

DEEP LEARNING OF AFTERSHOCK PATTERNS FOLLOWING LARGE
EARTHQUAKES

by

Feyzanur Tekbıyık

B.S., Mathematics, Boğaziçi University, 2022

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computational Science and Engineering
Boğaziçi University

2025

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my thesis advisor, Bertan Badur, for his knowledge, guidance, and patience throughout my thesis process. His support was invaluable in developing my academic perspective and adhering to the disciplined research process. I would also like to thank my co-advisor, Çağrı Diner, for his valuable contributions to the process with his constructive suggestions and guidance.

I am deeply grateful to my mother and my late father for their endless support and patience throughout my entire education life. I also thank my brothers for always being by my side and giving me strength. And a very special thanks to my beloved nephew, Doruk Alp, whose sweetness, pure love, and joy have always helped me forget my exhaustion and reminded me of what truly matters.

I am grateful to my friends and colleagues who made this process more meaningful and manageable, for always being there for me, giving me moral support, and providing support when I needed it.

Finally, I would like to thank my dear boyfriend, who is pursuing his graduate studies at the same time and experiencing this process alongside me. He has been there for me through this challenging time, giving me strength with his understanding, love, patience, and support. His presence made this process easier and more meaningful. Going through this time side by side, with mutual dedication and support, was truly invaluable to me.

ABSTRACT

DEEP LEARNING OF AFTERSHOCK PATTERNS FOLLOWING LARGE EARTHQUAKES

This thesis presents data-driven modeling strategies for predicting the spatial distribution, number, and magnitude of aftershocks following major earthquakes. To overcome the limitations of classical physics-based approaches, alternative model structures were developed for both classification and regression tasks using machine learning methods such as artificial neural networks, the XGBoost algorithm, and kernel-based probabilistic modeling. The study systematically evaluated various input combinations—including static components derived from stress tensors, neighborhood information, gradient-based derivative quantities, spatial coordinates, and distance to the epicenter—and additionally applied oversampling and undersampling strategies to mitigate class imbalance, while employing kernel-based continuous labeling to model the spatial continuity of aftershock probabilities. Through this comprehensive experimental analysis, the prediction performances of aftershock occurrence, count, and magnitude were compared across different data structures and modeling strategies, demonstrating the potential of stress-based machine learning models in aftershock forecasting. The findings indicate that kernel-based probabilistic models provide a more realistic and spatially coherent representation of aftershock regions, XGBoost enhances discrimination capability in classification tasks, and multilayer neural networks yield more balanced performance in numerical regression tasks. Overall, the results show that static features derived from stress tensors can successfully capture large-scale spatial patterns of aftershock behavior, yet modeling fine-scale variability will require richer temporal and geological datasets in future research.

ÖZET

BÜYÜK DEPREMLERİ TAKİP EDEN ARTÇI ŞOK ÖRÜNTÜLERİNİN DERİN ÖĞRENME İLE MODELLENMESİ

Bu tez, büyük depremleri izleyen artçı şokların mekânsal dağılımı, sayısı ve büyüklüğünü tahmin etmeye yönelik veri odaklı modelleme stratejileri sunmaktadır. Klasik fizik tabanlı yaklaşımların sınırlılıklarını aşmak amacıyla, yapay sinir ağları, XGBoost algoritması ve kernel tabanlı olasılıksal modelleme gibi makine öğrenimi yöntemleri kullanılarak hem sınıflandırma hem de regresyon görevleri için alternatif model yapıları geliştirilmiştir. Çalışmada, stres tensörlerinden türetilen statik bileşenler, komşuluk bilgileri, gradyan tabanlı türevsel nicelikler, mekânsal koordinatlar ve merkez üssüne olan mesafe gibi çeşitli girdi kombinasyonları sistematik biçimde test edilmiş; ayrıca sınıf dengesizliğini azaltmak için örnekleme stratejileri uygulanmış ve kernel tabanlı sürekli etiketleme yaklaşımıyla artçı şok olasılıklarının mekânsal sürekliliği modellenmiştir. Bu geniş kapsamlı deneysel analiz, farklı veri yapıları ve modelleme stratejilerine dayalı olarak artçı şokların oluşumu, sayısı ve büyüklüğüne ilişkin tahmin performanslarını karşılaştırmış ve stres tabanlı makine öğrenimi modellerinin artçı şok öngörüsündeki potansiyelini ortaya koymuştur. Elde edilen sonuçlar, kernel tabanlı olasılıksal modellerin artçı şok bölgelerini daha gerçekçi ve tutarlı biçimde temsil ettiğini, sınıflandırma görevlerinde XGBoost modellerinin ayırım gücünü artırdığını ve sayısal regresyon görevlerinde çok katmanlı yapay sinir ağlarının daha dengeli performans sunduğunu göstermektedir. Genel olarak bulgular, stres tensörlerinden türetilen statik niteliklerin artçı şok davranışına ilişkin geniş ölçekli uzamsal desenleri başarılı bir şekilde yansıttığını, ancak ince ölçekli varyasyonların modellenebilmesi için daha zengin zamansal ve jeolojik veri setlerine ihtiyaç bulunduğunu ortaya koymaktadır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZET	v
LIST OF FIGURES	viii
LIST OF TABLES	xi
LIST OF ACRONYMS/ABBREVIATIONS	xii
1. INTRODUCTION	1
2. LITERATURE REVIEW	4
2.1. Artificial Neural Networks: Classification and Regression	5
2.2. Studies for Seismic Prediction	7
2.3. Other Machine Learning Applications and Hybrid Methods	11
2.4. Overview	14
3. MATERIALS & METHODS	16
3.1. Problem Definition	16
3.2. Data Description	17
3.3. Data Preprocessing	19
3.3.1. Feature Selection and Engineering	19
3.3.2. Normalization and Scaling	22
3.3.3. Principal Component Analysis	22
3.4. Model Architecture	23
3.4.1. Artificial Neural Networks (ANN)	23
3.4.2. Convolutional Neural Networks (CNN)	25
3.4.3. eXtreme Gradient Boosting	26
3.5. Performance Metrics	27
3.5.1. Classification Metrics	27
3.5.2. Probabilistic Evaluation Metrics	28
3.5.3. Regression Metrics	29
4. EXPERIMENTS & RESULTS	31

4.1. Experimental Setup and Model Variants	31
4.2. Aftershock Occurrence Modeling	34
4.2.1. Binary Classification of Aftershock Occurrence	34
4.2.2. Kernel-Based Probabilistic Modeling of Aftershock Occurrence .	61
4.3. Aftershock Count Regression Modeling	66
4.3.1. Baseline Count Regression Models	66
4.3.2. Kernel-Based Count Regression Modeling of Aftershock Activity	81
4.4. Regression Modeling of Aftershock Magnitude	87
5. DISCUSSION	103
5.1. Aftershock Occurrence Modeling Results	104
5.2. Aftershock Count Regression Modeling Results	106
5.3. Regression of Magnitudes Results	108
5.4. General Evaluation and Limitations	109
5.5. Recommendations for Future Studies	110
6. CONCLUSION	112
REFERENCES	114
APPENDIX A: CODE REPOSITORY	121

LIST OF FIGURES

Figure 3.1.	Distribution of grid cells with and without aftershock.	19
Figure 3.2.	An example multilayer neural network architecture.	24
Figure 3.3.	An example UNET model architecture.	26
Figure 4.1.	Confusion matrix of BC model.	37
Figure 4.2.	Confusion matrix of BC-E model.	39
Figure 4.3.	Confusion matrix of BC-E-F model.	40
Figure 4.4.	Scree Plot for PCA of Face Neighbor Stress Components.	41
Figure 4.5.	Confusion matrix of BC-E-F-P model.	42
Figure 4.6.	Confusion matrix of BC-E-FE model.	44
Figure 4.7.	Scree Plot for PCA of Face and Edge Neighbor Stress Components.	45
Figure 4.8.	Confusion matrix of BC-E-FE-P model.	46
Figure 4.9.	Confusion matrix of BC-E-FEC model.	48
Figure 4.10.	Scree Plot for PCA of Face, Edge, and Corner Neighbor Stress Components.	49
Figure 4.11.	Confusion matrix of BC-E-FEC-P model.	50

Figure 4.12. Confusion matrix of BC-E-FE-SO model.	52
Figure 4.13. Confusion matrix of BC-E-FE-SU model.	54
Figure 4.14. Confusion matrix of BC-E-FE-SO-D model.	55
Figure 4.15. Confusion matrix of the BC-E-FE-SO-D-X model.	57
Figure 4.16. ROC curve of the BC-E-FE-SO-D-X model.	58
Figure 4.17. Confusion matrix U-Net model.	59
Figure 4.18. Actual and predicted voxel outputs of the U-Net model in 3D. . .	60
Figure 4.19. Training and validation loss curves (BCE) of the kernel-based probabilistic aftershock occurrence model.	62
Figure 4.20. Spatial distribution of aftershocks for the 1999 Izmit earthquake within the 3D domain.	63
Figure 4.21. Gaussian-kernel smoothing of aftershock occurrence for the 1999 Izmit earthquake.	63
Figure 4.22. Predicted aftershock occurrence probabilities for the 1999 Izmit earthquake with MLP model.	64
Figure 4.23. Training and validation loss curves of the kernel-based XGBoost probabilistic aftershock occurrence model.	65
Figure 4.24. Predicted aftershock occurrence probabilities for the 1999 Izmit earthquake with XGBoost model.	66

Figure 4.25. The relationship between the actual and predicted aftershock counts of the CR-E-X-C model.	80
Figure 4.26. Training and validation MAE curves of the kernel-based count regression model with MLP model.	84
Figure 4.27. Gaussian-kernel smoothing of aftershock counts for the 1999 Izmit earthquake.	84
Figure 4.28. Predicted aftershock count distribution for the 1999 Izmit earthquake obtained from the kernel-based regression model.	85
Figure 4.29. Training and validation RMSE curves of the kernel-based count regression model with XGBoost model.	86
Figure 4.30. Training and validation RMSE curves of the kernel-based count regression model with XGBoost model.	86
Figure 4.31. The relationship between the actual and predicted aftershock magnitudes of the R-E-X-C model.	101

LIST OF TABLES

Table 4.1.	Explanations of abbreviations used in model naming.	34
Table 4.2.	Comparative performance metrics of classification models.	60
Table 4.3.	Performance metrics of the kernel-based probabilistic aftershock occurrence model.	65
Table 4.4.	Comparative performance metrics of count regression models. . . .	81
Table 4.5.	Performance metrics of the kernel-based count regression models. .	87
Table 4.6.	Comparative performance metrics of magnitude regression models.	102

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial Neural networks
AUC	Area Under the Curve
CFS	Coulomb Failure Stress
CNN	Convolutional Neural Network
DNN	Deep Neural Network
ETAS	Epidemic-Type Aftershock Sequence
InSAR	Interferometric Synthetic Aperture Radar
KDTree	K-Dimensional Tree
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
XGBoost	eXtreme Gradient Boosting

1. INTRODUCTION

Earthquakes are natural disasters that occur as a result of sudden energy releases and can cause serious destruction. Predicting the temporal and spatial extent of aftershocks following major earthquakes is crucial for structural safety and disaster management. However, predicting the distribution of aftershocks is quite challenging due to the complex structure of the Earth's crust and the uncertainties inherent in stress fields.

While traditional models attempt to explain the distribution of aftershocks based on predefined physical assumptions such as fault plane geometry or Coulomb stress gradients, these approaches often fail to reflect the complex reality on the ground. In recent years, data-driven methods such as deep learning and machine learning have offered an alternative framework to overcome these limitations and are increasingly being used in aftershock prediction.

This study develops a sophisticated modeling approach using deep learning and modern machine learning methods such as XGBoost to predict the location, number, and magnitude of aftershocks following major earthquakes. The models are fed with physical and spatial attributes such as elastic stress components, distance to the epicenter, and neighborhood structure; The data set was systematically evaluated using strategies such as oversampling, undersampling, and distance filtering.

The data set is based on 199 finite fault models from the SRCMOD database, similar to the structure used by DeVries *et al.* [3], and aftershock records from the ISC-GEM catalogs. The volume defined around each mainshock was divided into small grid cells, and the elastic stress distribution was calculated, and this information was used as model input.

The main contribution of this thesis is that it offers a new methodological framework for aftershock prediction using a data-driven, learning-based approach, as an alternative to classical physical models. Furthermore, it makes a multifaceted contribution to the literature by comparing different modeling strategies, testing the effects of input attributes, and addressing classification and regression tasks together.

Limitations of this study include the absence of temporal evolution and dynamic stress effects in the modeling framework. Time-dependent stress accumulation and remote triggering mechanisms are not directly represented by the input variables employed, which are obtained from static stress components and their gradient- or norm-based quantities. Furthermore, the models' generalizability to other tectonic systems is limited because they were trained on a data set that was solely made up of major earthquakes. A continuous spatial representation of aftershock likelihood is made possible by the kernel-based probabilistic formulation; nevertheless, it is still susceptible to parameter selection and may induce smoothing bias in areas with high seismic density.

To further improve model performance and generalizability, future studies could expand the training data with earthquakes from various tectonic settings, optimize kernel parameters using physically informed criteria, and incorporate time-dependent stress components and dynamic triggering mechanisms into the model.

The remainder of this thesis is structured as follows:

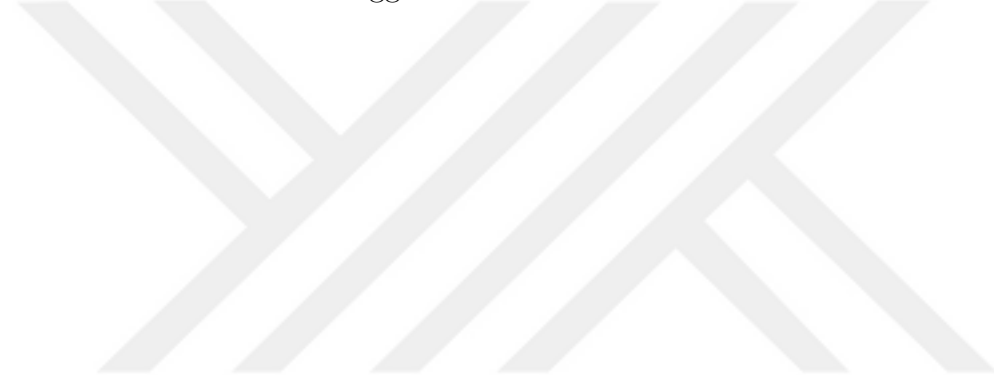
Chapter 2: Literature Review This chapter presents a comprehensive review of classical seismological models and modern machine learning approaches relevant to aftershock prediction.

Chapter 3: Materials and Methods This chapter describes the data set, pre-processing steps, feature engineering procedures, and the architectures of the neural network and XGBoost models used in this study.

Chapter 4: Experiments and Results This chapter details the experimental setups and reports the results for aftershock occurrence classification, kernel-based probabilistic modeling, count regression, and magnitude prediction.

Chapter 5: Discussion This chapter offers an integrated discussion of the findings, evaluating model behaviors, limitations, and methodological insights.

Chapter 6: Conclusion This chapter concludes the thesis with a summary of the main contributions and suggestions for future research directions.



2. LITERATURE REVIEW

This chapter provides the theoretical and empirical foundation for the modeling approaches developed in this thesis by reviewing key studies and concepts related to aftershock prediction. First, classical physics-based frameworks such as static Coulomb stress change and empirical laws including Omori–Utsu and Båth’s relations are discussed to outline the physical and statistical mechanisms behind aftershock triggering. The chapter then examines the growing body of research that applies artificial neural networks, deep learning architectures, traditional machine learning algorithms, and hybrid physics-data-driven methods to seismic forecasting. Finally, limitations of existing approaches—such as their reliance on simplified stress representations, lack of neighborhood information, or dependence on parametric assumptions—are highlighted to motivate the data-driven, stress-tensor-based modeling framework proposed in this thesis.

Aftershocks are smaller-scale earthquakes that occur after major earthquakes and usually occur in regions close to the rupture surface of the main shock. These tremors occur as a result of stress changes caused by the main shock and are usually concentrated in a certain time period and spatial area [4]. The sudden stress redistribution in the crust causes some regions to become more prone to rupture, while others are seismically suppressed. In this context, static stress change has become a frequently used physical parameter, especially to explain the spatial pattern of aftershocks.

The concept of Coulomb Failure Stress (ΔCFS) quantifies the rupture tendency by evaluating the shear stress increase on the fault plane and the normal stress decrease together. ΔCFS is a combination of the shear stress increase and the normal stress decrease (in the direction of fault opening, a positive value indicates compression) caused by the main shock at a specific observation point or on the fault plane. Simply put, if $\Delta CFS > 0$, the region is assumed to be moving closer to the fracture, and if $\Delta CFS < 0$, the rupture is assumed to propagate away from the fracture. The Coulomb

failure stress change, ΔCFS , is expressed as

$$\Delta CFS = \Delta\tau + \mu\Delta\sigma \quad (2.1)$$

where $\Delta\tau$ represents the shear stress change, $\Delta\sigma$ represents the normal stress change, and μ represents the friction coefficient [5].

The Coulomb model has successfully explained the distribution of aftershocks after many major earthquakes. In early studies conducted in this context, such as the 1995 Kobe earthquake, it was observed that aftershocks were mostly concentrated in regions with $\Delta CFS > 0$, while aftershock activity decreased in regions with $\Delta CFS < 0$ [6]. In addition, in the analyses conducted after the 1992 Landers (USA), 1999 İzmit (Turkey) and 2010 Maule (Chile) earthquakes, it was observed that the majority of aftershocks coincided with positive ΔCFS regions [7-9]. This shows that the model is also supported empirically. This situation shows that the Coulomb stress concept can be useful in understanding the general distribution of aftershocks in some cases. However, over time, it has become clear that this approach has limited predictive power, with significant performance differences across events. For example, it has been reported that the ΔCFS values calculated for the 1999 Chi-Chi earthquake were as successful as a random model in explaining the aftershock distribution, with the area under the ROC curve (AUC) remaining at only 0.47 [10].

2.1. Artificial Neural Networks: Classification and Regression

In contrast to the limited predictive power of such physics-based models, in recent years, artificial intelligence-based approaches, especially Artificial Neural Networks (ANN), have increasingly attracted attention in modeling the spatial, temporal, and magnitude distributions of aftershocks. ANN are computational systems with learning capabilities, inspired by the working principle of biological neurons in the human brain. These networks have a multi-layered structure consisting of a large number of interconnected nodes (neurons) [11,12]. Input data is transmitted to the first layer of the network; from here on, it is processed through weighted connections and nonlinear activation functions between the nodes in each layer and transferred to the next layer.

As a result, the network creates abstract representations and performs tasks such as classification, prediction, or pattern recognition through these representations [13].

Deep neural networks (DNN) are ANN structures that consist of a large number of hidden layers and can learn more complex patterns. These networks are used in tasks such as classification and regression. While classification tasks aim to separate data into specific categories, regression tasks aim to predict continuous values [14]. In seismological studies, these methods are used in various tasks such as predicting whether aftershocks will occur (binary classification), predicting the number or magnitude of aftershocks (regression) [15].

Deep learning models can provide high accuracy results by processing physical and spatial variables such as stress tensor components, Coulomb stress change, main shock parameters and location information [16,17]. Some common architectures used in this context include fully connected networks, convolutional neural networks (CNN), long-short term memory networks (LSTM) and hybrid models [18,19].

In recent years, especially in the field of aftershock prediction, multi-task learning structures that model regression and classification tasks together have become widespread. These models allow multiple target variables (e.g. both aftershock occurrence and magnitude estimation) to be processed in the same network [20].

Models used to explain aftershock triggering can be generally classified into two groups as static and dynamic stress models. While static models focus on permanent stress field changes caused by the main shock, dynamic models argue that seismic waves can trigger aftershocks, especially in distant regions, with transient but high-amplitude stress oscillations. Dynamic stress effects were found to be significant for some remote triggering events [21,22]. However, since modeling dynamic stress changes requires high-resolution seismic records and computationally complex scenarios, its applicability to large data sets remains limited. For this reason, the static stress approach is still more widely preferred in spatial prediction of aftershocks.

2.2. Studies for Seismic Prediction

However, not only static stress-based explanations are not enough, but also various approaches have been developed to explain the statistical regularities that aftershocks exhibit in time and magnitude dimensions. In this context, the temporal distribution of aftershocks is usually modeled by the Omori-Utsu law. This law predicts that the number of aftershocks following a mainshock decreases over time and this decrease follows an inverse exponential relationship [23]. On the other hand, Båth's law states that the average magnitude difference between the mainshock and the largest aftershock is generally constant (approximately 1.1 units) [24]. These two empirical laws can successfully explain the behavior of aftershock sequences in time and magnitude; however, they are insufficient when it comes to spatial distribution, and therefore require more complex, physics-based or machine learning-supported modeling approaches [3,25].

One of the most common statistical models developed in this context is the Epidemic Type Aftershock Sequence (ETAS) model. ETAS describes a process in which each earthquake produces its own aftershocks and assumes that the aftershock intensity decreases over time in accordance with the Omori-Utsu law. However, the spatial component of the model is generally based on proximity-based distribution functions; therefore, it does not directly account for real physical stress fields [26].

On the other hand, the static Coulomb stress analysis, a physics-based approach, aims to predict spatially in which regions aftershocks will be concentrated by calculating the stress changes in the crust after the main shock [5,27]. This model has been supported by observations that aftershocks are more frequent, especially in positive ΔCFS regions, but has been criticized for its limitations such as the fault plane assumption and the assumption that the crust is homogeneous [28].

To address this gap between classical statistical modeling and physics-based spatial analysis, a hybrid model was proposed by Baranov, Narteau, and Shebalin [29].

This approach combined the temporal productivity capability of the ETAS model with the spatial sensitivity of the Coulomb model based on stress maps. Thus, “real-time aftershock hazard maps” that can be updated in time and produce physically meaningful spatial estimates were created. This hybrid model offers significant contributions to disaster management and risk reduction applications by taking advantage of the flexibility of data-intensive computational techniques while preserving empirical regularities such as Omori-Utsu and Båth laws. However, even these hybrid models have some limitations, especially the dependence on parametric assumptions and the inability to fully reflect heterogeneous conditions.

In this context, deep learning models have increasingly gained attention in the literature as alternative methods that are data-driven and do not impose a priori assumptions on parameters.

One of the most important studies pioneering the application of deep learning in aftershock prediction was carried out by DeVries *et al.* [3]. The researchers aimed to predict the spatial distribution of aftershocks occurring after large earthquakes by learning entirely from the data, without assuming a predetermined fault plane orientation as in classical models. In this direction, 199 finite-fault rupture models obtained from the SRCMOD database were taken as basis and a three-dimensional volume of $100 \times 100 \times 50 \text{ km}^3$ was defined around each mainshock, divided into $5 \times 5 \times 5 \text{ km}^3$ cells. For each cell, the elastic stress change tensor produced by the mainshock was calculated with 6 independent components and a total of 12 input variables were created by evaluating the positive and negative aspects separately.

A deep neural network model with 6 layers and 50 neurons per layer was trained on this large classification data set ($>160,000$ cells) generated with input data. Each cell was labeled according to whether it contained at least one aftershock within a year, using information obtained from the International Seismological Center (ISC) catalog. After training, the model was evaluated on an independent test set of more than 30,000 mainshock-aftershock pairs; the area under the ROC curve (AUC) value was reported

as 0.849. This success is quite high compared to the $AUC \approx 0.583$ obtained in the predictions made using Coulomb ΔCFS maps on the same data.

Remarkably, this deep learning model not only produced highly accurate predictions, but also the internal workings of the model were found to be physically interpretable. More than 98% of the variance in the model outputs could be explained by only three scalar quantities as the maximum shear stress change, the von Mises criterion associated with the second invariant of the deviatoric stress tensor, and the sum of the absolute values of the six components of the stress tensor. This third quantity in particular is a new quantity that has not been proposed in the literature before and its physical interpretation is challenging since it does not remain constant under rotation; however, the network considered the regions where this quantity is high as critical for aftershocks. This situation revealed that the deep learning model can learn not only intuitive patterns but also physically meaningful quantities. DeVries *et al.* interpreted this result as “a static stress-based aftershock prediction rule developed without assuming a priori fault orientation” and emphasized that the study could provide scientific insights into aftershock triggering mechanisms.

Although the deep neural network model developed by DeVries *et al.* has had a significant impact on the literature, some subsequent studies have criticized this model for being overly complex. In particular, Mignan and Broccardo [30] made a comparison of “single neuron vs deep learning” and argued that the distribution of aftershocks can be predicted with similar success with only two simple parameters, such as the average shear magnitude of the main shock and the distance to the cell center. In this context, it has been shown that an artificial neural network with only one hidden layer or even a simple logistic regression model achieves comparable and even higher $AUC \approx 0.85$ performance compared to DeVries’ deep network in some tests. The researchers argued that the deep model actually learns these basic geometric and kinematic relationships, that is, uncomplicated physical regularities underlie the strong performance of the model. Stating that unnecessary model complexity reduces interpretability, Mignan and Broccardo emphasized that the simplest model compatible with the principles

should be preferred. They reiterated these views in their new study published in 2023 and showed that simpler artificial neural networks can also produce physically meaningful and sufficient predictions [30,31].

DeVries *et al.*'s [3] study focused on the prediction of the spatial distribution of aftershocks and limited the time dimension to only one year after the main shock. However, aftershock processes show an evolutionary structure depending on time; aftershock activity, which is intense in the early periods (first hours and days), thins out over time and the area affected during this thinning can also expand. This temporal change is called the "aftershock hysteresis effect" in the literature; that is, aftershocks can accumulate not only immediately after the main shock but also in different regions over time. In this context, Chen *et al.* [32] carried out an important study in order to model this effect by making the deep learning approach sensitive to time.

In this study published in the journal Natural Hazards and Earth System Sciences, finite fault models of many large earthquakes obtained from the SRCMOD database and the stress tensor components calculated from these models were used as input. However, unlike the study by DeVries *et al.* (2018), they modeled the aftershock distribution not only in a single time interval but also for five different time periods (1, 30, 90, 180 and 365 days later). A different neural network sub-model was trained for each time interval, thus creating a "hysteresis model" that can track the spatial spread of aftershocks over time.

In this study, AUC values are above 0.8 for all time intervals, some approaching 0.9. These results show that deep learning models can successfully predict not only long-term but also short- and medium-term aftershock distribution. Furthermore, when the model outputs were analyzed in detail, it was determined that the distribution of aftershocks expanded more slowly over time, initially exhibited rapid propagation, and then formed a fixed pattern over time. The spatial distribution of aftershocks showed an anisotropic, i.e., non-uniform directional structure in each time period. This implies that the area triggered by the main earthquake remained roughly constant, but

over time, aftershocks continued to fill this area, and new aftershocks did not travel far. The study was not limited to horizontal distribution only, but also presented important findings in depth. For example, while it was observed that aftershocks were most densely clustered at 10–15 km depth in the 2008 Wenchuan earthquake, it was reported that this depth range increased to 35–40 km in the 2011 Tohoku earthquake. This shows that deep neural networks can produce meaningful insights not only into surface seismicity but also into physical processes within the crust. Furthermore, in the analyses comparing the decreasing number of aftershocks and the propagation speed of the model with Omori’s law, it was revealed that the spatial propagation was relatively wider in periods when the number of aftershocks decreased rapidly, and that the propagation slowed down in later periods as the aftershock density decreased. This emphasizes the importance of evaluating spatial and temporal components together. This study showed that deep learning can successfully model not only static but also dynamic processes, and that incorporating the real fault geometry into the model, instead of the classical point source assumption, significantly increased the prediction performance. Thus, the study made significant contributions to predicting both where and for how long aftershocks will intensify in emergency response processes. This study demonstrated the potential of deep learning in aftershock prediction in a broader context and brought an innovative perspective to the field.

2.3. Other Machine Learning Applications and Hybrid Methods

In addition to deep learning approaches, there are studies in the literature where more traditional machine learning (ML) algorithms have demonstrated remarkable performance in aftershock prediction. In particular, these methods, which can integrate multiple data sources, offer the advantage of evaluating various factors such as fault geometry, stress change, and spatial relationships together. For example, Karimzadeh *et al.* [16] performed spatial prediction of aftershocks in the region using the slip distribution obtained from satellite (InSAR) data and the Coulomb stress change maps calculated accordingly for the 2017 Mw 7.3 Kermanshah (Western Iran) earthquake. In this study, different machine learning classifiers such as Naive Bayes, Support Vector

Machines, and Random Forest were tried; As inputs, features from multiple sources such as the amount of slip on the main fault, the Δ Coulomb value at that point, and the distance to neighboring active faults were used. The results showed that machine learning-based predictions provided higher accuracy compared to classical Coulomb stress maps. In particular, when detailed fault maps and stress data were available, the Random Forest model was able to successfully classify the distribution of aftershocks. However, the same study indicated that the performance of a simple logistic regression model was very close to advanced ML models. This indicates that, as in the deep learning discussion, more complex models provide limited advantage in aftershock prediction. Nevertheless, thanks to the flexibility of ML techniques, the ability to integrate multiple data layers (multi-source data) is a significant advantage.

Another remarkable study was conducted by Wang *et al.* [33]. In this study, the authors tried to estimate some features of aftershocks from mainshock parameters by using information from existing comprehensive earthquake catalogs. With the “Support Vector Regression” method they proposed, they aimed to establish a correlation between various parameters of a mainshock (magnitude, ruptured energy, seismic moment, main stress drop, etc.) and certain characteristics of the aftershock sequence (such as the largest aftershock magnitude, total number of aftershocks, time distribution of aftershocks).

This study also developed a software for real-time application, aiming to quickly estimate the aftershock hazard based on mainshock information. Although the results of this study have not been fully detailed in the literature, it is an attempt to predict the quantitative characteristics (number, time, magnitude) of the aftershock sequence with machine learning. For example, although Bath’s law gives a general average difference, there may be deviations for each event; Wang’s model perhaps tried to estimate the expected largest aftershock of a mainshock or the expected number of aftershocks in the first 24 hours with various physical quantities of the mainshock (stress drop, moment, etc.).

Indeed, Liu *et al.* [34] proposed a hybrid machine learning model that combines multiple data sources and focuses on explainability in generating aftershock hazard maps. In this model, seismological catalogs (ISC, USGS, etc.), geographic-physical data (e.g. regional fault lines, soil parameters) and classical stress calculations are used together to obtain more interpretable results in mapping aftershock risk. Such approaches, which are still new in the literature, aim to make transparent which factors contribute to the aftershock hazard to what extent by revealing the decisions of machine learning models with tools such as Shapley values. In this way, it is aimed to close the gap between data-based models and physical interpretation.

Not only where aftershocks will occur, but also how many and how large they will occur are important in terms of seismological analyses. Traditionally, the expected number of aftershocks after a mainshock is explained within the framework of Omori's law. However, it is not fully clear how the parameters of this law, such as productivity and decay slope, are related to the physical properties of the main earthquake. Recent studies have shown that other factors affecting the number of aftershocks are also important. In particular, it has been determined that elements such as the heterogeneity of the stress distribution, the amount of stress drop and the regional tectonic structure can directly affect aftershock productivity. In this context, Sharma *et al.* (2020) showed that scalar quantities such as von Mises and maximum slip give better results compared to the classical Coulomb criterion in estimations made with different stress measures.

Predicting the magnitude of aftershocks is still seen as a very challenging problem in the literature. According to Bath's law, the magnitude of the largest aftershock is on average 1.2 units lower than the main shock [24]. However, when looking at real data, this difference varies greatly from event to event. Currently, there is no model in the literature that accurately predicts the magnitude of the largest aftershock after a given main earthquake. The main reason for this uncertainty is that the magnitude of the aftershock is largely stochastic in nature. However, some studies attempt to provide indirect information about the possible magnitude structure of the aftershock

sequence, for example by analyzing the change in the Gutenberg-Richter b-value [35]. Although these methods provide statistical clues about whether the aftershocks will be large or small, they do not provide a definitive prediction.

2.4. Overview

In general, the location, number and magnitude of aftershocks are mostly considered separately in the literature. However, evaluating these three components together is valuable in terms of establishing a more holistic prediction system. Deep learning models have produced successful results in spatial predictions, especially by processing physical data such as stress components. However, deep learning studies focusing on number and magnitude estimation are limited.

Various machine learning and physics-based methods have been developed in the literature for aftershock prediction. Most existing studies treat spatial distribution, numerical density, and magnitude estimation as independent subproblems and typically build models based on a single target variable. Furthermore, stress field information is often represented only by local cell values; the influence of neighboring cells, stress gradients, or spatial correlations are not directly integrated into the model. This makes it difficult to fully capture the spatial continuity observed in the three-dimensional structure of the stress field.

This study addresses these limitations and also considers the local environmental context of the stress field. Each cell is evaluated together with the stress components of 26 neighboring cells in three-dimensional space, thus directly incorporating spatial interactions and gradient structures into the model. Furthermore, a kernel-based probabilistic formulation, largely overlooked in the literature, is proposed: aftershock labels are transformed into a probabilistic target by convolving them with a three-dimensional Gaussian kernel. This approach allows the model to go beyond deterministic classification or regression estimation to also represent the spatial uncertainties and probability distributions of aftershocks. Thus, the study offers a unique contribution to after-

shock modeling by integrating neighborhood, gradient, and kernel-based probabilistic components. These elements are currently missing in the existing literature.



3. MATERIALS & METHODS

In this chapter, we describe the methodological framework used to model aftershock behavior from static stress information. We first define the prediction problems – aftershock occurrence, count, and magnitude – and present the construction of the dataset, including finite-fault stress tensors, aftershock catalogs, and the 3D grid representation around each mainshock. We then detail the preprocessing pipeline, covering feature engineering from stress tensors, kernel-based target reformulation, neighborhood and distance features, normalization, and dimensionality reduction via Principal Component Analysis. Subsequently, we introduce the model architectures employed in this study, namely multilayer artificial neural networks, 3D U-Net, and XGBoost, together with the hyperparameter optimization strategy. Finally, we outline the evaluation metrics used for classification, probabilistic calibration, and regression, which provide a consistent basis for comparing all model variants.

3.1. Problem Definition

Understanding the spatial pattern of aftershocks following earthquakes is an important research area for modeling seismological processes and predicting potential future earthquakes. Seismic activity observed in regions where aftershocks occur is directly related to the accumulation and distribution of static stress in the Earth’s crust. This relationship can be analyzed more thoroughly by combining physics-based stress models and observational aftershock data. In particular, the decisive role of static elastic stress tensors in the formation of aftershocks provides a suitable basis for developing spatial prediction models.

This study aims to predict three fundamental characteristics of aftershocks occurring after major earthquakes: whether an aftershock occurred in a grid cell, the number of aftershocks, and their average magnitude. The models developed for this purpose are configured for both classification and regression tasks. During the predic-

tion process, the model considered not only the stress components of the target cell but also derived stress properties such as spatial derivatives of stress, gradient norms, stress trace, L1–L2–Linf norms, eigen-based quantities, maximum shear, and the von Mises equivalent stress. These variables were evaluated together with stress information obtained from neighboring cells, integrating the spatial dependence and environmental context of aftershocks into the model. This resulted in a more holistic and physically meaningful modeling approach that reflects not only the local stress field but also its environmental variations.

In this context, the main problem of the study is to predict the location, number, and average magnitude of aftershocks using stress tensors, derivative properties, and spatial neighborhood information.

3.2. Data Description

The data set used in this study covers major earthquakes occurring from 1968 to 2012 and the aftershocks that occurred within a year following these earthquakes. This allows the spatial pattern of short-term aftershock activity to be modeled and analyzed at high resolution. The data set was created by integrating observational seismological catalogs with elastic displacement models based on geophysical stress calculations. This structure allows each grid cell to be represented in detail in terms of both physical stress state and seismic activity. The aim is to create a robust and rich data base for estimating the spatial distribution and intensity of aftershocks. The data set consists of two main components. The first component consists of fault plane slip models for large-scale earthquakes occurring worldwide. These models, taken from the SRCMOD database, include parameters such as fault length, width, strike, slope, and surface slip for each earthquake. A total of 199 finite fault models correspond to 118 different mainshock events. The second component is aftershock data from the ISC-GEM catalogs. These data include the location, magnitude, and depth of each seismic event; only aftershocks occurring within the impact area of the relevant mainshock and occurring within a year following it are considered.

The area surrounding each mainshock was modeled as a three-dimensional analysis volume. This volume, extending to a depth of 100 kilometers horizontally and 50 kilometers vertically, was divided into equal-volume grid cells of $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$. This resulted in thousands of independent cells for each mainshock. Elastic stress calculations were performed at the center of each cell, yielding six independent stress tensor components.

The stress tensor is a symmetric tensor representing both normal and shear (*shear*) stresses along three axes. The σ_{xx} , σ_{yy} , and σ_{zz} components represent normal stresses perpendicular to the plane about the x , y , and z axes, respectively. Positive values indicate compression, while negative values indicate tension. The σ_{xy} , σ_{yz} , and σ_{xz} components represent shear stresses parallel to the plane; these components are directly related to the probable fracture orientations along the fault planes.

Various scalar quantities have been derived from these fundamental tensor components. These include:

- Maximum shear stress ($\Delta\tau_{\max}$) is a direct indicator of the fracture potential;
- Von Mises yield criterion ($\sqrt{\Delta J_2}$) is based on deviatoric components and reflects the plastic deformation potential;
- Total absolute stress ($\sum |\sigma_{ij}|$) is a density-based summary of all components of the stress tensor;
- Coulomb stress change (ΔCFS) is defined as

$$\Delta CFS = \Delta\tau - \mu \Delta\sigma_n \quad (3.1)$$

where μ represents the friction coefficient, usually taken as 0.4 [36].

The data set contains two important variables that define the seismic activity of each grid cell. The "grid_aftershock_count" column represents the annual number of aftershocks per cell, while the "Magnitudes" column lists the moment magnitude values for these aftershocks. This allows for cell-based analyses based on both frequency and magnitude distributions.

The data set contains a total of 6,120,930 grid cells. Only 101,727 of these cells contained at least one aftershock, while the remaining 6,019,203 cells had no aftershocks. The data set has a severely unbalanced distribution with approximately 1.7% positive labels. See the Figure 3.1

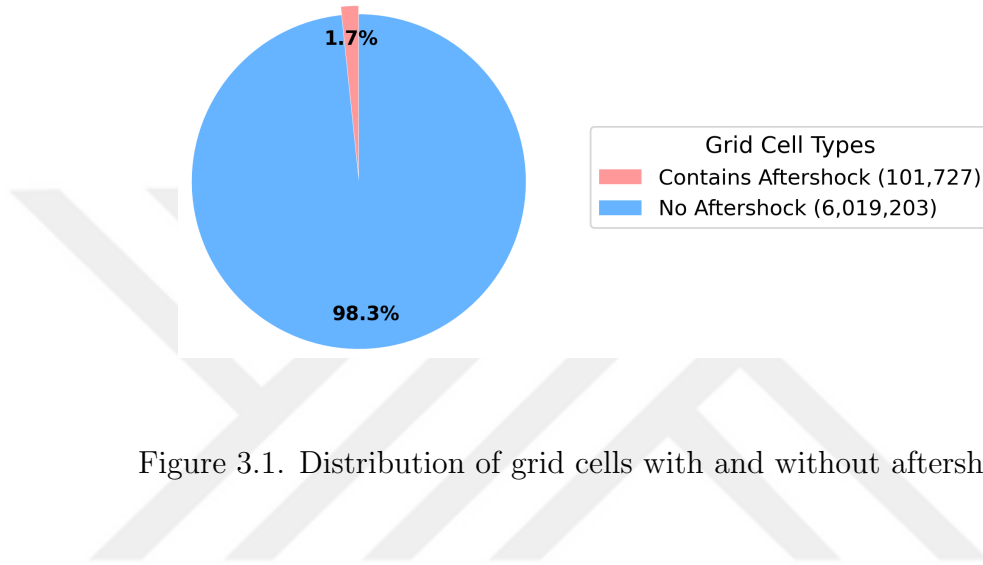


Figure 3.1. Distribution of grid cells with and without aftershock.

3.3. Data Preprocessing

The data preprocessing phase involves the systematic transformation and labeling of observational seismic records to form the basis of modeling. In this phase, the dataset was not only utilized in its raw form but also extended through the creation of target variables, the definition of neighborhood relationships, and the normalization of numerical attributes to make them ready for modeling.

3.3.1. Feature Selection and Engineering

The modeling process begins with the systematic labeling of observational seismic data. Aftershock information for each grid cell was matched with earthquake records from international catalogs, and the labels required for classification were generated accordingly. If a cell experienced at least one aftershock within one year following the mainshock, it was labeled as 1 (contains aftershocks); if no aftershocks were observed, it was labeled as 0 (no aftershocks). The *is_aftershock* column defined in this way is

the product of a purely observational labeling procedure and does not represent the output of any predictive model. This binary structure provides a directly usable target variable for classification tasks.

In addition to the binary presence of aftershocks, their intensity characteristics were also included in the modeling process. The magnitudes of the observed aftershocks were listed in a column named *Magnitudes*, and a unique value for each cell was obtained by averaging these magnitudes. These values were defined in a new column named *Average Magnitudes*, creating a continuous target variable reflecting the mean aftershock magnitude. Similarly, the *grid_aftershock_count* column, representing the total number of aftershocks in each cell, was used as a second target variable for regression models.

Kernel-Based Label Reformulation

In this study, a kernel-based target reformulation was implemented to extend beyond the classical 0/1 labeling framework while preserving the spatial pattern of aftershocks. A three-dimensional isotropic Gaussian kernel, parameterized by $\sigma = 10$ km and $R = 30$ km, was generated around each aftershock center, providing a distance-weighted contribution to neighboring cells. Consequently, each grid cell obtained a kernel density value reflecting the cumulative influence of surrounding aftershock centers.

This kernel-based density formulation was applied to both classification and regression targets. For the *is_aftershock* variable, the kernel-derived density was transformed into a smoothed probabilistic target within the range $[0, 1)$ using the exponential mapping as

$$y = 1 - \exp(-\alpha s) \quad (3.2)$$

where s denotes the cumulative isotropic Gaussian kernel intensity centered at each aftershock location, and α is a scaling parameter controlling the smoothness of the

exponential decay. The Gaussian kernel itself is defined as

$$K(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad (3.3)$$

where r represents the Euclidean distance from the aftershock center, and σ determines the spatial decay radius of the kernel. This transformation converts the original binary indicator into a continuous probabilistic representation, which was employed for probabilistic classification. For the *grid_aftershock_count* variable, the same kernel approach was used to produce a continuous, kernel-weighted count target representing the spatial distribution of aftershock intensity. This formulation preserves the discrete structure of the original labels while allowing the model to capture spatial uncertainties and smoothly varying density patterns across the stress field.

Extended Stress-Derived Features

The input feature space was not limited to the principal static stress components; numerous spatial derivatives and derivative-based summary quantities derived from these components were also incorporated into the model. In this context, the partial derivatives of each stress component with respect to the three spatial directions (x, y, z) were calculated, and the gradient norms of these derivatives were used to quantitatively represent orientation and amplitude variations in the stress field. Additionally, several scalar quantities summarizing the overall stress characteristics were computed, including the stress tensor trace, L_1 - L_2 - L_∞ norms, maximum shear magnitude, isotropy ratio based on the maximum-to-minimum eigenvalue ratio, and the von Mises equivalent stress. Together, these features represent the magnitude, anisotropy, and directionality of the local stress field.

Neighborhood and Distance Features

The dataset was further expanded to include not only physical and seismic information but also spatial context. Each grid cell has a total of 26 neighbors surrounding it, forming a cubic system with approximately 5 km spacing. This structure defines 6 surface neighbors (one forward and one backward along each axis: $\pm x, \pm y, \pm z$), 12 edge neighbors (points located along two axes), and 8 corner neighbors (all $\pm$ combinations of the three axes). Neighborhood relationships were calculated using the KDTree

algorithm based on each cell's coordinates, and only cells with exactly 26 neighbors were included in the dataset. Each neighbor was labeled with directional information and categorized into three main groups: surface, edge, and corner neighbors. Summary variables such as the number of aftershocks and average magnitudes were computed for each group and appended to the dataset. In addition, the Euclidean distance of each grid cell from the mainshock's centroidal (x, y) position was calculated and included as an additional feature to assess the spatial dependency of seismic activity.

3.3.2. Normalization and Scaling

In data mining and machine learning applications, particularly in gradient-based methods, the similarity of input variables can directly impact the model's learning process. Therefore, it is important to express numerical attributes on a common scale. In this study, Z-score standardization was applied to the continuous variables in the data set.

Z-score normalization is a method that scales each observation value by dividing its distance from the mean of its distribution by its standard deviation. This process is defined by the following formula

$$z = \frac{x - \mu}{\sigma} \quad (3.4)$$

where x represents the observation value, μ represents the mean of the relevant variable, and σ represents its standard deviation. As a result of this transformation, each variable is rescaled so that its mean is 0 and its standard deviation is 1. This makes the variables comparable and reduces the risk of biased weight updates in the modeling process [37].

3.3.3. Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that aims to represent a data set with fewer components by considering correlations between variables in high-dimensional data. This method is used to reduce the computational burden and to increase the interpretability of the modeling process [38].

In this thesis, in some model configurations, the stress tensor components of neighboring cells belonging to each grid cell were included in the data set. Because the neighborhood structure was considered in three categories (surface, edge, corner), the high dimensionality of the numerous input variables from these regions posed a potential problem for model performance. Therefore, these groups were considered separately and PCA was applied. A cumulative variance ratio plot (Scree Plot) was first generated for each group and the number of components accounting for 95% of the explained variance was used as the basis. A number of principal components determined according to this threshold were fed into the model as input instead of the original data. This process was implemented for dimensionality reduction and information condensation, particularly in extended model variants.

3.4. Model Architecture

This study utilizes deep learning and machine learning techniques. This section will examine various model architectures used in aftershock forecasting from a technical perspective. For each model, its structural features and operating principles will be explained in detail, and its basic architecture, learning process, and key hyperparameters will be discussed. Diagrams visually illustrating the architecture and learning process for each model are also presented. Among the models to be examined are Artificial Neural Networks (ANN) and eXtreme Gradient Boosting (XGBoost). This section will focus on the model architectures used and explain each model's learning mechanism in technical yet understandable terms.

3.4.1. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are multilayered and multinode computational models inspired by the nerve cells in the human brain [39]. An artificial neural network consists of units, or neurons (nodes), organized into layers. It typically has three types of layers: an input layer, one or more hidden layers, and an output layer. Nodes in the input layer receive features from the external world, while the output layer generates

the model's predictions. Hidden layers lie between these two and enable learning the complex relationship between input and output. Each artificial neuron receives values from the nodes in the previous layer connected to it via weights, calculates a weighted sum of these values, and applies an activation function to this sum. The resulting activation value is then passed to the next layer. This process generates output from the input data through forward computation (forward propagation). The Figure 3.2 shows the layered architecture of a simple feedforward artificial neural network.

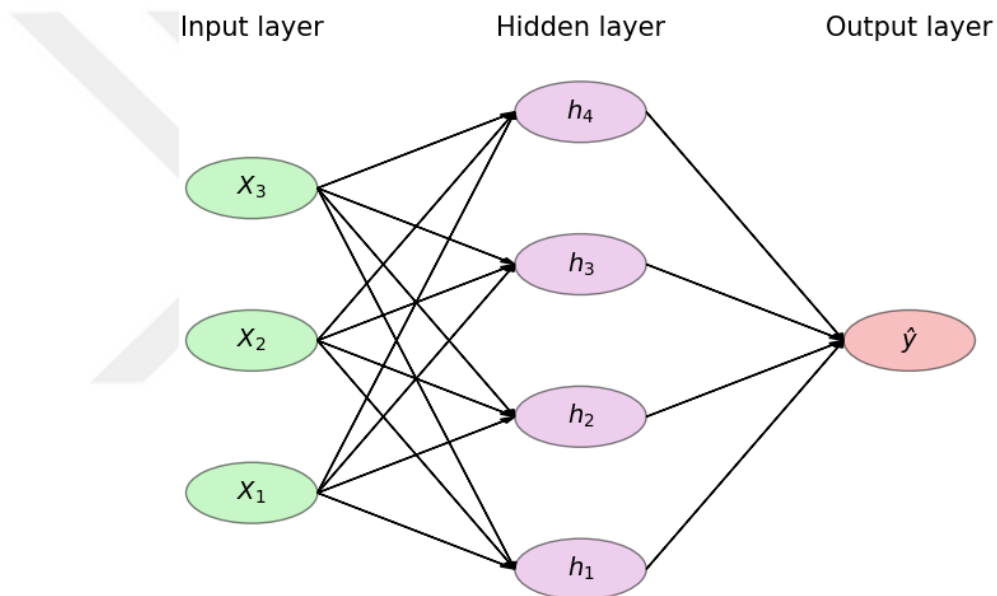


Figure 3.2. An example multilayer neural network architecture.

The key element that distinguishes an artificial neural network from linear models is the activation function applied by each neuron after the linear combination process. The activation function takes the sum of the linear inputs and transforms them into a nonlinear output. This allows the network to learn not only linearly separable problems but also complex and nonlinear relationships. For example, if no activation function is used (the unit function, i.e., $f(x) = x$), the network, regardless of the number of layers, still generates a linear mapping between input and output—in which case, hidden layers are useless.

The following equations represent the most widely used activation functions: sigmoid, tanh, and ReLU,

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (3.5)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad (3.6)$$

$$\text{ReLU}(x) = \max(0, x). \quad (3.7)$$

The sigmoid and tanh functions are “S-shaped” curves; sigmoid gives output in the range $[0, 1]$, while tanh gives output between $[-1, 1]$. The ReLU activation function is zero for negative inputs and linear (has unit slope) for positive inputs.

3.4.2. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN), one of the deep learning architectures widely used for image data analysis, provide particularly successful results in classification and segmentation tasks thanks to their ability to learn spatial relationships through convolutional layers [40]. CNN architectures enable the automatic processing of high-dimensional data by learning meaningful patterns in images. This structure, unlike classical fully connected artificial neural networks (ANN), reduces the number of parameters and is more effective at capturing local patterns.

In this regard, a CNN-based model called U-Net has been developed for use in problems requiring pixel-based location estimation, such as biomedical image segmentation. The UNet architecture is a CNN structure originally developed for biomedical image segmentation [41]. This structure follows an encoder-decoder scheme; The encoder extracts contextual information from the input data, while the decoder recovers spatial resolution, providing detailed segmentation. The key feature that distinguishes UNet from previous fully convolutional networks is the skip connections established between the encoder and decoder layers. These connections allow for the recovery of detailed spatial information lost during downsampling [41].

The 3D version of UNet was developed specifically for working with three dimensional volumetric data. In the 3D UNet architecture, 2D convolution and pooling operations are replaced with their 3D counterparts, allowing the model to simultaneously learn spatial features across width, height, and depth [42]. This feature makes it particularly suitable for working with multidimensional structures such as seismic volumetric data.

One of the most significant advantages of the UNet architecture is its ability to produce effective results even when the dataset is limited. This feature, combined with its structure suitable for data augmentation, has enabled UNet to be widely used not only in the medical field but also in different disciplines such as geoscience, environmental monitoring and autonomous systems [41]. The structure of the UNet model is illustrated in Figure 3.3.

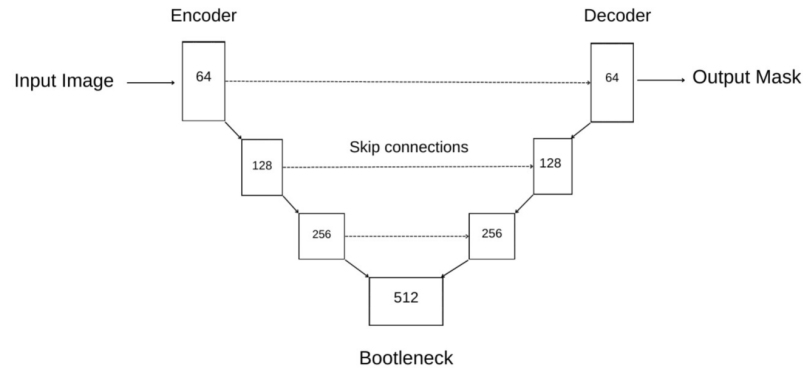


Figure 3.3. An example UNET model architecture.

3.4.3. eXtreme Gradient Boosting

The eXtreme Gradient Boosting is a powerful ensemble learning algorithm based on decision trees. It is based on gradient boosting, which is a method that generates short and simple decision trees from sequentially trained weak learners. Each tree is trained to compensate for the errors of the previous model, thus improving the

model’s predictive performance at each step [43]. XGBoost implements this approach by making it more efficient and regularized with an optimized objective function and techniques such as quadratic Taylor expansion [44].

A key contribution of XGBoost is its regularization terms that balance model complexity, its ability to handle missing data, and its high-speed processing of large data sets. Furthermore, technical options such as histogram-based tree growing and GPU acceleration increase computational efficiency [45]. For these reasons, XGBoost is a flexible and powerful modeling tool frequently preferred for both classification and regression problems.

3.5. Performance Metrics

3.5.1. Classification Metrics

Basic metrics such as accuracy, precision, recall, and F1-score were used to evaluate the performance of classification models. These metrics allow us to analyze not only the number of correct predictions but also the frequency of each type of error, helping to assess the model’s strengths and weaknesses in a multidimensional manner. Accuracy reflects the model’s overall success rate across all examples; however, it is not a sufficient indicator on its own, especially in data sets with imbalances between classes. For example, in a data set with very few positive classes, the model may achieve high accuracy by predicting only the negative class; this may conceal the fact that the model is not actually learning the positive class [46].

Precision measures the proportion of examples the model classifies as positive that are actually positive, while recall indicates how accurately the model can predict true positive examples. Considering these two metrics together allows for a more balanced assessment of the model’s false positive and false negative tendencies [47]. In this context, the F1 score, the harmonic average of precision and recall, offers a more reliable performance measure, especially for problems with imbalanced classes.

Because the F1 score considers the impact of both false positive and false negative results on model performance, it more effectively compensates for class imbalance.

The confusion matrix was used as a fundamental tool in the calculation and interpretation of all these metrics. The confusion matrix organizes the correct and incorrect classifications made by the model into four categories: true positive, false positive, true negative, and false negative. This structure makes it possible to visualize the classes in which the model made errors and analyze its classification behavior in more detail. A particularly high false negative rate can indicate that the model is inadequate in recognizing the positive class, which can pose critical risks in some application areas.

In addition, the study used the ROC curve (Receiver Operating Characteristic Curve) and the AUC (Area Under the Curve), which represents the area under this curve, to further evaluate model performance. ROC AUC is a comprehensive indicator that measures the model's ability to distinguish between positive and negative classes. An AUC value approaching 1 indicates that the model has high discriminatory power between classes, while values closer to 0.5 indicate that the model is making random predictions. In this respect, ROC AUC offers the opportunity to evaluate the model's overall classification performance without relying solely on a fixed threshold value, yielding more meaningful results than the classical accuracy metric, particularly in imbalanced data sets. Taking all of these metrics together, the performance of the classification models can be comprehensively analyzed both quantitatively and qualitatively, and the impact of the sampling strategies and architectural choices is clearly demonstrated.

3.5.2. Probabilistic Evaluation Metrics

For the kernel-based probabilistic classification model, additional metrics that quantify calibration and probabilistic accuracy were used alongside traditional classification scores. These include the Binary Cross-Entropy (BCE) loss and Brier Score.

Binary Cross-Entropy (BCE), also known as log-loss, measures the divergence between predicted probabilities p_i and soft target values $y_i \in [0, 1]$ defined by

$$\text{BCE} = -\frac{1}{N} \sum_i [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]. \quad (3.8)$$

Lower BCE values indicate that the predicted probabilities are closer to the true targets, implying a better probabilistic fit of the model. The Brier Score represents the mean-squared difference between predicted probabilities and observed outcomes by the formula

$$\text{Brier} = \frac{1}{N} \sum_i (p_i - y_i)^2 \quad (3.9)$$

It reflects both the *calibration* (how well predicted probabilities correspond to actual frequencies) and the *sharpness* (the concentration of predicted probabilities away from 0.5) of the model. Lower Brier scores indicate more reliable probability estimates.

In addition, ROC-AUC was used as a threshold-independent ranking measure, capturing whether positive samples systematically receive higher predicted probabilities than negative ones. Average Precision (AP), equivalent to the area under the precision-recall curve, was also computed, providing a more robust indicator than ROC-AUC in highly imbalanced data sets. Finally, the F1-score was computed across a range of decision thresholds, and the optimal threshold corresponding to the maximum F1 value was reported.

This combined evaluation — including BCE and Brier Score - provides a comprehensive assessment of both discrimination and calibration performance for the probabilistic formulation.

3.5.3. Regression Metrics

To evaluate the performance of regression models, three commonly used performance metrics Mean Squared Error (MSE), Mean Absolute Error (MAE), and Co-

efficient of Determination (R^2) were used in the study. These metrics allow for the analysis of the model's accuracy in numerical predictions from different perspectives.

One of the most commonly used metrics for evaluating the performance of regression models is the Mean Square Error (MSE). MSE is calculated by averaging the squared differences between the model's predicted values and the observed true values. This method penalizes large deviations, increasing the model's sensitivity to large errors. However, the squared unit can make the MSE difficult to interpret directly. Therefore, the Root Mean Square Error (RMSE), calculated by taking the square root of the MSE, is often preferred. RMSE allows errors to be interpreted in their original units and provides sensitivity, particularly when modeling large deviations [48,49]. Taken together, these two metrics reveal both the model's overall error level and its predictive performance at extreme values.

Mean Absolute Error (MAE) is a regression evaluation metric calculated by averaging the absolute differences between the model predicted values and the observed true values. By weighting each error term equally, MAE provides a measure insensitive to the direction of deviations. This feature makes it highly intuitive to interpret and directly carries unitary meaning. MAE is particularly suitable for evaluating model performance when the error distribution is homogeneous. Furthermore, because it is more tolerant of large errors than metrics such as MSE or RMSE, it provides a balanced reflection of the model's overall predictive accuracy [49,50].

The coefficient of determination (R^2 score) indicates the extent to which the model explains the total variability in the dependent variable. The R^2 value ranges from 0 to 1. As the R^2 value approaches 1, the model better fits the data. As it approaches 0, the model's explanatory power decreases. Negative R^2 values indicate that the model underperforms even with a fixed mean estimate. The R^2 score is a widely used and intuitive metric, particularly in comparative model evaluations [51].

4. EXPERIMENTS & RESULTS

In this chapter, the performance of the proposed models is evaluated through a comprehensive set of experiments designed to predict the spatial distribution, numerical intensity, and average magnitude of aftershocks. Using a unified dataset constructed from elastic-static stress tensors, derivative- and norm-based features, neighborhood information, and observational aftershock records, a wide range of model variants were systematically tested. All experiments were conducted on a consistent 80/20 train–test split, ensuring comparable conditions across classification, probabilistic, and regression tasks. The following sections detail the experimental setup and model configurations, and present the results obtained for aftershock occurrence, kernel-based probabilistic modeling, count regression, and magnitude regression.

The data set used in the study contains high-resolution stress and aftershock information from 199 different mainshock events. For the training and evaluation of all models, this data was split into an 80% training and 20% testing ratio, providing a controlled test environment to assess the generalization capacity of the models. Each model was built and systematically tested with different input structures based on the same training-test split.

4.1. Experimental Setup and Model Variants

Initially, a basic version of the model was created using only six principal stress tensor components as inputs, and this structure was evaluated as a reference to test the adequacy of local stress in explaining aftershocks. The model inputs were later expanded by adding the physical properties of 26 neighboring grid cells surrounding each cell. First, the stress components of 6 neighboring cells located on the surface plane were included, providing input for regions with direct interaction around the cell. This structure was then supplemented with stress components from 12 neighboring cells on the edge plane and 8 neighboring cells at corner locations.

To reduce the computational burden caused by this increase in input size and to increase the generalization capacity of the model, dimensionality reduction techniques were applied to the stress components of neighboring cells. Specifically, the inputs of surface, edge, and corner neighbors were evaluated as separate groups, and PCA was applied to these subsets to obtain low-dimensional but information-dense representations. Thus, instead of the original high-dimensional input set, a subset with high explanatory variance fed the model.

Following systematic testing of the input vectors, various sampling strategies were applied to prevent the significant imbalance observed between classes from negatively impacting model performance. First, due to the insufficient number of samples from the minority class (i.e., "aftershocks"), oversampling was applied to the training data, and a balanced distribution was achieved by multiplying the samples belonging to this class. This strategy was used to increase the model's sensitivity to the minority class and reduce learning errors resulting from imbalanced data. Conversely, in some experiments, to ensure class balance, the number of samples belonging to the majority class (i.e., "no aftershocks") was systematically reduced, thus creating a more balanced and compact training data set through undersampling.

In addition to this structure, the Euclidean distance variable, representing the distance of each grid cell from the epicenter, was included in the inputs. This variable was considered an explanatory variable that takes into account spatial interactions, particularly because aftershocks are more intense in areas close to the epicenter. The inclusion of distance in the model allowed for the stress distribution to be considered not only locally but also spatially.

In this study, not only deep learning-based artificial neural network and convolutional neural network models were used, but the XGBoost algorithm was also applied to classification and regression tasks as an alternative. In the XGBoost models, the same input structures were maintained, but the advantages of the tree-based approach were utilized. Specifically, due to class imbalance in the classification problem, appro-

appropriate hyperparameters were adjusted to increase the influence of the minority class. In the regression tasks, parameters such as learning rate, tree depth, and subsample ratio were optimized to increase the generalization success of the models.

As part of the study, hyperparameter optimization was performed to maximize the performance of the XGBoost algorithm in classification and regression tasks. For this purpose, Bayesian optimization, a more systematic and sample-sensitive search strategy, was used instead of traditional trial-and-error methods. Bayesian optimization, when directly observing the evaluation function is costly, approximates this function using a probabilistic model (usually Gaussian processes) and performs an efficient scan of the parameter space by selecting new sampling points from this representation. The optimization process was guided by the model’s error function on the validation set (e.g., ROC AUC for classification, MSE for regression), resulting in efficient results in terms of both model performance and the training process.

This framework constitutes the general methodological structure of the basic models (4.1.1, 4.2.1, and 4.3). However, unlike this general structure, the kernel-based models (4.1.2 and 4.2.2) do not include neighborhood-based inputs, PCA reduction, or sampling strategies. Instead, they provide a modeling framework based on probabilistic and density estimation using continuous objective functions obtained with a Gaussian kernel ($\sigma = 10$ km and $R = 30$), along with stress, differential component, and norm-based features.

The models described here test the relationship between stress distribution and aftershocks in different planes, allowing modeling of the impact of local and environmental characteristics on seismic activity in the context of deep learning and machine learning. Since multiple model variants are evaluated throughout this study, a systematic naming convention is used; the abbreviations applied in these model names are summarized in Table 4.1.

Table 4.1. Explanations of abbreviations used in model naming.

Model Naming Codes	Description
BC	Binary Classification model predicting aftershock presence built with only the base 6 stress components
CR	Count regression model for estimating the number of aftershocks built with only the base 6 stress components
R	Count regression model for estimating aftershock magnitude built with only the base 6 stress components
E	Euclidean distance between the grid and the epicenter
F	Stress components of neighboring cells in the Face direction
FE	Stress components of neighboring cells in the Face and Edge directions
FEC	Stress components of neighboring cells in the Face, Edge and Corner directions
P	PCA to all features except the basic 6 stress components and Euclidean distance
SO	Oversampling; balancing data by replicating minority class samples
SU	Undersampling; balancing data by reducing the majority class samples
D	Distance filter; only grids close to the earthquake center are included in the model
X	Model training with XGBoost
C	The absolute value of the grid Coordinates (x,y,z) as input

4.2. Aftershock Occurrence Modeling

4.2.1. Binary Classification of Aftershock Occurrence

The first step of the modeling process was performed in a space created by dividing a $100 \times 100 \times 50$ km cube into equal-volume grid cells of $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$ dimensions, with dimensions of $100 \text{ km} \times 100 \text{ km}$ horizontally and 50 km vertically. A model was developed to predict whether an aftershock occurred in each grid cell within this cubic structure through binary classification. In this classification model, the target variable is the observed aftershock presence for each cell: the target value is

labeled 1 if the cell hosted at least one aftershock within a year of the central shock, and 0 otherwise. Thus, the model aimed to predict the probability of an aftershock occurring in each cell within this three-dimensional volumetric space.

Throughout the modeling process, 12 different classification architectures were created, and various input vectors, layer numbers, and neuron configurations were systematically tested. All network structures were defined on the PyTorch platform and the ReLU activation function was used to capture non-linear relationships in hidden layers. The output layer, a single neuron, generates raw logit values without applying a sigmoid function. Therefore, BCEWithLogitsLoss was chosen as the loss function used in training; thus, the results obtained from the network were normalized in the range [0,1]. The Adam optimization algorithm was used to update the weights, and different learning rates were applied to the models. The training process continued for 300 epochs, and the current loss of the model was monitored at each epoch. Training data were presented to the models in mini batches of 32 or 64 samples, and the data was scaled using z-score normalization before training. To prevent overfitting, a dropout mechanism was also implemented in some network structures; different dropout rates were tested to evaluate the model’s generalization capacity.

To compensate for class imbalances observed in the data sets, some models applied oversampling or undersampling depending on the target variable. These strategies were comparatively evaluated for their impact on the overall performance of the model. Furthermore, some models only filtered grid cells within a specific distance range to make the training data more targeted. In this study, we used artificial neural network (ANN) and convolutional neural network (CNN, specifically 3D U-Net)-based models, as well as XGBoost, a powerful decision tree-based learning algorithm, for classification tasks. In the XGBoost models, hyperparameter optimization was performed using Bayesian optimization, aiming to improve the overall performance of the model.

The performance of the model was evaluated on the test set; output probabilities were converted to class labels using threshold values. Each architecture was evaluated

using different data splits and its performance was compared using classification metrics such as ROC, AUC, accuracy, and sensitivity. In this way, the effect of different structural configurations on classification accuracy was empirically demonstrated.

BC

This model is a reference ANN architecture using only six fundamental stress tensor components as inputs. The primary goal of this model is to predict whether an aftershock will occur in each grid cell using binary classification. The model was designed to test the extent to which stress components can explain aftershock behavior at a local scale in an isolated manner, and in this respect, it serves as a baseline structure that provides a benchmark for advanced models developed with more complex structures.

The model’s architecture is a fully connected deep learning architecture consisting of three hidden layers and one output layer. The input layer projects the six-dimensional input onto a 32-neuron representation plane. This is followed by two hidden layers, each consisting of 128 and 32 neurons, respectively. Thus, the model’s output is a continuous value representing the probability of an aftershock occurring for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. Training data was presented to the model in mini-batches of 32 samples. Probability outputs were converted to class labels using a threshold of 0.5.

The BC model is a simple structure based solely on the principal stress components. An ROC AUC score of 0.77 indicates that the model’s overall discrimination capacity is reasonable. However, the relatively low F1-score (0.08) and precision value (0.04) for the positive class reveal that the model produces a high number of false pos-

itive predictions. The high recall value (0.80) indicates the model’s tendency to detect aftershocks, but the accuracy of these predictions is poor. This suggests that while the weighted loss function applied to address class imbalance increases sensitivity, it may also negatively impact classification stability. Therefore, although the BC model provides high sensitivity with limited information, its high false alarm rate makes it insufficient for practical use. In this respect, it is considered a starting point, emphasizing the need for richer models with more spatial context. The confusion matrix corresponding to the BC model is presented in Figure 4.1

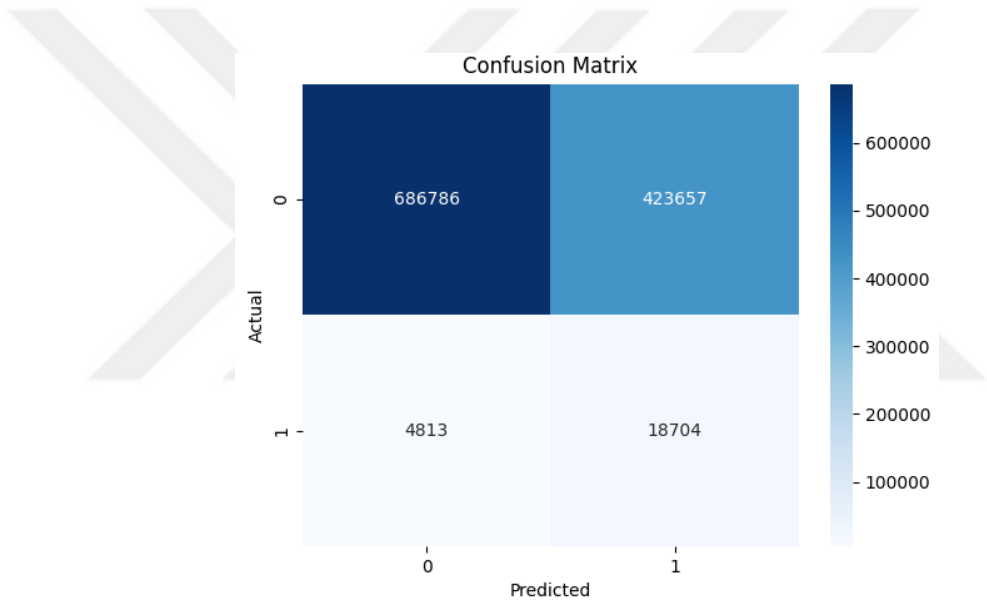


Figure 4.1. Confusion matrix of BC model.

BC-E

This model is an ANN with seven input features, consisting of the six fundamental stress tensor components of the BC model and the Euclidean distance between the epicenter and each grid cell. The model aims to capture spatial relationships influencing aftershock formation. Aftershocks generally cluster near the epicenter; therefore, the distance parameter was added assuming it could strengthen model performance.

The model’s architecture is a fully connected deep learning structure consisting of three hidden layers and one output layer. The input layer projects the seven-dimensional input onto a 32-neuron representation plane. This is followed by two hidden layers, each consisting of 128 and 32 neurons, respectively. Thus, the model’s output is a continuous value representing the probability of an aftershock occurring for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. Training data was presented to the model in mini-batches of 32 samples, and this process continued for a total of 300 epochs. Probability outputs were converted to class labels using a threshold of 0.5.

The BC-E model demonstrated a significant improvement over the baseline BC model by incorporating spatial distance information in addition to the stress tensor components. The ROC AUC score increased to 0.79, and the accuracy reached 0.72, demonstrating that the distance parameter contributes positively to the overall classification performance. Furthermore, the F1 score (0.10) and precision value (0.05) showed a slight increase compared to the BC model. However, the relatively low precision value indicates that the model continues to produce false positive predictions and does not fully address the inherent class imbalance problem. Nevertheless, the more balanced performance profile compared to the BC model confirms that Euclidean distance is an effective feature for representing the spatial context of aftershock occurrence. This result suggests that distance-based features help constrain the spatial extent of predicted aftershock regions, even when stress-based features alone are insufficient.

These results demonstrate that the model partially captures the distribution of aftershocks relative to epicentral distance and provides a meaningful step toward robust spatial modeling. The inclusion of distance information improves interpretability by linking aftershock likelihood to physically meaningful spatial proximity. The confusion matrix corresponding to the BC-E model is presented in Figure 4.2.

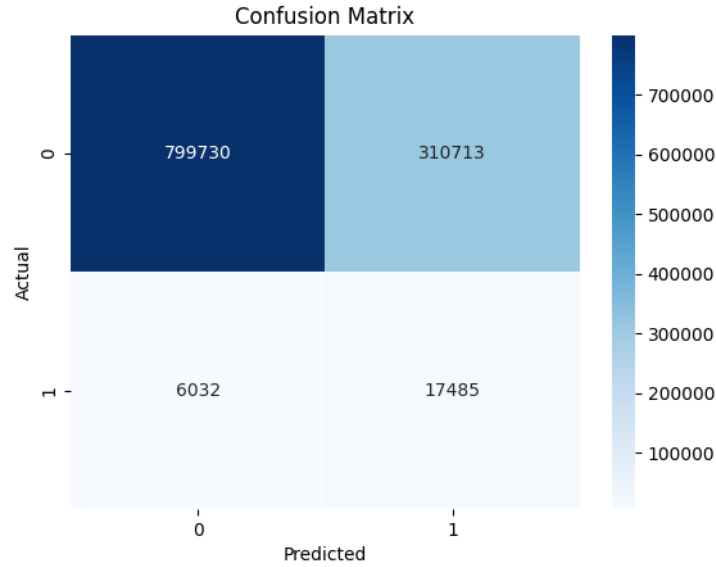


Figure 4.2. Confusion matrix of BC-E model.

BC-E-F

This model is an artificial neural network-based classification model with a total of 43 input features, in addition to the input features from the BC-E model. This model incorporates the stress distributions of not only the central cell but also its surface neighbors, contributing to the model's aftershock prediction. This approach aims to evaluate local stress patterns in a spatial context.

The model's architecture is a fully connected deep learning structure consisting of three hidden layers and one output layer. The input layer projects the 43-dimensional input onto a 64-neuron representation plane. This is followed by two hidden layers, each consisting of 128 and 32 neurons, respectively. Thus, the model's output is a continuous value representing the probability of an aftershock occurring for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. Training data was presented to the

model in mini-batches consisting of 32 samples, and this process continued for a total of 300 epochs. Probability outputs were converted into class labels using a threshold of 0.5.

The BC-E-F model expanded the spatial context by incorporating stress components from neighboring cells on the surface plane into the input. The ROC AUC score, at 0.79, remained at the same level as the BC-E model, indicating that the model’s overall discrimination ability was maintained. However, the F1 score (0.09) and particularly the low precision value (0.05) reveal that the model continues to produce a high number of false positives when detecting cells containing aftershocks. The high recall rate of 0.78 indicates that the model captured the vast majority of positive class examples; however, since a significant portion of these predictions were incorrect, the model’s reliability in distinguishing the positive class remains weak. Therefore, incorporating stress components from the surface plane was not effective in reducing the false alarm rate. The confusion matrix corresponding to the BC-E-F model is presented in Figure 4.3.

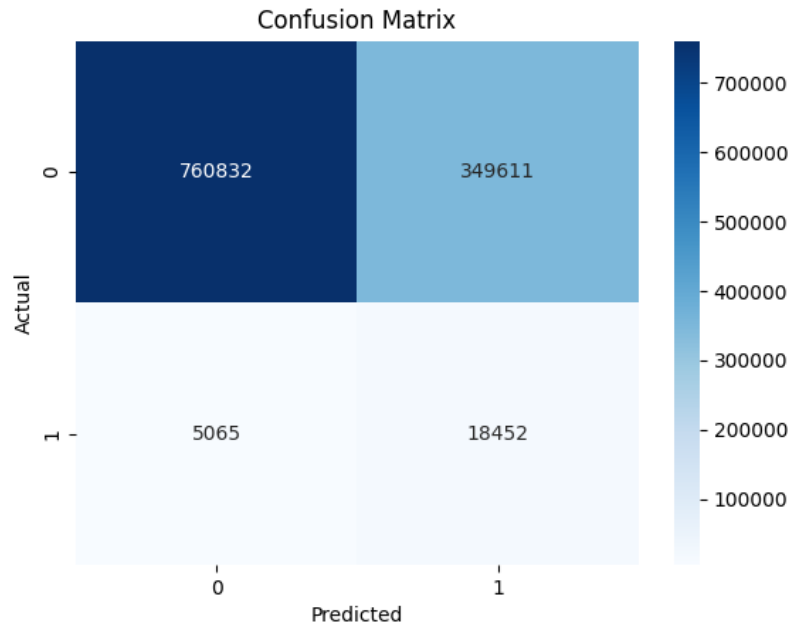


Figure 4.3. Confusion matrix of BC-E-F model.

BC-E-F-P

This model is an extended version of the previous structure, the BC-E-F model. However, here, the stress tensor components of the six neighboring grid cells on the surface plane surrounding the central cell were not directly introduced into the model; instead, this high-dimensional data was subjected to dimensionality reduction using PCA to achieve a more meaningful and compressed representation. This method created a highly information-dense feature set that can contribute to aftershock prediction.

The model's input vector consists of two main components. The first component consists of a total of seven main features, consisting of the six stress tensor components of the central grid cell and the Euclidean distance between each grid cell. The second component is a high-dimensional vector consisting of 36 stress components from the six neighboring cells on the surface plane. This vector was reduced to 25 dimensions by applying PCA, resulting in a total input dimension of 32. The number of PCA components was determined to ensure a cumulative variance ratio exceeding 95%. The variance explained by the PCA components, as determined through the elbow method. It is shown in Figure 4.4.

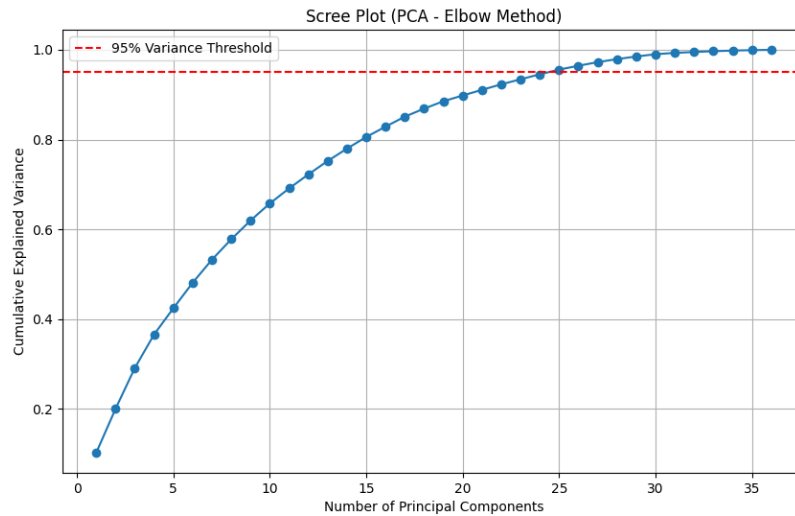


Figure 4.4. Scree Plot for PCA of Face Neighbor Stress Components.

The model employs a fully connected neural network architecture composed of three hidden layers and one output layer. The 32-dimensional input is first projected onto a 64-neuron latent space, followed by two hidden layers with 128 and 32 neurons, respectively. The final output represents the probability of aftershock occurrence for each grid cell.

To address class imbalance, a weighted loss function was applied during training, with the weight determined by the proportion of positive samples. Model parameters were optimized using the Adam optimizer with a learning rate of 0.001 over 300 epochs and a batch size of 32. A threshold of 0.5 was used to convert probability outputs into class labels. The BC-E-F-P model incorporates PCA to obtain a more compact representation of environmental stress information. Although ROC AUC (0.79) and accuracy (0.67) remained comparable to the previous model, the low F1 score (0.09) and precision (0.05) indicate persistent false positive predictions. While PCA improved computational efficiency, it did not sufficiently enhance class separation. The confusion matrix corresponding to the BC-E-F-P model is presented in Figure 4.5.

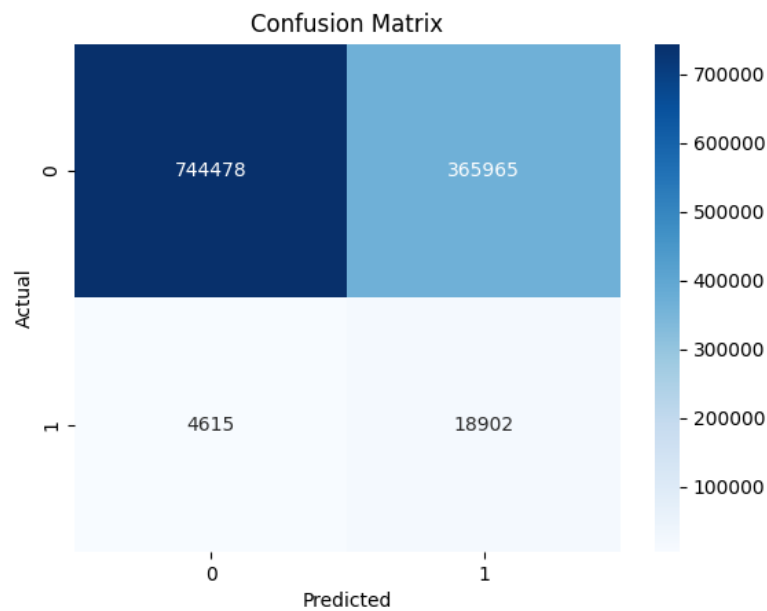


Figure 4.5. Confusion matrix of BC-E-F-P model.

BC-E-FE

This model is an artificial neural network-based classification model with a total of 115 input features, in addition to the input features in the BC-E-F model. 72 stress components from 12 neighboring cells overlapping on the edge plane are also added. The model's architecture is a fully connected deep learning structure consisting of three hidden layers and one output layer. The input layer projects the 115-dimensional input onto a representation plane of 128 neurons. This is followed by two hidden layers of 256 and 32 neurons, respectively. Thus, the model's output is a continuous value representing the probability of an aftershock for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The weight value was determined by considering the proportion of positive examples within the total samples, and this weight was integrated into the training process. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. Training data was presented to the model in mini-batches of 32 samples, and this process continued for a total of 300 epochs. Probability outputs were converted into class labels using a threshold of 0.5.

This model was structured to demonstrate the impact on classification performance of expanding environmental stress components beyond the surface plane to include edge planes. The results demonstrated that comprehensive consideration of spatial context in aftershock prediction provides significant benefits. The BC-E-FE model expanded the spatial context by incorporating the stress components of neighboring objects into the input. It maintained its overall discrimination capacity, with ROC AUC (0.79) and accuracy (0.70) values similar to the previous model. Despite a high recall rate (0.78), the precision remained at 0.05 and the F1 score at 0.10, indicating that the model did not achieve sufficient selectivity in distinguishing the positive class. While this structure increased environmental information, it did not significantly improve classification reliability. The confusion matrix corresponding to the BC-E-FE model is presented in Figure 4.6.

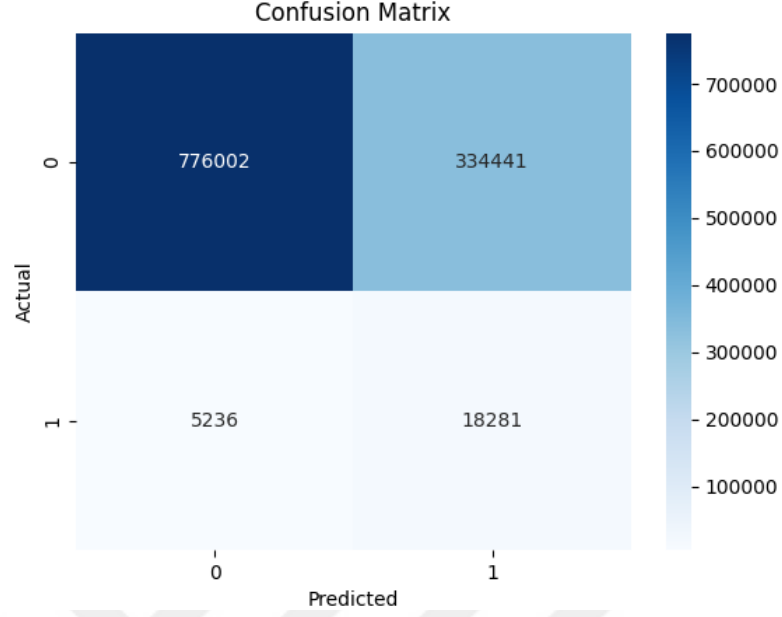


Figure 4.6. Confusion matrix of BC-E-FE model.

BC-E-FE-P

This model is an extended version of the BC-E-FE model, based on the idea of transforming the environmental context from a high-dimensional representation into a simpler, more information-intensive structure. The model aims to maintain classification performance and increase generalizability by using a summary representation vector extracted from the 108 stress tensor components of the 18 neighboring cells located on the surface and edge planes, rather than directly transferring them to the model.

The model's input vector consists of two main components. The first component consists of seven main features, consisting of the six stress tensor components of the center grid cell and the Euclidean distance between each grid cell. The second component is a high-dimensional vector consisting of 108 stress components of the 12 neighboring cells on the surface and edge planes. This vector was reduced to 75 dimensions by applying PCA, resulting in a total input dimension of 82. The number

of PCA components was determined to ensure that the cumulative variance ratio exceeded 95%. The variance explained by the PCA components, as determined through the elbow method, is shown in Figure 4.7.

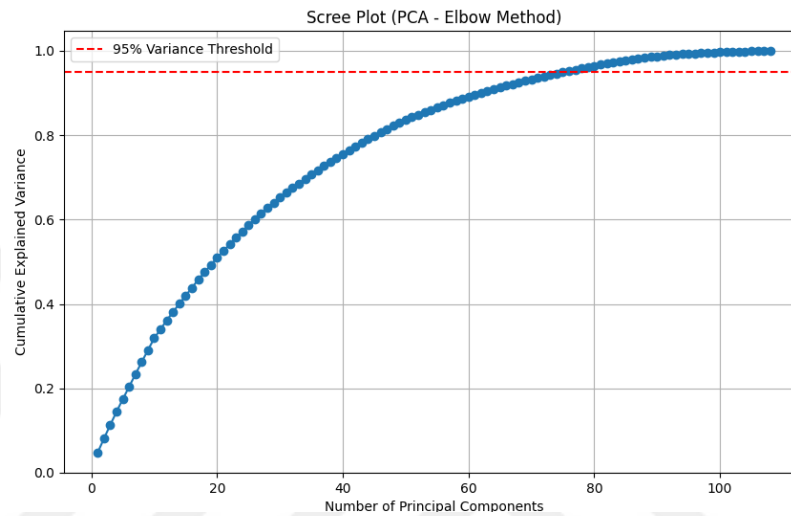


Figure 4.7. Scree Plot for PCA of Face and Edge Neighbor Stress Components.

The model's architecture is a fully connected structure consisting of four hidden layers and one output layer. The input layer projects 82-dimensional data to 128 neurons. This layer is followed by three intermediate layers consisting of 1024, 256, and 64 neurons, respectively. Thus, the model's output is obtained as a continuous value representing the probability of an aftershock for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The weight value was determined by considering the proportion of positive examples among the total examples, and this weight was integrated into the training process. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. The training data was presented to the model in mini-batches of 32 examples, and this process continued for a total of 300 epochs. The probability outputs were converted into class labels using a threshold of 0.5.

The BC-E-FE-P model reduced the high-dimensional structure of environmental stress components using PCA and performed classification based on these summary representations. The ROC AUC (0.79) and accuracy (0.72) values indicate that the model maintains its overall discrimination capacity at a level similar to previous models. While the recall rate of 0.74 indicates reasonable success in detecting positive classes, low values of precision (0.05) and F1 score (0.10) indicate that false positive predictions remain a significant problem. This suggests that the application of PCA does not dramatically improve classification performance. Therefore, while this model offers a balanced solution between information density and computational efficiency, it remains limited in terms of class discrimination performance. The confusion matrix corresponding to the BC-E-FE-P model is presented in Figure 4.8.

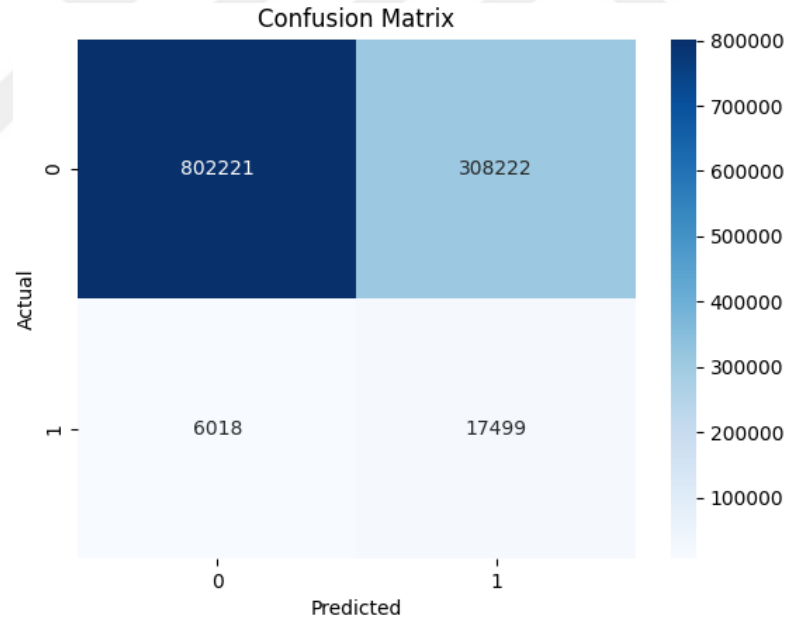


Figure 4.8. Confusion matrix of BC-E-FE-P model.

BC-E-FEC

This model is an artificial neural network-based classification model with a total of 163 input features, in addition to the input features from the BC-E-FE model.

This model aims to more comprehensively evaluate the spatial stress distributions that may affect aftershock formation. The artificial neural network architecture is a five-layer structure consisting of four hidden layers and one output layer. The input layer processes the 163-dimensional data set with 256 neurons, followed by layers of 256, 1024, 128, and 32 neurons, respectively. Thus, the model's output is a continuous value representing the probability of an aftershock for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The weight value was determined by considering the proportion of positive examples within the total samples, and this weight was integrated into the training process. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. Training data was presented to the model in mini-batches of 32 samples, and this process continued for a total of 300 epochs. Probability outputs were converted into class labels using a threshold of 0.5.

The BC-E-FEC model provides the most comprehensive representation of the spatial context by using all stress components from the surface, edge, and corner planes as input. This comprehensive structure partially maintained inter-class discrimination, with an ROC AUC score of 0.76. A recall value of 0.77 indicates that the model maintained its tendency to detect positive classes. However, low precision (0.05) and F1 score (0.09) values indicate that this success comes at the expense of significant false positives, and the model's predictive reliability remains low. The confusion matrix corresponding to the BC-E-FEC model is presented in Figure 4.9.

BC-E-FEC-P

This model is a more efficient version of the BC-E-FEC model, based on a dense representation of the environmental context, made more efficient through dimensionality reduction techniques. Instead of directly inputting the 156 stress tensor components of 26 neighboring grid cells located on the surface, edge, and corner planes, the BC-

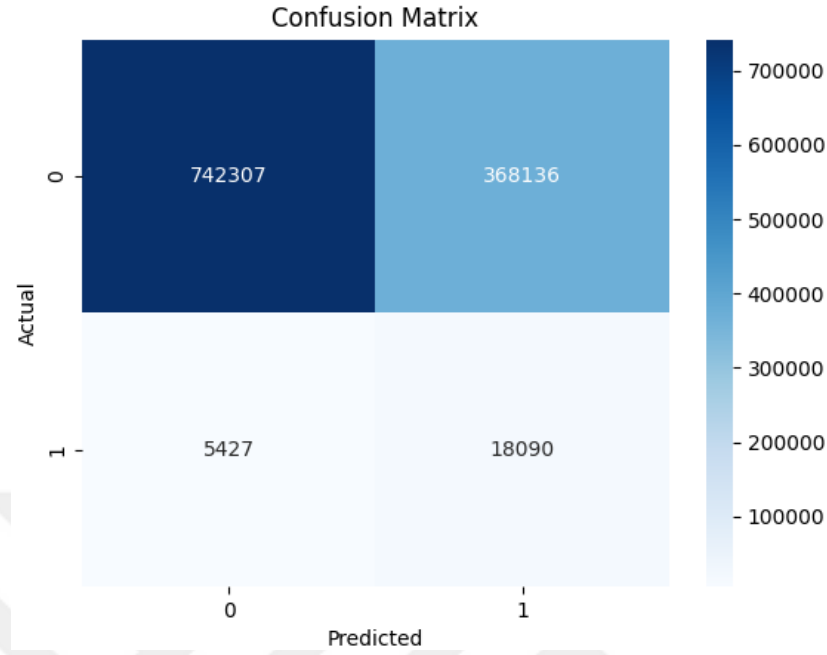


Figure 4.9. Confusion matrix of BC-E-FEC model.

E-FEC-P model aims to provide an information-dense yet computationally simpler representation of this high-dimensional structure. For this purpose, PCA was applied only to the stress components of neighboring cells.

The model's input vector consists of two main components. The first component contains seven fundamental features, consisting of the six stress tensor components of the center grid cell and the Euclidean distance between each grid cell. The second component is a high-dimensional vector consisting of 156 stress components of 26 neighboring cells on the surface, edge, and corner planes. This vector was reduced to 110 dimensions by PCA, resulting in a total input dimension of 117. The number of PCA components was determined to ensure that the cumulative variance ratio exceeded 95%. The variance explained by the PCA components, as determined through the elbow method, is shown in Figure 4.10.

The artificial neural network architecture consists of four hidden layers and one output layer. The input layer processes the 117-dimensional data set with 128 neurons,

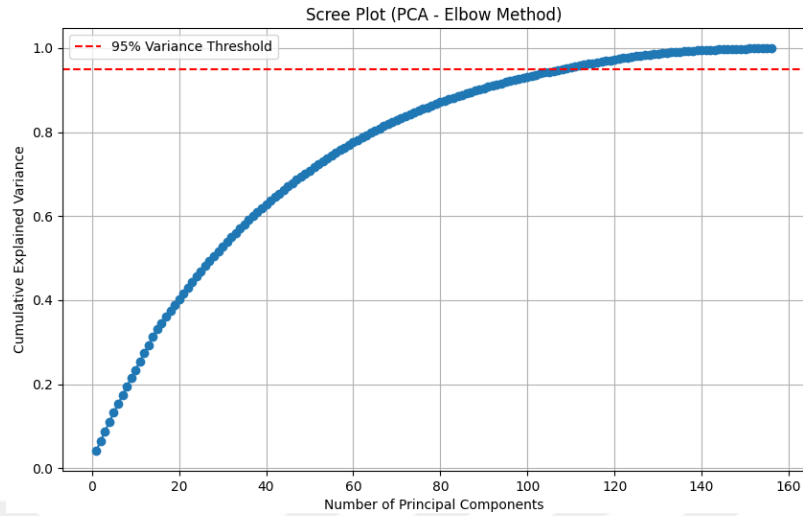


Figure 4.10. Scree Plot for PCA of Face, Edge, and Corner Neighbor Stress Components.

followed by intermediate layers consisting of 128, 1024, 256, and 64 neurons, respectively. Thus, the model's output is a continuous value representing the probability of an aftershock occurring for each grid cell.

During model training, a weighted loss function was applied for the positive class to compensate for class imbalance. The weight value was determined by considering the proportion of positive examples among the total examples, and this weight was integrated into the training process. The Adam optimization algorithm, with a learning rate of 0.001, was used to update the weights. The training data was presented to the model in mini-batches of 32 examples, and this process continued for a total of 300 epochs. The probability outputs were converted into class labels by a threshold of 0.5.

The BC-E-FEC-P model performed well in classifying surfaces, edges, and corners summarized using PCA. The ROC AUC score of 0.78 demonstrates that its discrimination ability is preserved compared to higher-dimensional versions. While accuracy (%70) and recall (0.77) are satisfactory, precision (0.05) and F1 score (0.10) reveal that errors due to class imbalance persist.

This model offers an advanced classification framework that aims to both increase computational efficiency and improve generalization ability by learning compact PCA representations of environmental stress components. The confusion matrix corresponding to the BC-E-FEC-P model is presented in Figure 4.11.

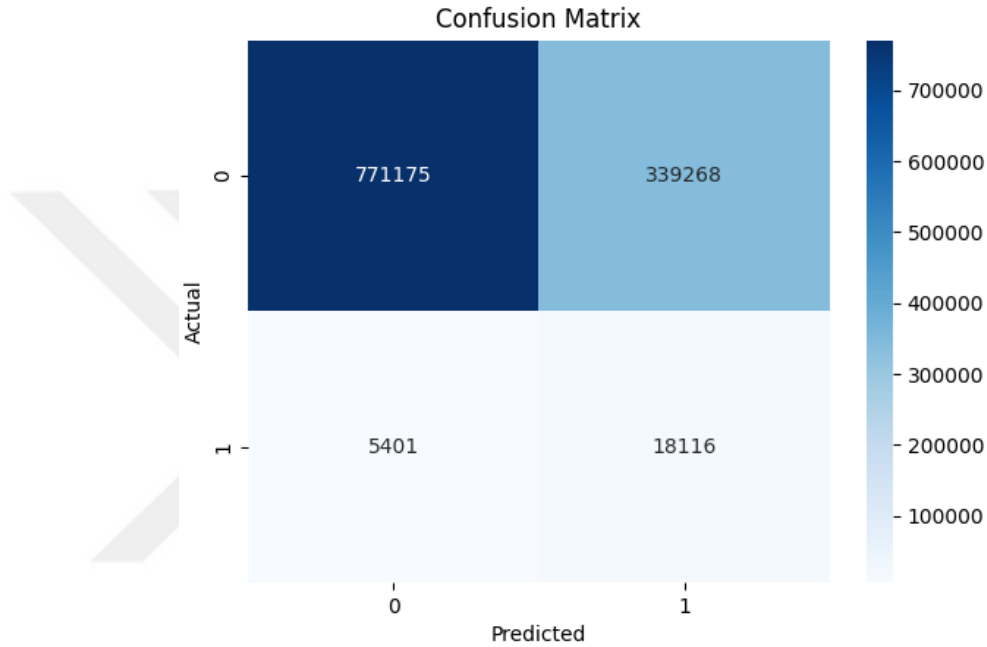


Figure 4.11. Confusion matrix of BC-E-FEC-P model.

BC-E-FE-SO

This model is an artificial neural network-based classification model that uses the same input structure as the BC-E-FE model, which yielded the most successful results in terms of input features. The input vector contains 115 features derived from the six principal stress tensor components, the Euclidean distance, and a total of 108 stress components from six neighbors coming from surfaces and twelve neighbors coming from edges in the three-dimensional grid system. While structurally unchanged from the BC-E-FE model, this model aims to reduce class imbalance by oversampling the training data.

The artificial neural network architecture consists of three hidden layers and one output layer. The input layer processes the 115-dimensional data set with 128 neurons, followed by intermediate layers of 256 and 64 neurons, respectively. Additionally, a dropout of 0.3 was applied to the model after each hidden layer to prevent overfitting.

Before training the model, due to class imbalance in the training data, oversampling was applied by multiplying the positive class samples, i.e., samples containing aftershocks, by half the number of negative class samples, i.e., samples without many aftershocks. This process aimed to create a more balanced distribution between positive and negative samples and to strengthen the model's ability to learn minority classes.

The Adam optimization algorithm, with a learning rate of 0.0005, was used to update the weights. The training data was presented to the model in mini-batches of 64 samples, and this process continued for a total of 300 epochs. Probability outputs were converted to class labels using a threshold of 0.65.

This model was designed to evaluate the impact of both the oversampling strategy and the decision threshold optimization; improvements were integrated with high-dimensional spatial inputs to address data imbalance and classification sensitivity.

The BC-E-FE-SO model was developed using the input features of the BC-E-FE model, which performed relatively well in previous stages, and by applying oversampling to reduce class imbalance. Test results of the model showed significant improvements over previous models in all key metrics, including accuracy (91%), ROC AUC (0.80), F1 score (0.16), precision (0.15), and recall (0.38). These results demonstrate that both detectability and overall discrimination ability are improved, particularly thanks to the more balanced representation of the positive class. In light of all these improvements, the BC-E-FE-SO model stands out as the structure that provides the highest classification performance among the models tested to date. The confusion matrix corresponding to the BC-E-FE-SO model is presented in Figure 4.12.

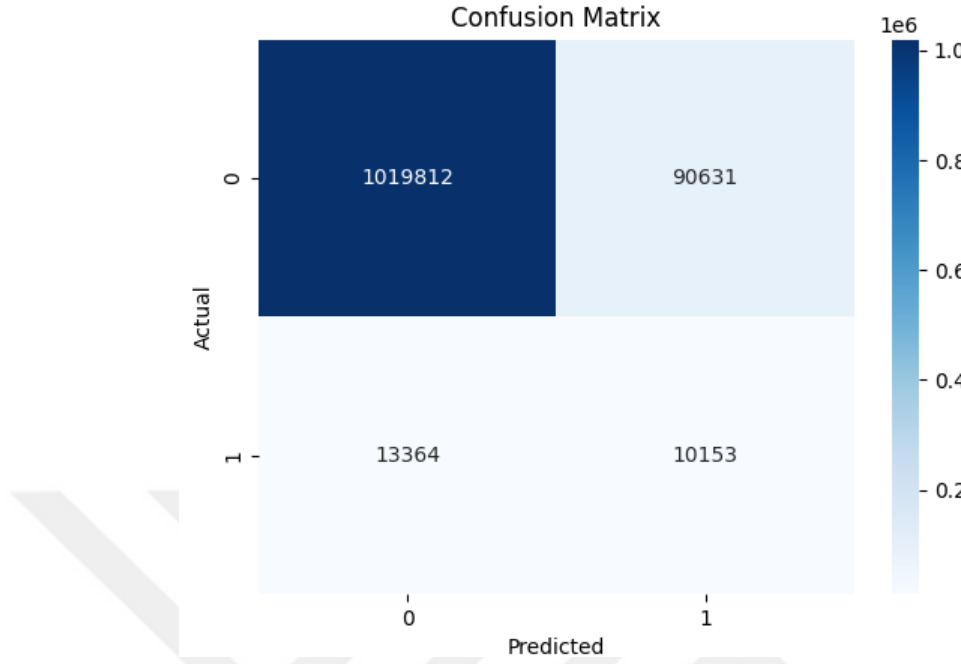


Figure 4.12. Confusion matrix of BC-E-FE-SO model.

BC-E-FE-SU

This model is an artificial neural network-based classification model that uses the same input structure as the BC-E-FE model, which has yielded the most successful results in terms of input features. The input vector contains 115 features derived from the six principal stress tensor components, Euclidean distance, and a total of 108 stress components from six neighbors coming from surfaces and twelve neighbors coming from edges in the three-dimensional grid system. While structurally unchanged from the BC-E-FE model, this model aims to reduce class imbalance in the training process. To achieve this, undersampling is applied to the majority class before training.

In the data preprocessing, undersampling was applied to balance the imbalanced class distribution. In this context, samples containing aftershocks, which constitute the positive class, were retained, while samples from the negative class, that is, those without aftershocks, were randomly selected to only receive twice the number of positive samples. This increased the model's sensitivity to positive classes.

The artificial neural network architecture consists of three hidden layers and one output layer. The input layer processes a 115-dimensional data set with 128 neurons, followed by intermediate layers consisting of 256 and 64 neurons, respectively. Additionally, a 0.3 dropout rate was applied after each hidden layer to prevent overfitting.

The Adam optimization algorithm, with a learning rate of 0.0005, was used to update the weights. The training data was presented to the model in mini-batches of 64 samples, and this process continued for a total of 300 epochs. The model's performance was evaluated using various classification metrics on the test set. Probability outputs were converted to class labels using a threshold of 0.5.

The BC-E-FE-SU model, trained with undersampling using BC-E-FE inputs, achieved 90% accuracy and an ROC AUC score of 0.81. The increase in recall rate to 0.46 indicates a significant improvement in positive class recognition, while precision (0.10) and F1 score (0.16) indicate that difficulties due to class imbalance persist. However, it presented a more balanced and successful structure compared to models without data augmentation or reduction, and performed similarly to models with oversampling, such as BC-E-FE-SO. The confusion matrix corresponding to the BC-E-FE-SU model is presented in Figure 4.13.

BC-E-FE-SO-D

This model used the 115 input variables from the highest-performing BC-E-FE-SO model and was additionally configured with a distance-based filtering strategy. Model training was conducted only with grid cells located less than 55 kilometers from the epicenter. In preliminary analyses, different distance thresholds (between 50 and 600 km) were tested, and the 55 km threshold achieved the highest classification accuracy. This approach is based on the assumption that distance is a determining factor in the spatial spread of aftershocks and aims to improve the model's learning quality by reducing noise originating from distant cells.

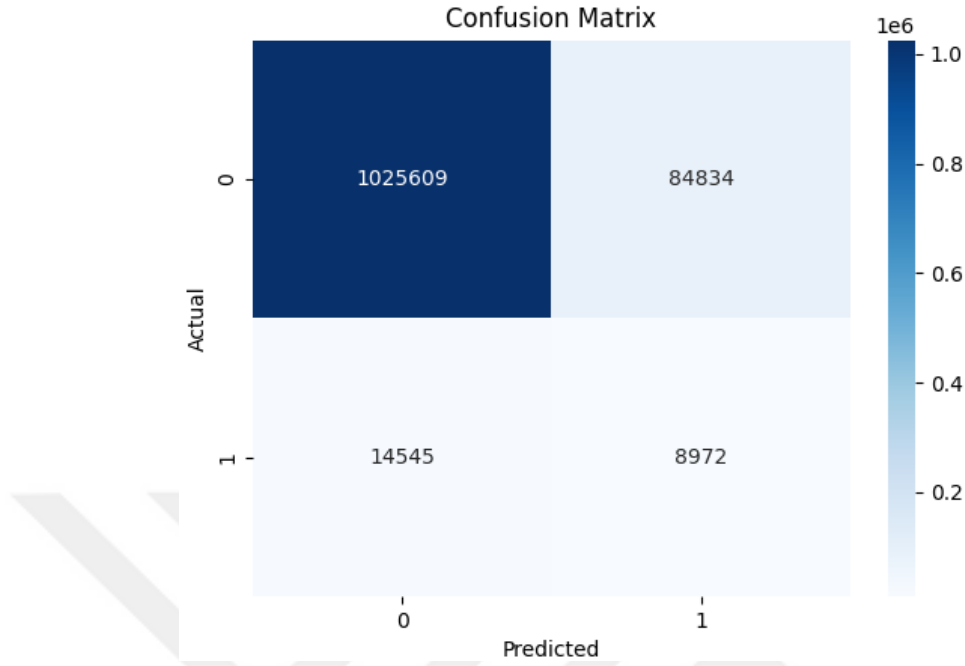


Figure 4.13. Confusion matrix of BC-E-FE-SU model.

During the data processing phase, only samples located closer than 55 kilometers to the epicenter were selected from the training data. Then, to address class imbalance, oversampling was applied by multiplying the positive class samples (i.e., samples containing aftershocks) by half the number of negative class samples (i.e., samples without many aftershocks). This process aimed to create a more balanced distribution.

The artificial neural network architecture is a deep learning structure consisting of three hidden layers and one output layer. The input layer contains 115 dimensional data, followed by three intermediate layers consisting of 128, 256, and 64 neurons, respectively. To prevent overfitting, a 0.3 dropout rate was applied after each layer.

The Adam optimization algorithm, with a learning rate of 0.0005, was used to update the weights. The training data was presented to the model in mini-batches of 64 samples, and this process continued for a total of 300 epochs.

This structure aims to both more reliably recognize the positive class and optimize classification performance geographically by controlling the effect of distance. The

BC-E-FE-SO-D model was created using the input variables from the best-performing BC-E-FE-SO model, filtering only cells closer than 55 km to the epicenter. This spatial pre-filtering, coupled with an oversampling strategy, aimed to increase the model's ability to recognize positive classes, particularly in the absence of a single cell. Test results indicated that the model significantly improved its classification performance with 89% accuracy, 0.50 recall, 0.10 precision, 0.16 F1 score, and 0.81 ROC AUC. This distance-based selective data structure significantly increased its accuracy, particularly in regions where aftershocks are more likely to occur. This provides a more balanced and geographically optimized classification structure than the BC-E-FE-SO model. The confusion matrix corresponding to the BC-E-FE-SO-D model is presented in Figure 4.14.

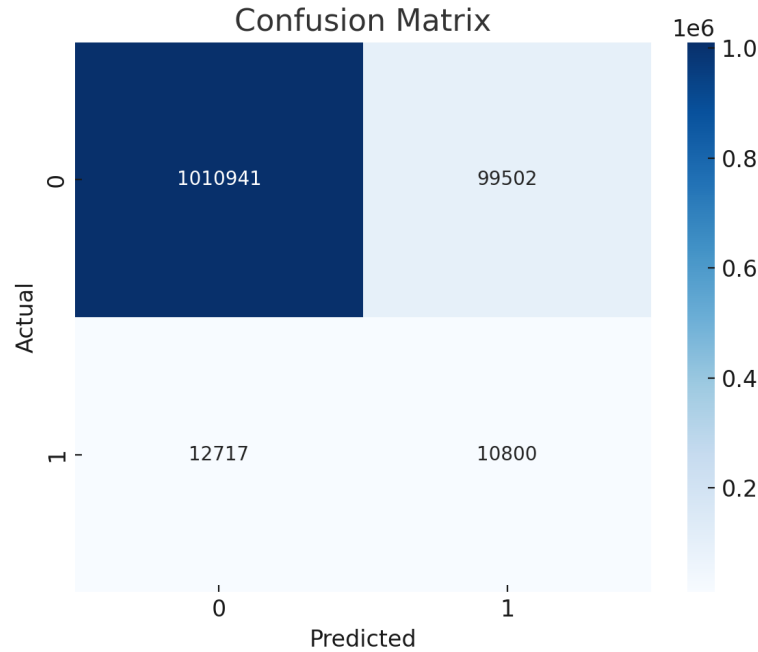


Figure 4.14. Confusion matrix of BC-E-FE-SO-D model.

BC-E-FE-SO-D-X

This model used the 115 input variables from the highest-performing BC-E-FE-SO-D model. Unlike this model, the classification process was performed using the

XGBoost algorithm. This aimed to develop a faster, more interpretable, and more lightweight model. The model’s goal was to predict the spatial distribution of aftershocks based on stress tensor components and spatial proximity metrics.

During the data processing phase, only samples located within 55 kilometers of the training data were selected. Then, to address class imbalance, oversampling was applied by multiplying the positive class samples (i.e., samples with aftershocks) by half the number of negative class samples (i.e., samples without many aftershocks). This process aimed to create a more balanced distribution between positive and negative samples and enhance the model’s ability to learn minority classes.

The XGBoost model was subjected to Bayesian Optimization during the hyperparameter optimization process. During this process, a total of six key hyperparameters were systematically adjusted based on ROC AUC scores to improve the model’s classification performance. The configuration, which determines the extent to which the model can branch in decision trees, was tested for values ranging from 3 to 10. The learning rate, which determines the impact of each tree on the model, was optimized to take continuous values between 0.01 and 0.3. Another setting, representing the total number of trees, was important for determining the model’s learning capacity, and different combinations were tested between 100 and 500.

In addition, the threshold value, which allows the model to branch more conservatively, was evaluated between 0 and 5. Furthermore, the ratios controlling the number of samples and the number of features to be used in training each tree were systematically tested between 0.5 and 1.0. These sampling strategies play a critical role in increasing the model’s diversity and generalizability, especially in high-dimensional data sets.

The optimization process was conducted using a total of 20 different parameter combinations, consisting of 5 starting points and 15 iterations, and the ROC and AUC scores obtained from each trial were compared. Ultimately, the parameter set

that yielded the highest AUC score was identified, the model was retrained with these settings, and the final evaluation was performed. This automated optimization process enabled the model to reach the most efficient configuration in terms of both accuracy and generalizability.

The BC-E-FE-SO-D-X model, a version of the BC-E-FE-SO-D model configured with the XGBoost algorithm, achieved the highest ROC AUC score (0.81) within the BC model family. This model achieved an F1 score of 0.17, a recall of 0.29, and a precision of 0.12, demonstrating a more balanced performance in positive class detection compared to previous models. The high ROC AUC score demonstrates the model's strong ability to discriminate between classes. These results demonstrate the effectiveness of the XGBoost algorithm in improving classification performance. Therefore, the model stands out as the most successful structure within the BC family in terms of both discrimination capacity and balance. The confusion matrix and ROC curve of the BC-E-FE-SO-D-X model are presented in Figures 4.15 and 4.16, respectively.

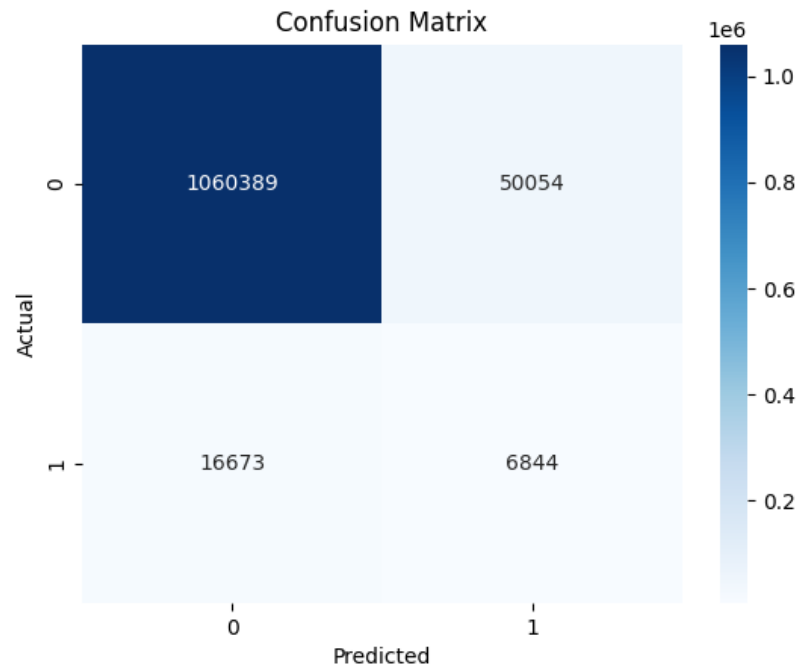


Figure 4.15. Confusion matrix of the BC-E-FE-SO-D-X model.

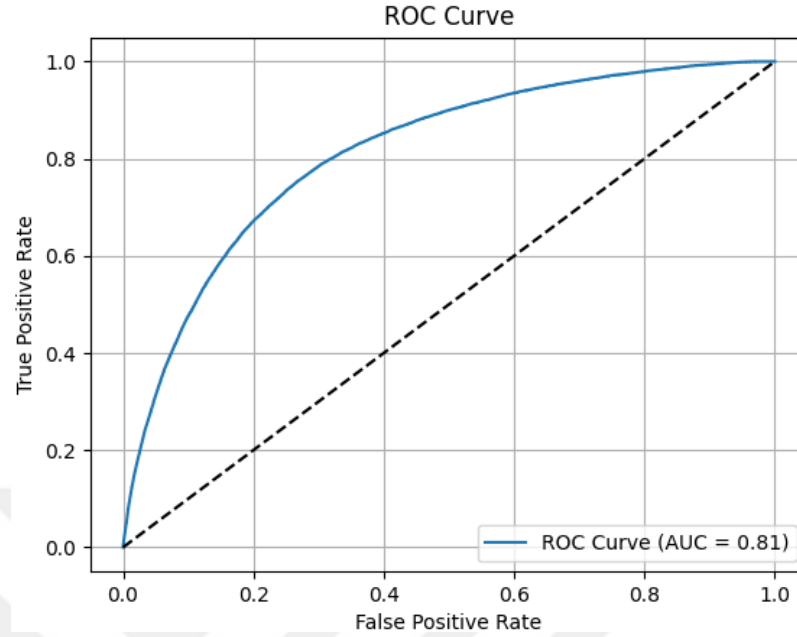


Figure 4.16. ROC curve of the BC-E-FE-SO-D-X model.

U-Net Model

The UNet model is a deep learning model developed to perform voxel-level aftershock prediction using three-dimensional stress tensors. The model is structured as a 3D extension of the classic U-Net architecture and consists of both encoder and decoder blocks. The input data for the model is a $310 \times 310 \times 10$ tensor composed of six basic stress components.

The encoder portion of the model contains eight convolution layers (Conv3D) and corresponding max-pooling layers. These layers are deepened by 6, 12, 24, 48, 96, 192, 384, and 768 channels, respectively. Pooling operations were performed in 3D, and in some layers, a reduced dimensionality ($2 \times 2 \times 1$) was applied for the z-axis. In the decoder section, a symmetric structure was followed with the encoder, and seven transpose convolution (ConvTranspose3D) layers were used. After each upsampling, a concatenation process was performed to match the activations from the encoder and the channel dimensions. The ReLU activation function was used in all intermediate layers, and the sigmoid activation function was used in the output layer. Thus, the output

was produced as a continuous probability map in the range $[0,1]$. The output of the model is a 3D voxel classification map with the same input dimensions ($310 \times 310 \times 10$). During training, this map was compared with the binary aftershock labels. Binary Cross Entropy (BCELoss) was used as the loss function.

The model was trained with a batch size of 16, and a total of 5 epochs were applied throughout the training process. Optimization was performed using the Adam algorithm with a learning rate of 0.001. The model's training and test data were split 80%–20% using a train-test split. After training, the model's accuracy was evaluated using a confusion matrix and classification report. Because predictions were made separately for each voxel, the high prevalence of negative classes in the dataset made interpreting the metrics difficult; therefore, we focused on precision, recall, and F1 scores. The confusion matrix corresponding to the U-Net model is presented in Figure 4.17 and the actual and predicted voxel outputs of the U-Net model in 3D are also presented in Figure 4.18.

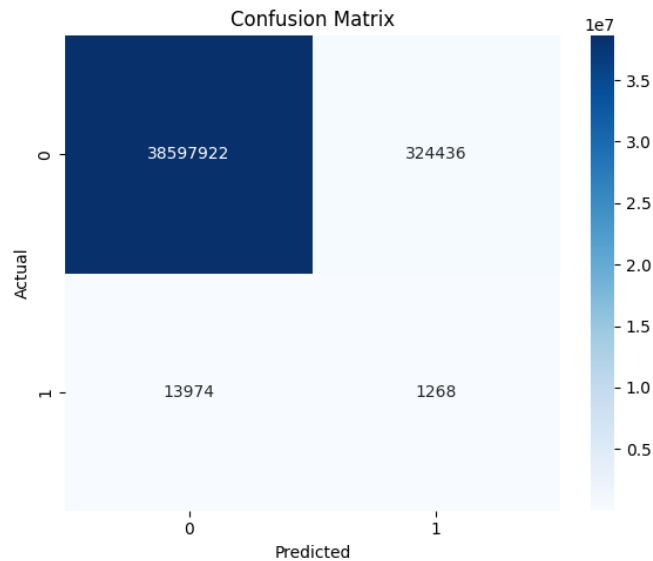


Figure 4.17. Confusion matrix U-Net model.

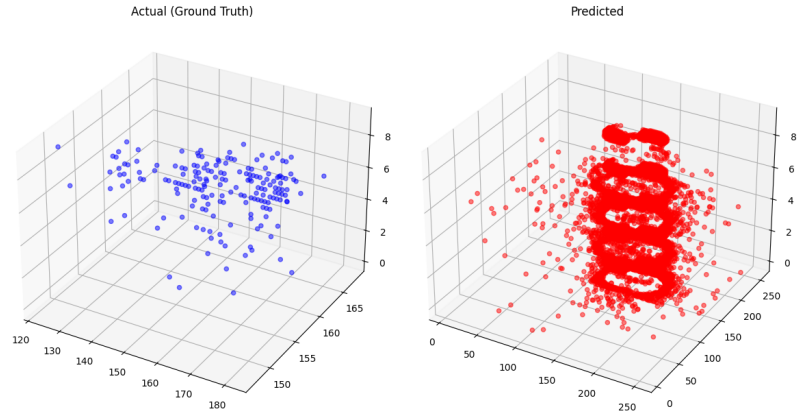


Figure 4.18. Actual and predicted voxel outputs of the U-Net model in 3D.

The performance results of all models developed for the binary classification task in this study are summarized in the Table 4.2.

Table 4.2. Comparative performance metrics of classification models.

Model	Accuracy	Recall	Precision	F1-score	ROC AUC
1. BC	0.62	0.80	0.04	0.08	0.77
2. BC-E	0.72	0.74	0.05	0.10	0.79
3. BC-E-F	0.69	0.78	0.05	0.09	0.79
4. BC-E-F-P	0.67	0.80	0.05	0.09	0.79
5. BC-E-FE	0.70	0.78	0.05	0.10	0.79
6. BC-E-FE-P	0.72	0.74	0.05	0.10	0.79
7. BC-E-FEC	0.67	0.77	0.05	0.09	0.76
8. BC-E-FEC-P	0.70	0.77	0.05	0.10	0.78
9. BC-E-FE-SO	0.91	0.43	0.10	0.16	0.81
10. BC-E-FE-SU	0.91	0.38	0.10	0.15	0.80
11. BC-E-FE-SO-D	0.89	0.50	0.10	0.16	0.81
12. BC-E-FE-SO-D-X	0.94	0.29	0.12	0.17	0.81
13. U-Net	0.99	0.04	0.02	0.02	0.06

4.2.2. Kernel-Based Probabilistic Modeling of Aftershock Occurrence

In this section, aftershock occurrence is modeled in a probabilistic framework. Instead of discrete labels, a continuous target is defined for each grid cell, derived from the sum of the surrounding aftershock events and weighted with a Gaussian kernel. The raw kernel density y_{raw} is mapped to the interval $[0, 1]$ by the following transformation

$$y_{\text{soft}} = 1 - \exp(-\alpha y_{\text{raw}}), \quad \alpha = 0.5 \quad (4.1)$$

where y_{raw} is obtained using a Gaussian kernel defined with parameters $\sigma = 10$ km and $R = 32$ km. The input features were generated from six elastic-static stress tensor components, their spatial derivatives, gradient norms, and tensor invariants (trace, L_1 , L_2 , L_∞ norms, maximum shear stress, eigenvalue ratio, and the von Mises equivalent stress). All numerical features were scaled using extreme value clipping and z-score standardization. Within this probabilistic setup, both a multilayer perceptron (MLP) and an XGBoost regressor were trained to learn the mapping from stress-based inputs to the soft aftershock probabilities.

Kernel-Based MLP Model

The model has a multilayer perceptron (MLP) structure with hidden layer widths of $[256, 256, 128, 64, 32]$. The layers follow the order of Linear–BatchNorm–Sigmoid–Dropout (0.10), and Sigmoid activation is used in the output layer. The training process was conducted with a Binary Cross-Entropy (BCE) loss, and optimization was performed with the AdamW algorithm ($lr = 2 \times 10^{-4}$, weight decay = 10^{-5}). The learning rate was dynamically adjusted with the ReduceLROnPlateau timer, and the early stopping criterion was set as the validation loss.

Model performance was evaluated not only with loss-based metrics but also with classification metrics. Probability calibration was measured with the Brier score, and overall discrimination was measured with the ROC AUC metric. Furthermore, the decision threshold on the probabilities ($t^* = 0.05$) was determined by maximizing the F1-score on the validation set; this threshold was applied to the test set.

The results show that the model successfully detected positive examples with a high Recall value (0.93). Although the Precision (0.68) value was somewhat low, the high F1 (0.79) and ROC AUC (0.82) scores demonstrate strong overall discrimination capacity. Furthermore, the low BCE (0.41) and Brier (0.06) values suggest that the model's probability estimates are well calibrated and have been learned in a balanced manner. The corresponding training-validation BCE loss curve are shown in Figure 4.19. The spatial distribution of observed aftershocks is illustrated in Figure 4.20, followed by the Gaussian-kernel smoothing of aftershock occurrence in Figure 4.21. The predicted probability field produced by the kernel-based probabilistic model is presented in Figure 4.22.

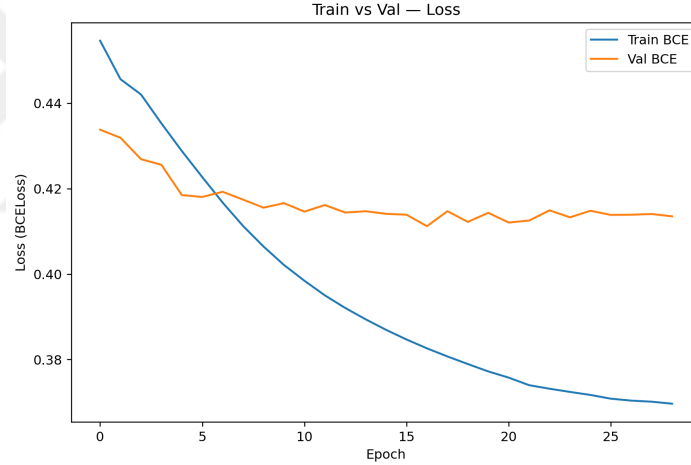


Figure 4.19. Training and validation loss curves (BCE) of the kernel-based probabilistic aftershock occurrence model.

Kernel-Based XGBoost Model

As an alternative to the neural-network architecture, a gradient-boosted decision tree model based on XGBoost was trained using the same kernel-based probabilistic targets. The input vector is identical to that of the MLP model and consists of the six

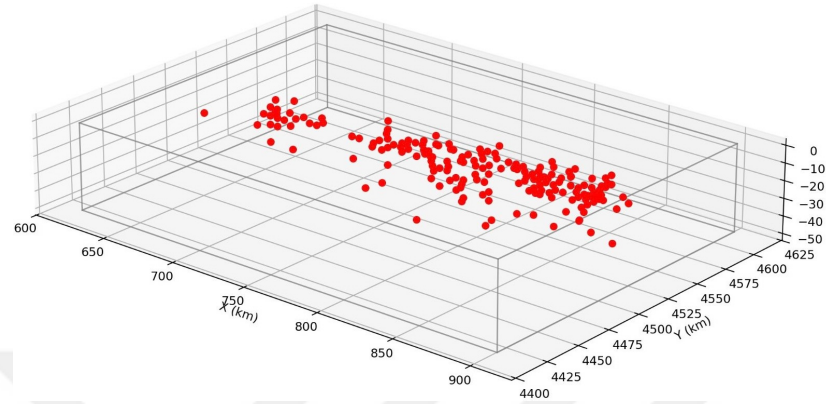


Figure 4.20. Spatial distribution of aftershocks for the 1999 Izmit earthquake within the 3D domain.

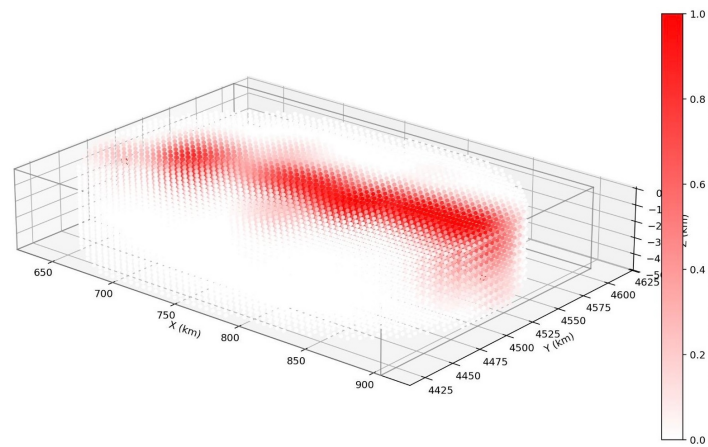


Figure 4.21. Gaussian-kernel smoothing of aftershock occurrence for the 1999 Izmit earthquake.

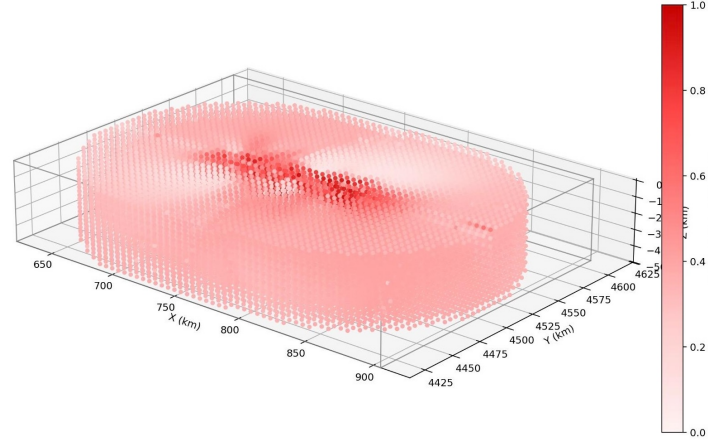


Figure 4.22. Predicted aftershock occurrence probabilities for the 1999 Izmit earthquake with MLP model.

elastic-static stress tensor components, their spatial derivatives, gradient norms, and tensor invariants (trace, L_1 , L_2 , L_∞ norms, maximum shear stress, eigenvalue ratio), together with the von Mises equivalent stress. All numerical features were preprocessed using extreme-value clipping and z-score standardization, and the target was defined as the soft probability y_{soft} obtained from the Gaussian kernel transformation. The XGBoost configuration employed 800 boosted trees with a maximum depth of 6, learning rate of 0.03, subsampling ratio of 0.8, column subsampling ratio of 0.8, and ℓ_2 regularization on the tree weights.

Model performance was again assessed in terms of both probabilistic calibration and classification quality. Probability calibration was quantified using the Brier score, while overall discrimination was measured with the ROC AUC metric. As in the MLP experiment, a decision threshold on the predicted probabilities (t^*) was determined on the validation set by maximizing the F1-score, and this threshold was then applied to the test set to compute Precision, Recall, and F1-score. The resulting metrics indicate that the XGBoost model preserves the high Recall (0.93) of the MLP configuration, while achieving a higher Precision (0.72). Consequently, the F1-score (0.81) and ROC AUC (0.85) are slightly improved relative to the neural-network model, and the Brier score (0.07) remains comparably low, indicating well-calibrated probability estimates.

The corresponding training-validation loss curve of the XGBoost model is shown in Figure 4.23 and the predicted probability field obtained from the kernel-based XGBoost model for the 1999 Izmit earthquake is presented in Figure 4.24. The results for kernel based probabilistic models are presented in the table 4.3.

Table 4.3. Performance metrics of the kernel-based probabilistic aftershock occurrence model.

#	Model	BCE	Brier	Recall	Precision	F1-score	ROC AUC
1	MLP	0.41	0.06	0.93	0.68	0.79	0.82
2	XGBoost	—	0.07	0.93	0.72	0.81	0.85

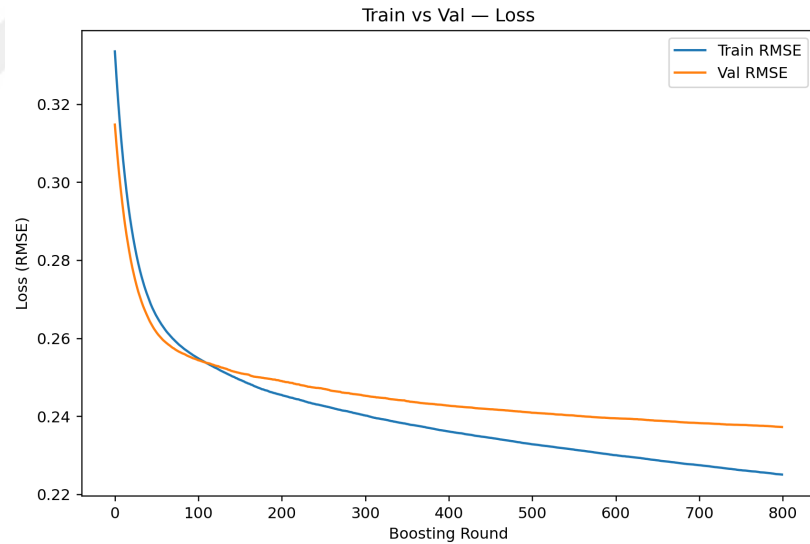


Figure 4.23. Training and validation loss curves of the kernel-based XGBoost probabilistic aftershock occurrence model.

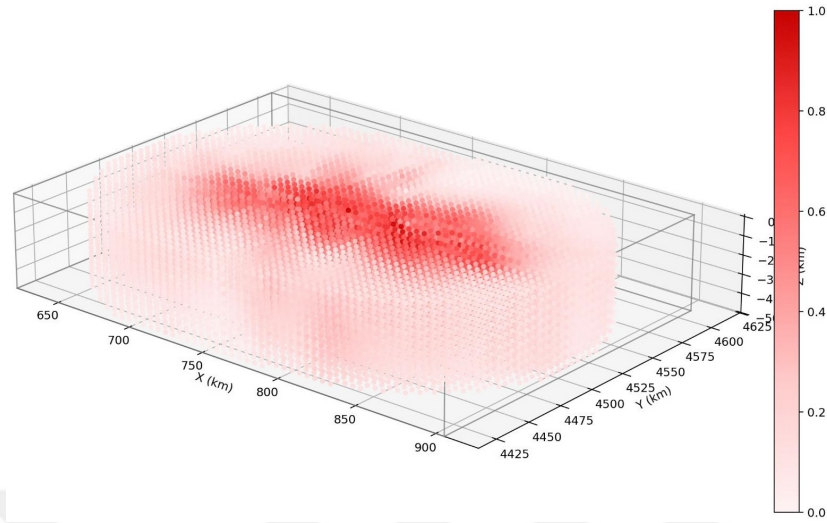


Figure 4.24. Predicted aftershock occurrence probabilities for the 1999 Izmit earthquake with XGBoost model.

4.3. Aftershock Count Regression Modeling

4.3.1. Baseline Count Regression Models

In the second phase of the modeling process, regression-based approaches were developed to estimate the number of aftershocks occurring in a specific time interval as a continuous variable within a three-dimensional cube area defined as $100 \text{ km} \times 100 \text{ km}$ horizontally and 50 km vertically, divided into grid cells of $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$ dimensions. In this context, numerical labels representing the number of aftershocks observed in each cell were used as target variables. Thus, both spatial resolution and numerical prediction accuracy were targeted.

In the study, various regression models based on both ANN and gradient boosting algorithms were constructed; different input vectors, layer architectures, and hyperparameter configurations were systematically tested in these models. ANN architectures were defined on the PyTorch platform, and different combinations of stress tensors (e.g., principal components only, stress values of neighboring cells, or PCA-reduced representations) were included in the input layers. All input data was scaled using

z-score normalization before model training to prevent weight imbalances that could arise from scale differences between variables.

The ReLU activation function was used to learn nonlinear patterns in the hidden layers. Also the output layer was terminated with ReLU to prevent negative predictions. Parameter updates were performed throughout the training process using the Adam optimization algorithm; in some architectures, a dropout mechanism was implemented at different rates to prevent overfitting. All ANN models were trained with mini-batches of 32 or 64 samples, and the training process for each architecture lasted 300 epochs.

To mitigate the imbalances observed in the data set, oversampling or under-sampling techniques were applied in some models, depending on the target variable distribution. The effects of these strategies on model performance were analyzed comparatively. Additionally, models were created that filtered according to a specific distance threshold, including only grid cells located within a certain distance from the epicenter.

In addition to artificial neural network-based models, regression models based on the XGBoost algorithm were also developed. Hyperparameter optimization in these models was performed using Bayesian Optimization instead of manual testing.

All regression models were evaluated on the test data set using basic regression metrics such as MAE, RMSE, and R^2 score. Thus, the prediction performances obtained with different structural and parametric choices were compared empirically.

CR

This model is an artificial neural network-based regression model that uses only six principal stress tensor components as inputs. Using these features, it continuously predicts the number of aftershocks that may occur in that cell as an output.

The model architecture has a fully connected structure consisting of four layers. The input layer passes the 6-dimensional stress vector to the first hidden layer with 32 neurons, followed by two hidden layers with 128 and 32 neurons. The final layer is defined as a single-neuron output unit. Consistent with the common modeling framework, ReLU activation was used in the hidden layers, and ReLU was also applied to the output layer to prevent negative values.

The model was trained with mini-batches of 32 samples. The Adam optimization algorithm and MSE loss function, with a learning rate of 0.001, were used throughout the training process; the model was trained for 300 epochs.

The CR model's MAE is quite low at 0.03, indicating generally accurate predictions. However, the RMSE of 0.29 suggests larger errors in some examples. While the model generally makes accurate predictions, it occasionally exhibits significant deviations.

CR-E

This model is an artificial neural network-based regression model with a total of seven input features, in addition to the six fundamental stress tensor components in the CR model, which also includes the Euclidean distance between the epicenter and each grid cell.

The model architecture consists of a total of four fully connected layers. The input layer transforms a 7-dimensional vector (6 stress components + 1 distance) into a first layer of 32 neurons. This is followed by two intermediate layers of 128 and 32 neurons, respectively. A single-neuron layer generates the predicted value at the output. Nonlinear relationships are modeled by applying the ReLU activation function to all hidden layers. The ReLU function is also used in the output layer to prevent the model from generating negative values.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

Adding the Euclidean distance feature to the CR-E model did not create a significant difference in performance; while the MAE and RMSE values of the model remained the same as the CR model.

CR-E-F

This model is an artificial neural network-based regression model with a total of 43 input features, in addition to the input features from the CR-E model. 36 stress components from six neighboring cells on the grid cell's surface plane are also added.

The model architecture features a fully connected neural network structure consisting of four layers. The input layer projects the 43-dimensional extended feature vector onto a representation space of 32 neurons. This is followed by two hidden layers of 128 and 32 neurons, respectively. The output layer consists of a single neuron and uses the ReLU activation function to prevent negative predictions. The ReLU function is also applied to all hidden layers to ensure the learning of nonlinear patterns.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

Despite a significant increase in input size, no improvement was observed in MAE and RMSE values. The addition of surface neighbors did not significantly improve model performance.

CR-E-F-P

This model uses the input features from the CR-E-F model; however, because this multidimensional neighborhood data can contain high correlation and data noise, PCA, a dimensionality reduction method, was applied to the 36 input features from the surface neighborhoods. Before PCA, all input features were normalized to the same scale, and then PCA was applied to the neighborhood stress components to select the 25 most significant components. Consequently, the model’s input vector became a 32-dimensional vector combining the normalized baseline features and the PCA-transformed neighborhood stress information.

The model architecture features a deep artificial neural network structure consisting of a total of six layers. The input layer receives this combined 32-dimensional vector and processes it through hidden layers consisting of 128, 1024, 256, and 64 neurons, respectively. The ReLU activation function is applied to each hidden layer to enhance the model’s ability to learn nonlinear patterns. The output layer is a single-neuron layer, and the ReLU activation function is also applied to prevent negative predictions.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

An examination of model performance reveals that, similar to previous CR models, the MAE (0.03) and RMSE (0.29) values were maintained. This suggests that, although the input space was made more compact and learnable through dimensionality reduction with PCA, the model’s success in explaining the number of aftershocks remained limited. In other words, even though neighborhood information was summarized and incorporated into the model, no significant improvement in predictive power was achieved.

CR-E-FE

This model is an artificial neural network-based regression model with a total of 115 input features, in addition to the input features from the CR-E-F model. 72 stress components from 12 neighboring cells in the edge plane are also added. The model architecture features a fully connected neural network structure consisting of four layers. The input layer receives this rich 115-dimensional feature vector and transfers it to the first hidden layer of 128 neurons. This is followed by two hidden layers of 256 and 32 neurons, respectively. The ReLU activation function is applied to all hidden layers to model nonlinear relationships. The output layer consists of a single neuron, and ReLU activation also prevents negative predictions.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm and MSE loss function were used with a learning rate of 0.001 throughout the training process, and the model was trained for 300 epochs.

An examination of the CR-E-FE model outputs reveals that the MAE (0.03) and RMSE (0.29) values remain the same as those of previous models. These results indicate that even though the data representation was expanded by increasing the input size and including neighboring neighbors, the model's ability to predict the number of aftershocks was not significantly improved. Therefore, while expanding the spatial context is theoretically meaningful, it has not directly impacted performance in practice.

CR-E-FE-P

This model uses the input features from the CR-E-FE model; however, in this model, 108 stress tensor components taken from 18 neighboring cells on the surface and edge planes were subjected to dimensionality reduction using PCA, reducing them to 75 principal components. Thus, the total input dimension was determined to be 82.

The model architecture consists of a deep artificial neural network structure consisting of four hidden layers and one output layer, starting from the input layer and containing 128, 1024, 256, and 64 neurons, respectively. Nonlinear relationships are modeled in each layer using the ReLU activation function. The output layer of the model is a single neuron, and negative predictions are also prevented using ReLU activation.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

This model reduces the computational load, especially by compressing multidimensional stress data via PCA, and provides a more balanced learning process by preserving meaningful variance information.

CR-E-FE-P model remained at the same level as previous models in MAE (0.03) and RMSE (0.29). despite using 108 neighborhood inputs simplified by PCA. This suggests that although the reduction in dimensionality reduced the complexity of the model and the computational overhead, the predictive performance was not significantly improved. PCA provided a more compact representation of the data structure, but did not directly contribute to predictive performance.

CR-E-FEC

This model is an artificial neural network-based regression model with a total of 163 input features, in addition to the input features of the CR-E-FE model. 48 stress components from 8 neighboring cells in the corner plane are also added.

The model architecture features a high capacity and deep artificial neural network. In the input layer, the 163-dimensional vector is first projected onto a representation plane of 256 neurons. Then it passes through three more layers of 1024, 128,

and 32 neurons, respectively. The ReLU activation function was applied to all hidden layers. The output layer consists of a single neuron, again limited by ReLU activation, to negative value generation.

The model was trained using mini-batches of 32 samples. The Adam optimization algorithm and MSE loss function were used with a learning rate of 0.001 throughout the training process, and the model was trained for 300 epochs.

The CR-E-FEC model expanded the spatial context by incorporating the environmental neighborhood most comprehensively; however, MAE (0.03) and RMSE (0.29) remained at the same level as previous models. Despite increasing input size and model capacity, no significant improvement in prediction performance was achieved. This suggests that the contribution of high-dimensional inputs is limited by overfitting or knowledge redundancy rather than by increasing data power.

CR-E-FEC-P

This model uses the input features from the CR-E-FEC model; however, in this model, 156 stress tensor components taken from 26 neighboring cells in the surface, edge, and corner planes were subjected to dimension reduction using PCA, reducing them to 110 principal components. Thus, including the basic 7 input features, the total input dimension was determined to be 117.

The model architecture is a fully connected deep learning structure. The input layer passes the 117-dimensional vector to the first hidden layer of 128 neurons. This is followed by three hidden layers, each containing 1024, 256, and 64 neurons, respectively. Nonlinear relationships are modeled using the ReLU activation function in all layers. The output layer consists of a single neuron and, again, includes ReLU activation, preventing the prediction of negative values. This structure allows the model to focus on predicting the number of aftershocks at or above zero.

The model was trained using mini-batches of 32 samples. The Adam optimization algorithm and MSE loss function were used with a learning rate of 0.001 throughout the training process; the model was trained for 300 epochs.

Despite having the largest input dimensionality, the CR-E-FEC-P model failed to improve on MAE (0.03) and RMSE (0.29) compared to previous models. While dimensionality reduction with PCA provides advantages in terms of computational efficiency, it did not improve the overall performance of the model.

CR-E-FE-SO

This artificial neural network-based regression model, which uses the same input features as the CR-E-FE model, takes as input a total of 115 input features derived from the six basic stress tensor components, Euclidean distance, and the stress components of six neighbors from surfaces and 12 neighbors from edges in the three-dimensional grid system. Unlike CR-E-FE, this model employs oversampling of the data set.

The model architecture is a fully connected neural network structure with a total of four hidden layers. The input layer takes a 118-dimensional vector and transforms it into a representation of 128 neurons in the first layer. This is followed by two more layers, each containing 256 and 64 neurons, respectively. The ReLU activation function is applied to all hidden layers to model nonlinear relationships. Additionally, dropout layers were added at 30%, 30%, and 20% ratios to enhance the model's generalization. The output layer of the model consists of a single neuron, and ReLU activation also prevents negative predictions.

Specifically for this model, due to class imbalance in the training data, the data set was balanced by oversampling cells with a 100% increase in the number of continuations greater than zero (by copying half as many positive examples as the existing negative examples). This increased the model's ability to learn positive values, particularly those with low frequencies.

The model was trained with mini-batches of 64 samples. The Adam optimization algorithm and MSE loss function, with a learning rate of 0.0005, were used throughout the training process. The model was trained for 300 epochs. At the end of each epoch, the RMSE performance of the model was monitored on the test data, and the model parameters were recorded when the best results were achieved. Although the CR-E-FE-SO model was trained on a balanced data set using an oversampling strategy, contrary to expectations, the MAE (0.27) and RMSE (0.48) values increased. This result suggests that the data augmentation strategy did not provide any improvement.

CR-E-FE-SU

This neural network-based regression model, which uses the same input features as the CR-E-FE model, takes as input a total of 115 input features derived from the six basic stress tensor components, Euclidean distance, and the stress components of six neighbors from surfaces and 12 neighbors from edges in a three-dimensional grid system. Unlike CR-E-FE, this model employs undersampling of the data set.

The model architecture features a fully connected artificial neural network structure consisting of five layers. The input layer transforms the 115-dimensional input vector into the first layer of 128 neurons. Subsequent layers contain 256, 64, and 1 neuron, respectively. Nonlinear relationships are modeled in the intermediate layers by applying the ReLU activation function. Dropout mechanisms are also used to prevent overfitting (at 0.3 and 0.2 rates). The ReLU function is also applied to the output layer to ensure only positive prediction values are obtained.

The model was trained using mini-batches of 64 samples. The Adam optimization algorithm and MSE loss function were used throughout the training process, with a learning rate of 0.0005. The model was trained for 300 epochs.

Although the CR-E-FE-SU model was trained on balanced data with the under-sampling strategy, the MAE (0.27) and RMSE (0.40) values remained high.

CR-E-FE-SU-D

This neural network-based regression model, which uses the same input features as the CR-E-FE-SU model, contains a total of 115 input features derived from the six principal stress tensor components, Euclidean distance, and the stress components of six neighbors from the surfaces and 12 neighbors from the edges in the three-dimensional grid system. In the model, consistent with the CR-E-FE-SU approach, it was assumed that the distance to the location of aftershocks would be a determining factor in the modeling process, and a distance filter was applied to the data set accordingly. After testing different threshold values, only observations located within a certain distance of the epicenter were included in the model. Thus, samples located further away and with a low probability of having aftershock characteristics were excluded from the model to achieve a more meaningful sample structure.

The model architecture features a fully connected neural network structure consisting of five layers. The input layer projects the 115-dimensional input vector onto a first hidden layer of 128 neurons. This is followed by two hidden layers of 256 and 64 neurons, respectively. Nonlinear relationships were learned by applying the ReLU activation function to all hidden layers. Furthermore, dropouts of 0.3, 0.3, and 0.2 were applied after each hidden layer, increasing the model's resistance to overfitting. The output layer consists of a single neuron, and ReLU activation prevents negative predictions.

The training data included only samples located within 55 km of the epicenter. Then, to reduce class imbalance, an undersampling process was performed by randomly selecting only twice as many samples from class labeled 0 as from class labeled 1. This balanced the training data set.

The model was trained with mini-batches of 64 samples. Throughout the training process, the Adam optimization algorithm with a learning rate of 0.0005 and the MSE loss function were used; the model was trained for 300 epochs. At the end of each epoch,

the RMSE performance of the model was monitored on the test data, and the model parameters were recorded when the best results were achieved. During the training process, the model weights that demonstrated the best validation performance were recorded, and this best model was used in the final evaluation.

The CR-E-FE-SU-D model was configured with spatial filtering (55 km) and undersampling strategies. However, the obtained performance values (MAE: 0.27, RMSE: 0.40) clearly indicate that the model does not adequately explain the target variable. These results indicate that distance-based sample selection and data reduction do not have the expected positive impact on model accuracy.

CR-E-SU-D-X

This model is an XGBoost-based regression model that aims to predict the number of aftershocks as a continuous output using a total of seven input features: six stress tensor components and the Euclidean distance. Compared to the CR-E-FE-SU-D model, the input dimension has been reduced; no features belonging to neighboring grid cells have been included in the model. Furthermore, the XGBoost algorithm, a tree-based gradient boosting method, was chosen over an artificial neural network-based approach.

The training data included only samples located within 55 km of the epicenter. Subsequently, an undersampling process was applied by selecting twice as many 0-labeled data as 1-labeled samples to reduce class imbalance. These samples were shuffled to create a balanced training set.

In the XGBoost-based regression model, a Bayesian optimization method was used to tune the hyperparameters. To increase the prediction accuracy and generalization capacity of the model, seven key hyperparameters were systematically optimized within specific ranges. In this context, the following were used: The depth of the decision tree, which directly affects the learning capacity of the model and the risk of

overfitting, was tested between 3 and 12, the learning rate was established between 0.0005 and 0.2, and the total number of trees was evaluated between 100 and 800. Furthermore, to increase the model's diversity and generalizability, the sample ratio and feature ratio used in training each tree were optimized between 0.4 and 1.0. The regularization parameters were also tested with values ranging from 0 to 5, making the model more resistant to overfitting. This model produced higher error values compared to previous models with MAE (0.29) and RMSE (0.40), obtained using the XGBoost algorithm with limited input.

CR-E-X

This model aims to predict the number of aftershocks through regression using a total of seven input features, consisting of six fundamental stress tensor components and the Euclidean distance to the target grid cell. Unlike the first model, the CR structure, this study employed the XGBoost algorithm, a tree-based gradient boosting method, instead of an artificial neural network.

No undersampling or oversampling techniques were applied to the training data; the model was trained to include samples from all grid cells.

In the XGBoost-based regression model, a Bayesian optimization method was used to tune the hyperparameters. To increase the model's prediction accuracy and generalization capacity, seven fundamental hyperparameters were systematically optimized within specific ranges. In this context, the decision tree depth, which directly affects the model's learning capacity and the risk of overfitting, was tested between 3 and 12, the learning rate was set between 0.0005 and 0.2, and the total number of trees was evaluated between 100 and 800. Additionally, to increase the diversity and generalizability of the model, the sample ratio and feature ratio used in training each tree were optimized between 0.4 and 1.0. Regularization parameters were also tested with values ranging from 0 to 5, making the model more resistant to overfitting.

The model was retrained with the best parameter set obtained through the Bayesian optimization process, and its performance was evaluated on the test data set. Because negative values in the model outputs were not significant, predictions were adjusted to avoid falling below zero.

According to the results of the model, MAE was obtained as 0.06 and RMSE as 0.28; these values are similar to the results of other models.

CR-E-X-C

This model is an XGBoost model that aims to predict the number of aftershocks through regression using a total of ten input features: the six basic stress tensor components, the Euclidean distance to the target grid cell, and the absolute location of the cell. Unlike the CR-E-X model, the absolute location information of the grid cell is also included in the model in this study.

No undersampling or oversampling techniques were applied to the training data; the model was trained to include samples from all grid cells.

In the XGBoost-based regression model, a Bayesian optimization method was used to tune the hyperparameters. To increase the model's prediction accuracy and generalization capacity, seven basic hyperparameters were systematically optimized within specific ranges. In this context, the decision tree depth, which directly affects the model's learning capacity and the risk of overfitting, was tested between 3 and 12, the learning rate was set between 0.0005 and 0.2, and the total number of trees was evaluated between 100 and 800. Additionally, to increase the diversity and generalizability of the model, the sample ratio and feature ratio used in training each tree were optimized between 0.4 and 1.0. Regularization parameters were also tested with values ranging from 0 to 5, making the model more resilient to overfitting.

After completing the Bayesian optimization process, the model was retrained with the best parameter combination obtained and evaluated on the test data. In this problem, where negative predictions were not significant, all model outputs were adjusted to avoid falling below zero.

The model demonstrated the highest accuracy and generalization performance among all previous XGBoost models, with an MAE of 0.05 and an RMSE of 0.25. These results demonstrate that the XGBoost algorithm is an effective approach for predicting the number of aftershocks, particularly when combined with Euclidean distance filters and location information. However, the neural network model performed better with an MAE of 0.03. See the Figure 4.25 to see the relationship between the actual and predicted aftershock counts of the CR-E-X-C model.

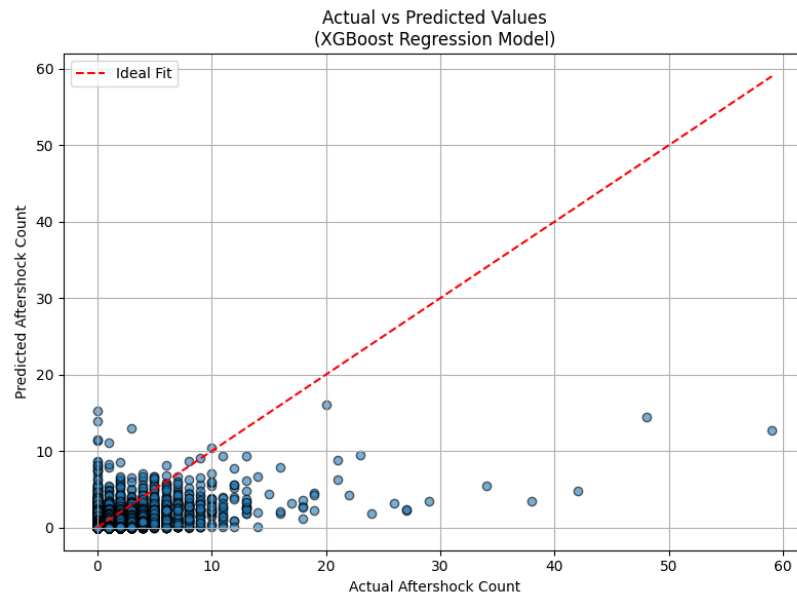


Figure 4.25. The relationship between the actual and predicted aftershock counts of the CR-E-X-C model.

The performance results of all models developed for the count regression task in this study are summarized in the Table 4.4.

Table 4.4. Comparative performance metrics of count regression models.

#	Model	MAE	RMSE
1	CR	0.03	0.29
2	CR-E	0.03	0.29
3	CR-E-F	0.03	0.29
4	CR-E-F-P	0.03	0.29
5	CR-E-FE	0.03	0.29
6	CR-E-FE-P	0.03	0.29
7	CR-E-FEC	0.03	0.29
8	CR-E-FEC-P	0.03	0.29
9	CR-E-FE-SO	0.27	0.48
10	CR-E-FE-SU	0.27	0.40
11	CR-E-FE-SU-D	0.27	0.40
12	CR-E-SU-D-X	0.29	0.40
13	CR-E-X	0.06	0.28
14	CR-E-X-C	0.05	0.25

4.3.2. Kernel-Based Count Regression Modeling of Aftershock Activity

In this section, a kernel-based regression framework was developed to predict the spatial distribution of aftershock counts within each grid cell. Unlike discrete classification of aftershock occurrence, this task aims to model the expected number of aftershocks as a continuous quantity derived from the Gaussian kernel-weighted sum of observed aftershock events. The kernel smoothing followed the same physical parameters as in Section 4.1.2, using $\sigma = 10$ km and $R = 32$ km to represent the spatial decay of seismic influence.

The raw kernel intensity y_{raw} is computed as the cumulative isotropic Gaussian kernel response over the set of aftershock locations, given by

$$y_{\text{raw}} = \sum_{i \in \mathcal{A}} \exp\left(-\frac{r_i^2}{2\sigma^2}\right) \mathbb{I}[r_i \leq R] \quad (4.2)$$

where r_i denotes the Euclidean distance to the i th aftershock, σ is the kernel bandwidth, R is the cutoff radius, and $\mathbb{I}[\cdot]$ denotes the indicator function. To stabilize the variance and reduce the influence of extreme values, a logarithmic transformation is applied as

$$y_{\text{log}} = \log(1 + y_{\text{raw}}). \quad (4.3)$$

The training process was conducted in this log-transformed space, and evaluation was performed in the raw space using the inverse transform $\hat{y}_{\text{raw}} = \exp(\hat{y}_{\text{log}}) - 1$. The input features of the model consist of six components of the elastic-static stress tensor (σ_{xx} , σ_{xy} , σ_{xz} , σ_{yy} , σ_{yz} , σ_{zz}); The spatial derivatives and gradient norms of these components on three axes are augmented with tensor invariants (trace, L_1 , L_2 , and L_∞ norms), the eigenvalue ratio, the maximum shear stress, and the von Mises equivalent stress. Non-finite values in all numerical columns were replaced with zero, extreme outliers were trimmed with a 99.5 percentile threshold, and data type conversions were ensured. Inputs were fit only to the training data and normalized using z-score standardization (StandardScaler), which was applied to the validation and test sets with the same parameters. This process brought all variables to a scale with zero mean and unit variance, balancing the effects of inputs of different sizes on the learning process, thus ensuring more stable model learning.

Data separation was performed deterministically; 16 CSV files in the training folder were reserved for validation, the remaining were used in training, and the data in the test folder were kept completely independent.

Kernel-Based Count Regression MLP Model

The model was built on a multilayer perceptron (MLP) architecture with layer widths of [256, 256, 128, 64, 32, 1]. Following linear transformation, batch normalization, ReLU activation, and 10% dropout were applied to each hidden layer. ReLU

activation was used in the output layer to produce positive and log-space consistent predictions. The loss function was the mean squared error (MSE) of the log1p transformation, and the optimization process was carried out with the AdamW algorithm. The learning rate was set to 2×10^{-4} , the weight decay value was set to 1×10^{-5} , the learning rate timer was set to ReduceLROnPlateau (factor=0.5, patience=3), and the early stopping mechanism (patience=12, criterion: validation log1p-MSE) was applied. The batch size was selected as 1024, and the gradient norm was limited to 5.0 during training. The log1p-MSE approach, which is more suitable for the continuous and smooth nature of the target variable, was preferred in this model.

During the experimental process, numerous experiments were conducted with different layer numbers, neuron widths, learning rates, and batch sizes to investigate the effects of the model architecture and hyperparameters. Within the scope of the search, node widths ranging from [128–512], different dropout rates (0.05–0.3), various optimization strategies (Adam, AdamW, RMSprop), and learning rate combinations were tested. The results showed that the configuration established with a layer order of [256, 256, 128, 64, 32, 1], a 10% dropout, a batch size of 1024, and a learning rate of 2×10^{-4} provided the most balanced and overall best performance.

Model performance was monitored in both log1p and raw space, with train–val and train–test metrics recorded throughout training. The final model was refitted with weights from the epoch that achieved the lowest validation log1p-MSE value and evaluated on the test set. Model performance was assessed using mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2).

The corresponding training–validation MAE loss curves are presented in Figure 4.26. The Gaussian-kernel smoothing of aftershock counts is illustrated in Figure 4.27, followed by the predicted aftershock count distribution obtained from the kernel-based regression model in Figure 4.28. Together, these figures provide a sequential view of the smoothing and prediction processes for aftershock count estimation.

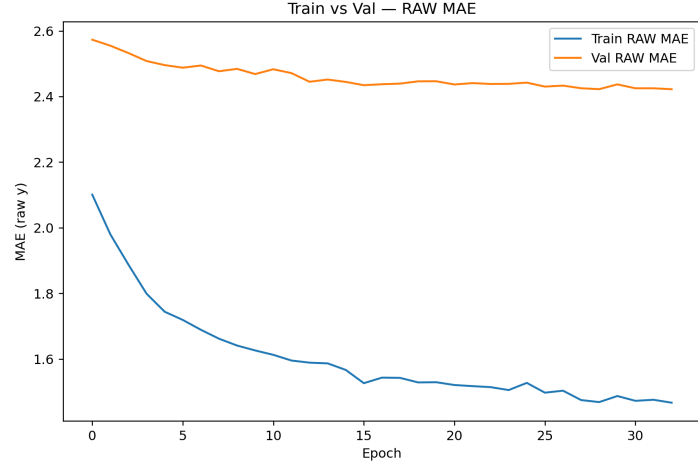


Figure 4.26. Training and validation MAE curves of the kernel-based count regression model with MLP model.

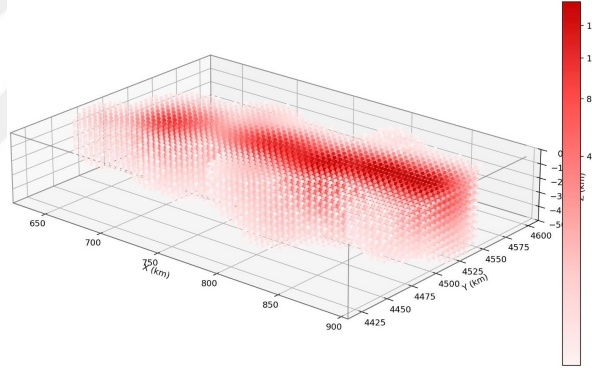


Figure 4.27. Gaussian-kernel smoothing of aftershock counts for the 1999 Izmit earthquake.

Kernel-Based Count Regression XGBoost Model

In addition to the neural network model, a gradient-boosted decision tree approach was implemented using the XGBoost algorithm to model the kernel-based aftershock counts. The same input feature set was used as in the MLP model, consisting of the six components of the elastic-static stress tensor, their spatial derivatives and gradient norms, tensor invariants (trace, L_1 , L_2 , and L_∞ norms), the eigenvalue ratio,

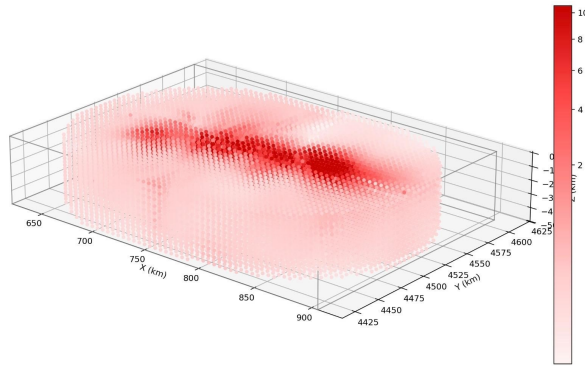


Figure 4.28. Predicted aftershock count distribution for the 1999 Izmit earthquake obtained from the kernel-based regression model.

the maximum shear stress, and the von Mises equivalent stress. All preprocessing steps — including clipping of extreme values at the 99.5th percentile, replacement of non-finite values with zero, and z-score standardization fitted only on the training set — were kept identical to ensure a fair comparison between the models. The deterministic train-validation-test split described above was also preserved.

The model was implemented as an `XGBRegressor` with 800 trees, a learning rate of 0.03, maximum tree depth of 6, and standard regularization parameters (L2 penalty $\lambda = 1.0$, L1 penalty $\alpha = 0.0$). Stochasticity was introduced through subsampling (0.8) and column subsampling (0.8), while the objective was set to squared error regression (`reg:squarederror`). During training, the root-mean-square error (RMSE) in log-space was monitored on both the training and validation sets. Also the evolution of these losses as a function of the boosting rounds is reported in Figure 4.29.

Model performance on the test set was quantified using the mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2). The XGBoost model achieved a test RMSE of approximately 6.42, an MAE of 2.36, and an R^2 of 0.25, as summarized in Table 4.5.

The spatial prediction of aftershock counts obtained from the XGBoost model is shown in Figure 4.30, which provides an alternative, tree-based representation of

the smoothed aftershock density field for the 1999 Izmit earthquake. Together with the MLP-based prediction map in Figure 4.28 and the kernel-smoothed counts in Figure 4.27, these results offer a comparative view of how different learning architectures approximate the same underlying kernel-based aftershock count distribution.

The results for kernel-based count models are presented in the table 4.5.

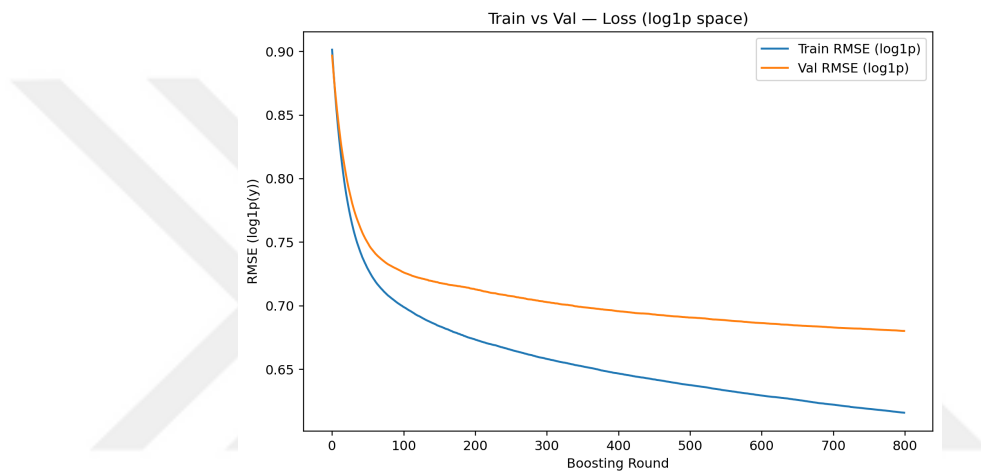


Figure 4.29. Training and validation RMSE curves of the kernel-based count regression model with XGBoost model.

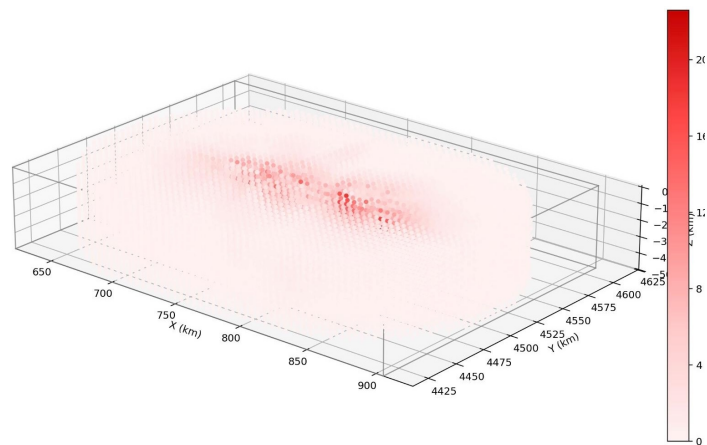


Figure 4.30. Training and validation RMSE curves of the kernel-based count regression model with XGBoost model.

Table 4.5. Performance metrics of the kernel-based count regression models.

#	Model	RMSE	MAE	R^2
1	MLP	6.03	2.19	0.34
2	XGBoost	6.42	2.36	0.25

4.4. Regression Modeling of Aftershock Magnitude

In the third phase of the modeling process, regression-based approaches were developed to estimate the magnitude of the aftershock occurring in each cell as a continuous variable within a three-dimensional cube area defined as $100 \text{ km} \times 100 \text{ km}$ horizontally and 50 km vertically, divided into $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$ grid cells of equal volume. In this context, numerical labels representing the highest aftershock magnitude observed in each cell were used as target variables. This allowed for the correlation of local magnitude estimates with spatial resolution.

In the study, various regression models based on ANN-based and gradient boosting algorithms were constructed. Different input vectors, layer architectures, and hyperparameter configurations were systematically tested in these models. The artificial neural network architectures were defined on the PyTorch platform, and different combinations of stress tensors (e.g., principal components only, stress values of neighboring cells, or PCA-reduced representations) were included in the input layers. All input data were scaled using z-score normalization before model training to prevent weight imbalances that could arise from scale differences between variables.

The ReLU activation function was used to learn nonlinear patterns in the hidden layers, and the output layer was also terminated with ReLU to prevent negative predictions. Parameter updates were performed throughout the training process using the Adam optimization algorithm; in some architectures, a dropout mechanism was implemented at different rates to prevent overfitting. All ANN models were trained

with mini-batches of 32 or 64 samples, and the training process for each architecture lasted 300 epochs.

To mitigate the imbalances observed in the data set, oversampling or under-sampling techniques were applied in some models, depending on the target variable distribution. The effects of these strategies on model performance were analyzed comparatively. Additionally, models were created that filtered according to a specific distance threshold, including only grid cells located within a certain distance from the epicenter.

In addition to artificial neural network-based models, regression models based on the XGBoost algorithm were also developed. Hyperparameter optimization in these models was performed using Bayesian Optimization instead of manual testing.

All regression models were evaluated on the test data set using basic regression metrics such as MAE, RMSE, and R^2 score. Thus, the size estimation performances obtained with different structural and parametric choices were compared empirically.

R

This model is an artificial neural network-based regression model that uses only six principal stress tensor components as inputs. Using these features, which correspond to each grid cell, it continuously predicts the aftershock magnitude that may occur in that cell as an output.

The model architecture has a fully connected structure consisting of four layers. The input layer passes the 6-dimensional stress vector to the first hidden layer with 32 neurons, followed by two hidden layers with 128 and 32 neurons. The final layer is defined as a single-neuron output unit. Consistent with the common modeling framework, ReLU activation was used in the hidden layers, and ReLU was also applied to the output layer to prevent negative values.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

The CR model's MAE is quite low at 0.08, indicating generally accurate predictions. However, the RMSE of 0.57 suggests larger errors in some examples. While the model generally makes accurate predictions, it occasionally exhibits significant deviations.

R-E

This model is an artificial neural network-based regression model with a total of seven input features, in addition to the six fundamental stress tensor components in the R model, which also includes the Euclidean distance between the epicenter and each grid cell. The model architecture consists of a total of four fully connected layers. The input layer transforms a 7-dimensional vector (6 stress components + 1 distance) into a first layer of 32 neurons. This is followed by two intermediate layers of 128 and 32 neurons, respectively, while a single-neuron layer generates the predicted value at the output. Nonlinear relationships are modeled by applying the ReLU activation function to all hidden layers. The ReLU function is also used in the output layer to prevent the model from generating negative values.

The model was trained using mini-batches of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs. Adding the Euclidean distance feature to the R-E model did not create a significant difference in performance; while the MAE and RMSE values of the model remained the same as the R model.

R-E-F

This model is an artificial neural network-based regression model with a total of 43 input features, in addition to the input features from the R-E model. 36 stress components from six neighboring cells on the grid cell’s surface plane are also added.

The model architecture features a fully connected neural network structure consisting of four layers. The input layer projects the 43-dimensional extended feature vector onto a representation space of 32 neurons. This is followed by two hidden layers of 128 and 32 neurons, respectively. The output layer consists of a single neuron and uses a ReLU activation function to prevent negative predictions. The ReLU function is also applied to all hidden layers to ensure the learning of nonlinear patterns.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs. Despite a significant increase in input size, no improvement was observed in MAE and RMSE values. The addition of surface neighbors did not improve model performance.

R-E-F-P

This model uses the input features from the R-E-F model; however, because this multidimensional neighborhood data can contain high correlation and data noise, PCA, a dimensionality reduction method, was applied to the 36 input features from the surface neighborhoods. Before PCA, all input features were normalized to the same scale, and then PCA was applied to the neighborhood stress components to select the 25 most significant components. Consequently, the model’s input vector became a 32-dimensional vector combining the normalized baseline features and the PCA-transformed neighborhood stress information.

The model architecture features a deep artificial neural network structure consisting of six layers. The input layer receives this combined 32-dimensional vector and processes it through hidden layers consisting of 128, 1024, 256, and 64 neurons, respectively. The ReLU activation function is applied to each hidden layer, enhancing the model's ability to learn nonlinear patterns. The output layer is a single-neuron layer, and the ReLU activation function is also applied to prevent negative predictions.

The model was trained using mini-batches of 32 samples. The Adam optimization algorithm and MSE loss function were used, with a learning rate of 0.001 throughout the training process; the model was trained for 300 epochs.

An examination of model performance reveals that, similar to previous R models, the MAE (0.08) and RMSE (0.57) values were maintained. This suggests that, although the input space was made more compact and learnable through dimensionality reduction with PCA, the model's success in explaining the number of aftershocks remained limited. In other words, even though neighborhood information was summarized and incorporated into the model, no significant improvement in predictive power was achieved.

R-E-FE

This model is an artificial neural network-based regression model with a total of 115 input features, in addition to the input features from the R-E-F model. 72 stress components from 12 neighboring cells in the edge plane are also added.

The model architecture features a fully connected neural network structure consisting of four layers. The input layer receives this rich 115-dimensional feature vector and passes it to the first hidden layer of 128 neurons. This is followed by two hidden layers of 256 and 32 neurons, respectively. The ReLU activation function is applied to all hidden layers to model nonlinear relationships. The output layer consists of a single neuron, and ReLU activation also prevents negative predictions.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

An examination of the R-E-FE model outputs reveals that the MAE (0.08) and RMSE (0.57) values remain the same as those of previous models. These results indicate that even though the data representation was expanded by increasing the input size and including neighboring neighbors, the model's ability to predict the number of aftershocks was not significantly improved. Therefore, while expanding the spatial context is theoretically meaningful, it has not directly impacted performance in practice.

R-E-FE-P

This model uses the input features from the R-E-FE model; however, in this model, 108 stress tensor components taken from 18 neighboring cells on the surface and edge planes were subjected to dimensionality reduction using PCA, reducing them to 75 principal components. Thus, the total input dimension was determined to be 82.

The model architecture consists of a deep artificial neural network structure consisting of four hidden layers and one output layer, starting from the input layer and containing 128, 1024, 256, and 64 neurons, respectively. Nonlinear relationships are modeled in each layer using the ReLU activation function. The output layer of the model is a single neuron, and negative predictions are also prevented using ReLU activation.

The model was trained using mini-batches consisting of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs. R-E-FE-P model remained at the same level as previous models in MAE (0.08) and RMSE (0.57). despite using 108 neighborhood inputs simplified by PCA. This suggests that

although the reduction in dimensionality reduced the complexity of the model and the computational overhead, the predictive performance was not significantly improved. PCA provided a more compact representation of the data structure, but did not directly contribute to predictive performance.

R-E-FEC

This model is an artificial neural network-based regression model with a total of 163 input features, in addition to the input features from the R-E-FE model. 48 stress components from 8 neighboring cells in the corner plane are added.

The model architecture features a deep and high-capacity artificial neural network. In the input layer, the 163-dimensional vector is first projected onto a representation plane of 256 neurons. It is then passed through three more layers of 1024, 128, and 32 neurons, respectively. The ReLU activation function is applied to all hidden layers. The output layer consists of a single neuron, again limited by ReLU activation, to the generation of negative values.

The model was trained using mini-batches of 32 samples. The Adam optimization algorithm, with a learning rate of 0.001, and the MSE loss function were used throughout the training process; the model was trained for 300 epochs.

The R-E-FEC model expanded the spatial context by incorporating the environmental neighborhood most comprehensively; however, MAE (0.08) and RMSE (0.57) remained at the same level as previous models. Despite increasing input size and model capacity, no significant improvement in prediction performance was achieved. This suggests that the contribution of high-dimensional inputs is limited by overfitting or knowledge redundancy rather than by increasing data power.

R-E-FEC-P

This model uses the input features from the R-E-FEC model; however, in this model, 156 stress tensor components taken from 26 neighboring cells in the surface, edge, and corner planes were subjected to dimension reduction using PCA, reducing them to 110 principal components. Thus, including the basic 7 input features, the total input dimension was determined to be 117.

The model architecture is a fully connected deep learning structure. The input layer passes the 117-dimensional vector to the first hidden layer of 128 neurons. This is followed by three hidden layers, each containing 1024, 256, and 64 neurons, respectively. Nonlinear relationships are modeled using the ReLU activation function in all layers. The output layer consists of a single neuron and, again, includes ReLU activation, preventing the prediction of negative values. This structure allows the model to focus on predicting the number of aftershocks at or above zero.

The model was trained using mini-batches of 32 samples. The Adam optimization algorithm and MSE loss function were used with a learning rate of 0.001 throughout the training process; the model was trained for 300 epochs.

Despite having the largest input dimensionality, the R-E-FEC-P model failed to improve on MAE (0.08) and RMSE (0.57) compared to previous models. While dimensionality reduction with PCA provides advantages in terms of computational efficiency, it did not improve the overall performance of the model. This suggests that conveying extended spatial information through a summary representation offers limited benefits.

R-E-FE-SO

This artificial neural network-based regression model, which uses the same input features as the R-E-FE model, takes as input a total of 115 input features derived from

the six basic stress tensor components, Euclidean distance, and the stress components of six neighbors from surfaces and 12 neighbors from edges in a three-dimensional grid system. Unlike R-E-FE, this model employs oversampling of the data set.

The model architecture is a fully connected neural network structure with a total of four hidden layers. The input layer takes a 118-dimensional vector and converts it into a representation of 128 neurons in the first layer. This is followed by two more layers, each containing 256 and 64 neurons, respectively. The ReLU activation function is applied to all hidden layers to model nonlinear relationships. Additionally, dropout layers were added at 30%, 30%, and 20% ratios to enhance the model's generalization. The output layer of the model consists of a single neuron, and ReLU activation also prevents negative predictions.

Specifically for this model, the data set was balanced by oversampling cells with a 100% increase in the number of continuations greater than zero. This increased the model's ability to learn positive values, particularly those with low frequencies. The model was trained using mini-batches of 64 samples. The Adam optimization algorithm and MSE loss function, with a learning rate of 0.0005, were used throughout the training process. The model was trained for 300 epochs. At the end of each epoch, the RMSE performance of the model was monitored on the test data, and the model parameters were recorded when the best results were achieved.

Although the R-E-FE-SO model was trained on a balanced data set using an oversampling strategy, contrary to expectations, the MAE (0.91) and RMSE (1.23) values increased. This result suggests that the data augmentation strategy did not provide any improvement in this structure.

R-E-FE-SU

This artificial neural network-based regression model, which uses the same input features as the R-E-FE model, takes as input a total of 115 input features derived from

the six basic stress tensor components, Euclidean distance, and the stress components of six neighbors from surfaces and 12 neighbors from edges in a three-dimensional grid system. Unlike R-E-FE, this model employs undersampling of the data set.

The model architecture features a fully connected artificial neural network structure consisting of five layers. The input layer transforms the 115-dimensional input vector into the first layer of 128 neurons. Subsequent layers contain 256, 64, and 1 neuron, respectively. Nonlinear relationships are modeled in the intermediate layers by applying the ReLU activation function. Dropout mechanisms are also used to prevent overfitting (at 0.3 and 0.2 rates). The ReLU function is also applied to the output layer to ensure only positive prediction values are obtained.

The model was trained using mini-batches of 64 samples. The Adam optimization algorithm and MSE loss function were used, with a learning rate of 0.0005 throughout the training process; the model was trained for 300 epochs. At the end of each epoch, the RMSE performance of the model was monitored on the test data, and model parameters were recorded when the best results were achieved. The R-E-FE-SU model has better MAE (0.08) and RMSE (0.57) values than the R-E-FE-SO model. It gives similar results to the non-sampling R models.

R-E-FE-SU-D

This neural network-based regression model, which uses the same input features as the R-E-FE-SU model, contains a total of 115 input features derived from the six principal stress tensor components, Euclidean distance, and the stress components of six neighbors from the surfaces and 12 neighbors from the edges in the three-dimensional grid system. In the model, consistent with the R-E-FE-SU approach, it was assumed that the distance to the location of aftershocks would be a determining factor in the modeling process, and a distance filter was applied to the data set accordingly. After testing different threshold values, only observations located within a certain distance of the epicenter were included in the model. Thus, samples located

further away and less likely to exhibit aftershock characteristics were excluded from the model to achieve a more meaningful sample structure.

The model architecture features a fully connected neural network structure consisting of five layers. The input layer projects the 115-dimensional input vector onto a first hidden layer of 128 neurons. This is followed by two hidden layers of 256 and 64 neurons, respectively. Nonlinear relationships were learned by applying the ReLU activation function to all hidden layers. Furthermore, dropouts of 0.3, 0.3, and 0.2 were applied after each hidden layer, respectively, to increase the model's resistance to overfitting. The output layer consists of a single neuron, and ReLU activation prevents negative predictions.

The training data included only samples located within 55 km of the epicenter. Then, to reduce class imbalance, an undersampling process was performed by randomly selecting only twice as many samples from class labeled 0 as from class 1. This balanced the training data set.

The model was trained with mini-batches of 64 samples. Throughout the training process, the Adam optimization algorithm with a learning rate of 0.0005 and the MSE loss function were used; the model was trained for 300 epochs. At the end of each epoch, the RMSE performance of the model was monitored on the test data, and the model parameters were recorded when the best results were achieved. During the training process, the model weights that showed the best validation performance were recorded and this best model was used in the final evaluation.

The CR-E-FE-SU-D model was configured with spatial filtering (55 km) and undersampling strategies. However, the obtained performance values (MAE: 0.13, RMSE: 0.60) indicate that the model does not adequately explain the target variable. These results indicate that distance-based sample selection and data reduction do not have the expected positive impact on model accuracy.

R-E-SU-D-X

This model is an XGBoost-based regression model that aims to predict the magnitude of the aftershock as a continuous output using a total of seven input features: six stress tensor components and the Euclidean distance. Compared to the R-E-FE-SU-D model, the input dimension has been reduced; no features belonging to neighboring grid cells have been included in the model. Furthermore, the XGBoost algorithm, a tree-based gradient boosting method, was chosen over an artificial neural network-based approach.

The training data included only samples located within 55 km of the epicenter. Subsequently, an undersampling process was applied by selecting twice as many 0-labeled data as 1-labeled samples to reduce class imbalance. These samples were shuffled to create a balanced training set.

In the XGBoost-based regression model, a Bayesian optimization method was used to tune the hyperparameters. To increase the model's prediction accuracy and generalization capacity, seven key hyperparameters were systematically optimized within specific ranges. In this context: The decision tree depth, which directly affects the model's learning capacity and risk of overfitting, was tested between 3 and 12, the learning rate was set between 0.0005 and 0.2, and the total number of trees was evaluated between 100 and 800. Furthermore, to increase the model's diversity and generalizability, the sample ratio and feature ratio used in training each tree were optimized between 0.4 and 1.0. Regularization parameters were also tested with values ranging from 0 to 5, making the model more resistant to overfitting.

The model was retrained with the best parameter set obtained from the Bayesian optimization process, and its performance was evaluated on the test data set. Because negative values in the model's output were not significant, the predictions were adjusted to avoid falling below zero.

This model yields lower results compared to the non-sampled models, with MAE (0.19) and RMSE (0.58) obtained using the XGBoost algorithm with limited input.

R-E-X

This model aims to predict the aftershock magnitude through regression using a total of seven input features, consisting of the six principal stress tensor components and the Euclidean distance to the target grid cell. Unlike the first model, the R structure, this study uses the XGBoost algorithm, a tree-based gradient boosting method, instead of an artificial neural network. No undersampling or oversampling techniques were applied to the training data; the model was trained to include samples from all grid cells.

In the XGBoost-based regression model, a Bayesian optimization method was used to tune the hyperparameters. To increase the model's prediction accuracy and generalization capacity, seven principal hyperparameters were systematically optimized within specific ranges. In this context, the following are included: The decision tree depth, which directly affects the model's learning capacity and risk of overfitting, was tested between 3 and 12, the learning rate was set between 0.0005 and 0.2, and the total number of trees was evaluated between 100 and 800. Furthermore, to increase the model's diversity and generalizability, the sample ratio and feature ratio used in training each tree were optimized between 0.4 and 1.0. Regularization parameters were also tested with values ranging from 0 to 5, making the model more resistant to overfitting.

The model was retrained with the best parameter set obtained through the Bayesian optimization process, and its performance was evaluated on the test data set. Because negative values in the model's output were not significant, the predictions were adjusted to avoid falling below zero.

The MAE value was calculated as 0.15, indicating that the model’s average prediction error was reasonable. However, the RMSE of 0.55 indicates that some predictions were far from the true values, and the overall model error remained high.

R-E-X-C

This model is an XGBoost model that aims to predict aftershock magnitude through regression using a total of ten input features: the six basic stress tensor components, the Euclidean distance to the target grid cell, and the absolute location of the cell. Unlike the R-E-X model, the absolute location information of the grid cell is also included in the model in this study.

No undersampling or oversampling techniques were applied to the training data; the model was trained to include samples from all grid cells. In the XGBoost-based regression model, a Bayesian optimization method was used to tune the hyperparameters. To increase the model’s prediction accuracy and generalization capacity, seven basic hyperparameters were systematically optimized within specific ranges. In this context, the decision tree depth, which directly affects the model’s learning capacity and the risk of overfitting, was tested between 3 and 12, the learning rate was set between 0.0005 and 0.2, and the total number of trees was evaluated between 100 and 800. Additionally, to increase the diversity and generalizability of the model, the sample ratio and feature ratio used in training each tree were optimized between 0.4 and 1.0. Regularization parameters were also tested with values ranging from 0 to 5, making the model more resistant to overfitting.

The model was retrained with the best parameter set obtained through the Bayesian optimization process, and its performance was evaluated on the test data set. Because negative values in the model’s output were not significant, the predictions were adjusted to avoid falling below zero.

Examining the model's performance metrics, MAE values of 0.14 and RMSE of 0.52 were obtained. These results indicate that the absolute error level of the model is lower than the previous model. See the Figure 4.31 to see the relationship between the actual and predicted aftershock counts of the R-E-X-C model.

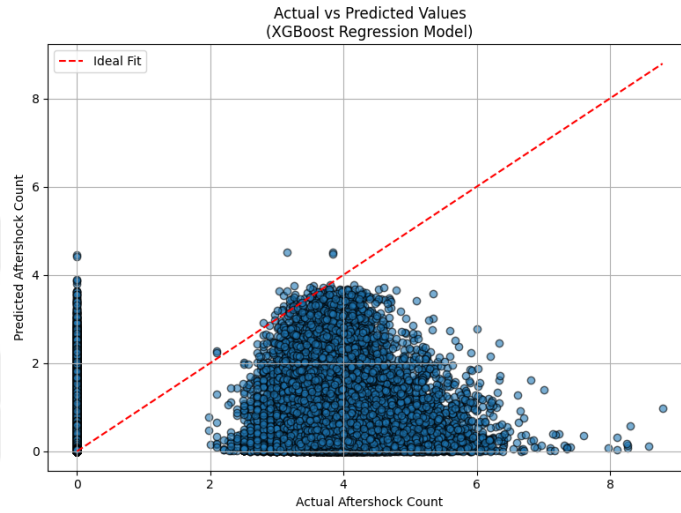


Figure 4.31. The relationship between the actual and predicted aftershock magnitudes of the R-E-X-C model.

The performance results of all developed for the magnitude regression task in this study are summarized in the Table 4.6.

Table 4.6. Comparative performance metrics of magnitude regression models.

#	Model	MAE	RMSE
1	R	0.08	0.57
2	R-E	0.08	0.57
3	R-E-F	0.08	0.57
4	R-E-F-P	0.08	0.57
5	R-E-FE	0.08	0.57
6	R-E-FE-P	0.08	0.57
7	R-E-FEC	0.08	0.57
8	R-E-FEC-P	0.08	0.57
9	R-E-FE-SO	0.91	1.23
10	R-E-FE-SU	0.08	0.57
11	R-E-FE-SU-D	0.13	0.60
12	R-E-SU-D-X	0.19	0.58
13	R-E-X	0.15	0.55
14	R-E-X-C	0.14	0.52

5. DISCUSSION

This chapter provides a comprehensive interpretation of the experimental findings by synthesizing the results of all classification, probabilistic, and regression models developed in this study. Beyond reporting numerical outcomes, the discussion evaluates what these results reveal about the relationship between static stress fields and after-shock behavior, compares the strengths and limitations of different modeling strategies, and explains why certain input structures, sampling techniques, or algorithms yielded better performance than others. The broader physical meaning of the learned patterns, systematic trends across models, and methodological constraints are highlighted, situating the findings within the wider seismological and machine learning literature.

Building on the experimental analyses presented in the previous chapter, the following sections interpret the behavior of each model family—binary classifiers, kernel-based probabilistic formulations, count regression models, and magnitude regression models—in terms of their predictive tendencies, sources of improvement, and remaining challenges. This chapter focuses on understanding what the results mean and how they advance the goal of stress-based aftershock prediction.

Within these tasks, different input combinations, sampling strategies, and modeling approaches are systematically evaluated. Additionally, as an alternative to classical classification and regression models, two new model families were developed using continuous labels generated by Gaussian kernels: (i) a kernel-based probabilistic classification model, which soft-labeled aftershock probability and extended binary classification to probability estimation, and (ii) a kernel-based numerical regression model, which represented aftershock intensity as a continuous, non-discrete intensity field. Taking all these models together, the study presents a comprehensive, multi-layered prediction that considers both local stress components and spatial interactions. This approach forms the basis of a data-driven seismic prediction system capable of modeling not only the occurrence trends of aftershocks but also their spatial distribution and density.

5.1. Aftershock Occurrence Modeling Results

The modeling of aftershock occurrence was approached through two complementary frameworks: (i) a binary classification setting using discrete labels (0 or 1), and (ii) a kernel-based probabilistic formulation that represented aftershock likelihoods as continuous soft probabilities. Both frameworks aimed to assess how effectively static stress components and their derived features could explain the spatial patterns of aftershock triggering.

The binary classification experiments began with the BC model, which utilized only the six fundamental stress tensor components to evaluate whether local stress information alone could predict the presence of aftershocks. Subsequently, additional features were incrementally introduced, including the Euclidean distance from the mainshock hypocenter (E) and the stress tensors of neighboring grid cells (F). Among these configurations, the BC-E-FE model, which incorporated both surface and edge neighborhood stresses, achieved the most balanced performance across accuracy, ROC AUC, and F1 metrics. This configuration was thus adopted as the representative baseline for subsequent comparisons.

Given the natural sparsity of aftershocks within the dataset, severe class imbalance emerged as a key challenge. To address this, oversampling was applied to augment the minority class (aftershock-positive cells), thereby improving the model’s sensitivity to rare events. Comparative analyses showed that oversampling yielded more balanced and reliable results than undersampling, particularly improving recall and overall AUC without excessively degrading precision. Additionally, a spatial filtering strategy was introduced: training was restricted to cells within a 55 km distance from the mainshock centroid, under the assumption that most aftershocks occur in the proximal stress field. This filtering reduced noise from seismically inactive regions and produced measurable improvements in ROC AUC and F1 scores.

The final classification architecture employed XGBoost, a gradient-boosted decision tree algorithm well-suited for handling nonlinearity and class imbalance. The BC-E-FE-SO-D-X configuration achieved the best overall performance, with an accuracy of 0.94 and a ROC AUC of 0.81. Nevertheless, the recall (0.29) and precision (0.12) remained limited, indicating that despite improved separation capability, the rarity of positive events still constrained predictive stability. Parallel experiments using a 3D U-Net convolutional neural network achieved markedly lower metrics (ROC AUC = 0.06), highlighting that under the current data representation and sampling scheme, convolutional volumetric models failed to generalize effectively.

to overcome the inherent limitations of discrete classification, a kernel-based probabilistic approach was developed. In this framework, binary labels were replaced with continuous soft probabilities derived from Gaussian kernel smoothing of observed aftershock locations. Within this setting, two alternative learning architectures were explored: a multilayer perceptron (MLP) and an XGBoost-based gradient boosting model. The MLP-based probabilistic model achieved strong calibration performance, with a validation BCE of 0.41 and a Brier score of 0.06, while the XGBoost model reached a comparable Brier score of 0.07. Both models preserved the smooth spatial decay of aftershock likelihood instead of enforcing hard class boundaries, thereby providing a more faithful representation of the underlying kernel-smoothed seismic response than the original binary classifier.

In terms of discrimination, XGBoost offered a slight advantage over the MLP. Although both models achieved the same high Recall (0.93), the XGBoost model yielded a higher Precision (0.72 vs. 0.68) and a slightly improved F1-score (0.81 vs. 0.79), together with a higher ROC AUC (0.85 vs. 0.82). These results indicate that the tree-based ensemble can exploit complex non-linear interactions in the stress-based feature space to refine the separation between high- and low-probability regions.

Overall, the kernel-based probabilistic framework — whether implemented with MLP or XGBoost — provided more stable, interpretable, and calibrated results than

the discrete classification models, effectively bridging the gap between classical binary estimation and continuous probability modeling in both identifying potential aftershock zones and mapping probabilistic hazard intensity.

5.2. Aftershock Count Regression Modeling Results

This section holistically evaluates the results of models developed to estimate the numerical intensity of aftershocks. In this context, both classical regression models based on principal stress components and spatial kernel-based probabilistic approaches are discussed.

First, ANN models built using principal stress components, derivative quantities, and neighborhood-based features were tested. CR Models aimed to estimate the number of aftershocks as a continuous variable and generally achieved low error levels. The mean absolute error (MAE) was approximately 0.03 and the root mean square error (RMSE) was around 0.29. These values indicate that the models accurately captured general trends but produced partial errors in high-variance aftershock clusters.

Sampling strategies were applied to examine the impact of data imbalance on model performance. Oversampling increased the sample size in regions where aftershocks were rarely observed, but it also increased the error rates (MAE: 0.27; RMSE: 0.48), leading to overfitting. In contrast, the undersampling strategy improved the model's generalization capacity by reducing dense negative examples and reduced the RMSE to 0.40. Furthermore, in experiments conducted with a data subset containing only cells within 55 km of the epicenter, error metrics remained similar, but the model was observed to perform more consistently in high-error regions.

Next, the artificial neural network-based approaches were compared with the XGBoost algorithm, which employs the gradient boosting method of decision trees. The XGBoost model (CR-E-X) increased the MAE to 0.06 and reduced the RMSE to 0.28, demonstrating that the model was able to make more accurate predictions in

high-variance areas, but that accuracy was slightly reduced with small errors. The CR-E-X-C model, which incorporated stress information from neighboring cells, further reduced the error rates (MAE: 0.05; RMSE: 0.25) and better captured the spatial continuity in the aftershock distribution.

Following these fundamental approaches, the kernel-based regression framework modeled aftershock counts as continuous targets smoothed through Gaussian kernels rather than discrete labels. Using kernel densities generated with parameters $\sigma = 10, \text{km}$ and $R = 32, \text{km}$ provides a more physically consistent representation of spatial transitions and focal concentrations in aftershock productivity. This smoothing enforces continuity in regions where raw count values fluctuate sharply, thus offering a more stable learning signal for regression models.

Two alternative architectures were examined within this formulation: a multi-layer perceptron (MLP) and a tree-based XGBoost regressor. The MLP model achieved $\text{RMSE} = 6.03$, $\text{MAE} = 2.19$, and $R^2 = 0.34$, outperforming the XGBoost model, which obtained $\text{RMSE} = 6.42$, $\text{MAE} = 2.36$, and $R^2 = 0.25$. These results indicate that although both models capture the large-scale structure of aftershock intensity, the neural network was more effective at learning the smooth, log-transformed target distribution imposed by the Gaussian kernel. In contrast, XGBoost introduced localized over- and underestimations in regions of high spatial variability, suggesting that gradient-boosted decision trees may be less suited to representing the continuous and low-noise behavior of the smoothed target field.

Overall, the findings presented in this section show that the kernel-based numerical regression approach can model aftershock intensity with moderate success but remains constrained by the informativeness of stress-derived input features. While Gaussian smoothing provides a physically meaningful bridge between discrete counts and continuous densities, the error levels demonstrate that these features alone cannot fully explain the spatial variability of aftershock numbers. This suggests that future work should incorporate additional spatial, geological, or temporal descriptors—such as

rupture geometry, heterogeneity of subsurface materials, or prior aftershock activity—to enhance predictive skill, particularly in complex high-variance fault segments.

5.3. Regression of Magnitudes Results

R models were developed to predict the magnitude of aftershocks as a continuous variable. The first eight models (R, R-E, R-E-F, R-E-F-P, R-E-FE, R-E-FE-P, R-E-FEC, R-E-FEC-P) include only the principal stress components and their versions in neighboring cells. These underlying structures were trained without any sampling strategy or distance filtering. The identical MAE (0.08) and RMSE (0.57) values for all these models indicate that this underlying structure has limited predictive capacity and that improvement is not possible unless new information or strategies are added.

The ninth model, R-E-FE-SO, incorporated the oversampling strategy (SO) into this underlying structure. Oversampling aimed to increase the model’s sensitivity to underrepresented high-magnitude aftershock samples by increasing their size. However, the model’s MAE increased to 0.91, and its RMSE increased to 1.23. This increase indicates that oversampling caused overlearning in this task or distorted the model’s overall sample distribution, resulting in large errors.

Undersampling, the opposite approach, was applied to the R-E-FE-SU model. This strategy aimed to achieve class balance by specifically reducing negative examples. In this model, the MAE and RMSE values returned to the baseline model at 0.08 and 0.57, respectively, meaning that the excessive distortion caused by oversampling was corrected. This demonstrates that undersampling at least maintained model balance.

In the R-E-FE-SU-D model, in addition to the undersampling strategy, a distance filter (D) was applied, which only considered cells within 55 km of the epicenter. After this filtering, the model’s MAE value became 0.13, and its RMSE was 0.60. This suggests that reducing information noise from distant regions provides some improvements. However challenges persist in predicting high-magnitude aftershocks.

The R-E-X model, in which the XGBoost algorithm is implemented, uses gradient boosting decision trees, unlike traditional artificial neural networks. This model yielded a MAE of 0.15 and an RMSE of 0.55. This demonstrates that XGBoost produces lower-variance and more stable predictions with stronger generalization capabilities.

When neighborhood information was added to XGBoost, i.e., when the R-E-X-C model was created, the model achieved the lowest error values of XGBoost models: MAE of 0.14 and RMSE of 0.52. However, these values remained higher than the performance of the baseline R model (MAE: 0.08, RMSE: 0.57) trained with only an ANN without any sampling strategy, distance filter, or XGBoost algorithm. This indicates that while XGBoost provides lower variance and more stable predictions, the baseline ANN model performs with lower average error in magnitude regression.

5.4. General Evaluation and Limitations

In this study, various classification and regression-based modeling strategies were developed to predict the spatial patterns, number, and magnitude of aftershocks following major earthquakes. The developed models evolved from simple structures containing only basic static stress components to more complex structures incorporating stress components derived from neighboring grid cells, spatial location information, distance to the epicenter, and probabilistic density features derived from Gaussian kernels.

Throughout the modeling process, the effects of different input combinations and sampling strategies (oversampling, undersampling) on model performance were systematically evaluated. However, performance gains achieved by simply changing input features or sampling techniques were limited. Class imbalance and the spatial sparsity of aftershocks reduced the variance explanation capacity, particularly in regression tasks, and increased prediction errors.

To mitigate these limitations, spatial filtering approaches were applied, including only grid cells within a certain distance from the epicenter. This approach made

the model's focus area more physically meaningful and reduced the influence of noisy regions. However, it was observed that it did not provide a significant increase in accuracy metrics, but offset the high error variances.

The most significant performance improvements were achieved with the implementation of the XGBoost algorithm, which is based on the gradient boosting method of decision trees. XGBoost significantly increased ROC AUC scores, particularly in classification tasks, and moved the coefficient of determination (R^2) from negative to positive values in regression tasks.

In addition, unlike traditional approaches, the kernel-based modeling framework defined the target variables in a continuous manner, using spatial densities smoothed with Gaussian kernels instead of binary or numerical labels. This method provided more consistent calibrations in probability estimates and more realistically represented spatial transitions. The use of kernel-based models produced more consistent results in explaining the spatial distribution of aftershocks compared to classical methods.

Limitations of the study include the sparse and unstable nature of the aftershock data, the fact that the input variables consisted of static (non-time-dependent) properties, and the study was conducted with a limited data set of large earthquakes.

5.5. Recommendations for Future Studies

The findings in this study demonstrate that the location, number, and magnitude of aftershocks can be predicted with a certain degree of accuracy using input features based on static stress tensors. However, the limited variance explanation, particularly in regression tasks, reveals that current approaches have room for improvement. Future studies should utilize more comprehensive and multidimensional methods to overcome these limitations and increase the generalizability of the models.

First, this study used data based solely on static stress fields. However, the evolution of aftershocks over time necessitates incorporating the dynamic components of the stress fields into the model. Therefore, the development of time-sensitive models can more accurately capture temporal changes in the intensity of aftershock occurrence.

Furthermore, while probabilistic approaches have been successfully implemented with a kernel-based modeling framework in this thesis, it is possible to enhance this method with more advanced spatial and temporal parameterizations in the future. Kernel functions defined at different scales, or multi-kernel approaches, can represent both spatial continuity and aftershock intensity gradients in greater detail.

On the other hand, the fact that the datasets used were limited to specific large earthquakes constituted a limitation in model generalization. Including broader datasets from different regions and diverse tectonic settings would increase the universal validity and robustness of the developed models.

Finally, this study established a prediction framework based solely on stress tensors. However, integrating geological parameters such as soil properties, fault types, local deformation zones, and lithological characteristics into the model could significantly enhance both the accuracy and the physical explanatory power of the predictions.

6. CONCLUSION

This study presents a comprehensive data-driven modeling framework based on components derived from static stress tensors to predict the spatial distribution, probability, number, and magnitude of aftershocks following major earthquakes. Aftershock behavior was considered in both discrete (probability of occurrence) and continuous (number and magnitude) dimensions, and a total of 45 models were systematically evaluated, comparing different input combinations, sampling strategies, spatial filtering approaches, and algorithmic structures.

In the binary classification experiments, incorporating neighboring stress components and constraining the analysis to grid cells within a physically meaningful distance provided moderate improvements, yet these enhancements remained limited by the sparse and imbalanced nature of aftershocks. More substantial gains were achieved with the introduction of the XGBoost algorithm, which significantly increased discrimination performance and yielded higher ROC AUC values (up to 0.85), despite persistent challenges in precision due to class rarity.

To address the shortcomings of sharp decision boundaries imposed by discrete classification, a kernel-based probabilistic formulation was developed, replacing binary labels with continuous soft probabilities obtained from Gaussian smoothing. This representation substantially improved probability calibration and spatial continuity. The MLP-based model achieved a BCE of 0.41, a Brier score of 0.06, and a ROC AUC of 0.82, while the XGBoost model produced similarly strong results with a Brier score of 0.07 and a higher ROC AUC of 0.85. Both models demonstrated high recall (0.93) and competitive F1-scores (0.79–0.81), indicating that the probabilistic formulation not only enhances calibration but also strengthens discrimination capability relative to classical classification. These findings show that kernel-based probabilistic modeling provides a more physically meaningful and statistically robust representation of aftershock likelihood.

Regression models for estimating the number of aftershocks generated continuous values using the principal components and derivative properties of stress tensors. While most models exhibited similar accuracy levels, the most stable results were obtained with the basic artificial neural network model, which did not include any sampling or spatial filtering (MAE: 0.03, RMSE: 0.29). The kernel-based numerical regression models, implemented with both a multilayer perceptron and an XGBoost regressor, achieved spatial continuity by representing aftershock intensities as a smoothed continuous field; however, their overall accuracy remained moderate, with the MLP configuration performing better (RMSE = 6.03, MAE = 2.19, $R^2 = 0.34$) than the corresponding XGBoost model. This demonstrates that predictions based solely on the static components of stress fields are inherently limited in their explanatory power, especially in high-variance and sparse regions.

Similarly, in regression models used for magnitude estimation, although the XGBoost based R-E-X-C model, enriched with neighborhood information, offers several advantages, the lowest error values were obtained with the artificial neural network model that uses only the principal stress components without any sampling or filtering (MAE: 0.08, RMSE: 0.52). This demonstrates that despite the structural advantages of XGBoost, the simple and straightforward artificial neural network model provides more successful results in magnitude estimation. However, it has been observed that even the best model fails to achieve high accuracy and has significant limitations, especially in the prediction of large-magnitude aftershocks.

Overall, this study demonstrates that static components derived from stress tensors have strong potential for predicting aftershock behavior. Kernel-based probabilistic modeling strategies contributed methodologically by more realistically representing aftershock spatial patterns. However, factors such as class imbalance, missing temporal components, and limited geological diversity continue to limit model accuracy. These findings indicate that stress-based, data-driven approaches offer a promising framework for aftershock prediction, but yield more robust and generalizable results when supported by comprehensive, time-sensitive, physically enriched datasets.

REFERENCES

1. SRCMOD, “Finite-Source Rupture Model Database”, 2012, <http://equake-rc.info/srcmod/>, accessed on November 5, 2025.
2. International Seismological Centre (ISC), “ISC-GEM Global Instrumental Earthquake Catalogue”, 1998, <https://www.isc.ac.uk/iscgem/>, accessed on November 5, 2025.
3. DeVries, P., F. Viégas, M. Wattenberg and B. Meade, “Deep Learning of Aftershock Patterns Following Large Earthquakes”, *Nature*, Vol. 560, No. 7720, pp. 632–634, 2018.
4. Liu, M. and S. Stein, “Earthquake, Aftershocks”, H. Gupta (Editor), *Encyclopedia of Solid Earth Geophysics*, pp. 192–194, Springer, Dordrecht, 2011.
5. King, G., R. Stein and J. Lin, “Static Stress Changes and the Triggering of Earthquakes”, *Bulletin of the Seismological Society of America*, Vol. 84, No. 3, pp. 935–953, 1994.
6. Toda, S., R. Stein, P. Reasenberg, J. Dieterich and A. Yoshida, “Stress Transferred by the 1995 Mw = 6.9 Kobe, Japan, Shock: Effect on Aftershocks and Future Earthquake Probabilities”, *Journal of Geophysical Research: Solid Earth*, Vol. 103, No. B10, pp. 24543–24565, 1998.
7. Kilb, D., J. Gomberg and P. Bodin, “Aftershock Triggering by Complete Coulomb Stress Changes”, *Journal of Geophysical Research: Solid Earth*, Vol. 107, No. B4, pp. ESE 2–1–ESE 2–15, 2002.
8. Papadimitriou, E., V. Karakostas and C. Papazachos, “Rupture Zones in the Area of the 17.08.99 Izmit (NW Turkey) Large Earthquake (Mw 7.4) and Stress Changes Caused by Its Generation”, *Tectonophysics*, Vol. 338, No. 3–4, pp. 217–234, 2001.

9. Toda, S., J. Lin and R. Stein, “Using the 2011 Mw 9.0 off the Pacific coast of Tohoku earthquake to test the Coulomb stress triggering hypothesis”, *Earth, Planets and Space*, Vol. 63, No. 1, pp. 725–730, 2011.
10. Sharma, S., S. Hainzl, G. Zöller and M. Holschneider, “Is Coulomb Stress the Best Choice for Aftershock Forecasting?”, *Journal of Geophysical Research: Solid Earth*, Vol. 125, No. 9, pp. 1–24, 2020.
11. Zheng, S., L. Qian, P. Li, C. He, X. Qin and X. Li, “An Introductory Review of Spiking Neural Network and Artificial Neural Network: From Biological Intelligence to Artificial Intelligence”, *Computer Science and Information Technology (CSIT Conference Series)*, pp. 12–145, Delhi, India, 2022.
12. Bishop, C., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
13. Goodfellow, I., Y. Bengio and A. Courville, *Deep Learning*, MIT Press, Cambridge, 2016.
14. Sengupta, S., S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, B. Ravi, M. Peters, M. Harris, P. Wang and A. Mitra, “A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends”, *Knowledge-Based Systems*, Vol. 194, No. 1, p. 105596, 2020.
15. Panakkat, A. and H. Adeli, “Neural Network Models for Earthquake Magnitude Prediction Using Multiple Seismicity Indicators”, *International Journal of Neural Systems*, Vol. 17, No. 1, pp. 13–33, 2007.
16. Karimzadeh, S., M. Matsuoka, J. Kuang and L. Ge, “Spatial Prediction of Aftershocks Triggered by a Major Earthquake: A Binary Machine Learning Perspective”, *ISPRS International Journal of Geo-Information*, Vol. 8, No. 10, p. 462, 2019.
17. Perol, T., M. Gharbi and M. Denolle, “Convolutional Neural Network for Earth-

- quake Detection and Location”, *Science Advances*, Vol. 4, No. 2, pp. 1–8, 2018.
18. Mousavi, S. M., W. Zhu, Y. Sheng and G. C. Beroza, “CRED: A Deep Residual Network of Convolutional and Recurrent Units for Earthquake Signal Detection”, *Scientific Reports*, Vol. 9, No. 1, pp. 10267–10277, 2019.
 19. Laurenti, L., E. Tinti, F. Galasso, L. Franco and C. Marone, “Deep learning for laboratory earthquake prediction and autoregressive forecasting of fault zone stress”, *Earth and Planetary Science Letters*, Vol. 598, No. 1, p. 117825, 2022.
 20. Yu, H. and R. Harrington, “Fluid-Injection-Induced Earthquakes Characterized by Hybrid-Frequency Waveforms Manifest the Transition from Aseismic to Seismic Slip”, *Nature Communications*, Vol. 12, No. 1, p. 6862, 2021.
 21. Hill, D., P. Reasenberg, A. Michael, W. Arabaz, G. Beroza, D. Brumbaugh, J. Brune, R. Castro, S. Davis, D. dePolo, W. Ellsworth, J. Gomberg, S. Harmsen, L. House, S. Jackson, M. Johnston, L. Jones, R. Keller, S. Malone, L. Munguia, S. Nava, J. Pechmann, A. Sanford, R. Simpson, R. Smith, M. Stark, M. Stickney, A. Vidal, S. Walter, V. Wong and J. Zollweg, “Seismicity Remotely Triggered by the Magnitude 7.3 Landers, California, Earthquake”, *Science*, Vol. 260, No. 5114, pp. 1617–1623, 1993.
 22. Gomberg, J., P. Reasenberg, P. Bodin and R. Harris, “Earthquake Triggering by Seismic Waves Following the Landers and Hector Mine Earthquakes”, *Nature*, Vol. 411, No. 6836, pp. 462–466, 2001.
 23. Utsu, T., Y. Ogata and R. Matsu’ura, “The Centenary of the Omori Formula for a Decay Law of Aftershock Activity”, *Journal of Physics of the Earth*, Vol. 43, No. 1, pp. 1–33, 1995.
 24. Båth, M., “Lateral Inhomogeneities in the Upper Mantle”, *Tectonophysics*, Vol. 2, No. 6, pp. 483–514, 1965.

25. Bach, C., D. Schorlemmer and S. Wiemer, “Improving Empirical Aftershock Modeling Based on Additional Source Information”, *Journal of Geophysical Research: Solid Earth*, Vol. 117, No. B4, pp. B05311–B05323, 2012.
26. Szabó, P., L. Tóth and J. Cerdà-Belmonte, “Hazard Model: Epidemic-Type Aftershock Sequence (ETAS) Model for Hungary”, *Applied Sciences*, Vol. 13, No. 5, pp. 2814–2842, 2023.
27. Freed, A., “Earthquake Triggering by Static, Dynamic, and Postseismic Stress Transfer”, *Annual Review of Earth and Planetary Sciences*, Vol. 33, No. 1, pp. 335–367, 2005.
28. Hardebeck, J. and E. Hauksson, “Crustal Stress Field in Southern California and Its Implications for Fault Mechanics”, *Journal of Geophysical Research: Solid Earth*, Vol. 106, No. B10, pp. 21859–21882, 2001.
29. Baranov, S., C. Narteau and P. Shebalin, “Modeling and Prediction of Aftershock Activity”, *Surveys in Geophysics*, Vol. 43, No. 2, pp. 437–481, 2022.
30. Mignan, A. and M. Broccardo, “One Neuron Versus Deep Learning in Aftershock Prediction”, *Nature*, Vol. 574, No. 7778, pp. E1–E3, 2019.
31. Mignan, A. and M. Broccardo, “A Deeper Look Into ‘Deep Learning of Aftershock Patterns Following Large Earthquakes’: Illustrating First Principles in Neural Network Physical Interpretability”, *Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, Proceedings, Part I*, pp. 3–14, Cham, 2019.
32. Chen, J., H. Tang and W. Chen, “Deep learning of the aftershock hysteresis effect based on the elastic dislocation theory”, *Natural Hazards and Earth System Sciences*, Vol. 20, No. 11, pp. 3117–3134, 2020.
33. Wang, Y., X. Li, Z. Wang and J. Liu, “Deep Learning for Magnitude Prediction

- in Earthquake Early Warning”, *Gondwana Research*, Vol. 123, No. 1, pp. 164–173, 2023.
34. Liu, B., H. Wen, M. Di, J. Huang, M. Liao, J. Yu and Y. Xiang, “Mapping and Interpretability of Aftershock Hazards Using Hybrid Machine Learning Algorithms”, *Journal of Rock Mechanics and Geotechnical Engineering*, Vol. 17, No. 8, pp. 4908–4932, 2025.
 35. Gulia, L. and S. Wiemer, “Real-Time Discrimination of Earthquake Foreshocks and Aftershocks”, *Nature*, Vol. 574, No. 7777, pp. 193–199, 2019.
 36. Verdecchia, A. and S. Carena, “Coulomb Stress Evolution in a Diffuse Plate Boundary: The Northern Walker Lane (California–Nevada Border)”, *Tectonophysics*, Vol. 677–678, No. 8, pp. 180–194, 2016.
 37. Han, J., M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2011.
 38. Jolliffe, I. and J. Cadima, “Principal Component Analysis: A Review and Recent Developments”, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 374, No. 2065, pp. 1–16, 2016.
 39. Montesinos López, O., A. Montesinos López and J. Crossa, “Fundamentals of Artificial Neural Networks and Deep Learning”, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pp. 379–425, Springer, Cham, 2022.
 40. LeCun, Y., Y. Bengio and G. Hinton, “Deep Learning”, *Nature*, Vol. 521, No. 7553, pp. 436–444, 2015.
 41. Ronneberger, O., P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015.

42. Çiçek, Ö., A. Abdulkadir, S. Lienkamp, T. Brox and O. Ronneberger, “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation”, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Cham, 2016.
43. Friedman, J., “Greedy Function Approximation: A Gradient Boosting Machine”, *Annals of Statistics*, Vol. 29, No. 5, pp. 1189–1232, 2001.
44. Chen, T. and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 2016.
45. Brownlee, J., *XGBoost With Python: Gradient Boosted Trees With XGBoost and Scikit-Learn*, Machine Learning Mastery, Victoria, 2016.
46. Powers, D. M. W., “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation”, *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37–63, 2011.
47. Sokolova, M. and G. Lapalme, “A Systematic Analysis of Performance Measures for Classification Tasks”, *Information Processing and Management*, Vol. 45, No. 4, pp. 427–437, 2009.
48. Chai, T. and R. Draxler, “Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments Against Avoiding RMSE in the Literature”, *Geoscientific Model Development*, Vol. 7, No. 3, pp. 1247–1250, 2014.
49. Willmott, C. and K. Matsuura, “Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance”, *Climate Research*, Vol. 30, No. 1, pp. 79–82, 2005.
50. Hyndman, R. and A. Koehler, “Another Look at Measures of Forecast Accuracy”, *International Journal of Forecasting*, Vol. 22, No. 4, pp. 679–688, 2006.

51. Montgomery, D., E. Peck and G. Vining, *Introduction to Linear Regression Analysis*, Wiley, Hoboken, 2012.



APPENDIX A: CODE REPOSITORY

All scripts used for data processing, model training, and result generation in this thesis are publicly available on GitHub at the following link:

<https://github.com/feyzanurtekbiyik/Deep-Larning-of-Aftershock-Thesis>

