Generating site-specific seismic motion attenuation relationship employing ML tools

Internship Report

by

Ankush Kumar Mandal (23B0675)

Nilesh Kumar Dhirhe (22B1541)

Rashi Jain (21D180036) - Team Leader

Tanvi Anand (23B4218)

Under the Supervision of

Soubhagya Karmakar

Saha Dauji

Indian Institute of Technology Bombay

2025

Report Approval

This report entitled **"Generating site-specific seismic motion attenuation relationship employing machine learning tools"** by **Ankush Kumar Mandal, Nilesh Kumar Dhirhe, Rashi Jain and Tanvi Anand** is hereby approved for the internship completion.

.....
Soubhagya Karmakar
(DCSEM, DAE, Mumbai, India)

.....
Saha Dauji
(Homi Bhabha National Institute, Mumbai, India)

Date: .....

Place: .....

Declaration

We declare that this written submission represents our ideas and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will cause disciplinary action by the Directorate and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

We declare that we have used artificial intelligence tools, including language models such as ChatGPT, to assist in writing code for experiments and rephrasing sections of this report. These tools were used responsibly to support our understanding, improve clarity, and enhance the presentation of our work.

Team Leader: Rashi Jain
Contribution:
- Introduction
- Literature Review
- Model Development

Nilesh Kumar Dhirhe
Contribution:
- Data Preprocessing
- Model Development
- Report Writing

Ankush Kumar Mandal
Contribution:
- Introduction
- Data processing
- Model Development
- Summary and conclusions

Tanvi Anand
Contribution:
- Introduction
- Literature Review
- Data Collection
- Mode Development

Name	Roll Number	Signature
Ankush Kumar Mandal	23B0675	
Nilesh Kumar Dhirhe	22B1541	
Rashi Jain	21D180036	
Tanvi Anand	23B4218	

Date: _____ Place: IIT Bombay

Acknowledgements

We are immensely grateful to supervisors Dauji Saha and Soubhagya Karmakar for their unwavering support, invaluable guidance, and motivation throughout our seminar work. Their enlightening ideas and clear directions played a significant role in the successful completion of our project.

Rashi Jain(Team Leader), Tanvi Anand,
Ankush Kumar Mandal, and Nilesh Kumar Dhirhe

Abstract

Accurate modeling of seismic ground motion attenuation is critical for site-specific seismic hazard assessment, particularly in regions with complex geological conditions. Traditional Ground Motion Prediction Equations (GMPEs), typically built on empirical regression, often fail to capture the non-linear and heterogeneous nature of seismic wave propagation across different sites. In this study, we propose a data-driven framework for generating site-specific seismic attenuation relationships using machine learning (ML) models. Utilizing over 275,000 ground motion records from the KiK-net database in Japan, we integrate key geological and seismological features—such as moment magnitude, rupture distance, focal depth, Vs30, fault type, and tectonic regime—to predict peak ground acceleration (PGA).

We benchmark four ML algorithms: Linear Regression, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Deep Neural Networks (DNN). Each model is trained and evaluated using 3-fold cross-validation with standardized and log-transformed features, and performance is assessed using R^2, RMSE, and MAE. The DNN model, despite moderate R^2 (0.54), exhibits robust generalization with low MAE (0.0479g), while XGBoost achieves superior overall performance ($R^2 = 0.64$, MAE = 0.0463g). Random Forest demonstrates the best bias-variance tradeoff but shows signs of overfitting, and Linear Regression underperforms due to its inability to capture non-linear site interactions.

We further apply SHAP-based interpretation and classification analysis (Low/Moderate/High PGA bins) to validate the physical relevance and predictive robustness of the models. Classification metrics indicate high precision for low-intensity events but expose challenges in capturing high-PGA occurrences due to class imbalance. Feature importance analysis confirms rupture distance and magnitude as the most influential predictors.

Our results highlight the potential of ML models—especially ensemble-based approaches like XGBoost—for building adaptive, interpretable, and physically consistent attenuation relationships. The proposed framework has significant implications for enhancing seismic design codes and hazard mitigation strategies in data-scarce and geologically complex regions. The code for reproducing the results are provide in this link: project-github

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The problem Addressed

Seismic hazard assessment fundamentally relies on accurate ground motion prediction equations (GMPEs), which have traditionally been developed using empirical regression techniques applied to recorded earthquake data. These conventional attenuation relationships often fail to capture the complex, non-linear interactions between seismic waves and local site conditions, resulting in significant uncertainties in hazard estimates (Banerjee, 2016). The propagation of seismic waves exhibits highly variable behavior due to regional geological heterogeneities, with waves attenuating non-uniformly in different directions based on the elastic properties of the propagation medium. This variability manifests in the three primary mechanisms of attenuation: geometrical spreading, scattering of seismic energy, and inelasticity of the propagation medium, which collectively contribute to site-specific amplification patterns that traditional models struggle to represent accurately.

1.2 Background & Context

The limitations of conventional GMPEs have significant implications for structural safety and regional planning, particularly in areas with complex geology or limited historical seismic data. Recent studies have demonstrated that tree-based machine learning algorithms like XGBoost and Random Forest consistently outperform traditional regression methods in predicting peak ground acceleration (PGA) and velocity (PGV), achieving accuracy rates exceeding 98 compared to conventional approaches Xiangqi Wang and Li (2023). The challenge lies in effectively incorporating the spatial heterogeneity of geological features, the non-linear behavior of soil during strong ground motion, and the complex

wave propagation phenomena into predictive models. Machine learning approaches offer promising solutions by identifying intricate patterns in high-dimensional seismic data without requiring explicit physical formulations, potentially revolutionizing site-specific hazard assessment in regions where traditional methods have proven inadequate.

This study develops a novel framework for site-specific seismic attenuation relationships using advanced machine learning techniques that adapt to local geological conditions. The approach integrates spatial data analysis with non-linear algorithms and incorporates mutual information-based feature selection and ensemble learning methods, which have demonstrated superior performance in recent seismic parameter prediction studies Anjom et al. (2024). More accurate site-specific attenuation relationships directly improve seismic design codes and risk mitigation strategies Xiangqi Wang and Li (2023). In regions with limited strong-motion recordings, these machine learning models can reduce the epistemic uncertainty that has traditionally affected seismic hazard assessments (Banerjee, 2016).

1.3 Challenges

Developing a site-specific seismic attenuation model is both scientifically interesting and technically challenging. The seismic response of a site is influenced by numerous nonlinear, uncertain, and interrelated parameters, such as input rock motion, local stratigraphy, shear wave velocity profiles, and damping characteristics (Frankel et al. (1996); McGuire and Toro (2008)).

Incorporating these uncertainties and relationships through traditional empirical models is complex and computationally expensive. Machine learning (ML) offers an efficient alternative by enabling models to learn complex, nonlinear mappings from data without relying on rigid assumptions. However, training effective ML models requires reliable data, robust validation, and careful interpretation to ensure physical plausibility and generalization.

1.4 Study Objective

The study aims to develop and evaluate machine learning-based models for predicting seismic motion attenuation parameters such as peak ground acceleration (PGA), peak ground velocity (PGV) on a site specific basis. The models will be developed using observed ground motion data from the KiK-net database in Japan, covering seismic events recorded between 1997 and 2017, a widely used source in recent ground motion modeling studies due to its high-quality paired surface and borehole recordings (Wu et al. (2023); Li and Gao (2024)). The predictive framework will incorporate key features including moment magnitude (Mw), epicentral distance, earthquake source depth, bedrock ground

motion, average shear wave velocity (Vs30), which have been identified as influential variables in both traditional GMPEs and recent machine learning-based studies Díaz et al. (2022). Four machine learning algorithms of Linear Regression, Random Forest (RF), XGBoost, and Deep Neural Networks (DNN) were will be trained and compared using standard performance metrics such as coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) which are commonly used in evaluating seismic prediction models Xiangqi Wang and Li (2023).

1.5 Relevance and Scope of the Report

This report is highly relevant for improving ground motion prediction by incorporating local site conditions. Traditional methods often fail to capture complex, site-dependent factors such as soft soils or varying fault distances Bhusal (2022). Machine learning techniques offer a powerful alternative, as they can model non-linear relationships without strict assumptions Hibert et al. (2024). Recent studies have shown that ML algorithms like XGBoost and Random Forest improve the accuracy and generalization of seismic parameter predictions Li and Gao (2024). This makes ML-based attenuation modeling valuable for seismic hazard assessments in high-risk or heritage-sensitive zones.

Chapter 2

Literature Review

2.1 Limitations of Generalized Attenuation Models

Traditional Probabilistic Seismic Hazard Analysis (PSHA) methods initially relied on deterministic ground motion prediction equations (GMPEs) based on physics-driven models and site classification. These methods incorporated site amplification factors applied to uniform hazard spectra on reference rock conditions, but required localized adjustment for soil behavior, which is often oversimplified or unavailable (James Kaklamanos and Booreb, 2011).

GMPEs frequently rely on VS30 (average shear-wave velocity in the upper 30 meters) as the primary site descriptor. However, this generalization does not fully capture the complex nonlinear behavior of soil layers or local geological features like stratigraphy and topography (Kamai et al. (2014); Shingaki et al. (2018)). As a result, there is a high degree of epistemic uncertainty in attenuation predictions in regions with limited instrumental data or non-standard site conditions.

Furthermore, regression-based models introduce bias due to region-specific earthquake records and limited spatial distribution of seismic stations, leading to underrepresentation of critical site effects Chiou and Youngs (2008). These models often neglect site-specific parameters (Dawood et al. (2011)). Therefore, ground motion variability due to site amplification remains underexplored in classical attenuation relationships.

2.2 Site-Specific Hazard Analysis (SSHA)

SSHA offers an improved framework over traditional PSHA by explicitly integrating local site response effects into seismic hazard quantification. It involves modifying hazard estimates on rock by incorporating amplification functions derived from detailed geotechnical

and geophysical site characterization (Boore (2004); (Kramer, 1996)). The modeling of vertically propagating shear waves in horizontally layered soil (1D site response) remains a common approach due to its simplicity and wide applicability.

The response analysis typically uses tools such as equivalent linear analysis (e.g., SHAKE software) or stochastic approaches like Random Vibration Theory (RVT) to estimate amplification across frequency bands (Rathje et al. (2010)). Parameters such as VS30, dynamic soil properties (G/Gmax, damping vs. strain curves), and layer-wise stiffness profiles are key inputs. However, these methods are data-intensive and demand detailed subsurface information—something that is often not feasible in many seismic regions.

Recent advances suggest that probabilistic site response analysis, incorporating uncertainties in both input motions and geotechnical properties (e.g., stratigraphy and Vs profiles), yields more realistic results (Dawood et al. (2011)).

2.3 Machine Learning in Seismology and Geotechnical Studies

The application of machine learning (ML) in seismology is growing rapidly. One of the most active areas is seismic signal classification and enhancement. ML has been used for noise filtering, dispersion curve picking, and microseismic signal identification (Zhu et al. (2019)).

ML has also shown promising results in predicting ground motion parameters such as PGA, PGV, PGD, and PSA from site-specific data. These models are trained on historical records and site characteristics and have been able to predict parameters with improved accuracy compared to classical GMPEs (Weatherill et al. (2024); Kotha and Traversa (2024)). For instance, ML-based models have been used to develop ground motion intensity maps with better resolution and local variability than traditional interpolation approaches.

In site response studies, ML models have demonstrated strong potential in predicting amplification factors directly from input motions and geotechnical properties Park and Lee (2024). These models often outperform physics-based approaches in both accuracy and generalizability, especially in the absence of complete subsurface data.

Additionally, ML is increasingly applied in subsurface characterization, estimating parameters such as shear strength, stratification, and seismic facies from borehole and geophysical data Sun et al. (2023). Neural networks have also been explored for full waveform inversion and structural classification Sun et al. (2023).

2.4 Common ML Techniques Used:

- Artificial Neural Networks (ANNs): Widely used for predicting ground motion parameters and identifying seismic signals. ANNs are effective in handling nonlinearities and can adapt to complex seismic datasets Saleem et al. (2024).
- XGBoost: This gradient boosting framework has become popular due to its high accuracy and interpretability. It has shown excellent performance in seismic duration prediction, outperforming several other models Li and Gao (2024).
- Random Forests: Extensively used for seismic risk analysis and prediction of dynamic soil parameters. Its ensemble nature provides robustness and reduces overfitting Mori et al. (2022).
- Gaussian Process Regression (GPR): Suitable for quantifying uncertainties and capturing heteroscedastic relationships in ground motion prediction. GPR has been applied for fragility curve development and probabilistic demand modeling Anjom et al. (2024).

2.5 Integration of ML with SSHA:

ML models can enhance SSHA by learning nonlinear site amplification behavior directly from data, reducing reliance on explicit site-response simulations. They allow for the prediction of amplification factors based on sparse and heterogeneous datasets, improving spatial coverage in regions with limited instrumentation Ortega (2024).

In recent work, ML models have also been used to predict spatially continuous site parameters like VS30 and fundamental frequency (f_0), which are critical for site-specific analysis Mori et al. (2022). These advances allow ML to serve as a bridge between detailed site response modeling and regional hazard analysis.

2.6 Identified Gaps in Literature:

Despite its potential, the use of ML in seismic hazard assessment still faces challenges:

- Limited Regional Applications: There is a lack of ML-based seismic studies in regions like Chile, where data sparsity hampers accurate model generalization Sun et al. (2023).
- Interpretability Issues: Many ML models, especially neural networks, act as "black boxes," making it difficult to understand how predictions are made. Techniques like SHAP are being explored to explain model behavior Mori et al. (2022).

- Data Limitations: Sparse sensor networks and limited ground motion records make it hard for ML models to generalize, especially in data-scarce regions. This often leads to overfitting Anjom et al. (2024).
- Lack of ML in Complex Soil-Structure Interactions: While ML is widely used for site parameter estimation, its application in detailed soil-structure seismic response modeling is still limited. More research is needed on failure mechanisms, nonlinearity, and coupled dynamic behavior Banerjee (2016).
- Narrow Geographic Scope: Many models are developed using datasets from a single country (e.g., Japan) and may not generalize well across different tectonic or geological environments. This calls for globally trained and validated models Li and Gao (2024).

Chapter 3

Data And Methods

3.1 Data

In this study, we utilized the updated KiK-net database developed by Bahrampouri et al. (2020), which builds upon the earlier work by Dawood et al. (2011). The ground motions were processed using a fully automated algorithm designed to ensure consistency and improve the quality of recordings. The processing steps include baseline correction, record tapering, zero padding, filtering, testing a high-pass corner frequency, iterative search for a suitable high-pass corner frequency, and finally, signal-to-noise ratio check). The database includes information of 6205 earthquakes from 8 October 1997 to 31 December 2017, yielding a comprehensive dataset of over 350,000 earthquake records. The database includes the latitude and longitude of the epicenter, Japanese Meteorological Agency (JMA) magnitude, moment magnitude, focal depth, focal mechanism. The database also includes observation station information such as station code, latitude and longitude of the station, time-averaged shear-wave velocities of soil deposits at particular depths (VSZ, where Z = 0, 5, 10, 20, 50, 100) and National Earthquake Hazards Reduction Program (NEHRP) site classes. In addition, the database presents various types of source-to-site distances, such as epicentral distance, rupture distance, horizontal distance to the fault, hypocentral distance as well as Joyner-Boore distance.

3.2 Data Distribution

Figure 3.1 shows the location distribution of the earthquake source and observation station for the dataset selected in the study. Previous research Bommer and Martínez-Pereira (2000) have shown that only ground motions with PGA more than 20 cm/s2 will have impact on our society, engineering structures and living environment. The PGA distribution of the data collected for this study is shown in Figure 3.2A. It illustrates that PGA for

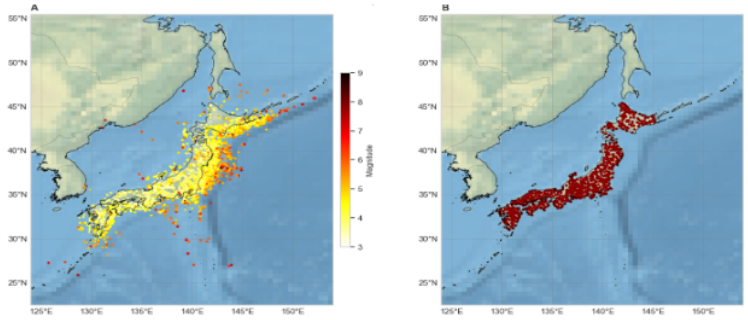

FIGURE 3.1: Location distribution of earthquake source and observation station. (A) Earthquake source station, (B) Observation station location.

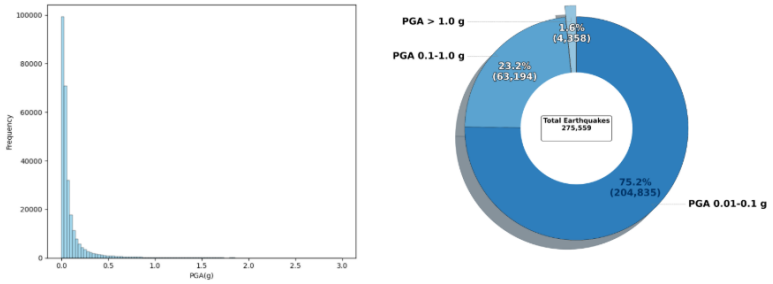

FIGURE 3.2: Record frequency distribution with PGA (A) Frequency distribution of PGA values in g, (B) Proportional distribution of PGA categories.

most records is below 0.5g. Figure 3.2B shows the distribution of PGA values, revealing that the majority of records (75.2%) fall within the low intensity range(0.01-0.1g) and a smaller proportion (23.2%) corresponds to moderate shaking (0.1-0.2g), while extreme events (>1.0g) are rare, representing only 1.6% observations.

Figure 3.3 shows the histogram distributions of magnitude, rupture distance, VS30 and depth. As can be seen VS30 mostly falls into the range of 250-700 m/s, and magnitude is mostly between 4 and 7, while r(rupture) is mostly less than 300 km and the source depth is mostly shallower than 70 km. Figure 3.4 shows the proportion distribution of magnitude as well as the magnitude-distance distribution. There is good coverage for magnitude and distance; however the majority of the data have magnitude less than 6.0.

3.3 Data Preprocessing

For this study, we selected 275,559 ground motion records from 697 stations, with a magnitude of 3.0 or higher and a depth of 0-620 km. Among these, the vast majority (87.1%) are from shallow earthquakes (0–70 km), followed by 11.2% from intermediate-depth

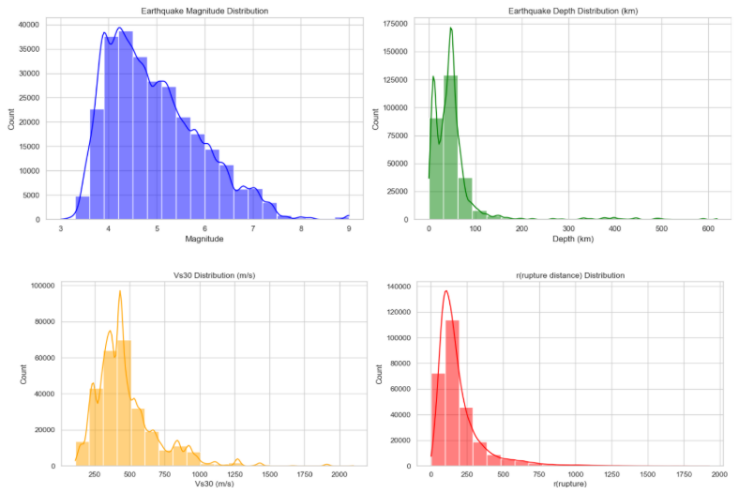

FIGURE 3.3: Histogram distributions of magnitude, rupture distance, VS30 and depth.

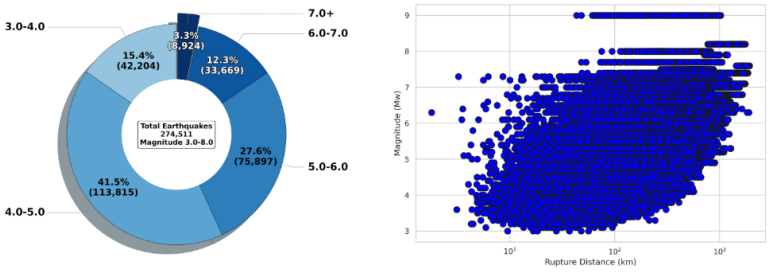

FIGURE 3.4: (A) Earthquake Magnitude Distribution (B) Magnitude versus Rupture distance (Km)

events (70–300 km) and only 1.6% from deep-focus earthquakes (>300 km). The epicentral distance(R) can be calculated via the latitude and longitude of the earthquake source and observation stations. Arias Intensity (Ia) for both surface and borehole motions is calculated as in Eq.1

$$Ia(T) = \sqrt{Ia(T)_{EW} \cdot Ia(T)_{NS}} \tag{3.1}$$

Where Ia(T)_EW, Ia(T)_NS are the arias intensity of the EW and NS component for period T.

There are numerous parameters recorded during an earthquake. To simplify the model, a selection of parameters is made based on findings from previous studies in this area. Additionally, a correlation matrix is presented to illustrate the relevance of each parameter to the observed PGA, helping to identify the most influential factors for prediction (Figure 3.5). Due to strong multicollinearity observed between r(rupture), r(epicenter) and r(hypocentral) as well as between ztor and Depth(Km), we retain only r(rupture) and

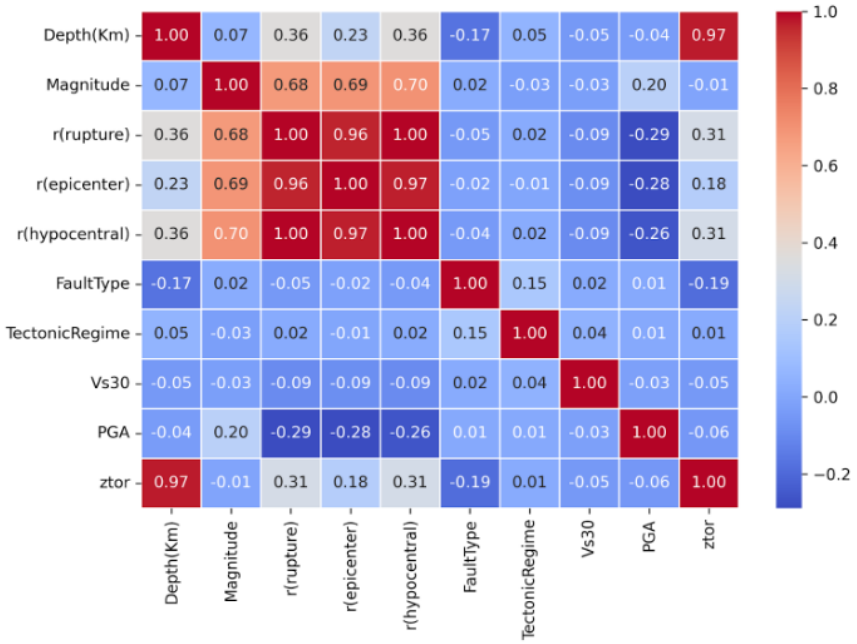

FIGURE 3.5: Correlation matrix of features and target parameter after transformation and standardization

Depth(Km) for further analysis. The input parameters chosen for the machine learning models include: Depth, Magnitude, Vs30, Rupture distance, Fault type, Tectonic regime.

Feature Engineering in this study involves the transformation and scaling of input data to optimize model performance and improve predictive accuracy. Preliminary data analysis revealed that the logarithmic transformation of certain input variables significantly improves model convergence and performance, particularly for variables that exhibit strong skewness. The input variables r(rupture), Depth(Km) and Vs30, along with the output PGA are transformed using the natural logarithm. This transformation addresses the positive skewness in these variables, making the distribution more suitable for machine learning. In addition to this, most machine learning algorithms are sensitive to the scale of numerical features, all parameters are standardized to have a mean of 0 and a standard deviation of 1. Without standardization, features with larger numerical ranges would dominate the model training, leading to biased or inefficient learning.

3.4 Evaluation Index

In this study, three evaluation indexes are used, namely variance (R^2), root mean square error (RMSE) and mean absolute error (MAE) to assess the effectiveness of the models.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{3.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.4}$$

Where n represents the sample size, y_i represents the actual observed value, $\hat{y}_i$ represents the predicted value and $\bar{y}$ represents the mean of the actual observed value.

The R^2 value ranges from 0 to 1 with a value closer to 1 indicating better prediction performance of the model. RMSE represents the square root of the average of the squared error between the predicted and actual values. MAE measures the average absolute error between the predicted and actual values. In comparison with other evaluation index such as RMSE, MAE is more robust to outliers.

3.5 Methodology:

- **Neural Network:** We employed a fully-connected feedforward Artificial Neural Network (ANN) consisting of three hidden layers with 128, 64, and 32 neurons, respectively. Each hidden layer uses the ReLU activation function to introduce non-linearity. To stabilize training and accelerate convergence, Batch Normalization is applied after the first and second hidden layers. Dropout regularization is used after each hidden layer with dropout rates of 0.3, 0.2, and 0.1 respectively, to reduce overfitting. The output layer consists of a single neuron with a linear activation, suitable for continuous-valued regression (PGA prediction). The network was trained using the Adam optimizer with a learning rate of 1e-3, using Mean Squared Error (MSE) as the loss function, and Mean Absolute Error (MAE) as an evaluation metric.

  To improve numerical stability and model sensitivity in the presence of highly skewed target values, the target variable PGA (Peak Ground Acceleration) was log-transformed using log1p before training. Predictions were later exponentiated back using expm1 for interpretability and evaluation.

  The model's performance was evaluated using 3-fold cross-validation, with each fold consisting of an 80/20 training-validation split. Prior to training, all input

features were standardized using z-score normalization (StandardScaler), and the target variable was log-transformed as described above. Model weights were initialized randomly, and training was conducted for up to 100 epochs using a batch size of 512. To prevent overfitting, early stopping with a patience of 10 epochs was applied, restoring the best weights based on validation MAE.

For each fold, we computed the following metrics on both the training and test sets: R^2 score (coefficient of determination), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Additionally, training curves of MAE vs. epochs were plotted for each fold to assess convergence behavior and generalization.

TABLE 3.1: Cross-Validation Performance Metrics for ANN Model

Dataset	**R^2 (mean $\pm$ std)**	**MAE (mean $\pm$ std)**	**RMSE (mean $\pm$ std)**
Train	0.5871 ± 0.0243	0.0474 ± 0.0005	0.2271 ± 0.0075
Test	0.5648 ± 0.0724	0.0478 ± 0.0008	0.2328 ± 0.0256

- **XGBoost:** XG boost is a powerful ensemble learning method that is based on gradient boosting decision trees. It was proposed by Chen and Guestrin Chen and Guestrin (2016). It builds upon the concept of sequential boosting. Here, each new tree is trained to correct the residual errors made by the ensemble of decision trees built before.

  In regression problems, the prediction output of XGBoost is the sum of the outputs from all individual trees. Mathematically, the ensemble model prediction can be expressed as: $F(x) = \sum_{i=1}^{n} f^i(x)$, where $F(x)$ is the final prediction, $f^i(x)$ is the prediction from the ith regression tree and n is the number of trees in the ensemble.

  Each successive tree in XGBoost is built to minimize a regularized objective function, which includes a loss function like squared error for regression and a regularization term that penalizes model complexity. This mechanism prevents overfitting and ensures model generalization.

  The model's algorithm uses gradient descent to optimize the model. It minimizes the first-order and second-order Taylor expansion of the loss function with respect to the model predictions. The tree splits are determined by selecting features and thresholds that maximize gain, i.e., the reduction in the regularized loss function.

  In this study, the XGBoost model was developed after several tuning iterations. It had 100 rounds of boosting with a maximum tree depth of 5 and learning rate or shrinkage factor of 0.1. The objective function used was square error loss for regression. The regularization parameters were set to default, no additional penalization was applied.

  Before training, the dataset was subject to scaling using standard normalisation and low-impact features were excluded based on domain knowledge.

- **Random Forest:** Random Forest (RF) algorithm is a classification prediction approach based on multiple decision trees by Breiman (2001), and it is an extension of decision tree algorithm.

  For classification problems, each tree outputs a classification result, and the random forest classification result is the mode of the classification results from each tree. For regression problems, each tree outputs a regression result, and the random forest regression result is the average of the regression results of all trees.

  $F(x) = mean(\sum_i^n f^i(x)$, where $F(x)$ is the output result, $f^i(x)$ is the result for tree i, and n is the number of trees in the forest.

  The RF algorithm traverses all nodes (trees) to be split, identifies the optimum split variable and its corresponding split threshold for maximum impurity reduction for all sub-nodes.The above process is repeated until the threshold value requirement is satisfied, thus resulting in generation of a forest of trees.

  In this study, a random forest model was established using default values of 100 base learners, a seed value of 42 for the random number generator, a maximum tree depth of 20, a minimum leaf node size of 1 and minimum samples split of 5.

- **Linear Regression:** Linear Regression is a fundamental statistical approach used for modeling the linear relationship between a dependent variable and one or more independent variables. It assumes that the change in the dependent variable (PGA in this case) can be explained as a weighted sum of the input features.

  The prediction function is expressed as $F(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$, where F(x) is the predicted output, x are the input features, and are the coefficients estimated by minimizing the residual sum of squares.

  In this study, this model was developed using the selected features. Before training, the data set was processed using feature scaling using standard normalization, and class imbalance in PGA values was addressed using up sampling.

  The model was evaluated using 5-fold cross-validation, and performance was assessed based on R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Diagnostic visualizations such as residual distribution, actual vs predicted scatter plot, correlation matrix, and confusion matrix (for binned PGA classes) were also employed to interpret and validate the model's effectiveness.

K-Fold Cross-Validation: To ensure robust estimation of model generalization performance, we employed K-fold cross-validation with K=3. This approach partitions the dataset into three equal-sized subsets, iteratively using two folds for training and one for testing. K-fold cross-validation helps mitigate the risk of overfitting to a single train/test split and provides more reliable performance estimates, especially in the presence of data heterogeneity. Metrics such as R^2, RMSE, and MAE were aggregated across folds to report average performance and variance.

Class Imbalance and Upsampling Strategy: The dataset exhibited a strong imbalance in the distribution of PGA values, with approximately 1.5% of the samples classified as high-PGA events ($PGA1.0g$). To address this issue, we initially experimented with upsampling the minority class using random replication to balance the training data distribution. Although effective in improving model sensitivity toward rare high-PGA samples, this approach introduced potential risks of overfitting. Therefore, subsequent experiments focused on using custom loss functions that inherently penalize errors on underrepresented PGA regions, allowing the model to learn without artificially modifying the data distribution.

Confusion Matrix and Discrete Error Analysis: Although the primary task is regression, we conducted post-hoc discretization of predicted and actual PGA values into ordinal classes (e.g., Low, Moderate, High) to analyze prediction reliability across different seismic intensity levels. The confusion matrix was used as a diagnostic tool to identify systematic biases in the model—particularly its tendency to underpredict high-PGA events. Derived metrics such as accuracy, precision, recall, and F1-score provided class-specific insights that complemented traditional regression metrics. This hybrid evaluation allowed us to better assess the model's practical applicability in seismic risk assessment contexts.

Unified Rank-Based Model Comparison: To enable fair and interpretable comparison across multiple machine learning models, we adopted a Unified Ranking strategy that aggregates performance across multiple evaluation metrics. Given the multidimensional nature of model performance—where different models may excel in different aspects (e.g., accuracy, robustness, error minimization)—traditional single-metric comparisons can be misleading or incomplete.

In our case, each model was evaluated using a set of regression metrics (R^2, RMSE, MAE) and classification-derived metrics (Accuracy, F1-score) based on discretized PGA levels. For each metric, models were ranked independently, with higher-is-better metrics (e.g., R^2, Accuracy, F1-score) ranked in descending order, and lower-is-better metrics (e.g., RMSE, MAE) in ascending order. The total rank for each model was computed as the sum of its ranks across all metrics, providing a scalar value that captures overall relative performance. The unified rank method enables fair model comparison by aggregating multiple performance metrics into a single interpretable score, helping to identify consistently strong models without bias toward any one metric.

float

Chapter 4

Results And Discussion

4.1 Neural Network model

Training and Validation Dynamics

The training progression of the neural network is depicted in Figure 4.1, which shows the Mean Absolute Error (MAE) across epochs for both the training and validation datasets. The model exhibits rapid convergence within the first few epochs and stabilizes after approximately 20 epochs. The use of early stopping based on validation MAE effectively prevented overfitting. Notably, the validation error consistently remained lower than the training error, indicating good generalization.

Regression Performance

To evaluate the generalization capability of the proposed Artificial Neural Network (ANN) model, a 3-fold cross-validation strategy was adopted. The performance was assessed using three key regression metrics: the coefficient of determination (R^2), Mean Absolute

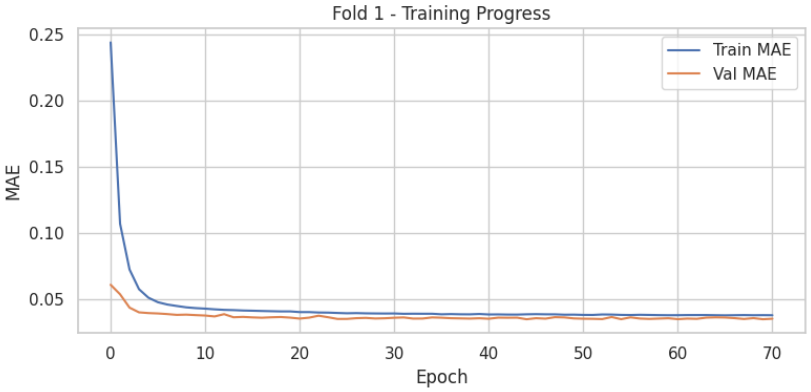

FIGURE 4.1: Training and Validation MAE across epochs for Fold 1

TABLE 4.1: Cross-validation performance metrics for regression

Metric	**Train (Mean ± SD)**	**Test (Mean ± SD)**
R^2 Score	0.5735 ± 0.0485	0.5444 ± 0.0337
MAE (g)	0.0474 ± 0.0011	0.0479 ± 0.0003
RMSE (g)	0.2307 ± 0.0160	0.2386 ± 0.0154

TABLE 4.2: Overall regression performance metrics across all folds

Metric	**Value**
R^2 Score	0.5426
MAE (g)	0.0479
RMSE (g)	0.2391

Error (MAE), and Root Mean Square Error (RMSE). As shown in Table 4.1, the model achieved an average training performance of $R^2 = 0.5735 \pm 0.0485$, MAE $= 0.0474 \pm 0.0011$g, and RMSE $= 0.2307 \pm 0.0160$g. The test results, while slightly lower, were consistent, with $R^2 = 0.5444 \pm 0.0337$, MAE $= 0.0479 \pm 0.0003$g, and RMSE $= 0.2386 \pm 0.0154$g. The closeness of these metrics across training and testing indicates that the model generalizes well without significant overfitting.

Moreover, when aggregating predictions across all folds, the overall performance metrics reached $R^2 = 0.5426$, MAE $= 0.0479$g, and RMSE $= 0.2391$g, as summarized in Table 4.2. These results suggest that the model maintains high accuracy and stability in PGA prediction. The relatively small difference between MAE and RMSE implies limited impact from outlier predictions. However, the moderate R^2 values reflect the inherent challenge of capturing complex seismic phenomena—especially under class imbalance where high-PGA events are underrepresented.

The R^2 values indicate that the model captures more than 54% of the variance in the test data. The MAE remains consistent across folds, reflecting robustness. However, RMSE and residual analysis suggest larger errors on high-magnitude events.

Actual vs. Predicted PGA

Figure 4.2 illustrates the actual versus predicted PGA values across all folds. While the model shows strong agreement for low and moderate PGA values, there is consistent underprediction in high PGA regions. The spread of points increases as the actual PGA values grow larger, revealing increased prediction error for high-magnitude events. This

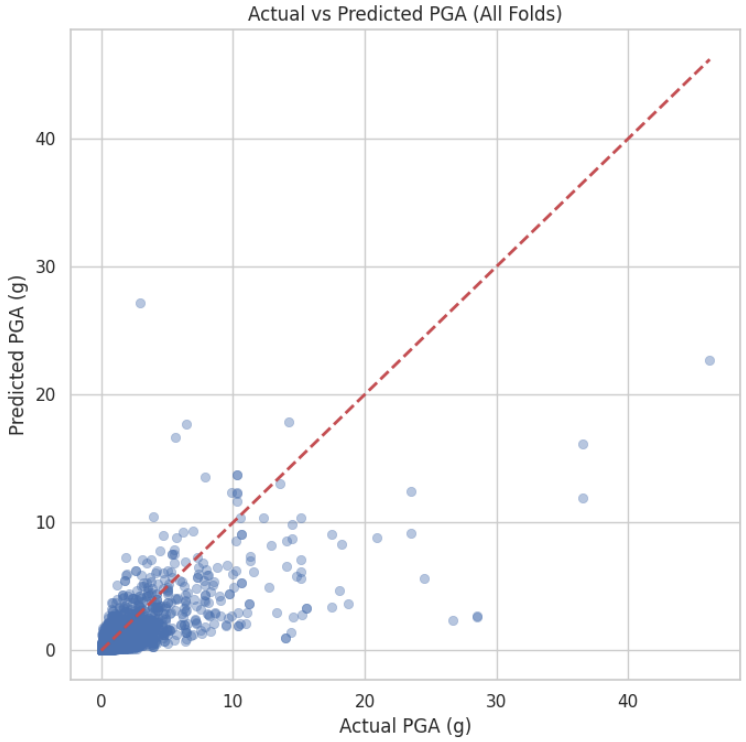

FIGURE 4.2: Actual vs. Predicted PGA values across all folds

TABLE 4.3: Residual statistics after ANN prediction

Statistic	**Value**
Mean	0.0098
Standard Deviation	0.2389
Skewness	31.7650
Kurtosis	4161.0838

dispersion highlights the model's tendency to underpredict higher PGA values, which is consistent with the residual distribution and classification recall for the "High" class.

Residual Analysis

The residual analysis reveals that the prediction errors are nearly unbiased on average, as indicated by the mean close to zero (0.0098). However, the residual plot in Figure 4.3 and the statistics below reveal a highly skewed residual distribution, with a concentration of small errors and some extreme outliers. The relatively high standard deviation (0.2389) further suggests variability in the model's predictions across different PGA ranges. Residual Statistics: Table 4.3

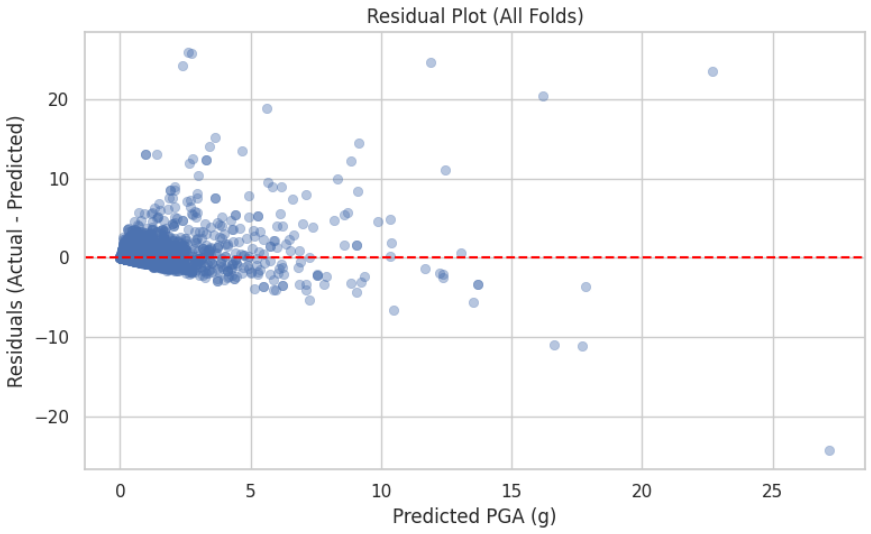

FIGURE 4.3: Residual plot (Actual - Predicted PGA) across all folds

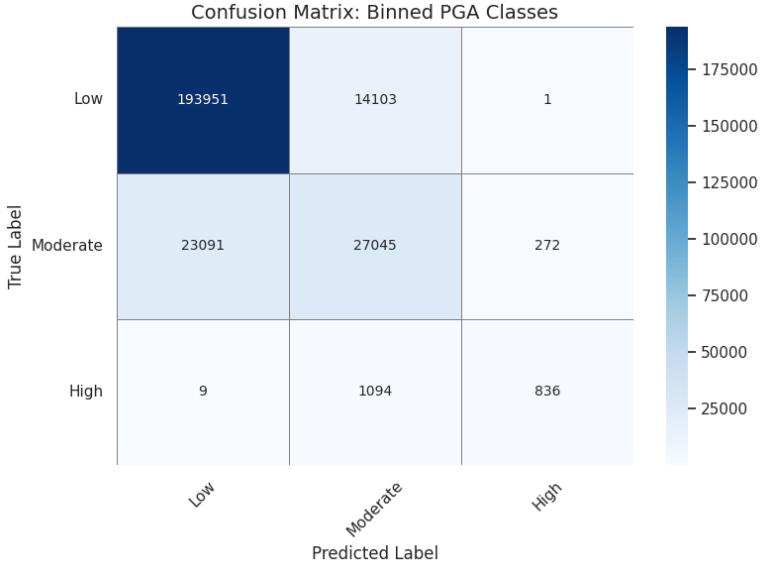

FIGURE 4.4: Confusion Matrix: Binned PGA Classes

Classification-Based Evaluation

To further analyze prediction behavior, PGA values were binned into **Low (0–0.1g)**, **Moderate (0.1–1.0g)**, and **High (>1.0g)** classes. The resulting confusion matrix is presented in Figure 4.4, with detailed precision, recall, and F1-scores provided in Table 4.4.

The classification metrics, based on binned PGA levels, demonstrate the model's strong ability to identify low-intensity ground motion events, achieving a high precision (0.89), recall (0.93), and F1-score (0.91) in the "Low" class. The "Moderate" class is captured

TABLE 4.4: Classification metrics based on binned PGA values

Class	Precision	Recall	F1-score	Support
Low	0.89	0.93	0.91	208056
Moderate	0.64	0.54	0.58	50408
High	0.75	0.43	0.55	1939
Accuracy		**0.8519**		
Macro Avg	0.76	0.63	0.68	260403
Weighted Avg	0.84	0.85	0.85	260403

with moderate precision and recall (F1-score: 0.58), whereas the "High" class remains the most challenging to predict accurately, with a recall of 43% despite a reasonable precision of 0.75. This imbalance in performance across classes is reflected in the lower macro-averaged F1-score (0.68) compared to the weighted average (0.85), highlighting the dominant influence of the majority class. The overall classification accuracy stands at a strong 85.19%, affirming the model's robust performance on common events but also underscoring the need for further refinement in detecting rare, high-impact PGA events.

Summary

The model performs well for the dominant low-PGA class with high accuracy and low MAE. However, it tends to underpredict high PGA values due to data imbalance. Residuals show high skewness and kurtosis, indicating difficulty in capturing rare high-magnitude events. Classification metrics further highlight the need to improve recall for the high-PGA class.

4.2 Linear Regression Model

Training Dynamics: The linear regression model demonstrated rapid convergence with minimal computational overhead, as expected for a parametric model with closed-form solution. Model fitting was straightforward without the need for hyperparameter tuning. The training and testing scores remained stable across 3-fold cross-validation, indicating consistent performance and minimal overfitting. The model demonstrated moderate predictive capability with limited variance capture, as evident in Table 4.5. The mean R^2 score across test folds is approximately 0.46, indicating that the model explains 46% of the variance in the test set. MAE and RMSE values remain relatively low and consistent across folds, suggesting stable but constrained performance due to the model's linear assumptions.

TABLE 4.5: Cross-validation performance metrics for regression

Metric	**Train (Mean ± SD)**	**Test (Mean ± SD)**
R^2 Score	0.4872 ± 0.0125	0.4631 ± 0.0354
MAE (g)	0.0571 ± 0.0012	0.0584 ± 0.0021
RMSE (g)	0.2668 ± 0.0083	0.2732 ± 0.0176

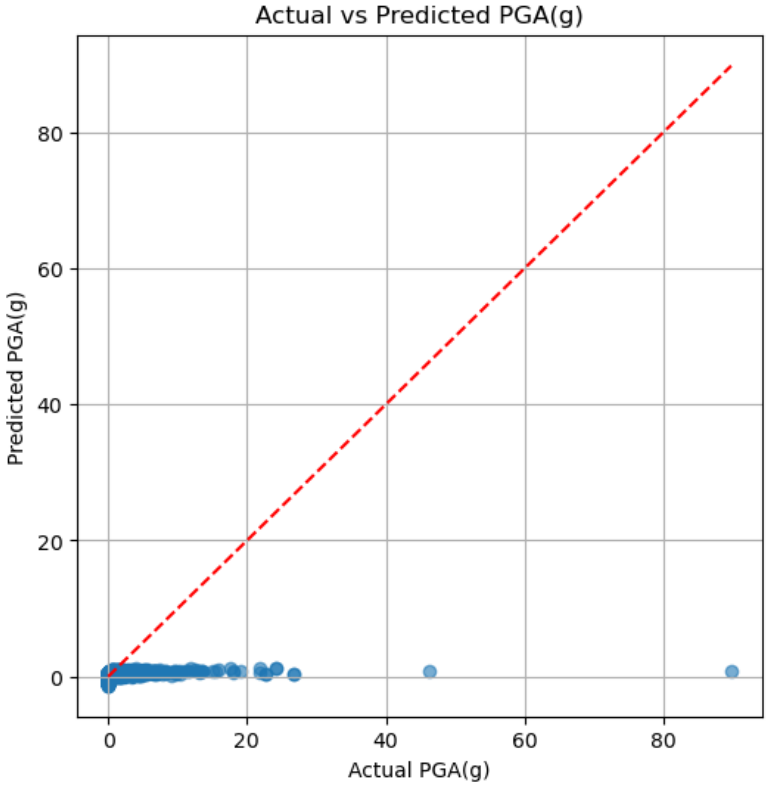

FIGURE 4.5: Actual vs. Predicted PGA values across all folds

Actual vs. Predicted PGA The scatter plot (Figure 4.5) reveals a strong alignment between actual and predicted PGA values in the lower intensity range. However, as PGA increases, predictions tend to cluster near the origin, indicating significant underestimation of high-magnitude events. The red reference line (y = x) emphasizes this bias, highlighting the model's inability to generalize beyond the dominant low-PGA region.

Residual Analysis The residual plot (Figure 4.6) exhibits a heteroskedasticity pattern, with error magnitude increasing for higher predicted PGA values. While most residuals remain close to zero in the low range, larger and more dispersed errors appear as PGA increases. This fan-shaped dispersion indicates model underfitting in the upper range and suggests a violation of the heteroskedasticity assumption inherent to linear regression.

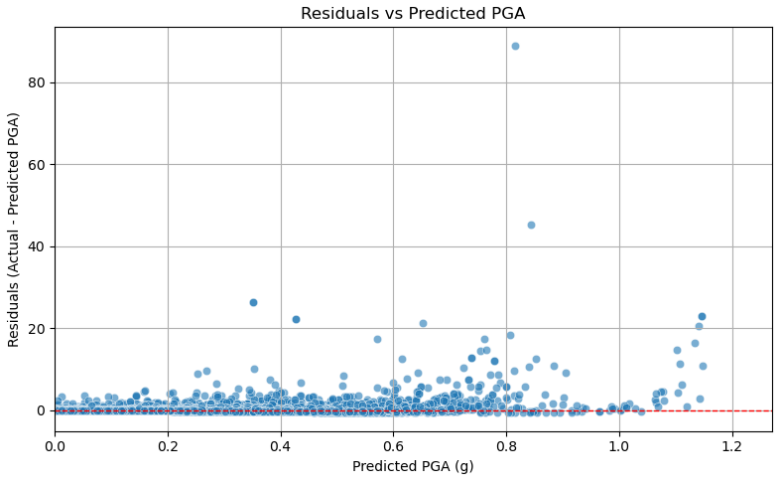

FIGURE 4.6: Residuals vs. Predicted PGA

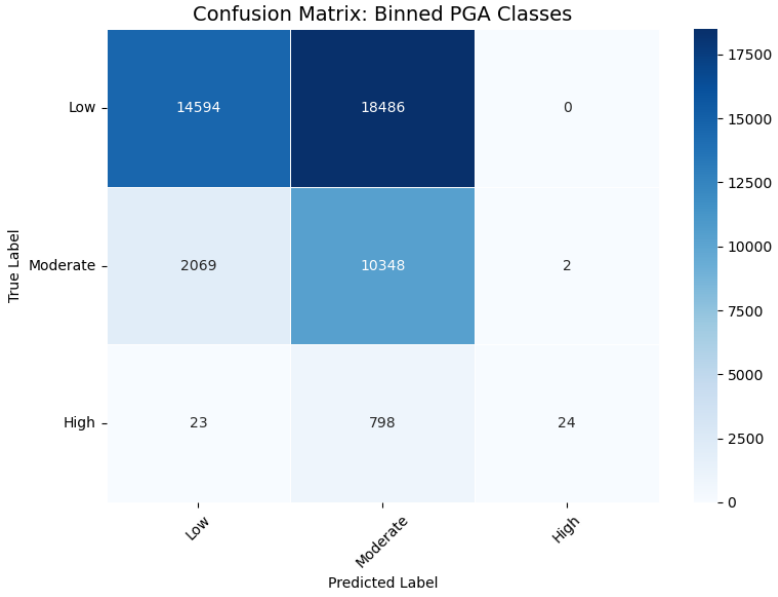

FIGURE 4.7: Confusion Matrix for Binned PGA Predictions

Classification-Based Evaluation To better understand the model's behavior across intensity levels, continuous PGA values were binned into classes. The confusion matrix (Figure 4.7) reveals high accuracy in predicting low PGA levels, but performance degrades significantly for moderate and high categories. Most high PGA events are misclassified as lower levels, demonstrating the model's limited sensitivity to rarer, more severe events. This skew is attributed both to the linear model's limited expressive power and the class imbalance present in the dataset.

Table 4.6: Classification report based on binned PGA values

Class	**Precision**	**Recall**	**F1-score**	**Support**
Low	0.87	0.35	0.50	41463
Moderate	0.35	0.81	0.49	12801
High	0.92	0.03	0.05	849
Accuracy		**0.4530**		

4.3 XGBoost Model

The XGboost model was evaluated using standard regression performance metrics on the test dataset. The results are summarized in Table 4.5.

Table 4.7: Model Evaluation Metrics (XGBoost)

Model	**R^2**	**RMSE**	**MAE**	**MAPE**
XGBoost	0.8123	0.1456	0.0924	0.1731

Interpretation of Final Model Evaluation Metrics $R^2$of 0.6343 indicates moderate explanatory power that the model explains about 63% of PGA variation. RMSE of 0.2138 suggests some larger prediction errors, possibly due to outliers. MAE tells us there is 4.63% average absolute error in predicting PGA. This is low, showing consistent average accuracy. MAPE is relatively high (70.18%) and may be misleading due to inflation by very small true PGA values. Overall, the model performs reasonably well, but could be improved with feature engineering and better treatment of small-magnitude events.

Fig 4.8 compares the relative feature importance of various parameters with respect to the PGA values as predicted by the model. We get the following insights:

1. Area intensity is the most predictive feature which makes sense because it's a direct measure of shaking.

2. Magnitude and epicentral/rupture distances are classic predictors for PGA and the model reflects that well.

3. Latitude and longitudinal parameters can be ignored as may be capturing unmodeled regional/geological variation.

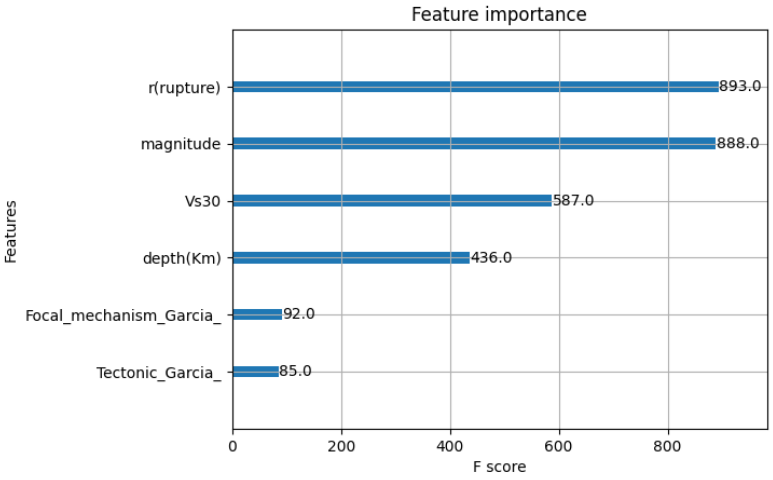

FIGURE 4.8: Feature Importance of Parameters in PGA prediction according to XG Boost

4. Tectonic setting and focal mechanism were least important according to the model. This could be due to the dataset lacking variation in these features, redundancy with other features or poor encoding.

TABLE 4.8: Cross-Validation Summary Metrics for XGBoost Model

Metric	**Train**	**Test**
R^2	0.7217 ± 0.0057	0.6355 ± 0.0269
MAE	0.0445 ± 0.0002	0.0463 ± 0.0005
RMSE	0.1865 ± 0.0038	0.2134 ± 0.0133

Cross Validation Summary The performance of the XGBoost model was evaluated using 3-fold cross-validation, and the results indicate a robust and stable predictive capability.

The coefficient of determination (R^2) was 0.7217 ± 0.0057 on the training sets and 0.6355 ± 0.0269 on the test sets, suggesting that the model is able to explain approximately 64% of the variance in unseen data. This modest drop between training and testing R^2 indicates that the model generalizes reasonably well without significant overfitting.

The mean absolute error (MAE) remained consistently low across folds, with 0.0445 ± 0.0002 on the training data and 0.0463 ± 0.0005 on the test data, demonstrating that the average magnitude of prediction errors is small and comparable between seen and unseen data.

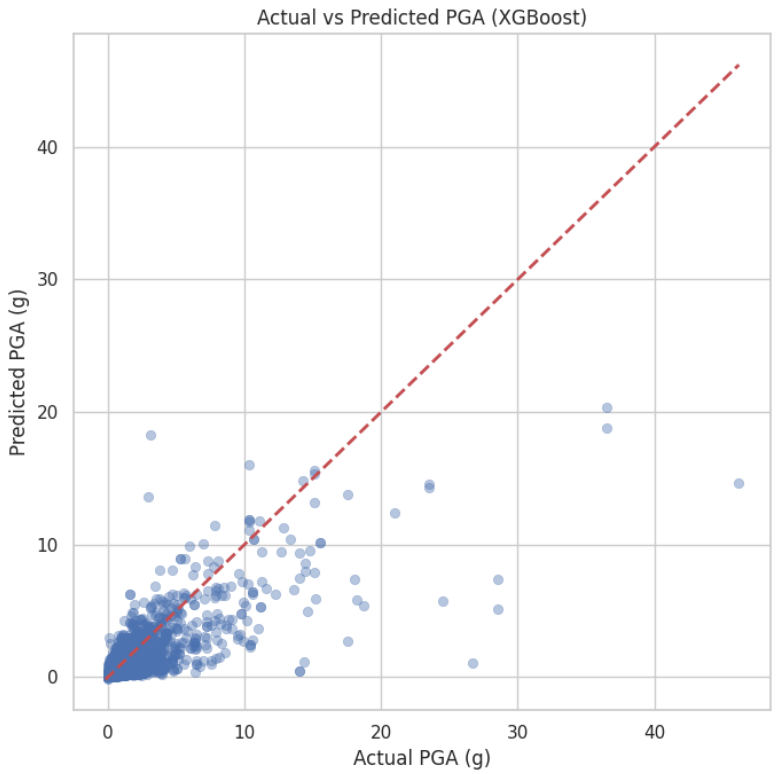

FIGURE 4.9: Actual vs predicted values of PGA using XGboost

The root mean squared error (RMSE), which penalizes larger deviations more heavily, was slightly higher on the test set (0.2134 ± 0.0133) compared to training (0.1865 ± 0.0038), implying the presence of some outliers or occasional large errors.

However, the low standard deviations across all metrics suggest that the model's performance is consistent and stable across different data splits. Overall, the results indicate that the XGBoost model achieves good predictive accuracy and generalization for the PGA estimation task.

Actual vs predicted plot Fig 4.9 is a plot that compares the PGA predicted by the model to the actual PGA. This is based on the testing dataset. We see that at lower PGA values, the model is roughly accurate with small deviations from the actual PGA values. But as it moves to larger PGA values, we see the points deviate away from the 45 degree line, indicating large predictive error. The model is unable to accurately predict PGA values at higher levels.

Residual Plot This residual plot in Fig 4.10 is a diagnostic tool used to evaluate the performance of a regression model. The residual is defined as the difference between the observed value and the predicted value. This visualization helps in assessing whether the model captures the underlying data pattern accurately, and whether systematic errors or biases exist.

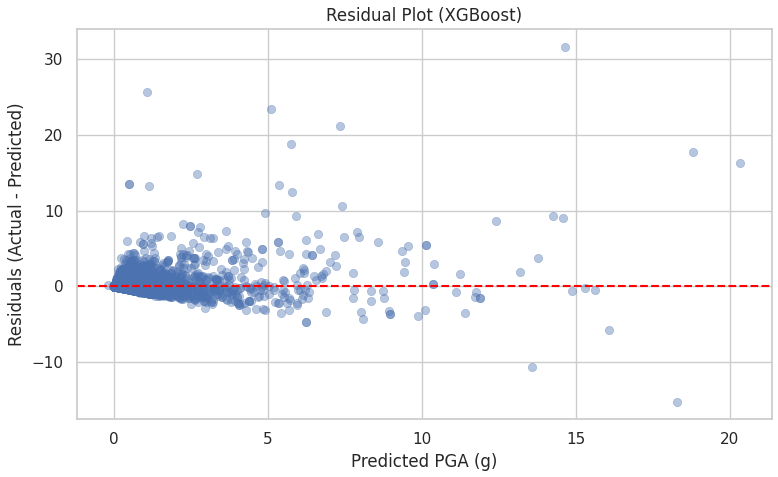

FIGURE 4.10: Residual Plot of PGA values by the XG boost model

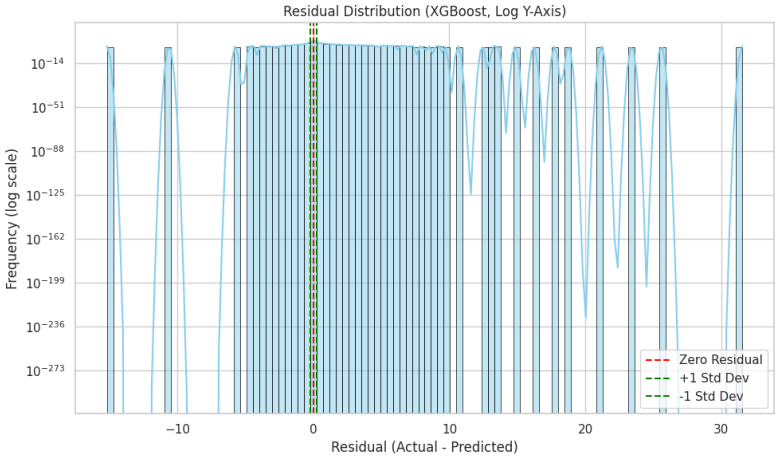

FIGURE 4.11: Residual Analysis distribution plot in lognormal scale for the XG boost model

Residual Analysis: Here are the results derived from the distribution of residuals in the Fig 4.11

- Mean : 0.0055 - This value is very close to zero. This indicates that the model does not exhibit systematic overprediction or underprediction. This suggests the model is unbiased on average.
- Std Dev : 0.2138 - This tells us that Most residuals cluster close to the horizontal axis at zero, with a standard deviation of approximately 0.2811, implying relatively tight error bounds for most predictions.
- Kurtosis : 4821.1213 - This is also extremely high, the typical kurtosis of a normal distribution is 3. Such extreme kurtosis implies a distribution with a very sharp

peak and very heavy tails. Indicates that most predictions are clustered near the mean, but a few massive outliers (errors) exist.

- Skewness: 46.2885 - This is extremely high and indicates a severely right-skewed distribution. Most of the residuals are small, but a few very large positive errors exist (i.e., underprediction of some large PGA values). This suggests that the model occasionally fails to capture high-end outliers.
- The plot shows increasing spread of residuals at higher PGA values, which indicates heteroscedasticity. This tells us that this model is less consistent in its predictions for high-magnitude shaking events, which may merit further investigation or transformation of the target variable.
- No strong patterns or curves are visible, which implies the model does not miss any major nonlinear relationships in the data.

The residual analysis confirms that the XGBoost model exhibits good generalization, with low bias and moderate variance. However, the presence of outliers and increasing variance at higher PGA predictions suggests that additional preprocessing or model refinement may further improve prediction accuracy, especially for high-magnitude events.

Table 4.9: Classification Report for PGA Binning (XGBoost Model)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Low	0.91	0.92	0.91	208,056
Moderate	0.63	0.60	0.62	50,408
High	0.77	0.44	0.56	1,939
Micro Avg	0.86	0.86	0.86	260,403
Macro Avg	0.77	0.65	0.70	260,403
Weighted Avg	0.85	0.86	0.85	260,403

Overall Accuracy: 0.8553

Classification Based Evaluation The confusion matrix in Fig 4.12 and classification report (Table 4.8) together provide a detailed picture of the XGBoost model's ability to classify Peak Ground Acceleration values into binned categories of Low, Moderate and High.

The model performs well on the majority class of **Low PGA** with high precision, recall and F1, all abobve 0.9. This is clearly reflected in the confusion matrix, where 191,562 out of 208,056 Low PGA instances were correctly predicted, and only 7.9% were misclassified as Moderate.

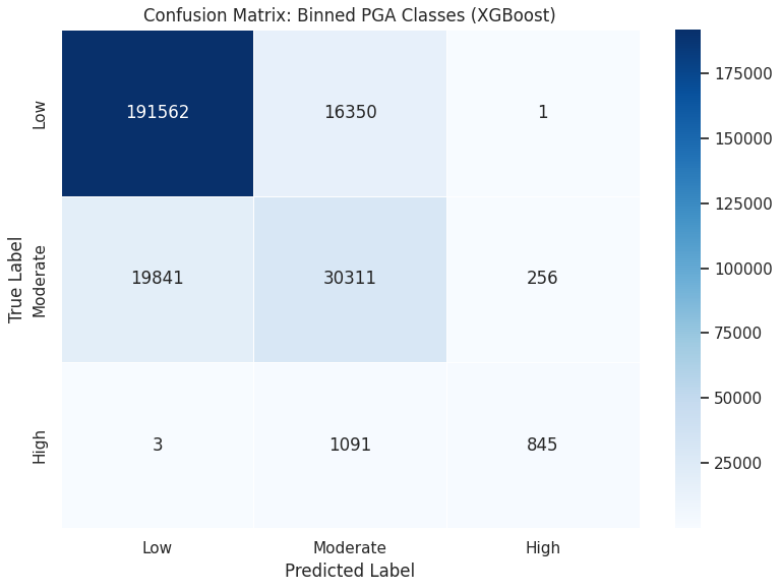

FIGURE 4.12: Confusion Matrix of the binned PGA values for the XG boost model

For the **Moderate class**, the classification report indicates moderate performance by all metrics. This is confirmed by the confusion matrix, where over 19,000 Moderate instances were misclassified as Low. This could be due to a common confusion due to their numerical proximity and possibly overlapping feature space. The model tends to undercall Moderate PGA , instead favouring the safer Low label.

The **High PGA** class presents the greatest error in recall although its precision is relatively high. The confusion matrix tells us that 1,091 High events were misclassified as Moderate and only 845 were correctly identified which is a skew typical in imbalanced datasets where the rare class lacks sufficient training representation, which is true for KiK net. The model shows class imbalance and limited learning signal for rare events.

The **macro average** is a simple mean shows moderate performance across all classes with F1 of 0.70, without accounting for class imbalance. The **weighted average** is dominated by "Low" class and hence F1 = 0.85 looks better but can mask poor recall in minority classes.

Conclusion The model's strength is in predicting the dominant "Low" class with moderate performance on "Moderate" and significant limitations in handling rare "High" PGA events.

4.4 Random Forest model

Training Dynamics

The Random Forest model is a robust and highly effective machine learning that improves prediction accuracy by aggregating the outputs of multiple decision trees. The model exhibited average performance across the evaluation metrics, with the detailed results summarized in Table 4.10.

The training set R^2 value of 0.88 indicates a strong fit to the training data, the substantial difference between the training and test performance suggests potential overfitting. Overfitting occurs when the model memorizes the training data rather than learning its underlying patterns, causing poor generalization to new data. The test set R^2 value of 0.71, which is notably lower than the training set score, further confirms this issue. Figure 4.13 illustrates the actual versus predicted PGA values of the training set for the model.

TABLE 4.10: Cross-validation performance metrics of RF model

Metric	**Train (Mean ± SD)**	**Test (Mean ± SD)**
R^2 Score	0.8803 ± 0.0064	0.7064 ± 0.0232
MAE (g)	0.0252 ± 0.0001	0.0418 ± 0.0006
RMSE (g)	0.1223 ± 0.0031	0.1916 ± 0.0127

Actual vs predicted plot

Figure 4.13 illustrates the actual versus predicted PGA values across all folds. We see that at lower PGA values, the model is roughly accurate with small deviations from the actual PGA values. But as it moves to larger PGA values, we see the points deviate away from the 45 degree line, indicating large predictive error. The model is unable to accurately predict PGA values at higher levels.

Residual Analysis

Ideally, residuals should be randomly scattered around the zero line (indicated by the red dashed line), with no clear pattern. However, in this plot, a funnel-shaped pattern is observed — the variance of residuals increases as the predicted PGA values increase. This pattern suggests heteroscedasticity, meaning the model's prediction errors are not constant across the range of outputs.

Specifically, the model tends to be more accurate for lower PGA values but shows greater dispersion and higher residuals as the predicted values increase.

Learning Curve Analysis Figure 4.15 illustrates the learning curve for the Random Forest model, showing the variation in Mean Absolute Error (MAE) for both the training and validation datasets as the size of the training set increases. The training MAE remains consistently low, increasing only marginally from approximately 0.0182 to 0.0192 as more

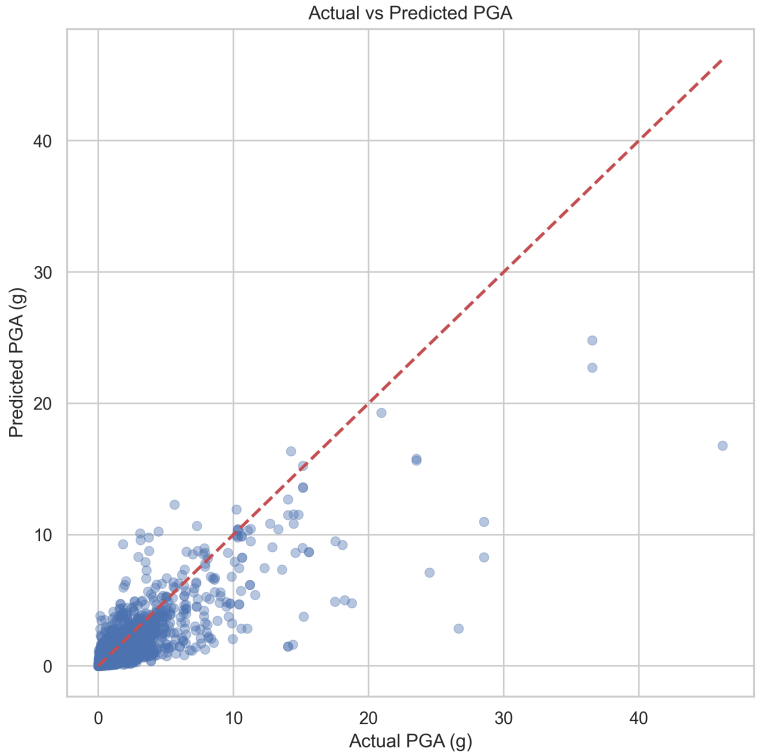

FIGURE 4.13: Actual vs. Predicted PGA values across all folds of RF model

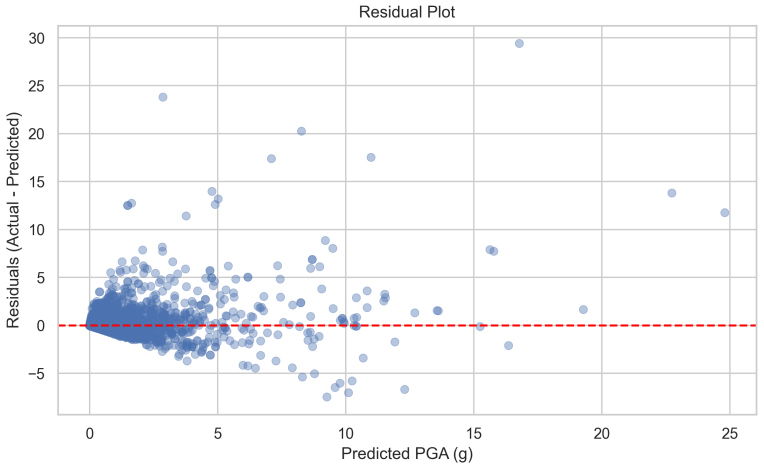

FIGURE 4.14: Residual plot across all folds

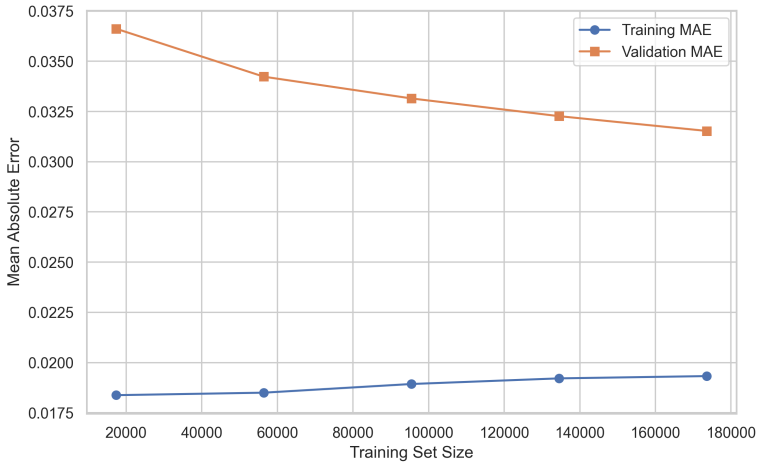

FIGURE 4.15: Learning Curve of RF model

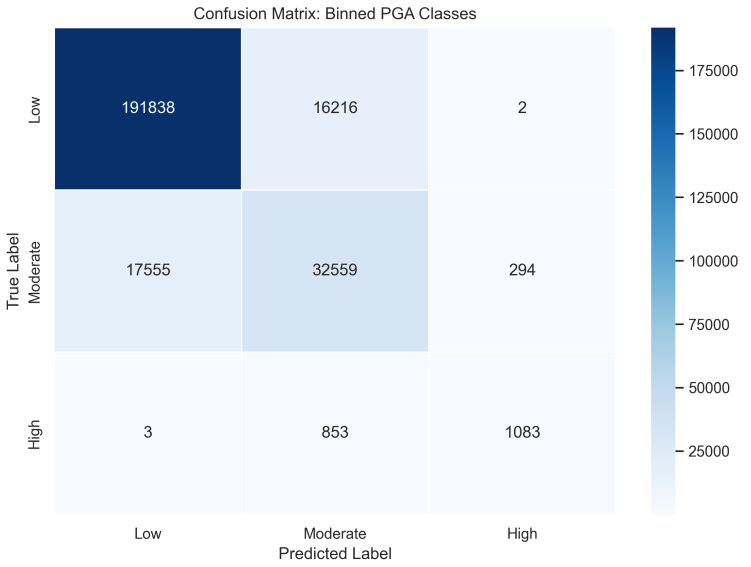

FIGURE 4.16: Confusion Matrix: Binned PGA Classes

data is introduced. This indicates that the model maintains a strong fit on the training data, regardless of sample size.

The validation MAE, on the other hand, exhibits a clear downward trend with increasing training set size—dropping from 0.037 to around 0.032. This decline in validation error suggests that the model's generalization performance improves as it is exposed to more data.

Classification-Based Evaluation As shown in the confusion matrix Figure 4.16, the model achieves high accuracy in classifying low PGA events. However, its performance

TABLE 4.11: Classification report based on binned PGA values

Class	**Precision**	**Recall**	**F1-score**	**Support**
Low	0.92	0.92	0.92	208056
Moderate	0.66	0.65	0.65	50408
High	0.97	0.56	0.65	1939
Accuracy		**0.87**		

drops notably for moderate and high categories, with many high-intensity events being misclassified as lower ones. This miss classification highlights the model's reduced sensitivity to less frequent, more severe ground motions.

Table 4.11 summarizes the classification performance on binned PGA values. The model performs well on the dominant "Low" class, with an F1-score of 0.92, but shows a noticeable drop for "Moderate" and "High" classes, both at 0.65.

Chapter 5

Summary And Conclusion

5.1 Introduction

Based on the observed ground motion between 1997 and 2017 and station information from the KiK-net, a large database was assembled that included both the site-specific and 6 feature parameters. Four regression models were developed and tested: Artificial Neural Network, XGBoost, Random Forest, and Linear Regression.

Each model was evaluated using three evaluation metrics: MAE, RMSE, and R^2 across training and testing phases. Our comprehensive assessment incorporates multiple performance metrics and analytical frameworks to rigorously compare model accuracy, generalization capability, and performance utility. A unified ranking approach was applied to assess the models' overall performance across all metrics.

5.2 Model Performance Analysis

Figure 5.1 presents a direct comparison of model performance based on three statistical metrics: MAE, RMSE, and R^2. Among all models, Random Forest achieved the lowest MAE and RMSE, indicating high predictive accuracy and robustness. XGBoost obtained the highest R^2 score, suggesting strong ability to explain variance in PGA, but with higher MAE than Random Forest. Neural Networks model demonstrated low MAE but comparatively higher RMSE and the lowest R^2, pointing to inconsistent performance. Linear Regression, while computationally efficient, consistently underperformed across all three metrics.

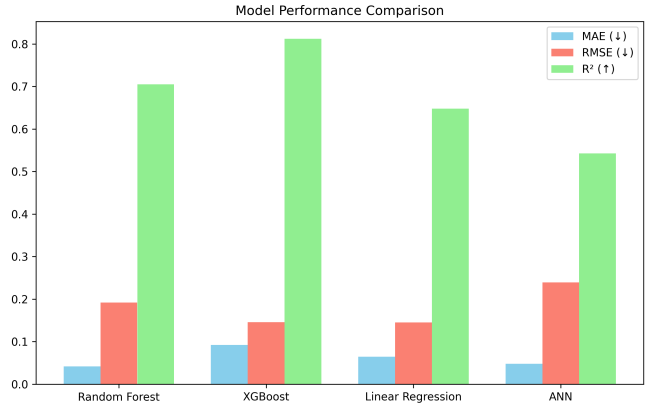

FIGURE 5.1: Comparison of model performance using three metrics: MAE, RMSE and R^2

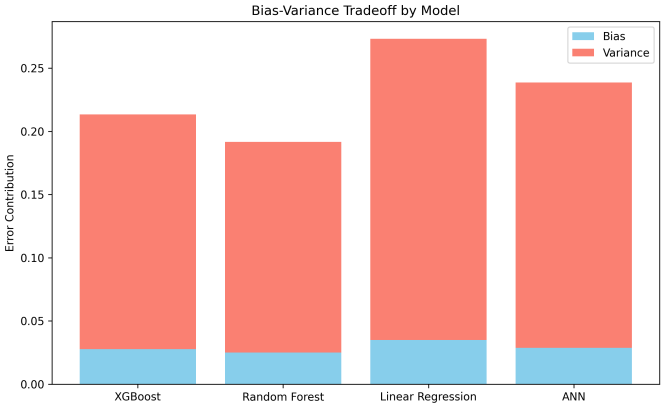

FIGURE 5.2: Bias-Variance Tradeoff plot

Figure 5.2 illustrates the bias–variance tradeoff for each model. Random Forest and XGBoost show a well-balanced tradeoff, with relatively low bias and moderate variance—indicative of good model generalization without overfitting. Linear Regression, however, exhibits high bias, which explains its inability to capture the complexity in the data. On the other hand, ANN suffers from higher variance, which likely led to its fluctuating performance across training and testing datasets.

Figure 5.3 compares the R^2 scores on the training and test sets. Random Forest displays a minimal drop between training and testing R^2, suggesting strong generalization. XGBoost also generalizes well, although a slightly higher gap indicates some tendency to overfit. ANN shows a smaller gap but at the cost of lower overall R^2, while Linear Regression exhibits both lower train and test R^2, confirming its underfitting behavior and limited capacity to learn nonlinear relationships.

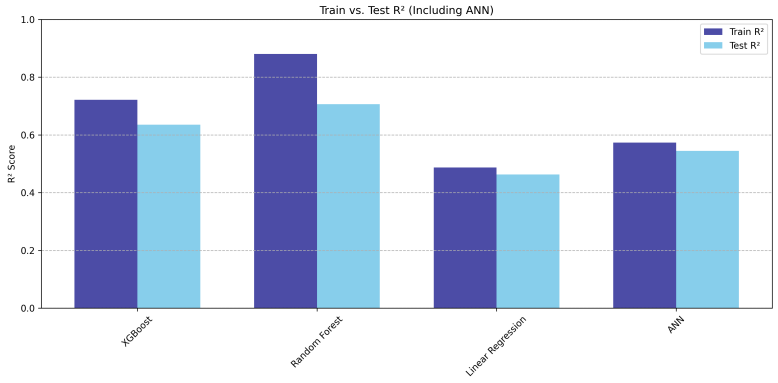

FIGURE 5.3: Train v/s Test R^2 comparison across models

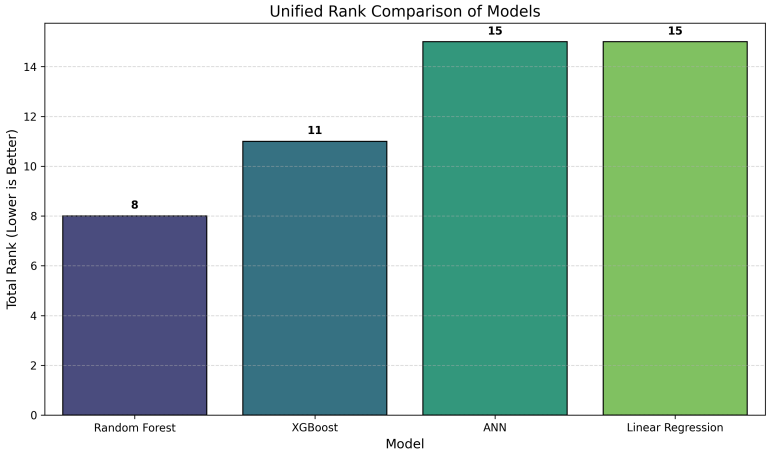

FIGURE 5.4: Unified ranking of models based on their aggregated performance

Figure 5.4 shows the unified ranking of all four models based on their performance across all metrics. Random Forest received the lowest total rank (8), confirming its consistent and balanced performance. XGBoost followed closely with a total rank of 11, validating its competitive accuracy despite slightly higher variance. Both ANN and Linear Regression scored the highest rank (15), indicating inferior performance in either precision or generalizability.

5.3 Conclusion

In terms of predicting Peak Ground Acceleration (PGA), machine learning models significantly outperform traditional linear regression. Among the four models evaluated, Random Forest achieved the most balanced performance in terms of accuracy and generalization, followed closely by XGBoost. Neural Networks showed moderate performance, while Linear Regression consistently underperformed across all metrics.

Of the six feature parameters, r(rupture) and magnitude have the largest impact on the prediction results, followed by depth(km), V_{s30}, Focal_mechanism_Garcia, and Tectonic_Garcia.

5.4 Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-2547/?version=2

References

Anjom, F.K., Vaccarino, F., Socco, L.V., 2024. Machine learning for seismic exploration: Where are we and how far are we from the holy grail? Geophysics 89, WA157–WA178. URL: https://doi.org/10.1190/GEO2023-0129.1, doi:10.1190/GEO2023-0129.1.

Bahrampouri, M., Rodriguez-Marek, A., Shahi, S., Dawood, H., 2020. An updated database for ground motion parameters for kik-net records. Earthquake Spectra 37. doi:10.1177/8755293020952447.

Banerjee, S., .K.A., 2016. Determination of seismic wave attenuation: A review. Disaster Advances , 10–11.

Bhusal, B., A.M.P.S..P.H.R., 2022. Site specific seismic hazard analysis of monumental site dharahara, kathmandu, nepal. Geomatics, Natural Hazards and Risk , 2674–2696URL: https://doi.org/10.1080/19475705.2022.2130109.

Bommer, J.J., Martínez-Pereira, A., 2000. Strong-motion parameters: Definition, usefulness and predictability, in: Proceedings of the 12th World Conference on Earthquake Engineering, Auckland, New Zealand. Paper no. 206.

Boore, D.M., 2004. Can site response be predicted? Journal of Earthquake Engineering 8, 1–41. U.S. Geological Survey, 345 Middlefield Road, MS 977.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. URL: https://doi.org/10.1023/A:1010933404324, doi:10.1023/A:1010933404324.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system., in: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (Eds.), KDD, ACM. pp. 785–794. URL: http://dblp.uni-trier.de/db/conf/kdd/kdd2016.html#ChenG16.

Chiou, B.S.J., Youngs, R.R., 2008. An nga model for the average horizontal component of peak ground motion and response spectra. Earthquake Spectra 24, 173–215. URL: https://doi.org/10.1193/1.2894832, doi:10.1193/1.2894832.

Dawood, H.M., Cotton, F., Rodriguez-Marek, A., Bonilla, F., Montalva, G.A., 2011. Contribution to ground motion uncertainty from site response of shallow surface deposits, in: Effects of Surface Geology on Seismic Motion, University of California, Santa Barbara.

Díaz, J.P., Sáez, E., Monsalve, M., Candia, G., Aron, F., González, G., 2022. Machine learning techniques for estimating seismic site amplification in the santiago basin, chile. Engineering Geology 306, 106764. URL: https://doi.org/10.1016/j.enggeo.2022.106764, doi:10.1016/j.enggeo.2022.106764.

Frankel, A., Mueller, C., Barnhard, T., Perkins, D., Leyendecker, E.V., Dickman, N., Hanson, S., Hopper, M., 1996. National Seismic Hazard Maps: Documentation June 1996. Open-File Report 96-532. U.S. Geological Survey. URL: https://pubs.usgs.gov/of/1996/532/CEUSatten.html.

Hibert, C., Noël, F., Toe, D., Talib, M., Desrues, M., Wyser, E., Brenguier, O., Bourrier, F., Toussaint, R., Malet, J.P., Jaboyedoff, M., 2024. Machine learning prediction of the mass and the velocity of controlled single-block rockfalls from the seismic waves they generate. Earth Surface Dynamics 12, 641–656. URL: https://doi.org/10.5194/esurf-12-641-2024, doi:10.5194/esurf-12-641-2024.

James Kaklamanos, S.M.EERI, L.G.B.M., Booreb, D.M., 2011. Estimating unknown input parameters whenimplementing the nga ground-motion prediction equations in engineering practice. Biometrics , 1220–1234.

Kamai, R., Abrahamson, N.A., Silva, W.J., 2014. Nonlinear horizontal site amplification for constraining the nga-west2 gmpes. Earthquake Spectra 30, 1223–1240.

Kotha, S.R., Traversa, P., 2024. A bayesian update of kotha et al. (2020) ground-motion model using résif dataset. Bulletin of Earthquake Engineering 22, 2267–2293. URL: https://doi.org/10.1007/s10518-023-01853-1, doi:10.1007/s10518-023-01853-1.

Kramer, S.L., 1996. Geotechnical Earthquake Engineering. Prentice Hall, Upper Saddle River, N.J. Available via Internet Archive.

Li, X., Gao, P., 2024. Significant duration prediction of seismic ground motions using machine learning algorithms. PLOS ONE 19, e0299639. URL: https://doi.org/10.1371/journal.pone.0299639, doi:10.1371/journal.pone.0299639.

McGuire, R.K., Toro, G.R., 2008. Site-specific seismic hazard analysis, in: Proceedings of the 14th World Conference on Earthquake Engineering, Beijing, China. URL: https://iitk.ac.in/nicee/wcee/article/14_07-0036.PDF.

Mori, F., Mendicelli, A., Falcone, G., Acunzo, G., Spacagna, R.L., Naso, G., Moscatelli, M., 2022. Ground motion prediction maps using seismic-microzonation data and machine learning. Natural Hazards and Earth System Sciences 22, 947–966. URL: https://doi.org/10.5194/nhess-22-947-2022, doi:10.5194/nhess-22-947-2022.

Ortega, R., 2024. Machine learning and seismic hazard: A combination of probabilistic approaches for probabilistic seismic hazard analysis, in: IntechOpen. URL: https://doi.org/10.5772/intechopen.1006533, doi:10.5772/intechopen.1006533. submitted: 16 July 2024; Reviewed: 30 July 2024; Published: 22 August 2024.

Park, D.P., Lee, Y., 2024. Machine learning based site amplification prediction models for shallow bedrock sites. Japanese Geotechnical Society Special Publication 10, 2203–2207. doi:10.3208/jgssp.v10.OS-47-07.

Rathje, E., Kottke, A., Trent, W., 2010. Influence of input motion and site property variabilities on seismic site response analysis. Journal of Geotechnical and Geoenvironmental Engineering - J GEOTECH GEOENVIRON ENG 136. doi:10.1061/(ASCE)GT.1943-5606.0000255.

Saleem, N., Mangalathu, S., Ahmed, B., Jeon, J.S., 2024. Machine learning-based peak ground acceleration models for structural risk assessment using spatial data analysis. Earthquake Engineering & Structural Dynamics 53, 152–178. URL: https://doi.org/10.1002/eqe.4021, doi:10.1002/eqe.4021.

Shingaki, Y., Goto, H., Sawada, S., 2018. Evaluation performance for site amplification factors: S-wave impedance vs. vs30. Soils and Foundations URL: https://doi.org/10.1016/j.sandf.2018.05.001, doi:10.1016/j.sandf.2018.05.001. accepted 15 May 2018, Available online 10 July 2018.

Sun, J., Innanen, K., Zhang, T., Trad, D., 2023. Implicit seismic full waveform inversion with deep neural representation. Journal of Geophysical Research: Solid Earth 128. doi:10.1029/2022JB025964.

Weatherill, G., Kotha, S.R., Danciu, L., Vilanova, S., Cotton, F., 2024. Modelling seismic ground motion and its uncertainty in different tectonic contexts: challenges and application to the 2020 european seismic hazard model (eshm20). Natural Hazards and Earth System Sciences 24, 1795–1834. URL: https://nhess.copernicus.org/articles/24/1795/2024/, doi:10.5194/nhess-24-1795-2024.

Wu, H., Zhou, Y., Qian, Y., Yang, F., Yu, G., Wu, G., Zhang, Y., 2023. Seismic-wave path attenuation and local site responses in eastern guangdong province revealed by the ground-motion spectral analyses. Frontiers in Earth Science 11. URL: https://doi.org/10.3389/feart.2023.1091785, doi:10.3389/feart.2023.1091785. section: Structural Geology and Tectonics.

Xiangqi Wang, Zifa Wang, J.W.P.M.H.D., Li, Z., 2023. Machine learning based ground motion site amplification prediction. frontiers , 1–14.

Zhu, W., Mousavi, S.M., Beroza, G.C., 2019. Seismic signal denoising and decomposition using deep neural networks. IEEE Transactions on Geoscience and Remote Sensing 57, 9476–9488. URL: http://dx.doi.org/10.1109/TGRS.2019.2926772, doi:10.1109/tgrs.2019.2926772.