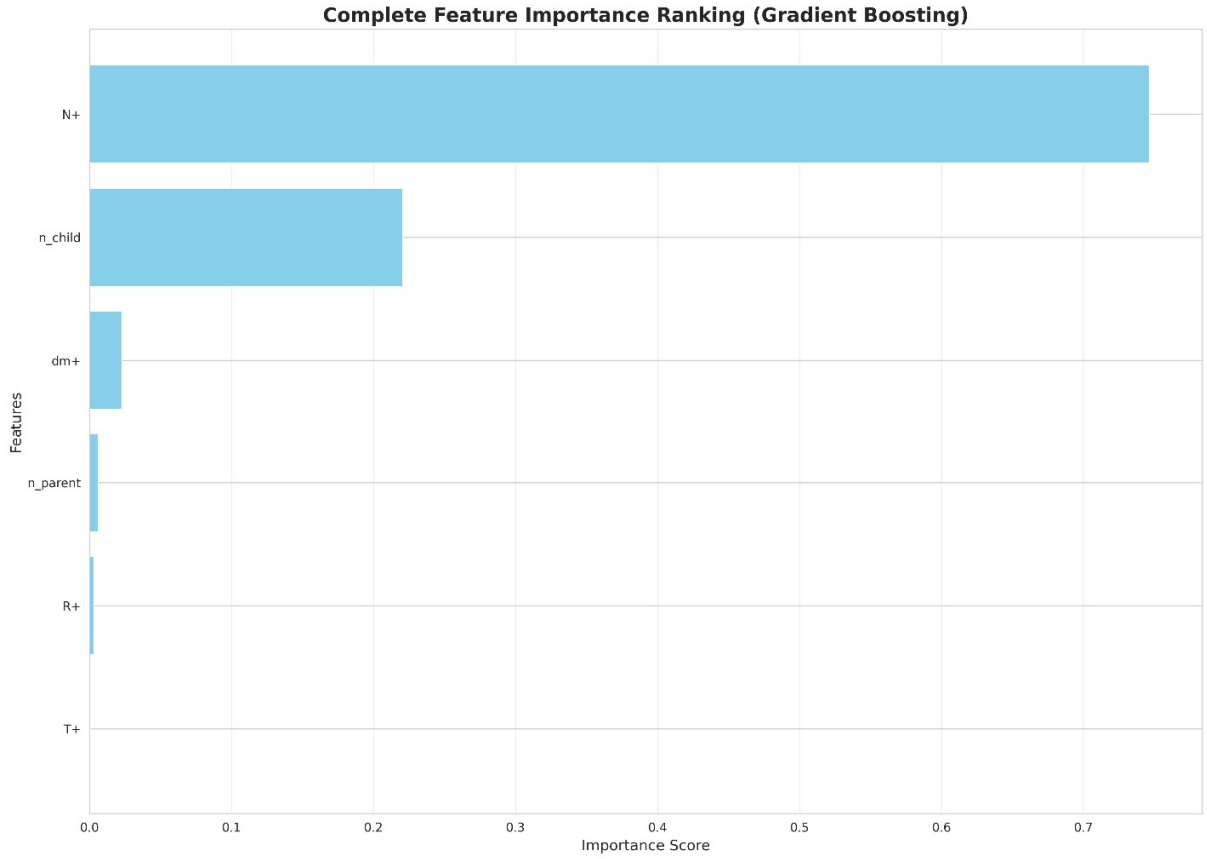
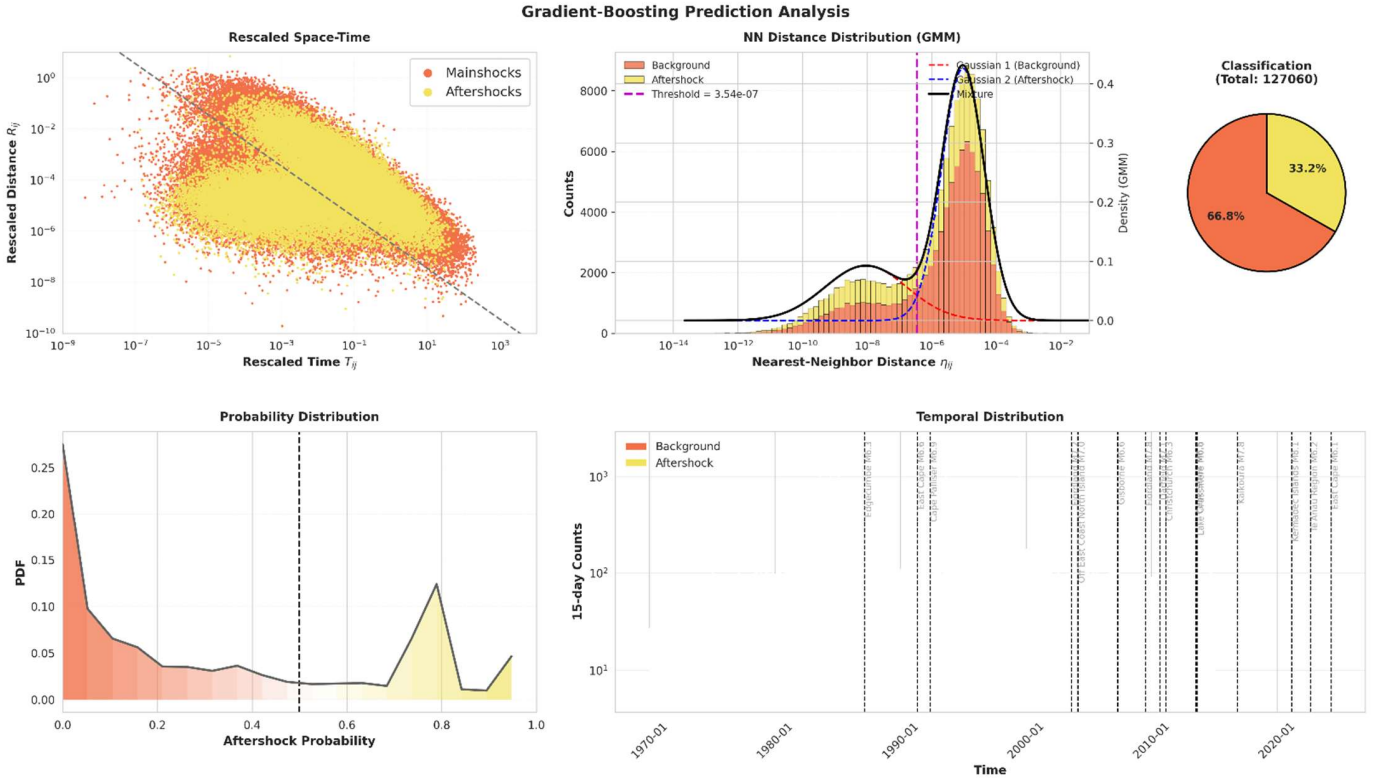


Figure 7.2b: *Feature importance of the Gradient Boosting model demonstrating the dominance of rescaled space–time features.*



7.1.3 XGBoost Model

Among all tested algorithms, **XGBoost** demonstrated the **best overall performance**, with 97.4% accuracy, precision = 0.9766, and recall = 0.9874.
The confusion matrix (Fig. 7.3a) shows that the model correctly classified 98.7 % of triggered events and 94.4 % of background events, with very few misclassifications.

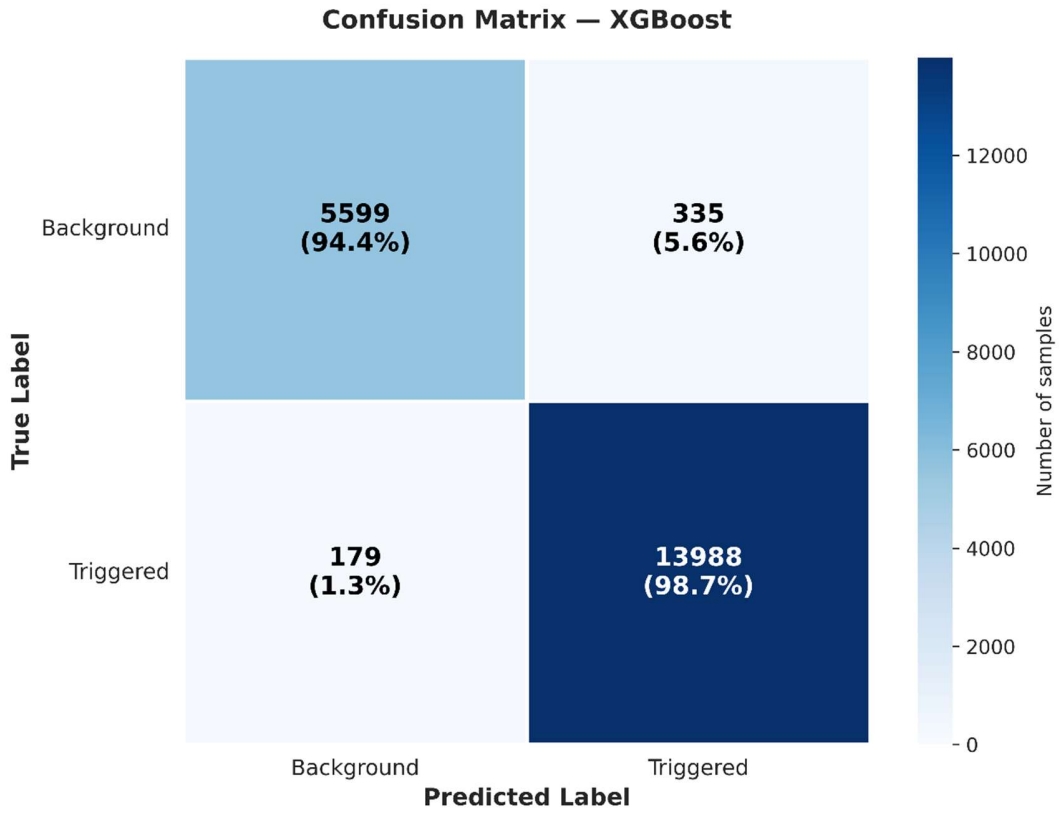


Figure 7.3a: *Confusion matrix for the XGBoost classifier showing high precision and minimal misclassification.*

The **feature-importance plot** (Fig. 7.3b) reinforces that *spatial (R)** and *temporal (T)** distances are the most influential factors, reflecting their critical role in determining triggering relationships.
Compared to the other ensemble models, XGBoost's regularisation and gradient optimisation helped maintain both accuracy and generalisation without overfitting.

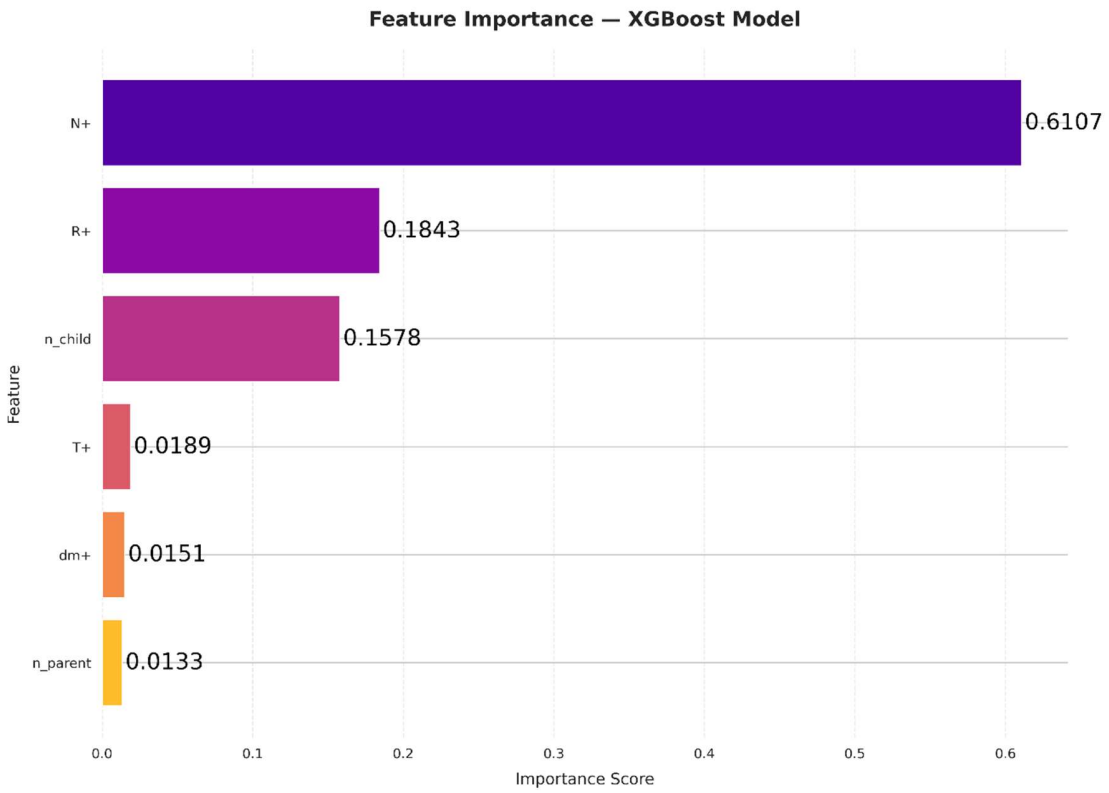


Figure 7.3b: *Feature-importance distribution of the XGBoost model, highlighting dominant contributions from spatio-temporal proximity features.*

7.2 Application to Real Earthquake Catalogue

All four classifiers—Random Forest, SVM, Gradient Boosting, and XGBoost—were applied to the actual New Zealand earthquake catalogue (1980–2024) following successful training and validation on synthetic ETAS data.

The main temporal and spatial clustering features of seismicity were replicated by every model, demonstrating the data-driven approach's generalizability outside of the simulated setting.

But throughout the whole region, the XGBoost model reliably demonstrated the highest stability and accuracy, accurately recognising background events and aftershock sequences.

Since XGBoost is the most accurate and physically consistent model among the examined classifiers, the ensuing subsections concentrate on a thorough analysis of the XGBoost-based results.

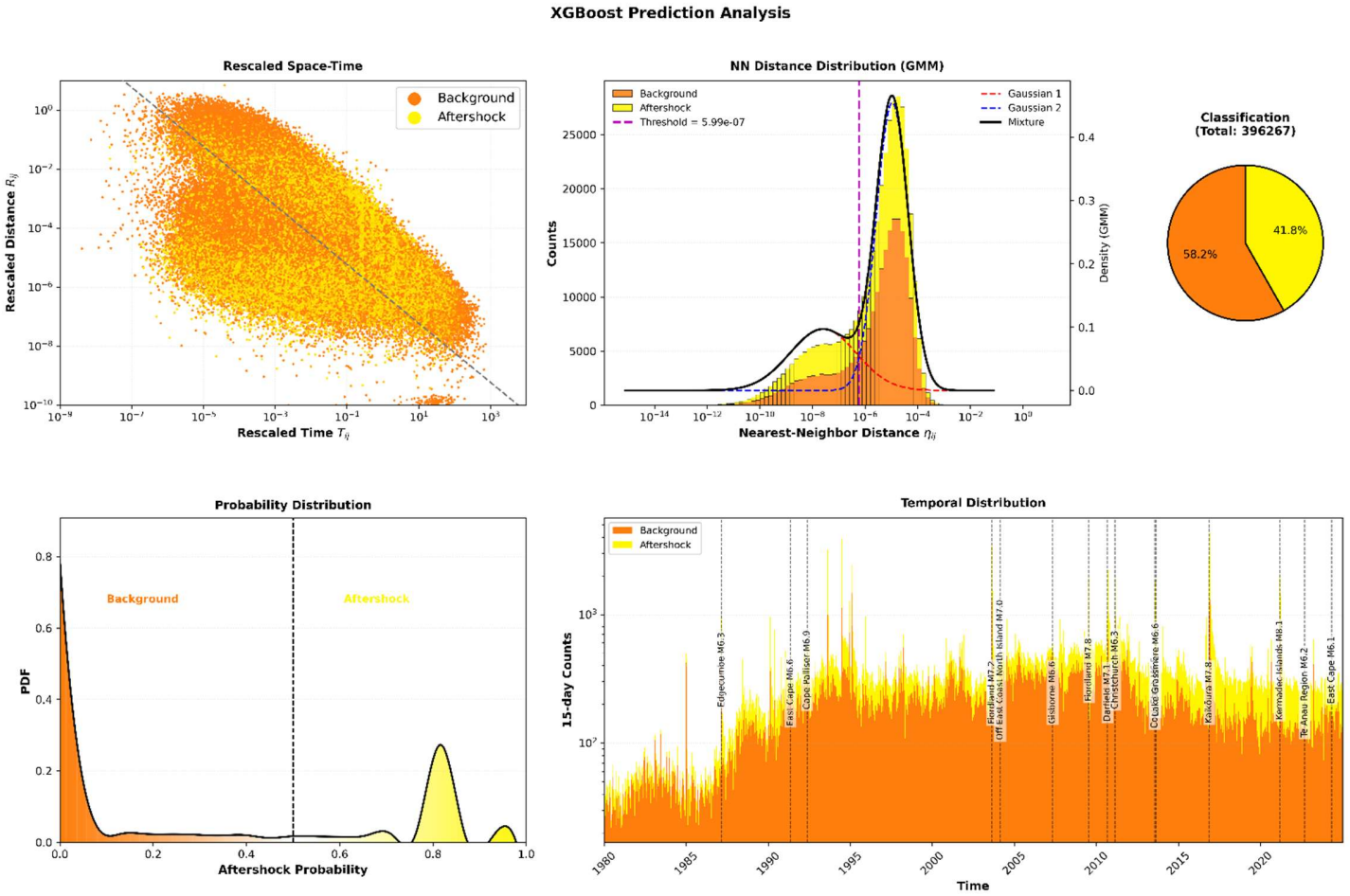


Figure 7.4: *XGBoost prediction analysis for the New Zealand earthquake catalogue (1980–2024). The plots show (top-left) rescaled space–time clustering, (top-right) nearest-neighbour distance distribution with Gaussian Mixture fit and classification ratio, (bottom-left) probability-density separation, and (bottom-right) temporal evolution of aftershock activity following major earthquakes.*

The above figure shows the XGBoost prediction analysis for the actual New Zealand earthquake catalogue (1980–2024). Aftershocks (yellow) cluster below the diagonal line, indicating tight spatial–temporal closeness to mainshocks, while background events (orange) are more dispersed, according to the rescaled space–time map (top-left). Two distinct peaks can be seen in the nearest-neighbour distance distribution (top-right) fitted with a Gaussian Mixture Model, indicating a clear distinction between background and triggered events. The total classification proportions, which are roughly 42% aftershocks and 58% background, are displayed in the pie chart. The temporal distribution (bottom-right) shows notable sequences, such as Edgecumbe (1987), Canterbury (2010), and Kaikōura (2016), with a decay pattern in line with the Omori rule. In contrast, the probability-density plot (bottom-left) emphasises substantial bimodality.

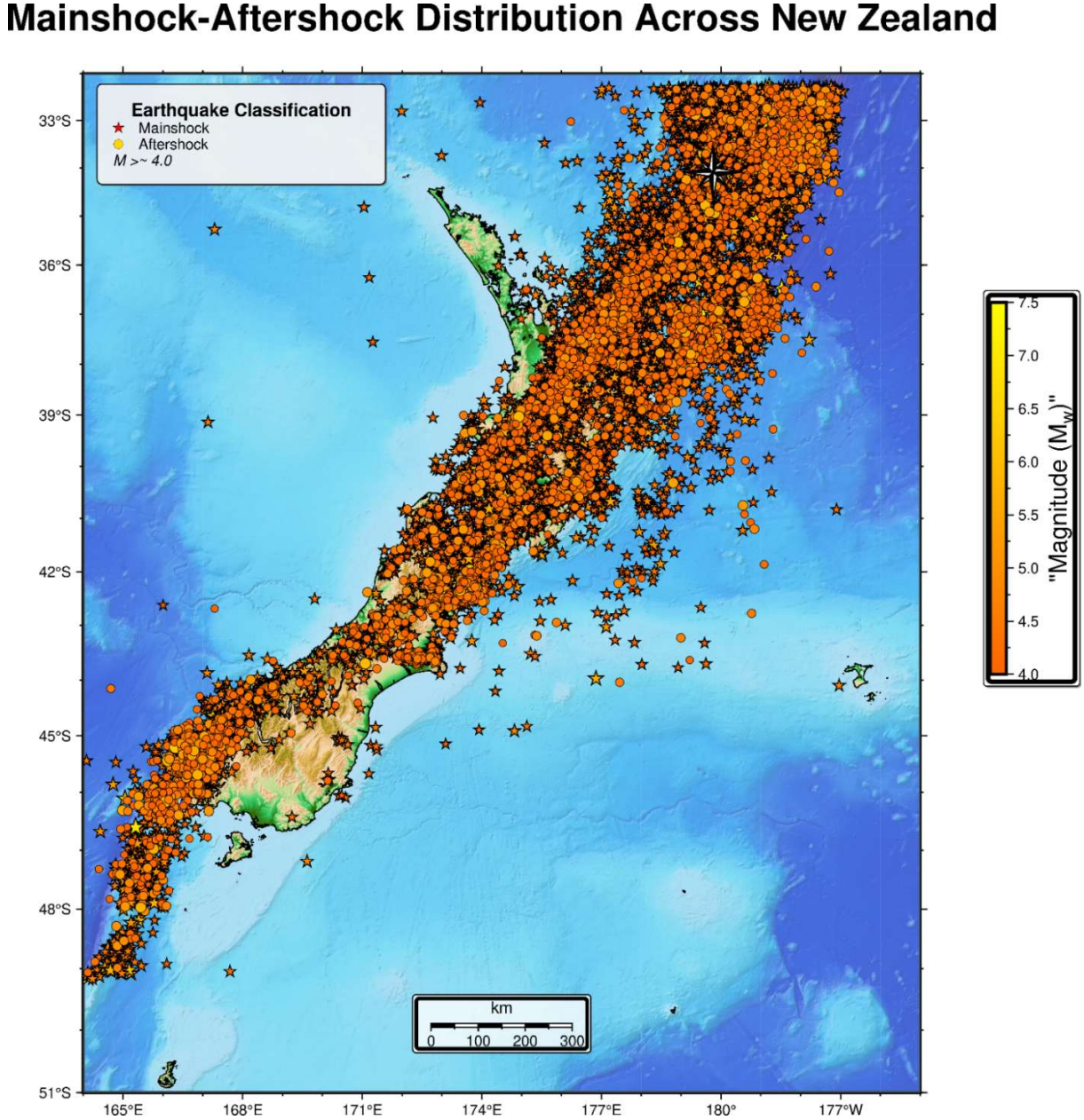


Figure 7.8: *Spatial distribution of mainshock and aftershock events ($M \geq 4.0$) across New Zealand*

Model	*Accuracy*	*Background Count*	*Triggered Count*
XGBoost	***0.9744***	***230758 (58.23%)***	***165509 (41.75%)***
Gradient-Boost	*0.9711*	*219362 (55.36%)*	*146905 (44.64%)*
SVM	*0.9436*	*289241 (72.69%)*	*107026 (27.01%)*
Random Forest	*0.9672*	*272671 (68.8%)*	*116339 (31.2%)*

Table 7.2: *Evaluation of four supervised machine-learning models showing overall accuracy and the number of events labelled as background.*

CONCLUSION

Declustering an earthquake catalogue remains a challenging task in seismology, even though it plays an important role in understanding seismic behaviour and improving hazard assessments. Several methods have been proposed over the years, but many of them still face difficulties when the catalogue contains mixed or overlapping sequences. Classical window-based approaches and nearest-neighbour techniques rely on fixed thresholds, and these values do not always reflect the actual behaviour of earthquakes in complex tectonic regions.

In this thesis, the aim is to resolve these problems by developing a machine-learning-based declustering approach. The main idea is to combine the physical insight provided by nearest-neighbour analysis with the pattern-recognition capability of supervised learning models. This allows background and triggered events to be separated in a way that reduces subjective choices and improves consistency across different seismic environments.

To build a reliable model, synthetic catalogues were first generated using the ETAS model. These ETAS simulations provided labelled background and aftershock events, which allowed supervised training. From each event pair, several nearest-neighbour-based features were extracted, including rescaled time (T), rescaled distance (R), magnitude difference, and the combined NND metric. These features helped describe how one earthquake relates to another in terms of space, time, and energy release. These features describe how earthquakes relate to one another in terms of space, time, and magnitude.

Multiple machine-learning models were tested, and **XGBoost** performed the best. It achieved an accuracy of approximately **97%** during validation and demonstrated a high recall for identifying triggered events. When applied to the real New Zealand catalogue, the model classified around **58.2% of the events as background and the remaining 41.8% as triggered.** The spatial patterns also made sense: background events mostly followed the major fault systems, while aftershocks formed tight clusters near well-known sequences such as **Canterbury and Kaikōura (2016)**. The temporal patterns of the triggered events also aligned with the timing of major historical earthquakes in the region.

A key outcome of this study is that the declustering does not depend on fixed thresholds. Instead, the separation is learned directly from ETAS physics and nearest-neighbour statistics, making the method more adaptable than classical approaches. The final declustered catalogue gives a clearer representation of the underlying seismic behaviour of New Zealand and can be used for hazard studies, forecasting work, and further statistical analyses

REFERENCES

1. **Aden-Antoniow, F., Frank, W. B., & Seydoux, L.** (2022). An adaptable Random Forest model for the declustering of earthquake catalogs. *Journal of Geophysical Research: Solid Earth, 127*(10), e2022JB024210. https://doi.org/10.1029/2022JB024210
2. **Aharoni, E., & Ben-Zion, Y.** (2023). Machine learning approaches for characterizing earthquake clusters. *Seismological Research Letters, 94*(5), 2772–2783.
3. **Aharoni, E., Ross, Z. E., & Ben-Zion, Y.** (2022). Machine learning and earthquake physics. *Nature Reviews Earth & Environment, 3*(7), 477–490. https://doi.org/10.1038/s43017-022-00285-y
4. **Benali, A., Jalilian, A., Peresan, A., Varini, E., & Idrissou, S.** (2023). Spatiotemporal analysis of the background seismicity identified by different declustering methods in Northern Algeria and its vicinity. *Axioms, 12*(3), 232. https://doi.org/10.3390/axioms12030232
5. **Davidsen, J., Gu, C., & Baiesi, M.** (2015). Generalized Omori–Utsu law for aftershock sequences in southern California. *Geophysical Journal International, 201*(2), 965–978. https://doi.org/10.1093/gji/ggv041
6. **ETH Zürich.** (2025). *Research Collection – Earthquake Seismology Datasets.* https://www.research-collection.ethz.ch
7. **Frontiers in Earth Science.** (2024). *Recent Advances in Earthquake Forecasting.* https://www.frontiersin.org
8. **Gardner, J. K., & Knopoff, L.** (1974). Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America, 64*(5), 1363–1367.
9. **GeoNet.** (2024). *New Zealand Earthquake Catalogue.* GNS Science. https://www.geonet.org.nz
10. **Guo, Y., Zhuang, J., & Zhou, S.** (2015). Improved space–time ETAS model for inverting rupture geometry from seismicity triggering. *Journal of Geophysical Research: Solid Earth, 120*(9), 6353–6369. https://doi.org/10.1002/2015JB012228
11. **Harte, D. S.** (2013). *Earthquake Sequences and the Epidemic-Type Aftershock Sequence Model.* Springer. https://doi.org/10.1007/978-94-007-6530-7
12. **Kothari, S., Shcherbakov, R., & Ben-Zion, Y.** (2023). Earthquake declustering using supervised machine learning. *Bulletin of the Seismological Society of America, 113*(3), 1854–1868. https://doi.org/10.1785/0120220324
13. **Land Information New Zealand (LINZ).** (2024). *Tectonic plate boundaries and geodetic data for New Zealand.* https://www.linz.govt.nz
14. **Mousavi, S. M., & Beroza, G. C.** (2023). Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences, 51*, 249–275. https://doi.org/10.1146/annurev-earth-030422-103813
15. **Ogata, Y.** (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the Royal Statistical Society: Series C, 37*(3), 315–328. https://doi.org/10.2307/2347569
16. **Ogata, Y.** (1998). Space–time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics, 50*(2), 379–402.

17. **Seal, A., Sahoo, S., Peresan, A., & Khan, P. K.** (2025). Statistical analysis of background seismicity of Southern California: Nearest neighbour declustering and network analysis. *Journal of Seismology, 29*(2), 201–220. https://doi.org/10.1007/s10950-025-10174-2
18. **Sharma, A., Nanda, S. J., & Vijay, R. K.** (2024). A spatio-temporal binary grid-based clustering model for seismicity analysis. *Pattern Analysis and Applications, 27*(5), 1–18. https://doi.org/10.1007/s10044-024-01210-2
19. **SpringerLink.** (2025). *Earthquake Science Articles and Statistical Models.* https://link.springer.com
20. **Uhrhammer, R. A.** (1986). Characteristics of northern and central California seismicity. *Bulletin of the Seismological Society of America, 76*(3), 1053–1077.
21. **Zaliapin, I., & Ben-Zion, Y.** (2013). Earthquake clusters in southern California I: Identification and stability. *Journal of Geophysical Research: Solid Earth, 118*(6), 2847–2864. https://doi.org/10.1002/jgrb.50179
22. **Zhuang, J.** (2006). Diagnostic analysis of space–time branching processes for earthquakes. *Lecture Notes in Statistics, 187*, 1–43. https://doi.org/10.1007/978-0-387-32885-7_1
23. **Zhuang, J., Ogata, Y., & Vere-Jones, D.** (2002). Stochastic declustering of space–time earthquake occurrences. *Journal of the American Statistical Association, 97*(458), 369–380. https://doi.org/10.1198/016214502760046925