Seismic data evaluation using machine learning algorithms

By

Raisa Suleman

Fall 2019-MSCS-9-00000318849

Supervisor

Dr. Pakeeza Akram

Masters of Computer Science

School of Electrical Engineering and Computer Sciences (SEECS)

National University of Sciences and Technology, NUST H-12,
Islamabad

November 2021

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Seismic data evaluation using machine learning algorithms" written by RAISA SULEMAN, (Registration No 00000318849), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature:

Name of Advisor: Pakeeza Akram

Date: 06-Dec-2021

Signature (HOD): ______________________

Date: ________________________________

Signature (Dean/Principal): ______________

Date: ________________________________

Approval

It is certified that the contents and form of the thesis entitled "Seismic data evaluation using machine learning algorithms" submitted by RAISA SULEMAN have been found satisfactory for the requirement of the degree

Advisor: Pakeeza Akram

Signature:

Date: 06-Dec-2021

Committee Member 1: Dr. Muhammad Khuram Shahzad

Signature: MKhurram.

Date: 02-Dec-2021

Committee Member 2: Dr. Abdul Wahid

Signature: Awahid.

Date: 02-Dec-2021

Committee Member 3: Dr. Rabia Irfan

Signature: Irfan

Date: 02-Dec-2021

DEDICATION

Dedicating this thesis to my parents and elder brother for their constant emotional and financial support.

Certificate of Originality

I hereby declare that this submission titled "Seismic data evaluation using machine learning algorithms" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: RAISA SULEMAN

Signature:

ACKNOWLEDGEMENT

I am highly thankful to Allah Almighty for guiding me throughout this journey. I would also like to show gratitude towards my supervisor Dr. Pakeeza Akram for her continuous support and guidance. I would also like to thank my committee members, Dr. Abdul Wahid, Dr. Rabia Irfan, and Dr. Khuram Shahzad for their constant supervision and support.

Moreover, I would like to thank USGS for making the catalog publicly accessible for research.

TABLE OF CONTENTS

List of Abbreviations

ABBREVIATION	Explanation
AR picker	Auto Regressive Picker Algorithm
CART	Classification and Regression Tree
EDA	Exploratory Data Analysis
GPC	Gaussian Process Classifier
KNN	K nearest Neighbor
LDA	Linear Discriminant Analysis
LSTM	Long short-term memory
RELU	Rectified linear activation function
SVM	Support vector machine
SVM-HNN	Support vector machine-Hybrid neural network
SVM-PSO	Support vector machine- Particle swarm optimization
USGS	United States Geological Survey
XGBOOST	Extreme Gradient Boosting Algorithm

List of Tables

List of Figures

Abstract

Earthquakes are one of the devastating natural disasters which cause significant damage to property due to their destructive nature. Seismic stations around the globe record data continuously to make it available for research and information purpose. An enormous amount of research has been done in this regard in the past as well but generally, the research is done on the seismic regions only. This identifies that there is limited work done on the data analysis for country-wise seismic data. This thesis specifically analyzes and evaluates collective country-wise seismic data through machine learning algorithms. From a geological perspective, Pakistan is located on three tectonic plates. The historic seismic activity of Pakistan along with its neighboring countries including China and Afghanistan is considered for an efficient evaluation. For an unbiased comparative analysis, two evaluation techniques are considered that include threshold-based binary seismic classification and magnitude categorization based on the Mercalli intensity scale for determining magnitude destructive nature. Decision tree, Random forest, XGBoost, Adaboost, and KNN are implemented on three country-wise seismic datasets. Among the five applied algorithms, two algorithms including Random forest and XGB performed exceptionally well in the selected evaluation methods.

The proposed evaluation methods can be applied to other natural hazardous data as well to evaluate the performance of applied algorithms on selected evaluation criteria. All the algorithms are compared on the basis of selected comparative metrics that provides an insight into the quality of the algorithm performance on country-wise seismic historic activity.

CHAPTER 1

Introduction

Earthquakes are among one of the major destructive natural hazards that harm properties and cause huge loss of lives all around the globe. Geographically Pakistan is located on three active tectonic plates including Eurasian, Arabian, and Indian plates [1]. Asian countries including Pakistan, China, and Japan are prone to disastrous high-intensity earthquakes[2]. The worldwide statistics show almost 1433 earthquakes of magnitude greater than or equal to 5.0 are recorded in 2020[3]. The deadliest earthquake of 7.6 magnitude hits the northern area of Pakistan on 8th October 2005 at 8:50 am that affected the neighboring countries as well. An analysis stated that almost nighty thousand people died and seventy-nine were injured whereas at least 3.5 million were homeless[4]. The major shock was followed by more than 1200 aftershocks in a one-month time span.

Seismology is the interpretation of seismic waves which occurred due to the movement of the materials within the soil which causes fault slips, avalanches, and explosions[5]. The seismic activity takes place due to the sudden breaking of rocks underground and results in a fault. The instant release of energy through seismic waves causes the earth's vibration[6]. Earth's crust is made up of tectonic plates which are continuously moving. The boundary of the plates slips within each other creating fault zones, this is known as tectonic earthquakes[7]. Other than tectonic earthquakes seismologist identifies other types of earthquakes which include volcanic earthquakes, explosive earthquake, and collapse earthquake. Volcanic earthquakes are caused due to fault occurrence near the volcano[8]. Explosion earthquakes are caused due to nuclear explosions[9]. Collapse earthquakes are caused due to rock explosions in mines and caves[10].

Tectonic plates are huge blocks of rocks that move around the earth's lithosphere and slide right on the top of the earth's mantle. There are seven tectonic plates that cover up almost 95% of the surface of the earth. These tectonic plates include Eurasian, South American, North American, Antarctic, African, Pacific, and Indo-Australian plates [11]. These tectonic plates continually move around which causes the plate's boundaries to strike into each other. Tectonic plate movements are further categorized into three types based on their interactability with each other. The types are categorized as convergent, divergent, and transform boundaries [12]. Convergent boundaries happen when two tectonic plates collide with each other at the same point. Divergent boundaries occur when tectonic plate glides apart and move away from each other. Whereas transform boundaries slide side by side in an opposite direction causing an earthquake.

According to the global seismic hazard assessment program constructs a major seismic zones map where scientists split the map into 20 seismic regions based on the historic seismic activities [13]. These regions were identified from prime seismic active zones which include Asia, South America, North America, Africa, Antarctica, Europe, and Oceania. Pakistan is divided into five active zones based on the severity of the earthquakes[14]. The zones are categorized from lowest to highest magnitude scale.

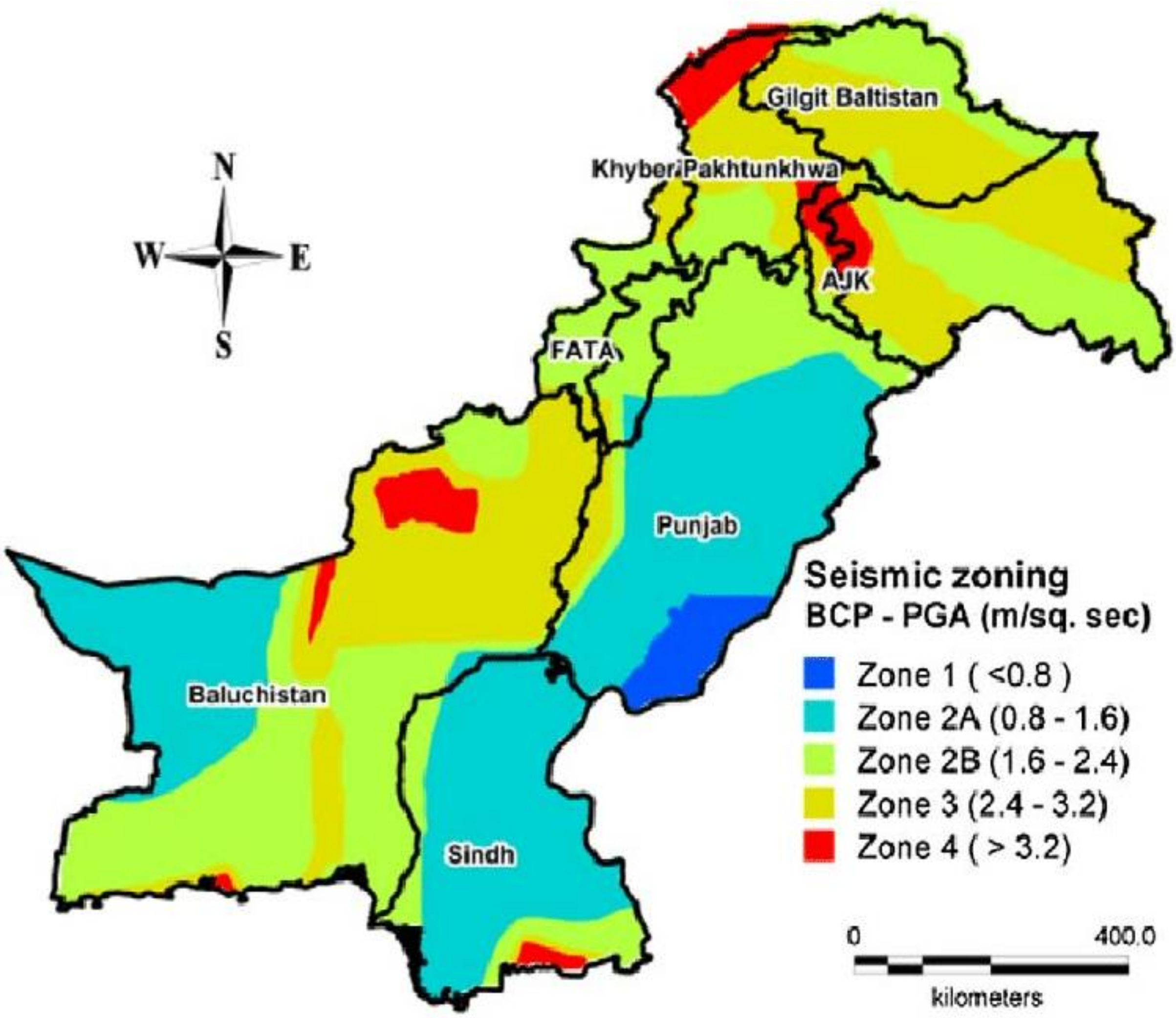


Figure 1: Seismic zones of Pakistan

Figure 1 [14] shows the active seismic zones in all the four provinces of Pakistan divided on the basis of recorded magnitude in all the highlighted regions.

Moreover, Japan is considered the most active seismic country due to its opaque seismic network as compared to the rest of the world. Japan lies beside the pacific ring of fire where the majority of volcanic and earthquake activity takes place [15].

1.1 Background and Motivation

The thesis mainly focused on the tectonic plate's movement which causes the plate's boundaries to collide and emits seismic wave's energy in all directions. Major seismic terminologies include a brief description of earthquake occurrence, types of seismic waves, calculating the magnitude and the depth, locating the epicenter, categorization of earthquake magnitude in multiple classes as well as the correlation of recorded depth with the destructions caused on the surface of the earth.

1.1.1. Seismic activity

Earthquakes are mainly caused due to sudden ground movement which is caused due to the movement of tectonic plates [16]. The plates of the earth's surface move which results in fault that occurred due to the release of strong seismic waves beneath the earth's surface which is the hypocenter and the location directly above the surface are known as the epicenters [17].

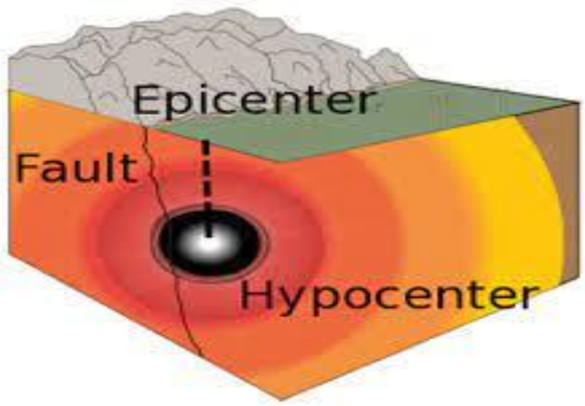


Figure 2: Origin of seismic activity

Figure 2 clearly shows the origin of seismic wave activity occurrence [18]. The distance between the hypocenter and the epicenter is the recorded depth covered by an earthquake in kilometers. Distance is categorized into three categories which include shallow earthquakes which range from 0 to 70 km, intermediate earthquakes that cover a distance from 70 to 300 km, and deep earthquakes range from 300 to 700 km from the epicenter [19].

1.1.2. Types of seismic waves

The seismic waves are categorized into two types which are listed below:

1. Body waves.

2. Surface waves.

Body waves include primary and secondary wave which is denoted as P-waves and S-waves. The primary waves are the initial waves that are captured by the seismographs and are compressional in nature which could travel through the liquid and solid objects [20]. These primary waves travel way faster than the surface waves due to the energy transmit ability as the earth's internal component is incompressible. Secondary waves are always the second waves that are recorded after the arrival of the primary waves because they are slower in speed. The S-wave is also called shearing waves as it can only move on solid objects.

Surface waves arrived at the very end of body waves and they are also sub-categorized into two types which include Love and Raleigh waves known as L-waves and R-waves [21]. Both the surface waves cause serious destruction and have different amplitude properties. These surface waves can be easily distinguished from the seismogram reading due to lowered captured frequency. Love waves travel horizontally on the surface and Raleigh waves travel in all directions of the surface [22]. The amplitude intensity in the love wave is greater as they horizontally travel which is interconnected with the depth. Amplitude intensity for shallow earthquakes is quite greater on the seismographs reading and it disperses as the traveled depth of the seismic wave kept on increasing with the passage of time.

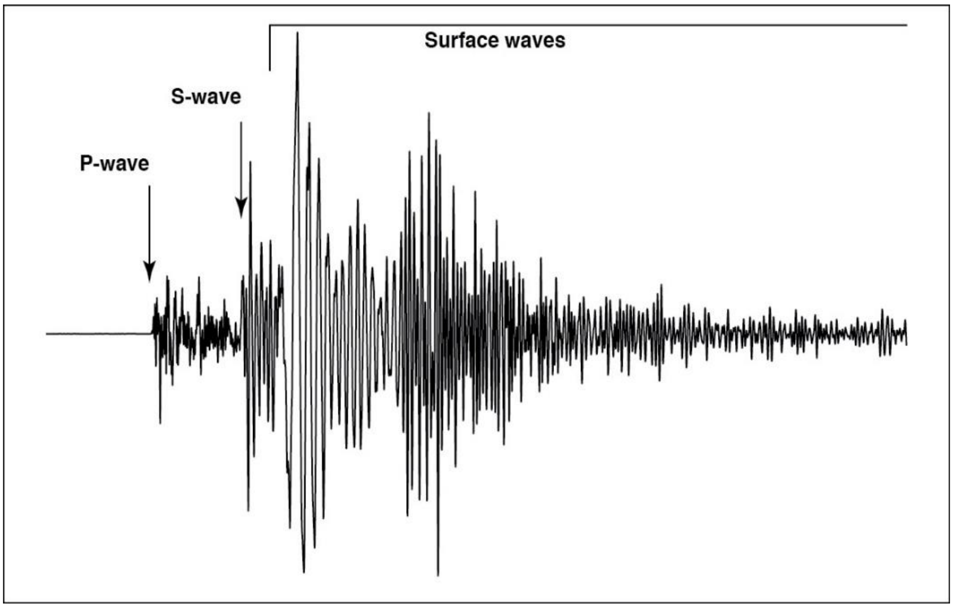


Figure 3: Seismic waves recording

Figure 3 [23] distinguished both the waves and it also depicts the amplitude length as well as the frequency at which all these waves travel through the body and the surface of the earth. Initially, both the body waves are recorded. Through the peaks of the recorded waves, the categorization is taken place. As the figure shows high amplitude waves at first and at the very end the intensity is reduced due to the dispersion of the seismic waves' energy due to the high depth covered along the way towards the bottom of the earth.

1.1.3. Magnitude calculation

The Richter scale used for magnitude calculation was initially developed in 1935 by Charles Francis Richter [24]. It calculates the earthquake magnitude by applying the logarithm of the seismograph recorded wave's amplitude.

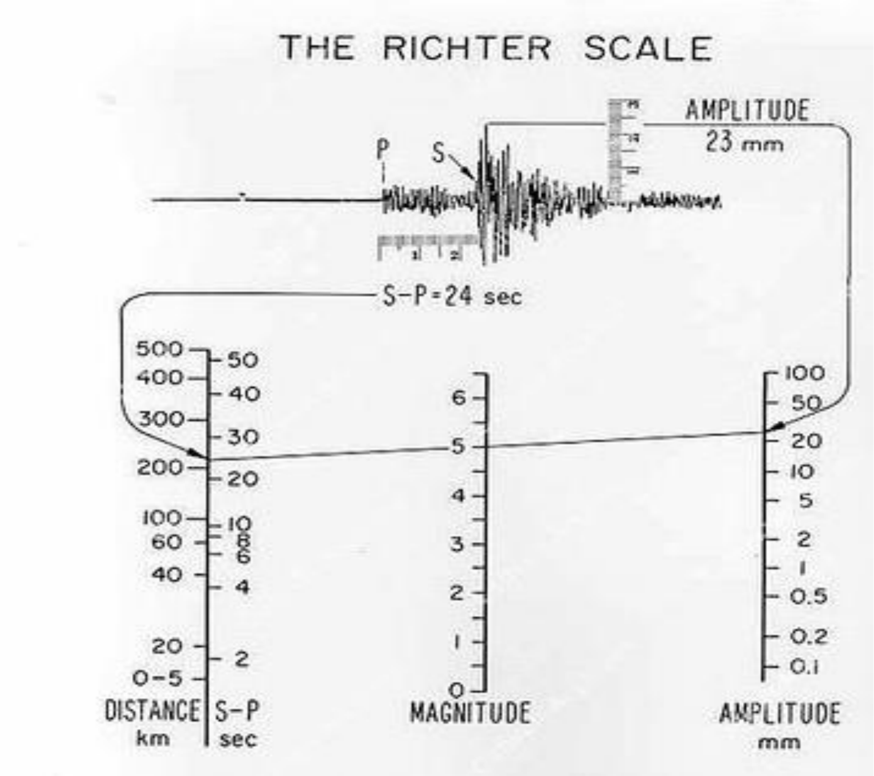


Figure 4: Richter scale

Figure 4 [25] is the Richter scale which was first developed for magnitude calculation but now it's used for small-scale surface earthquakes. The calculation procedure through the Richter scale to record local earthquakes denoted by ml after recording the seismic waves is:

1. The distance of the recorded P-wave and S-wave is calculated in seconds.
2. Measuring the total height of the overall wave which is denoted as amplitude.
3. Pointing out the time on the left and the amplitude placed on the right; a line is drawn which passes through the middle magnitude scale.
4. Through this procedure the total distance recorded in kilometers is also identified along with the total wave travel time.

5. The identified local magnitude value is equal to 5.0.

The Richter scale is replaced by the moment magnitude scale developed in 1970 by Thomas C. Hanks and Hiroo Kanamori which is efficient in calculating large scale earthquakes [26]. It mainly covers the physical aspect of the earthquake including the total energy released and the total covered area of the fault along with the covered distance.

1.1.4. Epicenter Location

As the seismographs capture the seismic waves, it doesn't update on the location of the recorded earthquake waves [27]. Seismic stations after receiving the seismic data find out the distance and the magnitude but not the actual location where the fault occurs. For finding the right direction of the earthquake triangulation method is applied [28].

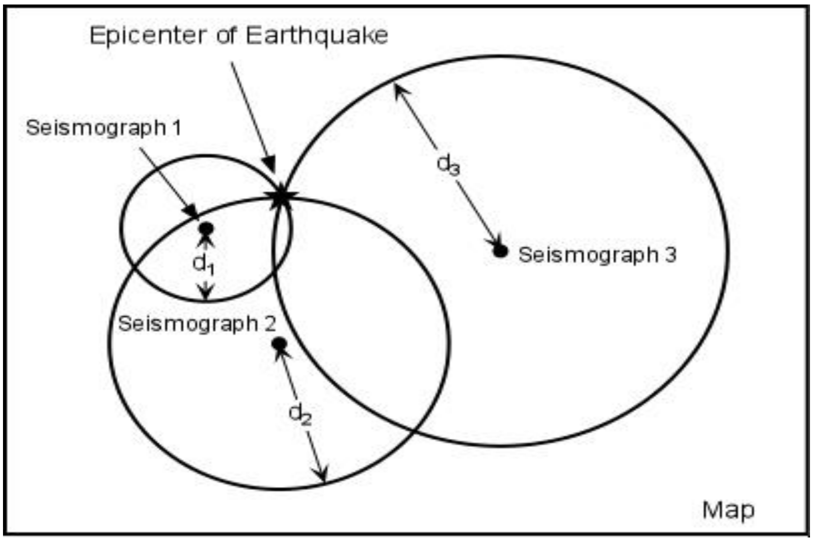


Figure 5: Locating epicenter

Figure 5 [29] shows the triangulation method used by the station in order to locate the epicenter on the map. This method is quite simple; the seismic stations draw a circle with a radius equal to the total distance covered in kilometers. The point where these three circles intersect together is the epicenter point on earth. For getting the exact location of the fault, the triangulation method implemented by three seismic stations is mandatory.

1.1.5. Magnitude categorization

The magnitude categorization is based on the level of destruction it can cause on the surface. On the scale, as the number increases so does the intensity as well as the energy of the seismic waves gets stronger [30]. The seismic wave's amplitude can easily differentiate between small-scale and large-scale earthquakes. If the recorded wave is shorter; it's considered as a local earthquake that is easily calculated by the Richter scale but if the wave is longer in nature it depicts the total fault area as well as the intensity.

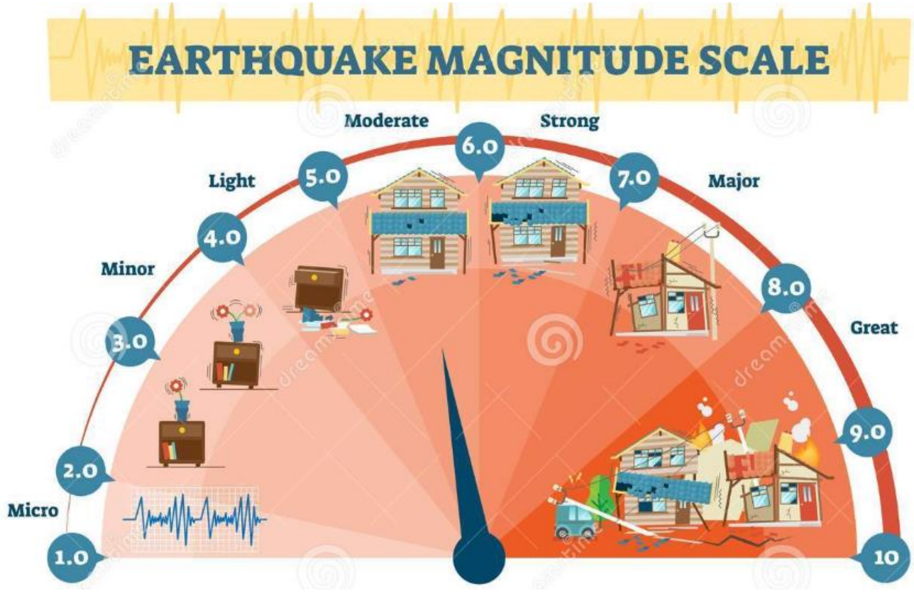


Figure 6: Magnitude scale division

Figure 6 [31] shows the division of magnitude in separate classes starting with micro magnitude equal to 1.0 till great earthquakes with a magnitude value of 10. The destructive effect of recorded magnitude on the surface is categorized into different classes. Table 1 [32] shows the magnitude division along with the damage caused on the surface.

Magnitude Division	Effect
0-1.9	It's not felt by humans
2.0-3.0	It's hardly felt on the surface
3.1-3.9	Felt by very few individuals
4.0-4.9	It causes slight damage to buildings
5.0-5.9	Significant damage on the surface
6.0-6.9	Great damage to poorly constructed buildings
7.0-7.9	It causes huge destruction on the surface
8.0-8.9	It results in massive destruction
9.0-10	Destroy everything on the surface of earth

Table 1: Intensity of magnitude scale division

1.1.6. Region vs country-wise seismic analysis

A region specifies a specific part of land whereas a country is a separate territory sharing borders with other countries and is controlled by a government. Multiple studies have been done on different regions to evaluate machine learning classifiers performance. Forecasting earthquake intensity through seven supervised machine learning algorithms on six different regions of India is performed for efficient comparison [33]. Country-wise comparative analysis of global seismic data analysis through magnitude discretization to predict the earthquake occurrence through KNN and Random forest classifier is performed [34].

1.2. Problem Statement

Comparative analysis of earthquake occurrences of Pakistan with its neighboring countries and then comparing Pakistan with the earthquake hot spot country Japan. The analysis is performed to not only evaluate country-wise earthquake data but also to classify the severity of the earthquakes as well.

1.3. Objectives

The objective of this research is to perform:

1. Country-wise seismic exploratory data analysis of Pakistan along with its neighboring countries which share borders such as Afghanistan and China.
2. Performing comparative study of supervised machine learning algorithms' performance of Pakistan's seismic data.
3. Lastly, to compare the outcomes with the most vulnerable country with respect to earthquake occurrences in the same continent.

1.4. Thesis Contribution

Following is the thesis contribution:

1. Best to our knowledge limited research is done regarding Pakistan and its neighboring countries with respect to evaluating seismic activities. We used machine learning techniques in this least explored domain.
2. The applied analysis criteria aren't limited to a single dataset but instead, multiple country-wise datasets are taken to critically analyze the quality of evaluation methods. The proposed framework of the comparative analysis can be adopted to perform efficient analysis of any real-world disastrous events.

3. This research critically analyzes the algorithm’s performance on a standard time-series seismic dataset to predict the earthquake type based on the Mercalli intensity scale. The proposed evaluation methods are feasible in nature and can be utilized in the future to evaluate the severity of natural hazards on people and the infrastructure. This research focuses on an in-depth analysis of the machine learning algorithm’s performance for evaluating collective country-wise historic seismic data.

CHAPTER 2

Literature Review

A comparison of multiple machine learning algorithms' performance for predicting earthquakes on Indonesia's seismic data was performed [35]. Prediction of multiple features including depth, location, and magnitude is done. Three algorithms are used on multiple combinations of datasets which are SVM, Naïve Bayes, and multinomial logistic regression which got accuracy ranging between 72 to 92%. On both the collective grouped and ungrouped data SVM shows good performance as compared to the other two algorithms. Authors in [36] performed extensive exploratory data analysis on Pacific and Australian plate boundaries through line plots and pie charts which provides an insight into performing analysis on historic time series seismic datasets. Authors in [37] classify bridge damage caused due to strong seismic waves in which supervised machine learning algorithms performed well in analyzing damage caused on the surface. Five algorithms including Random Forest, Decision Tree, Logistic Regression, KNN, and XG Boost were applied to the dataset acquired from Github. Out of the applied algorithms, the decision tree predicted bridge damage with a mean accuracy of 96% along with higher precision and recall rate. The results were significantly better than the previous research done on the same dataset.

Another study provides a methodology to detect aftershocks after an earthquake of 7.3 magnitude hits the western part of Iran [38]. A binary approach is applied to accurately classify an aftershock and non-aftershock by considering two faults location out of four that got smaller Euclidean distance from the epicenter. Four machine learning algorithms including Naïve Bayes, Logistic Regression, KNN, and RBF were applied. Naïve Bayes got the highest accuracy of 78% whereas logistic regression obtained a higher AUC value equal to 0.85. In [39], volcanic eruption classification is performed to find useful hidden patterns from Nevado del Ruiz and Telica Volcano in order to differentiate a volcanic activity as eruptive or non-eruptive. Four supervised machine learning algorithms including SVM, Logistic Regression, Random Forest, and Gaussian Process Classifier were applied. SVM got 82.6% accuracy in Nevado data whereas Gaussian Process Classifier obtained 90.5% accuracy in the Telica dataset. The authors in [40] forecast earthquakes in Indian Subcontinent to avoid damage in the early stages through acquiring the dataset from USGS and the Indian meteorological department. Support vector regressor and random forest regressor along with ensemble stacking, bagging, and boosting techniques were applied. The applied ensemble techniques obtained accuracy that ranges between 74%-83%. However, ensemble stacking got the highest accuracy of 83%.

In [41], an earthquake-damaged building’s architectural safety is classified through Random forest and CART. Both algorithms were applied to predict whether the building is safe or not for living. The overall accuracy of both classifiers ranges between 90%- 91%. The patterns obtained from the proposed methodology could be utilized in order to get prior confirmation of the damaged building's safety status. In [42], laboratory-created acoustic experimented data that imitate earthquakes is predicted through the XGB algorithm. A six-fold cross-validation strategy was adopted to find out good mean absolute error for selecting optimal features. The result concludes that three features which include acoustic data, first amplitude value on sliding window and amplitude obtained the lowest mean absolute error value of 1.913 as compared to other feature combinations. Another interesting approach is presented in [43] to analyze residential building conditions after an earthquake. The engineers had inspected each building and assigned three tags based on the overall damage caused to the buildings. Unsafe buildings are assigned red and safe buildings are assigned green whereas yellow tag is given to buildings that can be reoccupied. For accurate classification linear discriminant analysis, KNN, Decision Tree, and Random Forest were applied. Random Forest got the highest accuracy of 66%. A similar approach for classifying damage can be applied to other natural disaster data to detect potential harm in the early stages.

The authors in [44] identified posttraumatic stress disorder (PTSD) in children due to earthquakes. Combinations of multiple factors including sleep cycle, analyzing mood, earthquake experience faced, and daily activities are utilized to identify potential posttraumatic stress disorder. The study identifies female adolescents are more vulnerable to stress due to emotional thinking mechanisms. The XGBoost classifier obtained an AUC value equal to 0.80 along with 74% accuracy. Authors in [45] presented similar work to this paper through magnitude conversion into binary classification through setting threshold greater than equal to 5.0 for differentiating positive and negative magnitude instances and selecting best-performing algorithms. Eight algorithms including Random forest, SVM, Naïve Bayes, Logistic regression, Adaboost, KNN, Multilayer perceptron, and CART are applied. Random forest got the highest accuracy of 76.97% whereas KNN got 75.53%. Another similar binary approach in [46] was applied to Hindukush region seismic activities after acquiring the dataset from USGS. Binary magnitude conversion is performed by setting a feasible threshold value. Tree-based algorithms are implemented which include decision tree, random forest, rotboost, and rotation forest. Rotation forest gained maximum performance by obtaining AUC and precision values as 95.9% and 90.5% respectively.

Similar binary seismic classification analysis performed on the global seismic dataset highlights the importance of distinguishing medium or big earthquakes after magnitude discretization [34]. The conversion was performed by considering magnitude ranging from 0 to 5.8 as medium and 5.8-10 as big scale earthquakes. Random forest and KNN were implemented for accurate comparative analysis. Random forest obtained 99% accuracy whereas KNN only obtained 55% accuracy.

The authors in [33] conducted related multiclass magnitude division through forecasting types of earthquakes mainly divided into fatal, moderate, and mild earthquakes. Considering regional magnitude value greater than 5.5 categorized as fatal whereas range from 4.5-5.5 as moderate and 2.5-4.5 as mild earthquakes. The study was conducted to avoid major disasters in India's six different regions by implementing seven algorithms including Random forest, Bayes net, Logistic regression, Simple logistic, Random tree, ZeroR, and LMT. Simple logistic regression and LMT achieved accuracy ranging from 98.18%-99.94% on multiple regions of India.

2.1 Limitations

After doing an in-depth literature review, some of the major observed limitations in performing evaluation on seismic data is:

1. Most of the machine learning algorithms applied on multiple small regions of a country are taken into account instead of considering the overall country's historic seismic occurrences. Overall country-based analysis can provide an in-depth insight into the overall magnitude intensity through active tectonic plates.
2. Only limited work is done on Pakistan and its neighboring countries' seismic analysis performed through implementing machine learning algorithms.
3. As evident from the literature review, cutoff magnitude is set based on the highest magnitude occurrence in a dataset for binary analysis. This technique ignores other major magnitude values causing major destruction on the surface of the earth which damages the infrastructure.
4. Most of the work is only focused on forecast and prediction performed through magnitude conversion to binary class or multiclass. However, we can also categorize

magnitude value in terms of the level of severity and the destruction capability each earthquake holds based on the Mercalli intensity scale.

CHAPTER 3

Methodology

This chapter provides a detailed explanation of the proposed methodology used for this research. Two seismic evaluation techniques are utilized. The chapter is divided into two sections; the first section provides the working of applied machine learning algorithms and the second section includes the adopted proposed methodology for seismic evaluation.

3.1 Applied Algorithms Description

For a fair comparison of machine learning algorithms on the acquired seismic data from the United States geological survey; five supervised machine learning algorithms are applied. The selected supervised machine learning algorithms working are given below.

3.1.1 Decision Tree

Decision tree is the simplest algorithm used for both classification and regression [47]. The decision tree split is dependent on split purity calculated through entropy, information gain, or Gini impurity [48]. Purity denotes that the split chosen has data samples belonging to only one class. The leaf nodes denote the class labels. Figure 7 [49] shows the general architecture of the decision tree algorithm.

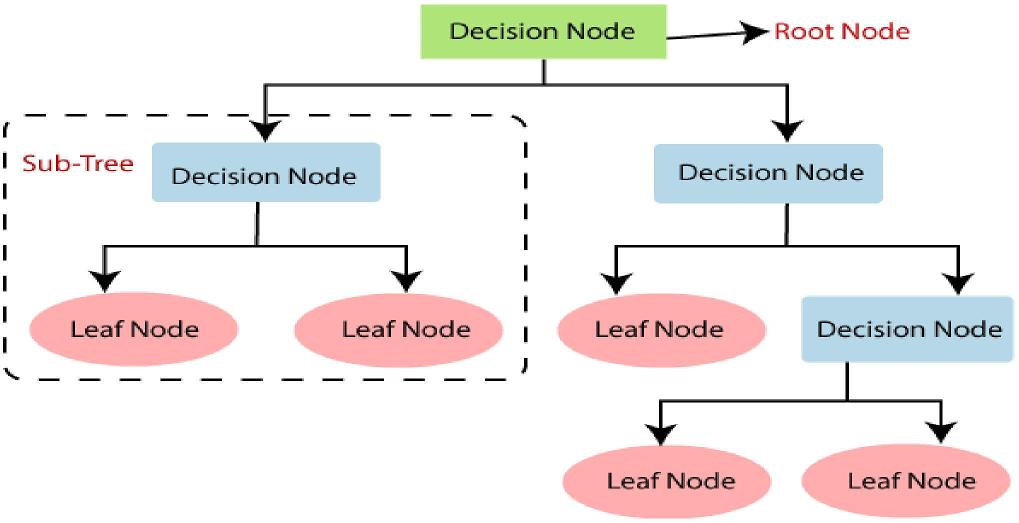


Figure 7: Decision Tree workflow

3.1.2 Random Forest

Random forest is an ensemble-based algorithm also known as a bootstrap aggregation that simply builds multiple decision trees in order to avoid high variance.

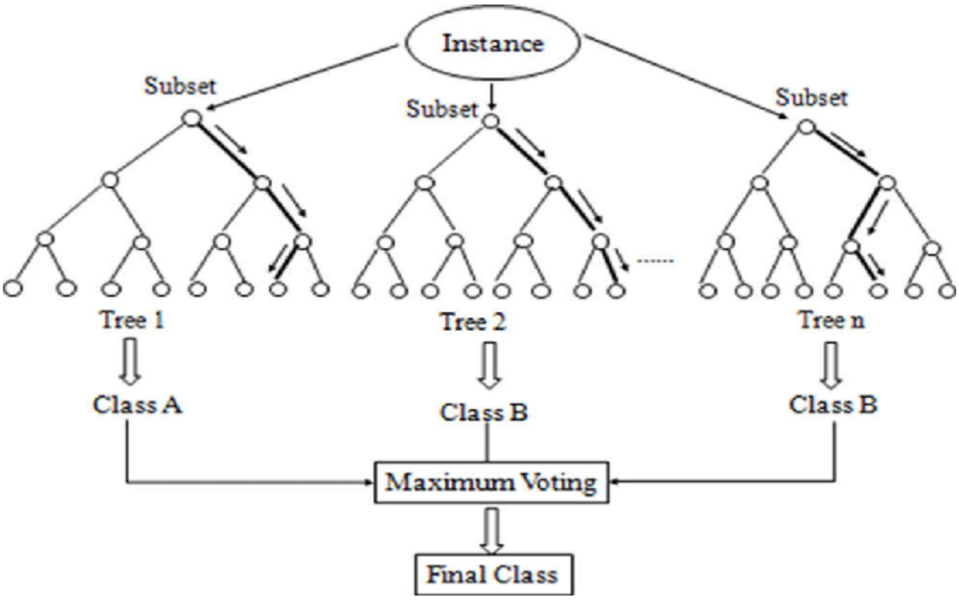


Figure 8: Random forest architecture

Figure 8 [50] depicts the overall working architecture of the algorithm. The training data instances are provided to multiple decision trees. The leaf node of all the decision trees contains the prediction of a specific class. In classification, maximum voting is considered for assigning a class label.

3.1.3 Adaboost Classifier

The adaptive boosting algorithm merges weak learners into a stronger ones. The basic idea behind the adaptive boosting is creating multiple decision stumps corresponding to each feature in the training data. After selecting the initial base model; number of incorrect observations are noted and total error value is calculated for the stump performance. Weights are updated according to the obtained performance which will eventually assign more value to incorrectly classified points and lesser values to correctly classified points.[51]

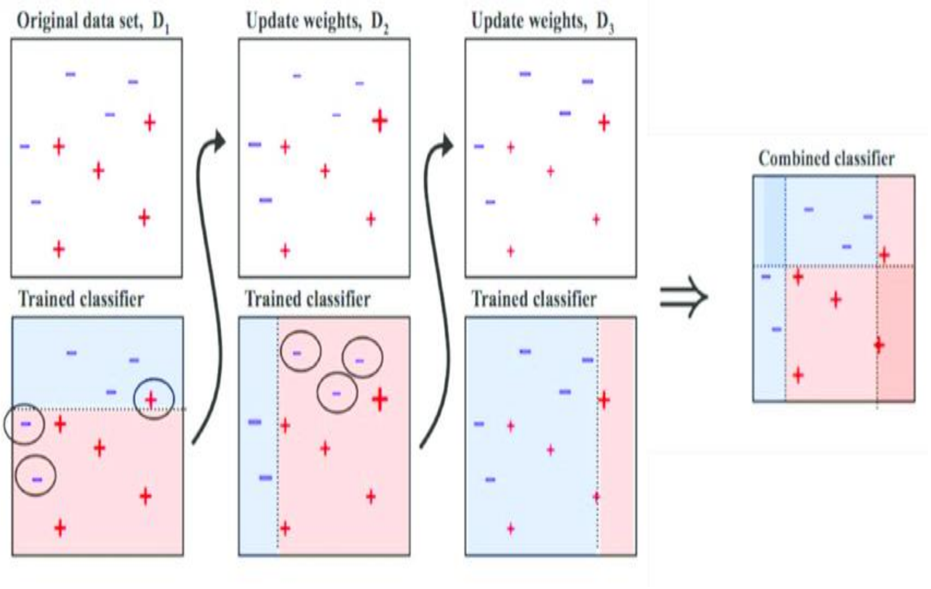


Figure 9: Adaboost classifier workflow

Figure 9 [52] demonstrates the sequential trained classifier along with the updated weights. The misclassification error obtained from the initial model is updated with new weights, and the process continues until every class is classified accurately.

3.1.4 XGBoost Classifier

Extreme gradient boosting classifier is one of the most powerful gradient boosting algorithm for regression and classification. It is an ensemble of decision trees that optimizes the loss function. Multiple decision trees are constructed as base learners to classify the dependent feature by calculating the similarity weight and gain of each split. The output of multiple base models is combined together for efficient prediction. Regularization parameter controls overfitting [53].

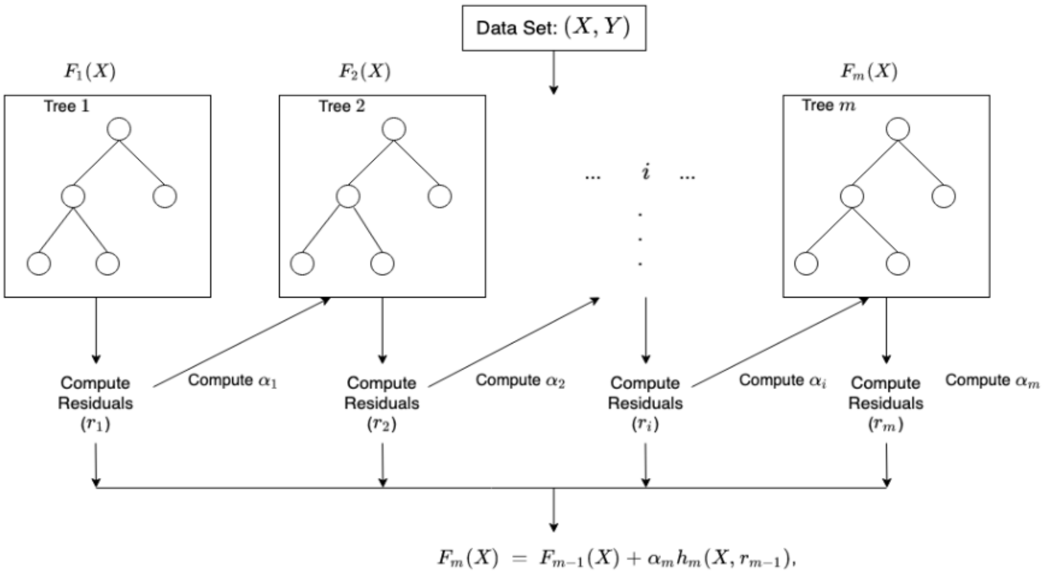

Data Set: (X, Y)
$F_1(X)$
Tree 1
$F_2(X)$
Tree 2
$F_m(X)$
Tree m
... i ...
Compute Residuals (r_1)
Compute α_1
Compute Residuals (r_2)
Compute α_2
Compute Residuals (r_i)
Compute α_i
Compute Residuals (r_m)
Compute α_m
$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$

Figure 10: XGB classifier workflow

Figure 10 [54] illustrates the overall working of the XGBoost algorithm which constructs the decision tree sequentially whereas each decision tree generates residual errors that denote the total loss of each base model.

3.1.5 K-nearest Neighbors

K-nearest neighbors is the simplest supervised classification algorithm based on the simplest nonparametric approach that emphasizes that the same class data points are close to each other. It can be used for binary along with multi-class classification problems due to its versatile nature. By selecting an optimal value of K, the data points are assigned to their nearest neighbor that is at a minimum distance.

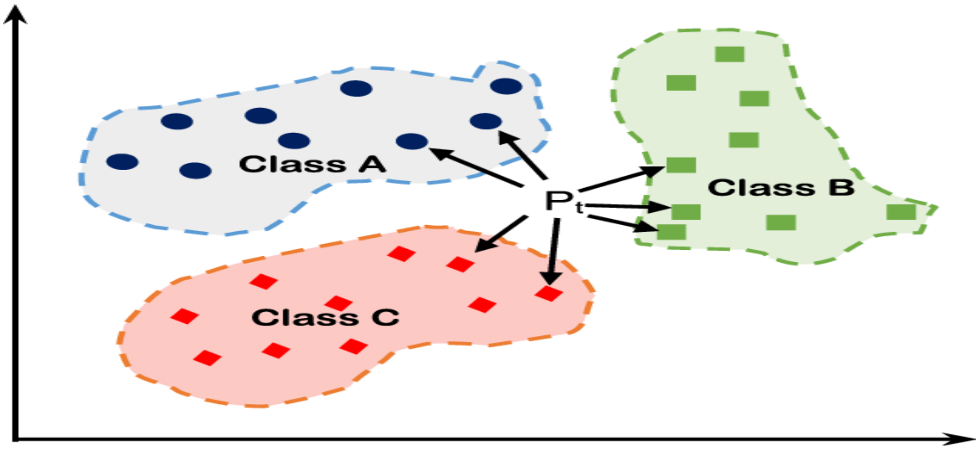


Figure 11: KNN workflow

Figure 11 [55] shows a multiclass classification problem having data points distributed in three classes. Accuracy is mostly improved in KNN if standardization or normalization is performed.

3.2 Proposed Approach

In this section, the proposed approach for seismic data evaluation is discussed along with the acquired seismic dataset and applied supervised machine learning algorithms. The proposed methodology includes six major stages, each containing multiple steps. Figure 12 illustrates all the necessary steps taken in every stage.

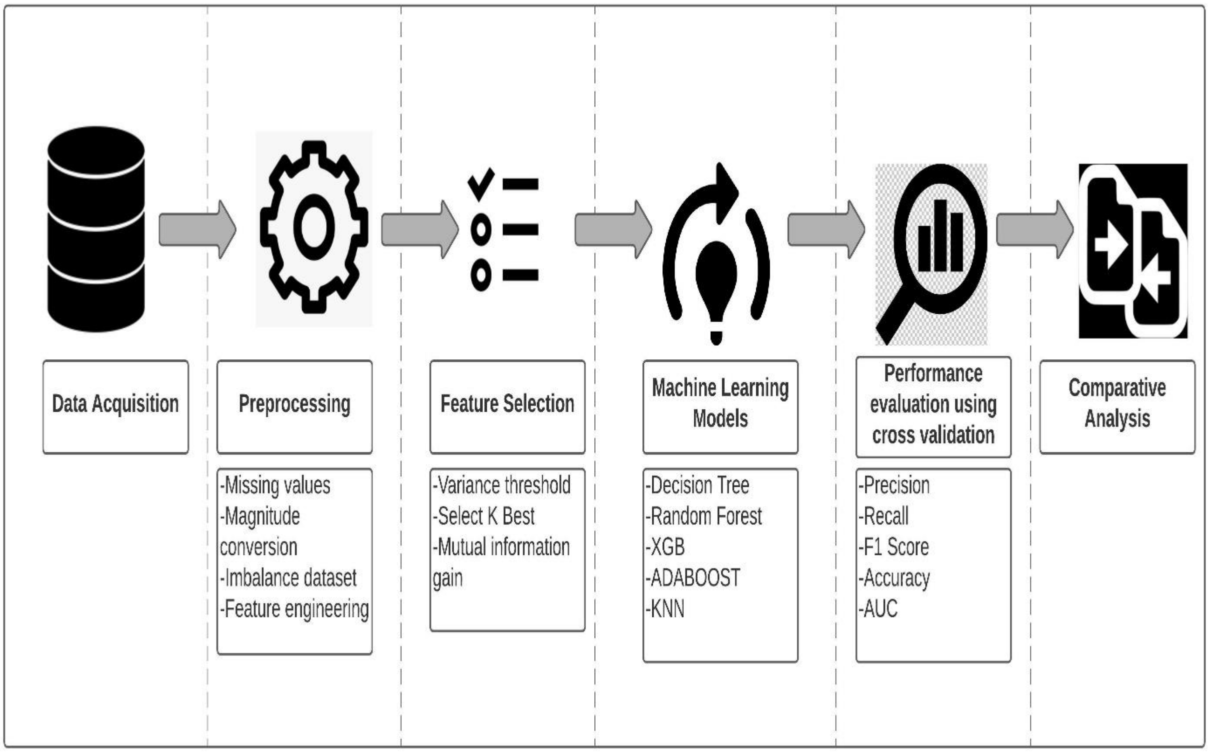


Figure 12: Proposed methodology

3.2.1 Data Acquisition

The seismic dataset is acquired from the United States Geological Survey (USGS) earthquake catalog [56]. The historic seismic activity records range from 1990 to 2020. The major countries include Pakistan, Afghanistan, China, and Japan.

3.2.2 Data pre-processing Data pre-processing is an essential step in machine learning as data quality directly influences the applied algorithm learning capability. Real-world datasets have missing values and categorical features that must be cleaned and formatted to obtain good accuracy [57]. Another essential step in pre-processing is scaling. Features scaling is an important aspect as the range of features in USGS data varies in their units of measurement as well as their magnitude.

3.2.2.1 Dealing with outliers

The interquartile range is used to detect outliers and visualize the spread of data as well. Outliers are mostly deleted from the dataset but since we have smaller datasets; treating the outliers by imputing it is a better option as to not lose any data [58].

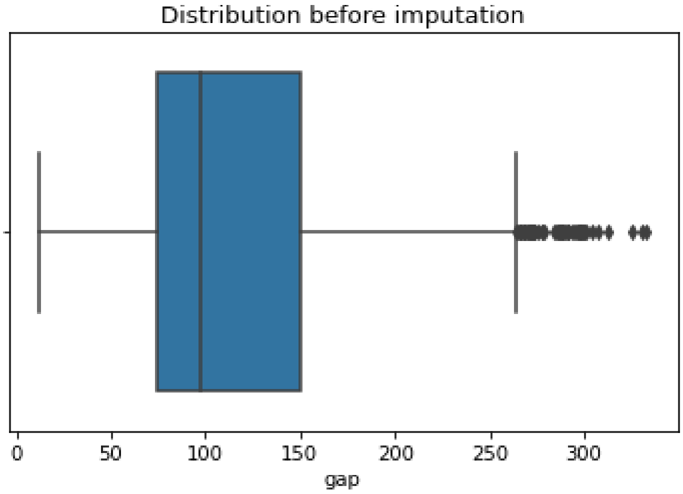


Figure 13: Boxplot of gap feature

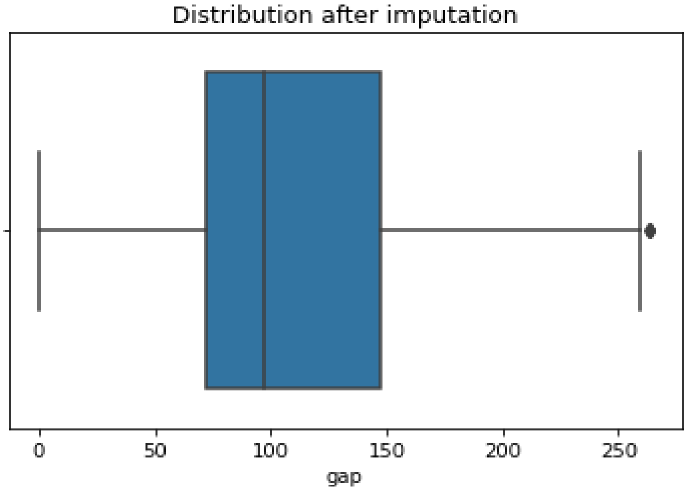


Figure 14: Boxplot after imputation

Figure 13 is the boxplot of the gap feature containing outliers after the max value whisker. This feature is imputed with zero value because its original format recorded by the catalog is in degrees ranging from 0.0 to 180 respectively. Figure 14 shows the distribution of data after imputation.

3.2.2.2 Handling missing values

There might be various reasons behind missing values which includes data entry error or problem during data gathering. However, the performance of the algorithms is highly dependent on how we handle missing values present in the dataset [59]. To not lose any information from the seismic dataset, data imputation methods were applied.

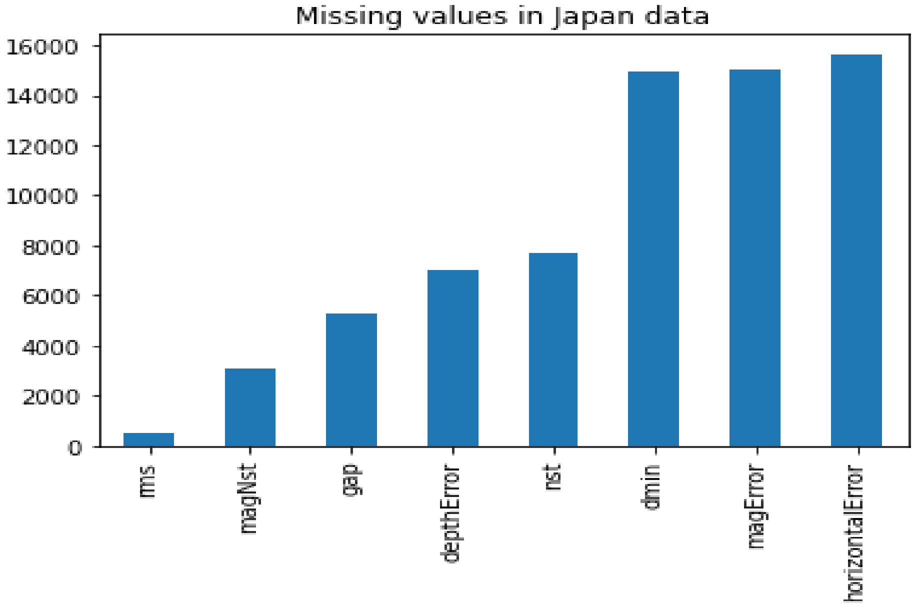


Figure 15: Missing values histogram

Figure 15 shows the bar plot of missing features values in the Japan dataset. These are the eight features that contain missing values in all the six datasets utilized in this thesis. The applied methods for data imputation are as follows:

1. Mean imputation: It is the simplest method to deal with missing values. It substitutes the mean of the available values on missing values. The advantage of mean imputation is that it maintains the size of the dataset.
2. Mode imputation: This method substitutes the missing values with the most frequent value present in the missing features columns. The mode imputation is a better method as it considers the maximum values present in the data.

3.2.2.3 Country wise evaluation criteria

The two seismic data evaluation criteria adopted in the thesis are as follows:

1. Binary classification: The magnitude value ranges from 3.5 to 10.

2. Multiclass classification: Categorization of the magnitude values into their appropriate class that corresponds to the strength and effect of the seismic waves on the surface. For better categorization, the seismic magnitude range is from 0-10 which contains body waves as well as surface waves.

3.2.2.3.1 Evaluation methods implementation

1. Setting the average magnitude value as a threshold for conversion into binary classification. The magnitude value greater than and equal to the obtained average is set to 1 else 0.
2. The minimum magnitude in collective data is set as the initial starting point for categorization. The largest magnitude varies in all the datasets that’s why it’s different from each other. The example of categorization in the Pakistan and China dataset is set as values ranging from 2.8-4.0 are mild earthquakes. Whereas magnitude greater than 4.0 till 4.4 are major earthquakes. And lastly values greater than 4.4 till 7.9 are destructive earthquakes.

Both the binary and multi-class categorization will provide an in-depth evaluation of country-wise seismic datasets.

3.2.2.4 Handling categorical features

Handling categorical features of a dataset plays a significant role in preprocessing step. Category label encoding is applied to convert the object feature into its appropriate numeric machine-understandable format.

3.2.2.5 Handling imbalanced dataset

Imbalanced dataset problem arises after magnitude categorization in multiclass japan evaluation. Handling imbalanced datasets is a crucial component of seismic evaluation to avoid bias towards the majority class. Random oversampling technique is adopted to handle the imbalanced dataset. In random oversampling, the minority class in the data is balanced by randomly replicating the samples in the training dataset [60]. By applying the oversampling technique to balance the class distribution, the results obtained were significantly better than other available balancing techniques.

3.2.3 Feature Selection

Feature selection is one of the fundamental steps in machine learning as the performance of the applied models is heavily dependent on it. Selecting fewer features from the dataset while achieving good performance is the goal of feature selection. The adopted algorithms for feature selection are as follows:

1. Removal of constant features by applying a feature selector variance threshold algorithm to get rid of all those features present in the dataset that have zero variance. As the features having less variance carry little to no information, it contains exact same value in the entire dataset.
2. Mutual info classification algorithm is applied which calculates the dependency among the features. The value is zero when the features are independent whereas the greater value depicts high dependability. Mutual information is an interchangeable term of information gain which calculates how much information is obtained from a variable given another.
3. Top ten features are selected by specifying k value equal to 10 in SelectKBest along with mutual information algorithm scoring function.

3.2.4 Model training and selection

The applied machine learning algorithms are trained on 80% of the dataset and the remaining 20% is set as testing data. K-fold cross-validation technique is applied to obtain a generalized model. Only those models are selected which obtained the best cross-validation accuracy, precision, recall, f1-score, and AUC.

Machine learning algorithm’s performance is highly dependent on selecting the best hyperparameters. Hyperparameter tuning is performed through Randomized Search which selects random combinations from the parameter distribution. In this thesis, Randomized Search is used for tuning hyperparameters instead of Grid Search as it gets better results in the lower-dimensional datasets [61].

3.2.5 Evaluation Metrics

The obtained results are evaluated based on multiple evaluation metrics which include area under the receiver operating characteristics, precision, recall, and f1-score. These performance metrics will provide an insight into the performance of machine learning algorithms. Precision indicates the percentage of actual positive values predicted by each classifier. Whereas recall indicates how many positive values are predicted correctly. Precision takes into account the type 1 error that is false positive rate whereas recall considers type 2 error that is false-negative rate. The formula for calculating precision and recall is given in figure 16.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Figure 16: Precision and Recall formula

To get good precision and recall score, the number of false-positive and false-negative should be minimum.

Another metric which is the f1-score considers both false positives and false negatives. It’s basically an amalgamation of recall and precision and takes a harmonic mean of both recall and precision as seen in figure 17.

$$\text{F1 Score} = \frac{2 \text{ x } (\text{Precision x Recall})}{\text{Precision + Recall}}$$

Figure 17: F1 score formula

The f1-score is a more practical measure as compared to accuracy and is considered mostly for measuring the quality of the classifier. Classifiers that obtained a higher f1-score is considered good in term of classification capability.

The most significant component for comparing the classifier’s performance is through plotting the AUC-ROC curve in binary as well as multiclass classification. It basically denotes the capability of an applied model in distinguishing classes at multiple threshold values.

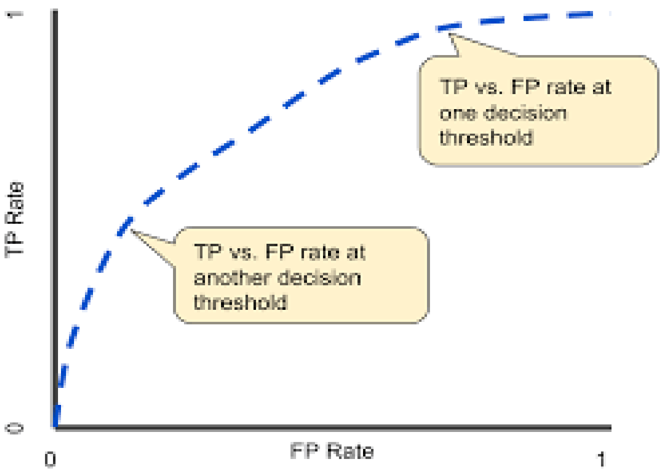

1
TP Rate
0
TP vs. FP rate at one decision threshold
TP vs. FP rate at another decision threshold
0
FP Rate
1

Figure 18: AUC-ROC curve structure

In Figure 18, the x-axis denotes the false positive rate whereas the true positive rate is on the y-axis. The rate of true positive value and false-positive value of the applied algorithm at multiple threshold levels is denoted by the curve. An optimal model covers more area under the curve and distinguishes between the classes efficiently by keeping the rate of true-positive higher than the false-positive rate.

3.2.6 Comparative Analysis

The end goal is to perform a comparative analysis of all the seismic datasets to compare the performance of applied supervised machine learning algorithms. The results are based on multiple metrics which involve area under the curve, accuracy, precision, recall, and f1-score obtained through 5-fold cross-validation. Through these evaluation metrics, the outcome of applied algorithms can be better analyzed and won’t be dependent on only a single performance measure.

CHAPTER 4

Results

This chapter provides seismic evaluation results obtained from the selected countries' datasets.

4.1 Dataset Description

The dataset is acquired from the United States Geological Survey (USGS) Agency located in the United States that collects natural hazards data. The data can be easily acquired from the geographical earthquake USGS catalog by mentioning the countries as well as the range of magnitude value required [56]. The time series seismic record consists of a total of 22 features and their description is provided in table 2.

Features	Description
time	The earthquake occurrence time recorded in date and time both in UTC format. The initial rupture time also known as the origin time recorded through the seismographs.
latitude	To locate an earthquake epicenter coordinates the latitude and longitude is used. It's basically division of the earth from the equator in to two parts which is north and south. The degrees ranges from -90 to +90 which differentiate the two hemispheres.
longitude	The division of globe in to east and west is done from the center vertically which is known as the prime meridian. The range of the longitude is from -180 to +180, in which the eastern part is taken as positive and western as negative.
depth	The recorded seismic activity recorded depth in kilometers. It denotes how deep the earthquake is from the epicenter which signifies the intensity. Shallow earthquakes which ranges from 0-70 kilometers which are considered destructive. Intermediate earthquakes which ranges from 70 to 300 km and deep earthquakes which ranges from 300 to 700 km. The USGS earthquake catalog records depth from 0 to 1000 km.
mag	The recorded Richter's magnitude scale of earthquake that ranges from 0 to 10.

magType	The selected method for calculating the earthquake magnitude.
Nst	The total sum of earthquake stations near the epicenter which reported the location.
Gap	The distance recorded as an azimuthal gap between the stations. The smaller the reported degree, more reliable is the reported location.
dmin	The minimum distance recorded in degrees between the epicenter and the nearby stations. The total distance of 1 degree is equal to 111.2 km approximately.
rms	Root mean square which is the surplus value obtained in seconds and basically denotes the observed fit of the reported and predicted time of earthquake.
Net	The original network source of the reported seismic activity.
Id	The identifier assigned to the event which differentiates every natural hazard accurately.
updated	It basically denotes the latest updation done on the recorded event in case if there's any error in the reported time initially.
place	The reported place of the earthquake.
type	The type of hazardous event which in this case is either earthquake or a quarry.
horizontalError	The seismic location uncertainty recorded in kilometers ranging from 0 to 100.
depthError	The seismic depth uncertainty recorded in kilometers ranging from 0 to 100.
magError	The seismic magnitude uncertainty recorded in the data.
magNst	The total sum of stations which calculates the event magnitude.
status	The status represented by three values which are deleted, automatic and reviewed that basically denotes whether the reported seismic event is verified by a human or it's recorded automatically.
locationSource	The authorized source of the epicenter location.
magSource	The authorized source of the magnitude of the seismic activity.

Table 2: Dataset description

Other than the described features in table 2, four more features are extracted from the time feature which are month, year, weekday, and day. These four features are retrieved to perform exploratory data analysis by plotting graphs and pie charts to visualize the relationship between the features. Moreover, month-wise and day-wise percentages of seismic activity are visualized to compare which country encounters more earthquakes in the past 31 years of historic data.

Another reason for discarding the retrieved feature is to avoid overfitting since the time feature is already been selected during feature selection. Table 3 shows the selected ten features for evaluation.

No.	Features
1	time
2	depth
3	Nst
4	Gap
5	Rms
6	Depth error
7	Mag error
8	Mag nst
9	Mag type
10	Mag source

Table 3: Selected features

4.1.1 Specified Region

The collective datasets include four main countries includes Pakistan, Afghanistan, China, and Japan. Collective 31 years of seismic datasets include combined data of China and Pakistan, Afghanistan and Pakistan, and the Japan dataset is taken alone.

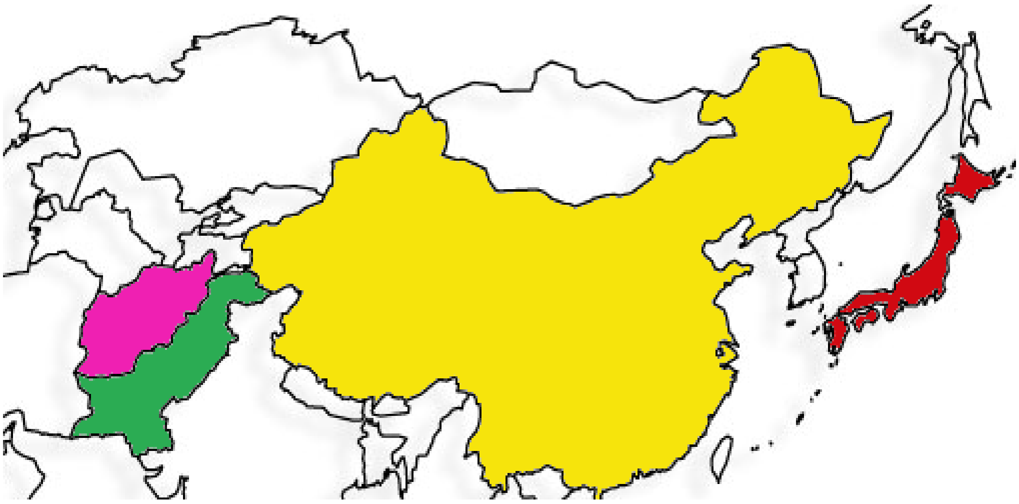


Figure 19: Selected countries for seismic evaluation

Figure 19 highlights all the main countries selected for historic earthquake data evaluation.

4.1.2 Country-wise seismic data distribution

The data size for each country in binary classification seismic data is presented in table 4.

Country	Total Volume	Distribution
Pakistan and Afghanistan	6917	Pakistan:2177 Afghanistan:4740
Pakistan and China	9785	China:7371 Pakistan:2414
Japan	19090	___

Table 4: Country-wise binary data distribution

Country-wise data size in multiclass seismic evaluation is given in table 5.

Country	Total Volume	Distribution
Pakistan and Afghanistan	7399	Pakistan:2343 Afghanistan:5056
Pakistan and China	8848	China:6540 Pakistan:2308
Japan	18628	___

Table 5:Country-wise multiclass data distribution

The collective dataset is acquired to analyze the total seismic activity of Pakistan's neighboring countries. And since the border sharing countries face more seismic activity in all regions on yearly basis, combined time-series data provides insight into the total tectonic movement in Pakistan as well.

4.2 Exploratory Data Analysis

This section presents an in-depth country-wise exploratory data analysis. Seismic time series data provides descriptive statistics about the tectonic movement over a period of time. Earthquake data is interpreted with line plots, pie charts, and histograms to visualize the underlying patterns.

4.2.1 Binary classification seismic data

Richter scale magnitude average value is taken as the standard threshold for binary classification. Magnitude greater than average is considered as class 1 and others as class 0. Through binary conversion, we can get a frequent average magnitude that hits each country in the past 31 years of the historic seismic record. The average earthquake threshold obtained in all the specified countries ranges between 4.29-4.49 that is of medium intensity according to the Mercalli intensity scale and carries enough potential to cause minor damage.

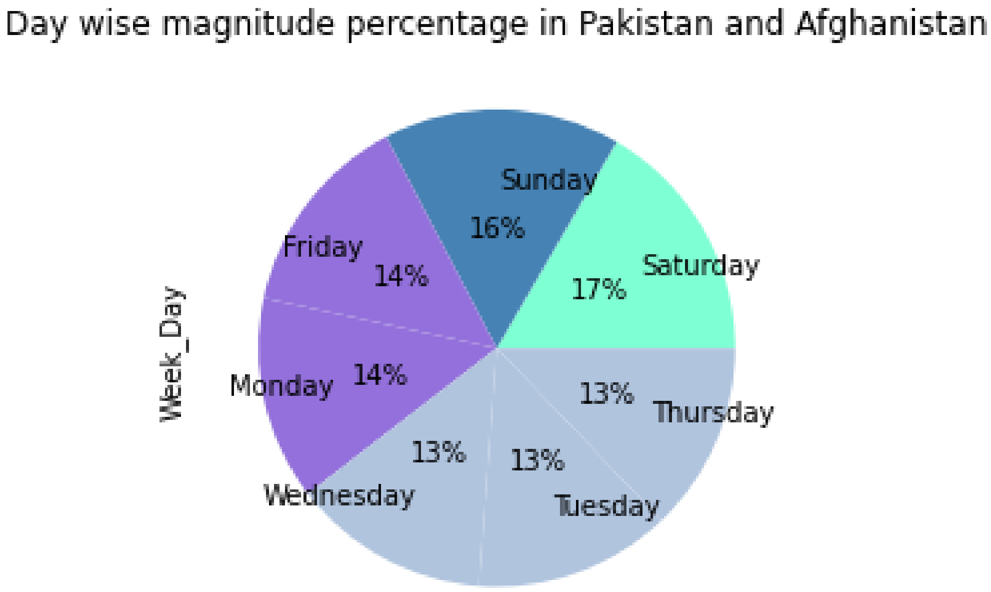


Figure 20: Day wise Pie-chart of Pakistan and Afghanistan

Figure 20 shows the day-wise seismic activity percentage in Afghanistan and Pakistan in 31 years of historic data. Approximately 17% of earthquake hits on Saturday whereas the least amount of seismic activity occurred on Tuesday, Wednesday, and Thursday with a total of 13% only.

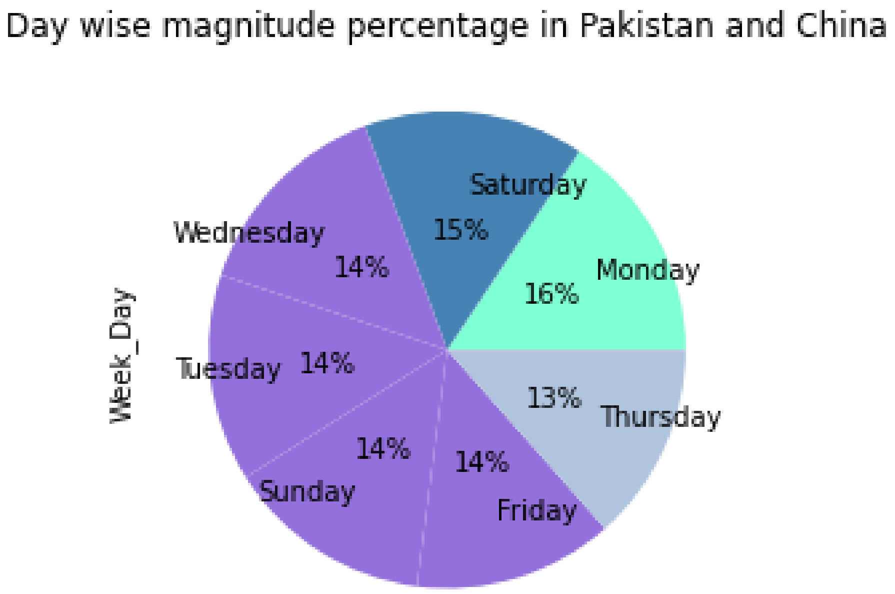


Figure 21: Day wise Pie-chart of Pakistan and China

Figure 21 shows most of the seismic activity towards the northeast of Pakistan occurs on Monday with the highest 16% percentage. Only 13% of seismic waves are recorded on Thursday.

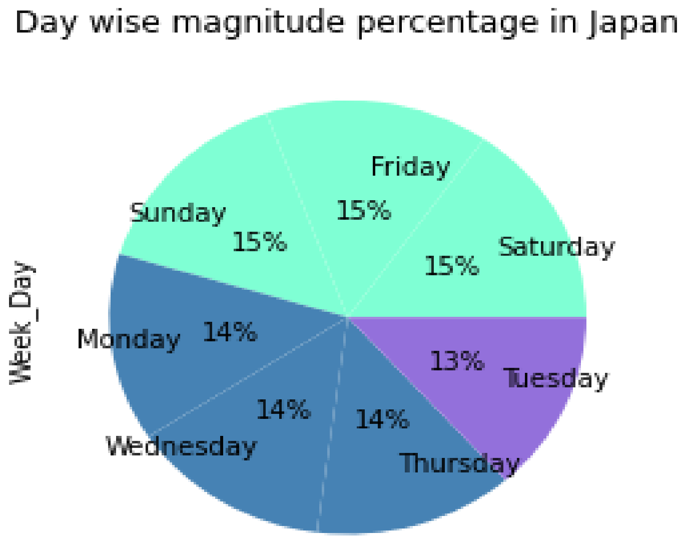


Figure 22: Day wise Pie-chart of Japan

Figure 22 depicts Japan’s highest seismic activity is recorded on Friday, Saturday, and Sunday with a total of 15% seismic occurrences all over the country.

These pie charts further depict that frequent seismic activity is mostly recorded on Saturdays.

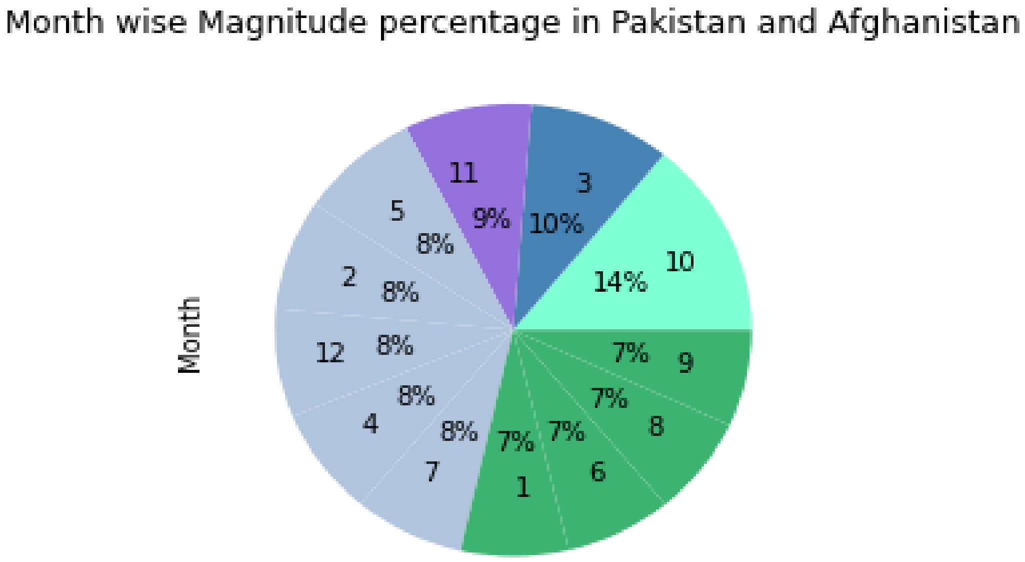


Figure 23: Month wise pie-chart of Pakistan and Afghanistan

Figure 23 presents month-wise seismic activity through which we have further gained insight that 14% of earthquake hits in October followed by March with 10% of the total seismic record.

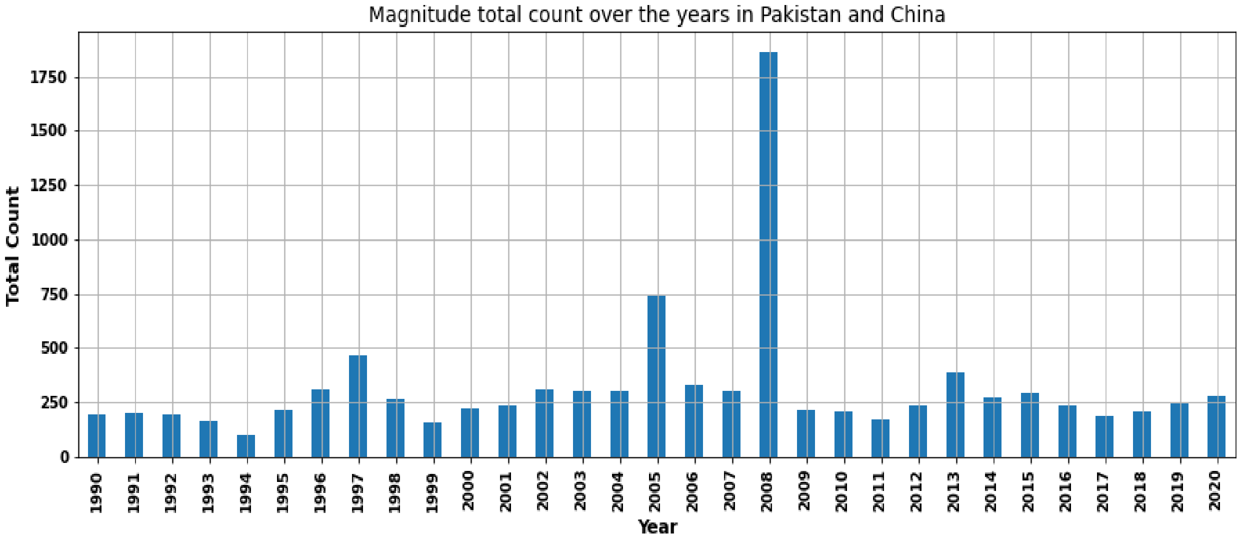


Figure 24: Year wise magnitude count in Pakistan and China

Figure 24 shows year-wise magnitude count in Pakistan and China seismic data. The highest magnitude count is recorded in the year 2008 with more than 1750 earthquakes.

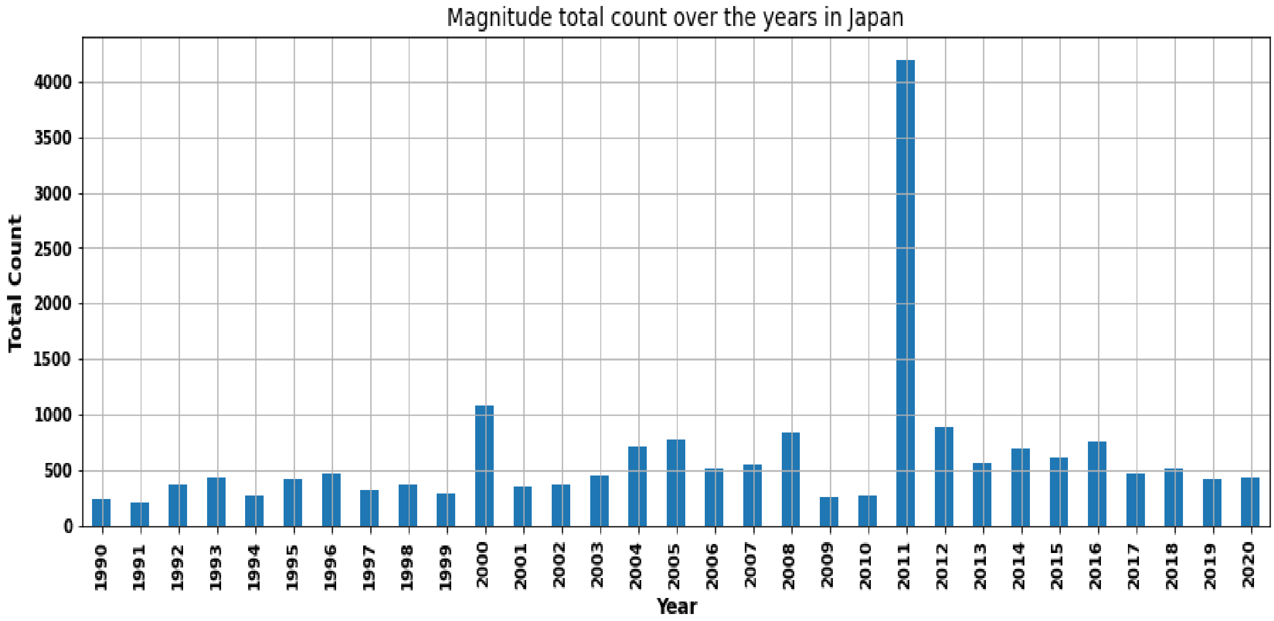


Figure 25: Year wise magnitude count in Japan

In Japan, more than four thousand earthquakes occurred in 2011 as shown in Figure 25.

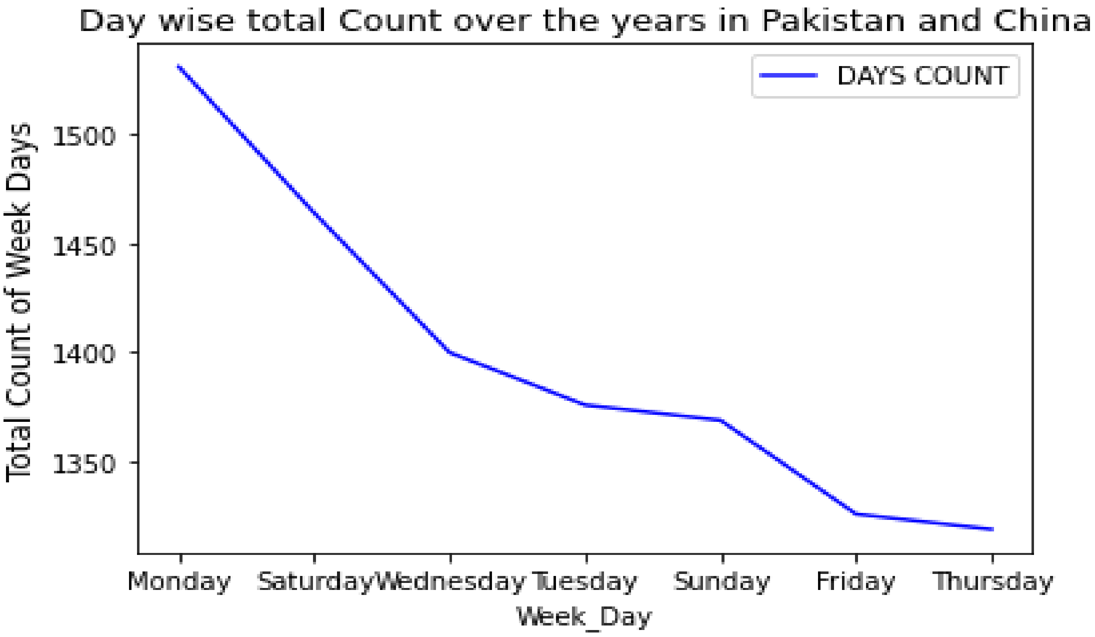


Figure 26: Day wise seismic count in Pakistan and China

A line plot is retrieved to visualize total day-wise seismic occurrences shown in Figure 26. This is an extension to the pie chart shown in Figure 21, instead, here we visualize the total count on the y-axis. Monday got most of the seismic occurrences followed by Saturday.

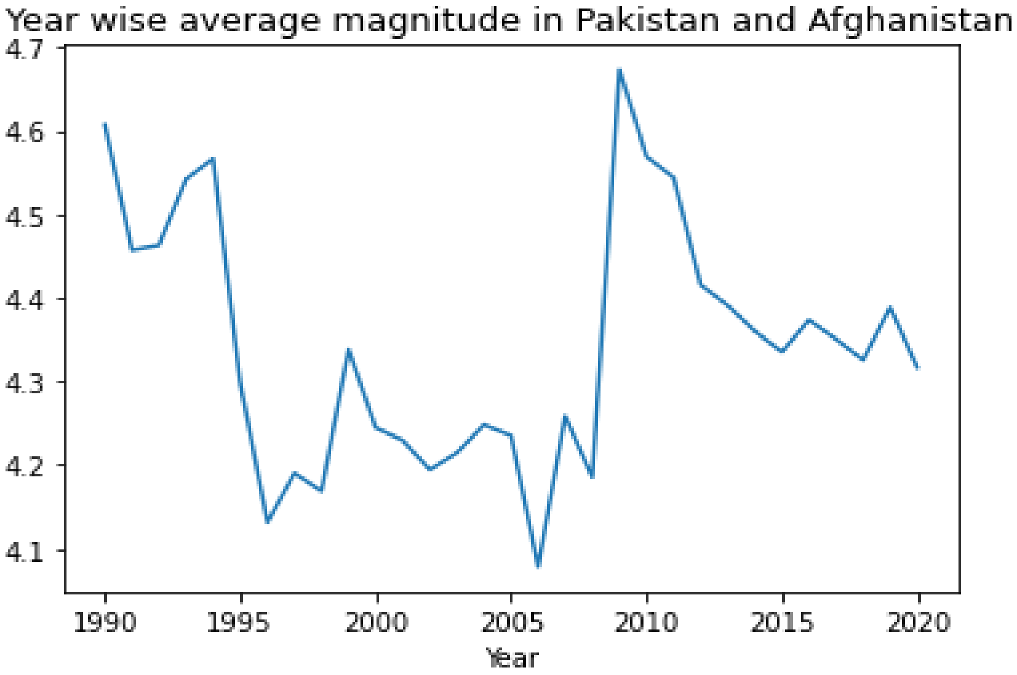


Figure 27: Average year wise magnitude in Pakistan and Afghanistan

Figure 27 shows the average magnitude with respect to every year that ranges between 4.1-4.7. In all the binary seismic datasets, the highest average is recorded in the years 1990 and 2010.

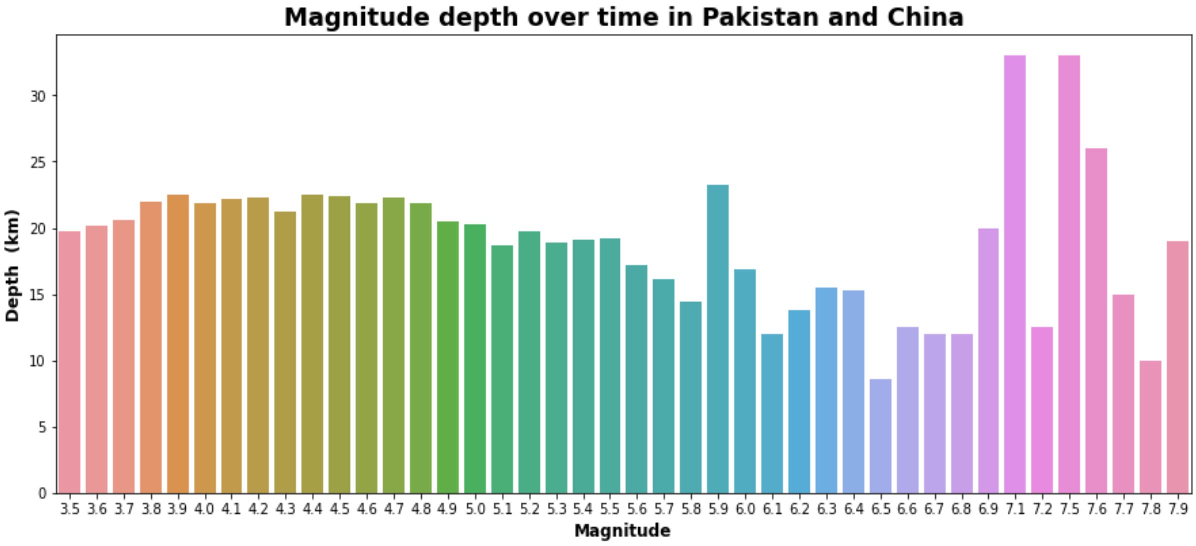


Figure 28: Depth of recorded magnitude in Pakistan and China

Depth is recorded in kilometers which shows how far the seismic waves travel from the epicenter. Figure 28 shows that the depth of 7.1 and 7.5 magnitude earthquake covers more than 35 kilometers. It's categorized as a shallow destructive earthquake because the seismic wave's intensity doesn't disperse farther into the earth's surface. The least distance from the epicenter to the hypocenter, the more destruction is caused on the surface due to its strong seismic intensity [62].

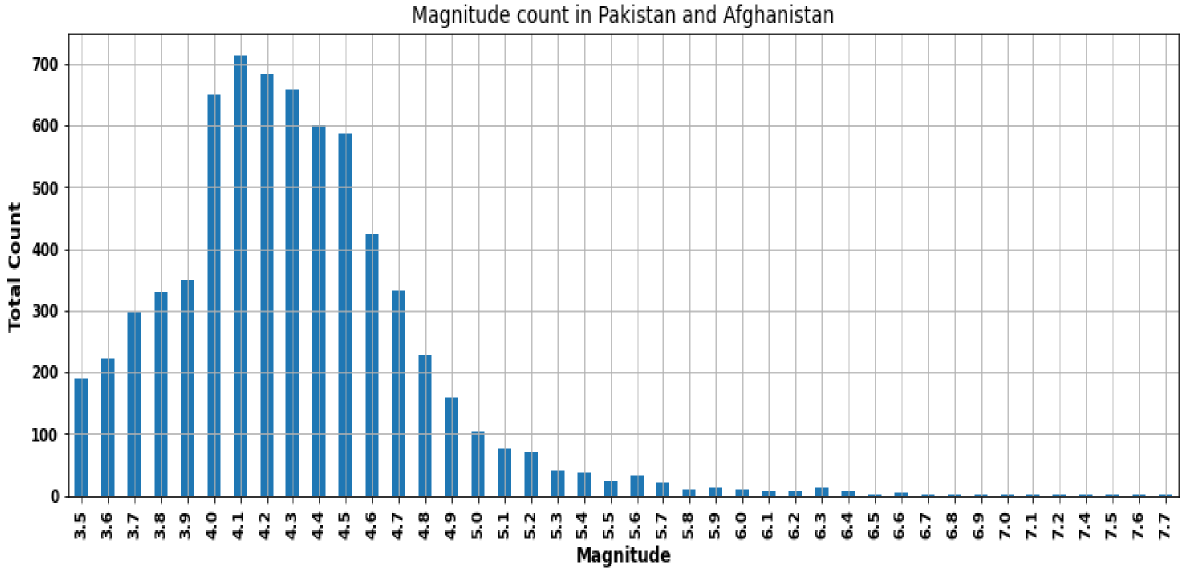
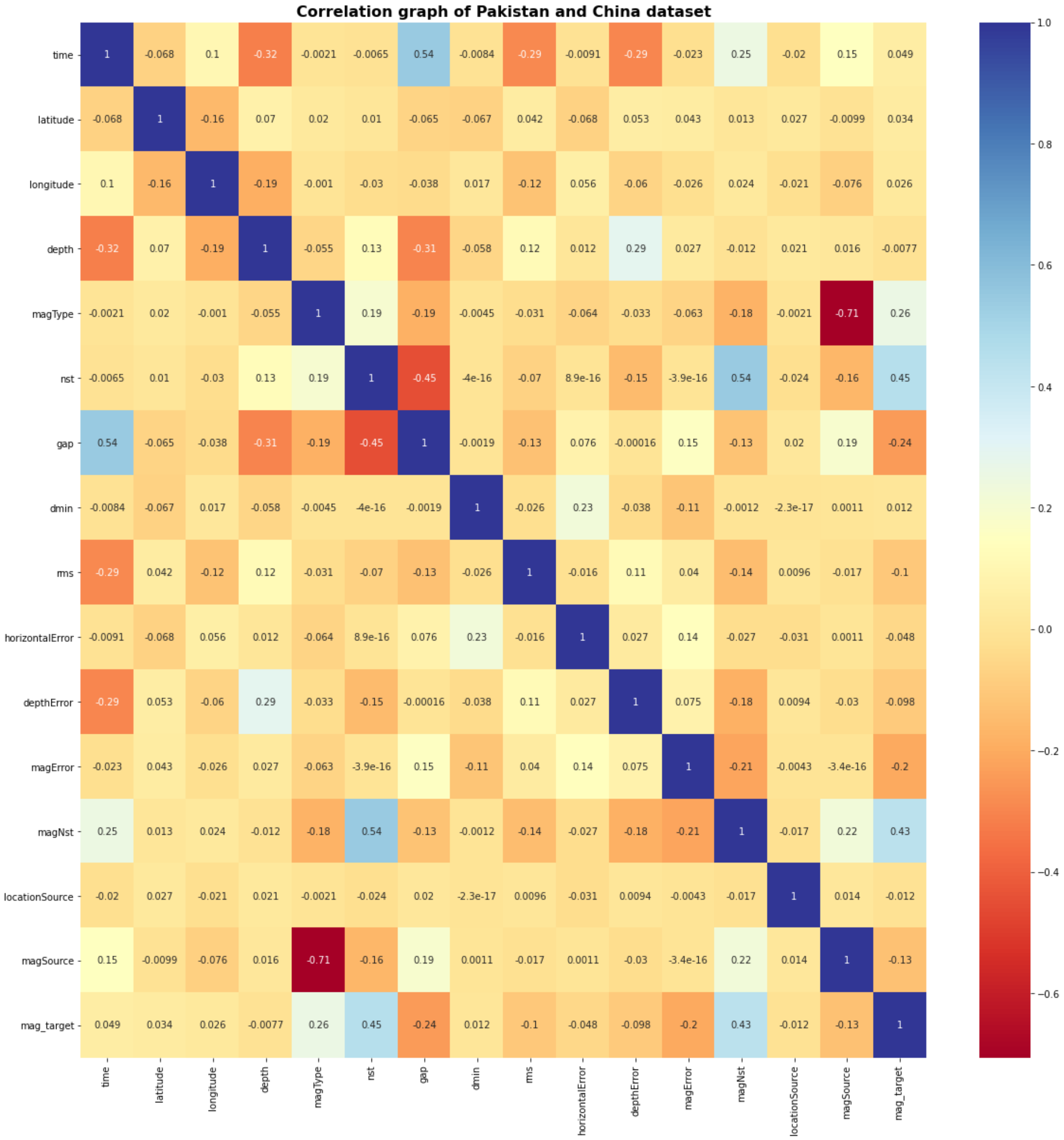


Figure 29: Magnitude count plot of binary Pakistan and Afghanistan data

The histogram in figure 29 shows the magnitude values on the x-axis and their corresponding count on the y-axis. As the binary threshold evaluation magnitude starts from 3.5, it's the same for all binary datasets whereas the last highest recorded magnitude varies. The ending magnitude in Pakistan and Afghanistan data is 7.7 as shown in figure 28, whereas in the other two datasets the highest recorded magnitude is equal to 7.9. This plot further depicted that both the neighboring countries of Pakistan and Japan faces a few damaging seismic waves in past.



	time	latitude	longitude	depth	magType	nst	gap	dmin	rms	horizontalError	depthError	magError	magNst	locationSource	magSource	mag_target
time	1	-0.068	0.1	-0.32	-0.0021	-0.0065	0.54	-0.0084	-0.29	-0.0091	-0.29	-0.023	0.25	-0.02	0.15	0.049
latitude	-0.068	1	-0.16	0.07	0.02	0.01	-0.065	-0.067	0.042	-0.068	0.053	0.043	0.013	0.027	-0.0099	0.034
longitude	0.1	-0.16	1	-0.19	-0.001	-0.03	-0.038	0.017	-0.12	0.056	-0.06	-0.026	0.024	-0.021	-0.076	0.026
depth	-0.32	0.07	-0.19	1	-0.055	0.13	-0.31	-0.058	0.12	0.012	0.29	0.027	-0.012	0.021	0.016	-0.0077
magType	-0.0021	0.02	-0.001	-0.055	1	0.19	-0.19	-0.0045	-0.031	-0.064	-0.033	-0.063	-0.18	-0.0021	-0.71	0.26
nst	-0.0065	0.01	-0.03	0.13	0.19	1	-0.45	-4e-16	-0.07	8.9e-16	-0.15	-3.9e-16	0.54	-0.024	-0.16	0.45
gap	0.54	-0.065	-0.038	-0.31	-0.19	-0.45	1	-0.0019	-0.13	0.076	-0.00016	0.15	-0.13	0.02	0.19	-0.24
dmin	-0.0084	-0.067	0.017	-0.058	-0.0045	-4e-16	-0.0019	1	-0.026	0.23	-0.038	-0.11	-0.0012	-2.3e-17	0.0011	0.012
rms	-0.29	0.042	-0.12	0.12	-0.031	-0.07	-0.13	-0.026	1	-0.016	0.11	0.04	-0.14	0.0096	-0.017	-0.1
horizontalError	-0.0091	-0.068	0.056	0.012	-0.064	8.9e-16	0.076	0.23	-0.016	1	0.027	0.14	-0.027	-0.031	0.0011	-0.048
depthError	-0.29	0.053	-0.06	0.29	-0.033	-0.15	-0.00016	-0.038	0.11	0.027	1	0.075	-0.18	0.0094	-0.03	-0.098
magError	-0.023	0.043	-0.026	0.027	-0.063	-3.9e-16	0.15	-0.11	0.04	0.14	0.075	1	-0.21	-0.0043	-3.4e-16	-0.2
magNst	0.25	0.013	0.024	-0.012	-0.18	0.54	-0.13	-0.0012	-0.14	-0.027	-0.18	-0.21	1	-0.017	0.22	0.43
locationSource	-0.02	0.027	-0.021	0.021	-0.0021	-0.024	0.02	-2.3e-17	0.0096	-0.031	0.0094	-0.0043	-0.017	1	0.014	-0.012
magSource	0.15	-0.0099	-0.076	0.016	-0.71	-0.16	0.19	0.0011	-0.017	0.0011	-0.03	-3.4e-16	0.22	0.014	1	-0.13
mag_target	0.049	0.034	0.026	-0.0077	0.26	0.45	-0.24	0.012	-0.1	-0.048	-0.098	-0.2	0.43	-0.012	-0.13	1

Figure 30: Correlation graph

Figure 30 is the correlation graph of Pakistan and China earthquake data. The correlation matrix shows that there’s a positive correlation between the features time and gap, nst and magnst, depth and deptherror. Whereas little negative correlation between magtype and magSource, nst and gap, rms and magnst, magSource and magtype. And few of the features in the dataset have no correlation which is close to 0.

4.2.2 Multiclass classification seismic data

This section includes a histogram plot of multiclass seismic data. Most of the plots in multiclass seismic evaluation are the same as the binary evaluation because of the exact year range. The only difference between the two evaluation methods is the magnitude range that's why only one plot is included which varies from the previous magnitude plot figure 28.

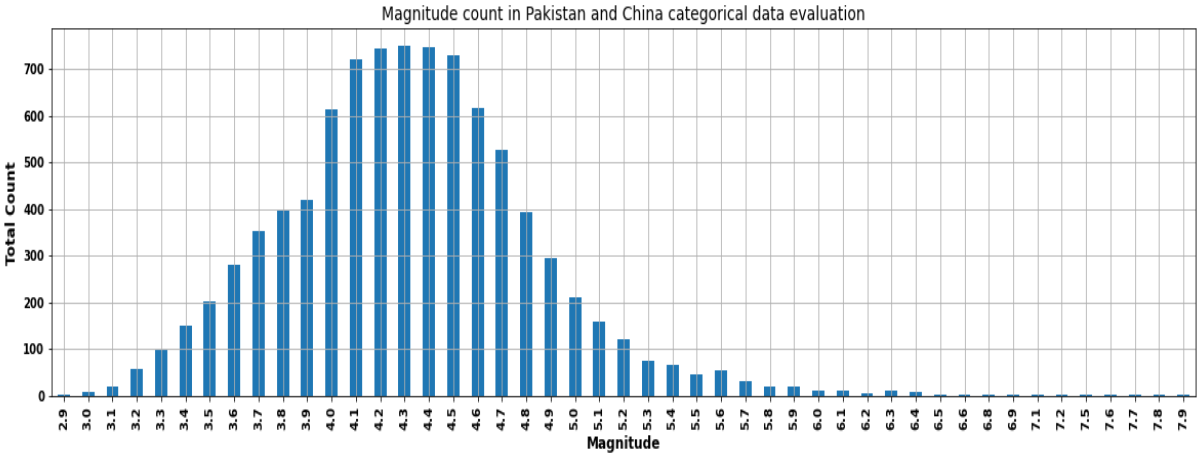


Figure 31: Magnitude count plot of categorical evaluation

Figure 31 shows the magnitude count histogram of Pakistan and China categorical evaluation data which ranges from 2.9 to 7.9. As the selected time is the same, the highest recorded magnitude is identical too.

4.3 Binary seismic evaluation results

For obtaining an appropriate evaluation of the retrieved seismic data, multiple machine learning algorithms have been applied such as Random Forest, Decision Tree, K Nearest Neighbor, Extreme gradient boosting, and Adaptive boosting. The derived results from each country-wise data are given in tables [6-11] along with the selected metrics for comparison that are obtained through applying 5-fold cross-validation technique.

ROC curves obtained through 5 fold cross-validation are derived for each classifier in order to compare the classifier performance. The comparison through ROC curves provides a better representation of the applied seismic evaluation techniques. Average roc is computed of each 5-fold cross-validation to get a stable value for performance comparison. Out of 30 roc curves; a few are attached below. For the collective countries dataset, the roc curves are titled according to the initial alphabet of each country to distinguish the curves.

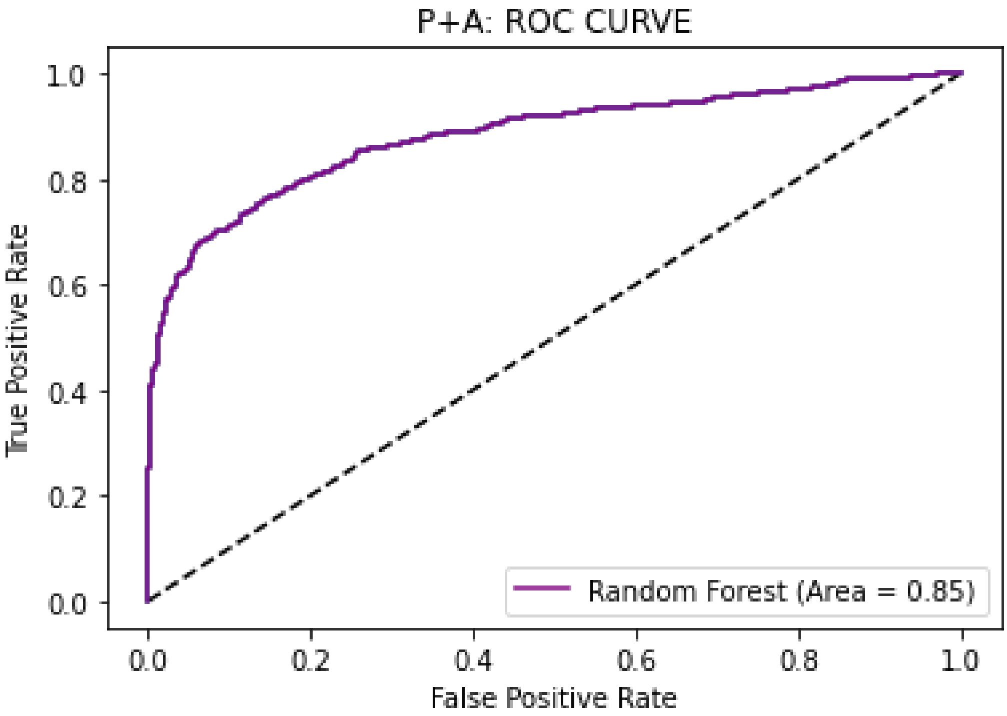


Figure 32: Random forest roc of Pakistan and Afghanistan data

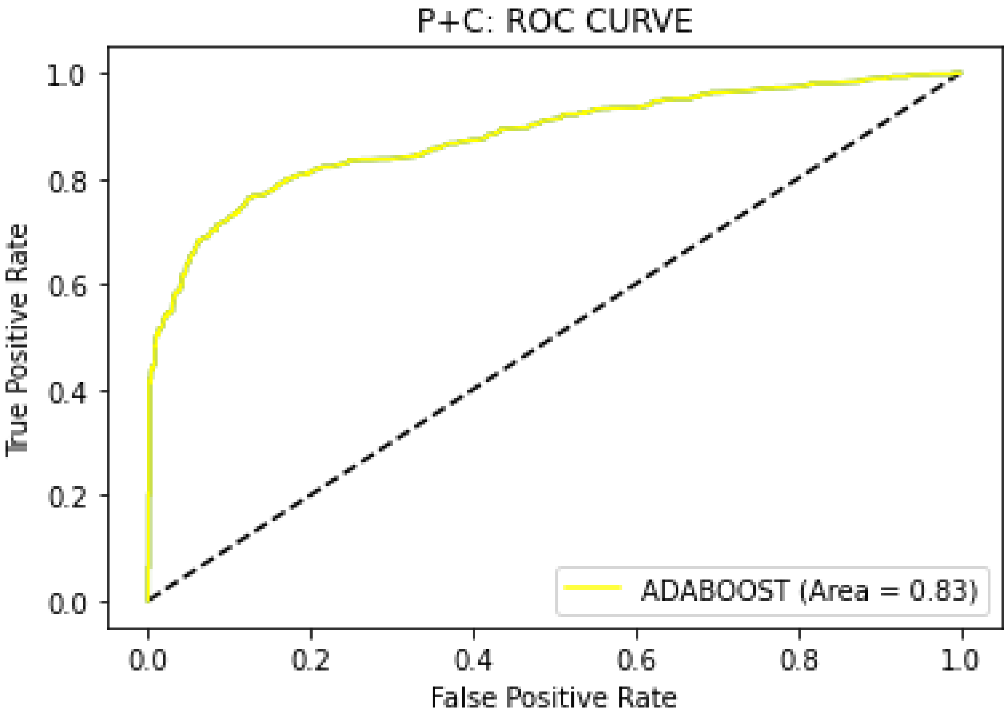


Figure 33: Adaboost roc of Pakistan and China data

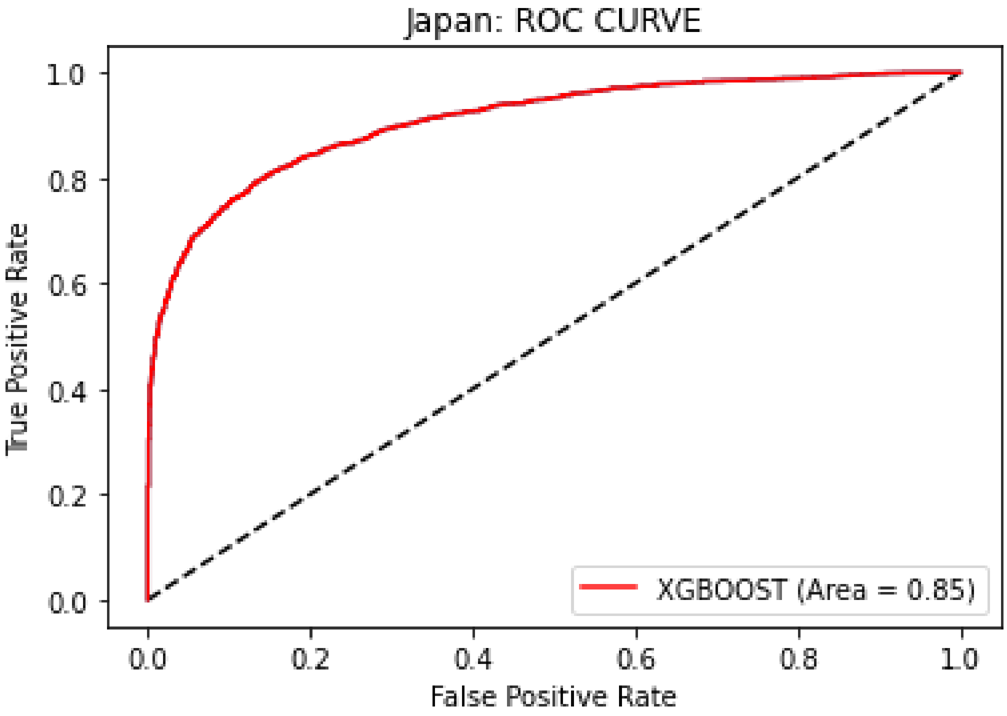


Figure 34: XGboost roc of Japan data

4.3.1 Pakistan and Afghanistan dataset

Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Random Forest	0.85	0.82	0.71	0.76	77
Decision Tree	0.79	0.76	0.66	0.70	72
AdaBoost	0.83	0.84	0.61	0.70	74
XGBoost	0.84	0.80	0.69	0.73	75
KNN	0.50	0.41	0.55	0.45	53

Table 6: Binary evaluation results of Pakistan and Afghanistan data

Table 6 shows the binary cross-validated evaluation results of the Pakistan and Afghanistan seismic dataset. Random Forest achieved the highest cross-validated accuracy of 77%, AUC of 0.85, 0.71 recall value, and F1- Score of 0.76 respectively. Out of the applied five algorithms, KNN performed the worst with 53% accuracy.

4.3.2 Pakistan and China dataset

Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Random Forest	0.86	0.82	0.73	0.76	77
Decision Tree	0.82	0.83	0.65	0.72	76
AdaBoost	0.83	0.84	0.70	0.75	78
XGBoost	0.85	0.87	0.67	0.75	78
KNN	0.51	0.37	0.67	0.48	51

Table 7: Binary evaluation results of Pakistan and China data

Table 7 shows the binary evaluation results of the Pakistan and China dataset. Random Forest classifier obtained the highest AUC of 0.86. Random Forest obtained a 0.73 recall value, f1-score of 0.76. XGBoost algorithm obtained the highest precision rate of 0.87.

4.3.3 Japan dataset

Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Random Forest	0.85	0.75	0.81	0.75	73
Decision Tree	0.77	0.74	0.73	0.72	71
AdaBoost	0.83	0.75	0.75	0.73	72
XGBoost	0.85	0.79	0.74	0.73	73
KNN	0.56	0.47	0.47	0.40	59

Table 8: Binary evaluation results of Japan data

Table 8 shows Japan’s binary evaluation results. Random forest and XGB algorithm obtained the highest accuracy of 73% and AUC as 0.85 respectively. Whereas Random Forest obtained 0.81 recall along with a 0.75 f1-score.

4.4 Multiclass seismic evaluation results

The applied algorithm’s performance obtained from the country-wise multiclass seismic evaluation is discussed in this section. The selected roc curves for multiclass seismic evaluation are attached below.

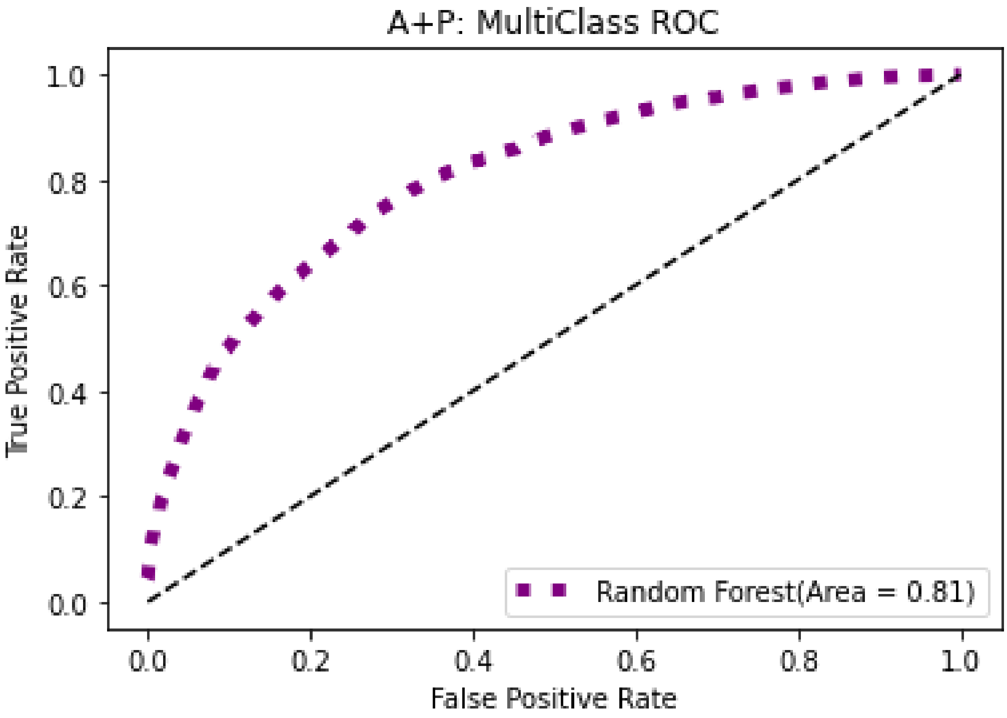


Figure 35: Random forest roc of multiclass Pakistan and Afghanistan data

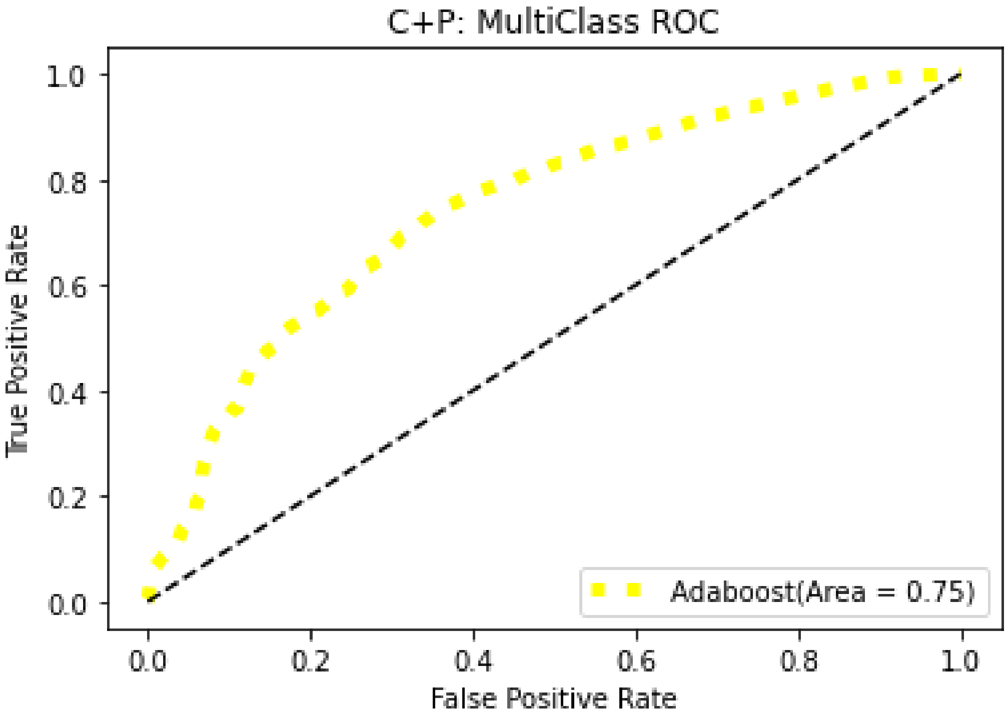


Figure 36: Adaboost roc of multiclass Pakistan and China data

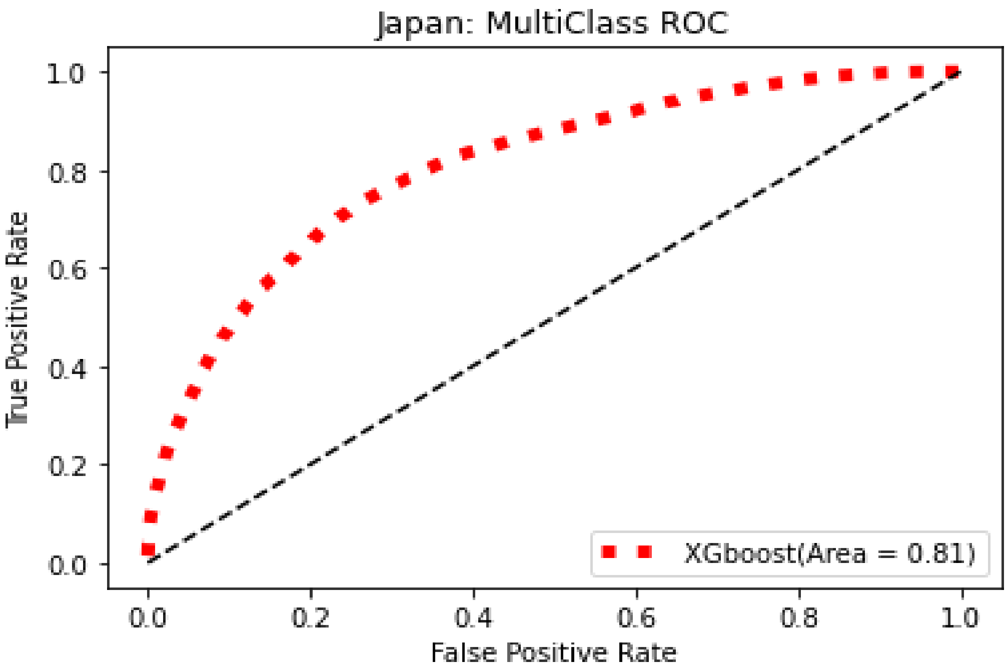


Figure 37: XGboost roc of multiclass Japan data

4.4.1 Pakistan and Afghanistan dataset

Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Random Forest	0.81	0.60	0.60	0.59	64
Decision Tree	0.77	0.57	0.57	0.56	60
AdaBoost	0.77	0.60	0.59	0.57	63
XGBoost	0.82	0.66	0.61	0.61	65
KNN	0.59	0.43	0.43	0.44	44

Table 9: Multiclass evaluation results of Pakistan and Afghanistan data

Multiclass seismic evaluation results of Afghanistan and Pakistan are shown in table 9. Among the applied algorithms, XGBoost remains the best-performing supervised algorithm with an AUC of 0.82, 0.66 precision, 0.61 recall value along with the highest F1-Score and accuracy score of 0.61 and 65%. However, KNN performed the worst amongst the applied algorithms with 44% accuracy only.

4.4.2 Pakistan and China dataset

Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Random Forest	0.81	0.62	0.58	0.57	63
Decision Tree	0.79	0.59	0.54	0.59	59

AdaBoost	0.75	0.61	0.57	0.56	62
XGBoost	0.82	0.63	0.58	0.57	63
KNN	0.60	0.38	0.33	0.23	38

Table 10: Multiclass evaluation results of Pakistan and China data

The multiclass seismic evaluation result of China and Pakistan is shown in table 10. The evaluation results show that XGB got the highest AUC of 0.82, 0.63 precision value, 0.58 recall value, and 63% accuracy. The highest F1 score of the Decision tree is 0.59. KNN is the only classifier that obtained the lowest evaluation results in all performance metrics.

4.4.3 Japan dataset

Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Random Forest	0.81	0.65	0.65	0.64	65
Decision Tree	0.67	0.60	0.61	0.61	61
AdaBoost	0.71	0.64	0.61	0.61	61
XGBoost	0.81	0.67	0.54	0.53	67
KNN	0.62	0.53	0.54	0.53	53

Table 11: Multiclass evaluation results of Japan data

Japan’s historic seismic data evaluation results are shown in Table 11. XGB classifier got 67% accuracy, 0.67 precision, and AUC equals 0.81. Whereas Random Forest got the best recall and F1-Score of 0.65 and 0.64 respectively. KNN obtained the lowest accuracy of 53% only.

CHAPTER 5

Discussion

Machine learning algorithms play a significant role in identifying yearly and monthly seismic activity patterns from collective neighboring country data. Analyzing seismic activity based on its intensity along with an evidence-based threshold further aids in evaluating the classifier's performance. Depending on the nature and complexity of the selected evaluation methods; each machine learning algorithm performs differently.

Authors in [46] performed seismic analysis on the Hindukush region through cut-off magnitude binary classification. As the seismic records data source is similar to this thesis and the time period range overlaps with our specified historic records as well. The comparison of the applied Decision tree and Random forest performance is performed with our evident-based binary threshold. The results obtained through 10-fold cross-validation are given below.

Classifier	AUC	Precision	Recall	F1-Score
Decision Tree	0.861	0.785	0.778	0.861
Random Forest	0.854	0.791	0.803	0.797

Table 12:10-fold CV results of Hindukush region

As compared to the results derived in the paper [46], our cross-validated binary evaluation through setting an evident average magnitude obtained higher precision value in collective country-wise seismic records. In collective Afghanistan and China data, the Random forest algorithm obtained precision value equal to 0.82 which is greater than 10-fold cross-validated precision of 0.79, whereas the Decision tree applied on the China dataset got higher precision value of 0.83. Moreover, Random forest implemented on collective Afghanistan and China data got a similar AUC of 0.85 respectively.

Through evaluating historic country-wise data, it's been observed that machine learning plays a major contribution in the seismic field. The final conclusion derived from the above-mentioned results is that we have to apply multiple machine learning algorithms initially for evaluation in order to select the best one's among them by comparing multiple performance metrics. Every country's performance metrics accuracies differ as each seismic data feature varies.

CHAPTER 6

Conclusion

The seismic datasets are constantly recorded for keeping us updated about the recent activity that happened anywhere in the globe. Retrieving the publically available dataset and evaluating it holds significant importance. This thesis presents two major techniques for analyzing magnitude intensity and evaluating the performance of multiple machine learning algorithms. Through machine learning algorithms earthquake datasets can be analyzed further based on the unit of measurement through which the natural hazard is analyzed for its intensity and potential risk to human lives. Based on the proposed methodology any natural hazardous event can be evaluated through machine learning algorithms and their performance can be compared. Furthermore, the proposed approach evaluated the publically available dataset through two methods, and it’s then compared to most seismic active country Japan; which gives us further assurance of both the selected methods for evaluation. As the country-wise comparative evaluation hasn’t been done in the past especially concerning Pakistan and its border sharing countries; this particular thesis plays a significant role.

6.1 Future Work

In future work, the proposed evaluation can be applied to any country’s seismic record that has contributing features that could result in improved performance of applied classifiers. Moreover, the seismic analysis can be applied to the Eurasian continent as both Europe and Asia lies on the same tectonic plate. Historic seismic events analysis of transcontinental countries including Kazakhstan, Turkey, Russia, Georgia, and Azerbaijan can be considered as it’ll provide insight into seismic activities of both European and Asian countries. Similar evaluation through machine learning can be applied to public hurricane and tornado datasets in the future. Enhanced Fujita scale division comparative analysis through machine learning algorithms can contribute to analyzing and predicting the damage caused by tornadoes. Similarly, the Saffir-Simpson division from light to strong wind intensity can be divided into its appropriate five categories and analyzed accordingly as well. Another perspective could be performing geological tectonic plate-wise earthquake analysis.

References

[1] I. Mahmood, A. A. Kidwai, S. N. Qureshi, M. F. Iqbal, and L. Atique, “Revisiting major earthquakes in Pakistan,” *Geol. Today*, vol. 31, no. 1, pp. 33–38, Jan. 2015, doi: 10.1111/gto.12085.

[2] editor, “Asia’s Most Quake-Prone Countries | Asian Geographic Magazines.” https://www.asiangeo.com/environment/asias-most-quake-prone-countries/ (accessed Sep. 30, 2021).

[3] “Number of earthquakes globally 2000-2020,” *Statista*. https://www.statista.com/statistics/263105/development-of-the-number-of-earthquakes-worldwide-since-2000/ (accessed Sep. 30, 2021).

[4] S. T. Maqsood and J. Schwarz, “Analysis of Building Damage during the 8 October 2005 Earthquake in Pakistan,” *Seismol. Res. Lett.*, vol. 79, no. 2, pp. 163–177, Mar. 2008, doi: 10.1785/gssrl.79.2.163.

[5] S. Earle, “9.1 Understanding Earth through Seismology,” Sep. 2015, Accessed: Sep. 15, 2021. [Online]. Available: https://opentextbc.ca/geology/chapter/9-1-understanding-earth-through-seismology/

[6] “Why Do Earthquakes Happen? | UPSeis,” *Michigan Technological University*. https://www.mtu.edu/geo/community/seismology/learn/earthquake-cause/ (accessed Sep. 15, 2021).

[7] “What Happens During an Earthquake?,” *Caltech Science Exchange*. http://scienceexchange.caltech.edu/topics/earthquakes/what-causes-earthquakes (accessed Sep. 15, 2021).

[8] “Volcanic Earthquakes,” *Pacific Northwest Seismic Network*. https://pnsn.org/outreach/earthquakesources/volcanic (accessed Sep. 15, 2021).

[9] “Earthquake | UN-SPIDER Knowledge Portal.” https://www.un-spider.org/disaster-type/earthquake (accessed Sep. 15, 2021).

[10] “New Page 1.” https://people.uwec.edu/jolhm/eh/toivonen/types.htm (accessed Sep. 15, 2021).

[11] “How Many Tectonic Plates Are There?,” *WorldAtlas*, Aug. 12, 2020. https://www.worldatlas.com/articles/major-tectonic-plates-on-earth.html (accessed Nov. 16, 2021).

[12] “CEA - Understanding Plate Tectonic Theory.” https://www.earthquakeauthority.com/Blog/2020/Understanding-Plate-Tectonic-Theory (accessed Nov. 16, 2021).

[13] E. S. B. A., “The World’s Major Earthquake Zones,” *ThoughtCo*. https://www.thoughtco.com/seismic-hazard-maps-of-the-world-1441205 (accessed Nov. 16, 2021).

[14] M. S. Siddique and J. Schwarz, “Elaboration of Multi-Hazard Zoning and Qualitative Risk Maps of Pakistan,” *Earthq. Spectra*, vol. 31, no. 3, pp. 1371–1395, Aug. 2015, doi: 10.1193/042913EQS114M.

[15] P. H. Ltd, “Japan and Earthquakes: Why They Happen and How to Scale Them,” *PLAZA HOMES*. https://www.realestate-tokyo.com/living-in-tokyo/emergency-disaster/earthquake-scale/ (accessed Nov. 16, 2021).

[16] “Reading: Seismic Waves | Geology.” https://courses.lumenlearning.com/geology/chapter/reading-seismic-waves/ (accessed Sep. 09, 2021).

[17] “The Science of Earthquakes.” https://www.usgs.gov/natural-hazards/earthquake-hazards/science/science-earthquakes?qt-science_center_objects=0#qt-science_center_objects (accessed Sep. 15, 2021).

[18] “Seismic Network.” http://redsismica.uprm.edu/English/education/earthquakes/information.php (accessed Sep. 15, 2021).

[19] “Determining the Depth of an Earthquake.” https://www.usgs.gov/natural-hazards/earthquake-hazards/science/determining-depth-earthquake?qt-science_center_objects=0#qt-science_center_objects (accessed Sep. 15, 2021).

[20] “What is the difference between body waves and surface waves, and between P-waves and S-waves?,” *Earth Observatory of Singapore*. https://earthobservatory.sg/faq-on-earth-sciences/what-difference-between-body-waves-and-surface-waves-and-between-p-waves-and-s (accessed Sep. 15, 2021).

[21] “Surface waves » Seismic Resilience.” http://www.seismicresilience.org.nz/topics/seismic-science-and-site-influences/earthquake-energy/surface-waves/ (accessed Sep. 09, 2021).

[22] “Types of Waves | Geology.” https://courses.lumenlearning.com/wmopen-geology/chapter/outcome-types-of-waves/ (accessed Sep. 15, 2021).

[23] “Seismology | UPSeis,” *Michigan Technological University*. https://www.mtu.edu/geo/community/seismology/learn/seismology-study/ (accessed Sep. 09, 2021).

[24] M. Bellis, “Meet the Man Who Invented the Earthquake Richter Scale,” *ThoughtCo*. https://www.thoughtco.com/charles-richter-and-richter-magnitude-scale-1992347 (accessed Sep. 15, 2021).

[25] “Earthquake Glossary.” https://earthquake.usgs.gov/learn/glossary/?term=richter%20scale (accessed Sep. 09, 2021).

[26] “Earthquake Magnitude, Energy Release, and Shaking Intensity.” https://www.usgs.gov/natural-hazards/earthquake-hazards/science/earthquake-magnitude-energy-release-and-shaking-intensity?qt-science_center_objects=0#qt-science_center_objects (accessed Sep. 09, 2021).

[27] “How Can I Locate the Earthquake Epicenter? | UPSeis | Michigan Technological University.” https://www.mtu.edu/geo/community/seismology/learn/earthquake-epicenter/ (accessed Sep. 09, 2021).

[28] “What is triangulation?” https://www.qrg.northwestern.edu/projects/vss/docs/navigation/1-what-is-triangulation.html (accessed Sep. 09, 2021).

[29] "EARTHQUAKES & TSUNAMI AND HAZARDS, EARTH'S INTERIOR | Environmental Issues and Resources." https://mediakron.bc.edu/environmentalissues/topic-9-plate-tectonics-earthquakestsunami-and-their-hazards/map-location-id-of-earthquake-using-3-seismograph-stations (accessed Sep. 15, 2021).

[30] "CEA - Earthquake Measurements: Magnitude vs Intensity." https://www.earthquakeauthority.com/Blog/2020/Earthquake-Measurements-Magnitude-vs-Intensity (accessed Sep. 09, 2021).

[31] "Earthquake Magnitude Levels Vector Illustration Diagram Stock Vector (Royalty Free) 1056375386," *Shutterstock.com*. https://www.shutterstock.com/image-vector/earthquake-magnitude-levels-vector-illustration-diagram-1056375386 (accessed Sep. 15, 2021).

[32] "Reading: Magnitude versus Intensity | Geology." https://courses.lumenlearning.com/geo/chapter/reading-magnitude-versus-intensity/ (accessed Sep. 15, 2021).

[33] P. Debnath *et al.*, "Analysis of Earthquake Forecasting in India Using Supervised Machine Learning Classifiers," *Sustainability*, vol. 13, no. 2, p. 971, Jan. 2021, doi: 10.3390/su13020971.

[34] A. A. V. L. Sruthi, R. Bhargavi, and V. R. Gospati, "Analysis of Seismic data using Machine Learning Algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1070, p. 012042, Feb. 2021, doi: 10.1088/1757-899X/1070/1/012042.

[35] I. M. Murwantara, P. Yugopuspito, and R. Hermawan, "Comparison of machine learning performance for earthquake prediction in Indonesia using 30 years historical data," vol. 18, no. 3, p. 12, 2020.

[36] M. F. A. Azis, F. Darari, and M. R. Septyandy, "Time Series Analysis on Earthquakes Using EDA and Machine Learning," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia, Oct. 2020, pp. 405–412. doi: 10.1109/ICACSIS51025.2020.9263188.

[37] Y. Garg, A. Masih, and U. Sharma, "Predicting Bridge Damage During Earthquake Using Machine Learning Algorithms," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, Jan. 2021, pp. 725–728. doi: 10.1109/Confluence51648.2021.9377100.

[38] S. Karimzadeh, M. Matsuoka, J. Kuang, and L. Ge, "Spatial Prediction of Aftershocks Triggered by a Major Earthquake: A Binary Machine Learning Perspective," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 10, p. 462, Oct. 2019, doi: 10.3390/ijgi8100462.

[39] G. F. Manley *et al.*, "Understanding the timing of eruption end using a machine learning approach to classification of seismic time series," *J. Volcanol. Geotherm. Res.*, vol. 401, p. 106917, Sep. 2020, doi: 10.1016/j.jvolgeores.2020.106917.

[40] P. Bangar, D. Gupta, S. Gaikwad, B. Marekar, and J. Patil, "Earthquake Prediction using Machine Learning Algorithm," vol. 8, no. 6, p. 5, 2020.

[41] Y. Zhang, H. V. Burton, H. Sun, and M. Shokrabadi, "A machine learning framework for assessing post-earthquake structural safety," *Struct. Saf.*, vol. 72, pp. 1–16, May 2018, doi: 10.1016/j.strusafe.2017.12.001.

[42] M. N. Brykov *et al.*, “Machine Learning Modelling and Feature Engineering in Seismology Experiment,” *Sensors*, vol. 20, no. 15, p. 4228, Jul. 2020, doi: 10.3390/s20154228.

[43] S. Mangalathu, H. Sun, C. C. Nweke, Z. Yi, and H. V. Burton, “Classifying earthquake damage to buildings using machine learning,” *Earthq. Spectra*, vol. 36, no. 1, pp. 183–208, Feb. 2020, doi: 10.1177/8755293019878137.

[44] F. Ge, Y. Li, M. Yuan, J. Zhang, and W. Zhang, “Identifying predictors of probable posttraumatic stress disorder in children and adolescents with earthquake exposure: A longitudinal study using a machine learning approach,” *J. Affect. Disord.*, vol. 264, pp. 483–493, Mar. 2020, doi: 10.1016/j.jad.2019.11.079.

[45] R. Mallouhy, C. A. Jaoude, C. Guyeux, and A. Makhoul, “Major earthquake event prediction using various machine learning algorithms,” in *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, Paris, France, Dec. 2019, pp. 1–7. doi: 10.1109/ICT-DM47966.2019.9032983.

[46] K. M. Asim, A. Idris, F. Martinez-Alvarez, and T. Iqbal, “Short Term Earthquake Prediction in Hindukush Region Using Tree Based Ensemble Learning,” in *2016 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, Dec. 2016, pp. 365–370. doi: 10.1109/FIT.2016.073.

[47] “A Survey on Decision Tree Algorithms of Classification in Data Mining,” *Int. J. Sci. Res. IJSR*, vol. 5, no. 4, pp. 2094–2097, Apr. 2016, doi: 10.21275/v5i4.NOV162954.

[48] H. H. Patel and P. Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/ijcse/v6i10.7478.

[49] “Machine Learning Decision Tree Classification Algorithm - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (accessed Sep. 15, 2021).

[50] “Figure 2. Random forest classification.,” *ResearchGate*. https://www.researchgate.net/figure/Random-forest-classification_fig2_325303084 (accessed Sep. 15, 2021).

[51] T. Chengsheng, L. Huacheng, and X. Bing, “AdaBoost typical Algorithm and its application research,” *MATEC Web Conf.*, vol. 139, p. 00222, 2017, doi: 10.1051/matecconf/201713900222.

[52] “Understanding AdaBoost for Decision Tree | by Valentina Alto | Towards Data Science.” https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851 (accessed Sep. 15, 2021).

[53] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[54] “How XGBoost Works - Amazon SageMaker.” https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html (accessed Sep. 15, 2021).

[55] “Fig. 2 Example on KNN classifier,” *ResearchGate*. https://www.researchgate.net/figure/Example-on-KNN-classifier_fig1_331424423 (accessed Sep. 15, 2021).

[56] “Earthquakes.” https://www.usgs.gov/natural-hazards/earthquake-hazards/earthquakes (accessed Sep. 13, 2021).

[57] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data Preprocessing for Supervised Leaning,” vol. 1, no. 1, p. 8, 2006.

[58] S. K. Kwak and J. H. Kim, “Statistical data preparation: management of missing values and outliers,” *Korean J. Anesthesiol.*, vol. 70, no. 4, p. 407, 2017, doi: 10.4097/kjae.2017.70.4.407.

[59] H. Kang, “The prevention and handling of the missing data,” *Korean J. Anesthesiol.*, vol. 64, no. 5, p. 402, 2013, doi: 10.4097/kjae.2013.64.5.402.

[60] “ON METHODS FOR IMPROVING THE ACCURACY OF MULTICLASS CLASSIFICATION ON IMBALANCED DATA,” *Inform. Appl.*, Mar. 2020, doi: 10.14357/19922264200109.

[61] K. Maladkar, “Why Is Random Search Better Than Grid Search For Machine Learning,” *Analytics India Magazine*, Jun. 14, 2018. https://analyticsindiamag.com/why-is-random-search-better-than-grid-search-for-machine-learning/ (accessed Sep. 16, 2021).

[62] A. C. A. S. Writer, “Difference between shallow, deep earthquakes,” *Clinton Herald*. https://www.clintonherald.com/news/difference-between-shallow-deep-earthquakes/article_adb7067e-6b8b-11e6-a8ab-4be3fd07f666.html (accessed Sep. 17, 2021).