

University of Bahrain

College of Science

Postgraduate Programs



Earthquake Predictions Using Machine Learning Approaches

Thesis submitted in partial fulfilment of the requirements for the degree
of M.Sc. In Big Data Science and Analytics

Prepared by:

Fatehan Tael Shajrah

20082472

Supervised by:

Prof. Nabil Hewahi

Professor

University of Bahrain

Kingdom of Bahrain

March 2024



Dissertation Deposit Form

Author Name:	Fatehan Tael Shajrah		
Email:	fathanshajera@gmail.com	Contact No:	33252200
College:	Science		
Department:	Mathematics		
Thesis Title:	Earthquake Predictions Using Machine Learning Approaches		
Year:	2020-2021		

Author's Declaration: I agree to the following conditions:

The above thesis will be available (in the University of Bahrain library and externally) and reproduced as necessary at the discretion of the University of Bahrain. It may also be digitized by the University of Bahrain and made available via the university's repository or via other parties on the Internet.

This condition shall apply to ALL copies including electronic copies.

The above thesis has been provided on the understanding that it is copyright protected material and that no quotation from this thesis may be published without proper acknowledgement.

Signature of the Author:		Date:	07-Jul-24
--------------------------	---	-------	-----------

Defense Committee's Approval

University of Bahrain
Deanship of Graduate Studies
And scientific Research



جامعة البحرين
عمادة الدراسات العليا والبحث العلمي

Thesis Discussion Committee Decision

Student Name: FATEHAN TAEI SHAJRAH	Student ID: 20082472
Department: Choose an item.	College: Science
Program: M.Sc. in Big Data Science and Analytics	
Thesis Title: Earthquake Predictions Using Machine Learning Approaches	
Supervisor Name: Prof. Nabil Hewahi	Student GPA: 3.45

The Discussion Committee met on 24-Jun-24 In S41-1002. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade **PASS** Choose an item. in the thesis subject.

- ☐ Without making any changes to the thesis.
- ☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name

1. **Prof. Nabil Hewahi**
2. **Prof. Faisal Alkhateeb**
3. **Dr. Saeed Sharif**

Signature

(Advisor) *Nabil Hewahi*
(Internal Examiner) *[Signature]*
(External Examiner) *Saeed Sharif*

Abstract

An effective and precise earthquake prediction technique could potentially save numerous lives and economies. Bearing this goal in mind, this thesis conducts a comparative study that explores the application of machine learning (ML) approaches in predicting earthquake characteristics. It utilizes the United States Geological Survey (USGS) dataset, which includes earthquakes of magnitude 3.5 and above from the last year (2023), and the Los Alamos National Laboratory (LANL) datasets for a comprehensive analysis. The USGS dataset is employed to predict the magnitude of earthquakes through regression analysis, while the LANL dataset focuses on forecasting the timing of forthcoming seismic events. A diverse array of machine learning algorithms, including Linear Regression (LR), Support Vector Regression (SVR), Random Forest Regressor (RFR), XGBoost Regressor (XGB), and advanced deep learning models such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), are meticulously evaluated for their predictive capabilities. The assessment of model performance leverages several evaluation metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Among the models evaluated, the RFR demonstrated outstanding performance on the USGS dataset, achieving an MAE of 0.167, an MSE of 0.055, and an impressive R^2 score of 0.729. This highlights its robust capability in modeling complex relationships within seismic magnitude data. In contrast, the GRU model excelled in the LANL dataset, achieving the best performance with an MAE of 1.777, an MSE of 4.192, and an R^2 score of 0.610. Future efforts should optimize feature engineering and selection processes by integrating diverse indicators such as seismic activity, historical patterns, and geological data to develop more sophisticated tools for earthquake preparedness and risk mitigation.

Acknowledgements

I would like to thank the University of Bahrain for offering the Big Data Science and Analytics master's program.

My deepest appreciation goes to Prof. Nabil Hewahi, whose guidance was invaluable during my thesis work.

I am also grateful to Dr. Sawsan Hilal for her continuous commitment and support during the program.

Finally, I wish to thank all the program instructors for their commendable efforts in conducting the courses.

Table of Contents

Defense Committee's Approval	A
Abstract	B
Acknowledgements	C
Table of Contents	D
List of Tables	G
List of Figures	H
List of Abbreviations	J
Chapter One Introduction	2
1.1. Introduction	2
1.1.1. Background on Earthquakes and Their Impact	2
1.1.2. Importance of Earthquake Prediction	3
1.2. Research Problem	5
1.3. Research Questions	6
1.4. Research Objectives	7
1.5. Research Significance	7
1.6. Research Scope and Limitations	8
1.7. Thesis Structure	8
Chapter Two Background and Literature Review	11
2.1. Background	11
2.1.1. Implemented Models	11
2.1.2. Review of Conventional Methods for Earthquake Prediction	15
1) Animal Behaviour	15

2) Historical Seismic Data Analysis	15
3) Radon Gas Emission	16
4) Seismic Gap Hypothesis.....	16
5) Geodetic Measurements.....	17
2.1.3. Limitations of Traditional Approaches	17
2.2. Literature Review	19
Chapter Three Methods and Procedures.....	32
3.1. Adopted Methodology	32
3.2. Tested Approaches.....	33
3.3. Performance Metrics.....	33
Chapter Four Results and Discussion.....	37
4.1. Data Sources	37
4.1.1. USGS Dataset	37
4.1.2. LANL Dataset.....	47
4.2. Results.....	57
4.2.1. USGS Dataset Models Performance	57
4.2.2. LANL Dataset Models Performance	63
4.3. Comparison with other studies	80
4.4. Results Discussion	84
4.4.1. USGS Dataset	85
4.4.2. LANL Dataset.....	85
Chapter Five Conclusion and Future Work.....	89
5.1. Conclusion	89
5.2. Future Work.....	90

5.3. References.....	92
الملخص	100

List of Tables

Table 1 : Summary of Related Works	27
Table 2 :USGS dataset Description	38
Table 3: Magnitude Descriptive Statistic	45
Table 4: Selected Features Descriptive Statistics	47
Table 5: Performance Metrics for Earthquake Magnitude Prediction Models on USGS Dataset	57
Table 6: LSTM Model's Performance	62
Table 7: GRU Model's Performance	63
Table 8: Models performance (LANL)	64
Table 9: LSTM Model Architectures	68
Table 10: LSTM Performance for LANL dataset	69
Table 11: GRU Architecture Used for LANL dataset.....	75
Table 12:GRU Performance (LANL).....	76

List of Figures

Figure 1: Adopted Methodology Used	33
Figure 2: First 5 row of USGS dataset.	39
Figure 3: Correlation Matrix (USGS).....	43
Figure 4: Mutual Information Scores for All USGS features.	43
Figure 5:Earthquake Magnitude Histogram	45
Figure 6: Magnitude over Time.....	46
Figure 7: 1% from Total Acoustic Data	49
Figure 8: Acoustic Data and Time to Failure (1% data)	49
Figure 9: Density for 25k samples of Acoustic Data	50
Figure 10:density plot for 25k instances of "time_to_failure"	50
Figure 11: Boxplot of Acoustic Data (25k samples)	51
Figure 12: Boxplot of Time to Failure (25k).....	52
Figure 13: LANL Feature set (58).....	53
Figure 14: LANL feature set (37).....	55
Figure 15: MAE for LR, SVR, RFR and XGBOOST Regressor....	58
Figure 16:MSE for LR, SVR, RFR and XGBoost Regressor	58
Figure 17:R-squared for LR, SVR, RFR and XGBoost Regressor ..	59
Figure 18: RFR Actual vs Predicted Magnitudes	60
Figure 19: MAE for LR, RFR, XGBoost and SVR (LANL)	65
Figure 20: MSE For LR, RFR, XGBoost and SVR (LANL)	66
Figure 21: LSTM MAE metrics	71
Figure 22: LSTM MSE metrics	72
Figure 23: LSTM R-squared metrics.....	73
Figure 24: Model 5 architecture prediction performance (LANL) ..	74
Figure 25: MAE Metrics for GRU (LANL)	77
Figure 26:MSE metrics For GRU (LANL)	78
Figure 27:R-squared metric for GRU (LANL)	79

Figure 28: GRU Model Prediction	80
---------------------------------------	----

List of Abbreviations

[illegible]

Chapter One

Introduction

1.1. Introduction

1.2. Research Problem

1.3. Research Questions

1.4. Research Objectives

1.5. Research Significance

1.6. Research Scope and Limitations

1.7. Thesis Structure

Chapter One

Introduction

1.1. Introduction

Seismic tremors represent one of the most unforeseeable calamities in nature, causing extensive damage to human lives and economies. According to the USGS, the Earth experiences over 500,000 earthquakes annually; however, most are too minor to register (Elnashai et al., 2008). The challenge of predicting seismic activity has captured the attention of researchers, geophysicists, and scientists, all aiming to diminish the resulting havoc. Continuous efforts are being made to understand the characteristics of earthquakes and the dynamics between seismic risks and the environment to issue timely alerts. Despite extensive research, the precise prediction and forecasting of seismic events remain elusive. Earthquake forecasting is complex (Rundle, et al., 2021) Nonetheless, the recent advancements in ML technologies offer new possibilities. This dissertation will delve into the evolution from traditional forecasting methods to sophisticated ML techniques, examining current research developments and future possibilities.

1.1.1. Background on Earthquakes and Their Impact

The Earth's lithosphere, constantly in motion, generates immense stress over extended periods, culminating in seismic events. The lithosphere consists of tectonic plates that move slowly some just a few centimetres yearly. These plates interact through converging, diverging, or sliding past each other at their boundaries, causing sudden energy releases

manifesting as earthquakes. Divergent movements create new crusts, while sliding plates generate friction (Gao et al., 2022). Such tectonic activities produce seismic waves leading to ground tremors. Moreover, human interventions like underground nuclear tests and fracking can alter ground viscosity, contributing to tectonic stress. Understanding these factors is crucial for developing effective early warning systems and earthquake prediction models, utilizing the data to construct precise forecasts and mitigate seismic risks.

1.1.2. Importance of Earthquake Prediction

The primary goal of researching earthquake prediction is risk mitigation, aiming to minimize the immediate and long-term effects of these natural disasters. Although preventing natural calamities is unfeasible, developing an early warning system for detecting precursory signs is essential. Accurate predictions, including timing, epicenter identification, and aftermath projections, can significantly enhance the preparedness of disaster response teams and the execution of search-and-rescue missions during crises (Jordan, 2006). Additionally, residents in vulnerable areas can greatly benefit from well-structured evacuation plans based on predictive data. Furthermore, a focused approach on seismic-prone regions can lead to a deeper understanding of geological features specific to compression zones.

Overview of ML in Earthquake Prediction, allows computers to identify patterns and make predictions autonomously, without explicit programming (Bergen et al., 2019). This field has seen significant advancements, attracting attention from various scientific domains, including geoscience, to leverage these innovative solutions. Although the

concept of ML was introduced in 1950, it has recently become a prominent method for analyzing seismic hazards and predicting earthquakes (Kulkarni, 2012). The growing popularity of ML in geoscience is evident, with its successful applications in diverse areas due to new algorithmic developments, the availability of extensive datasets, and enhanced computational efficiency. Recent advancements have led to the application of both established and novel ML tools in addressing various geoscientific challenges, such as detecting seismic waves, identifying phases and locations, analyzing geological structures, providing earthquake early warnings, monitoring volcanoes, denoising InSAR data, imaging tomography, characterizing reservoirs, and more. It's important to note the substantial work over the past decades in utilizing these techniques to understand fault physics and anticipate fault failures.

The advancements in ML within the realm of artificial intelligence have been quite impressive recently. These advancements find applications in diverse fields such as image and voice recognition, natural language processing, and autonomous navigation (Hassin & Abbood, 2021). The surge in computational capabilities, coupled with the widespread availability of ML libraries, has democratized access to these technologies. This accessibility is paving the way for innovative applications in predicting various natural events like weather conditions (Meenal et al., 2022) and seismic activities. Notably, the adoption of cutting-edge ML techniques, including probabilistic modelling, is under scrutiny, especially in scenarios demanding high safety standards where even minimal errors can be critical.

In recent times, there's been a significant influx of data concerning seismic activity rates and locations. This data is instrumental in

revolutionizing the methods of seismic hazard identification and integration into risk assessment frameworks. However, there's a noticeable gap in discussions regarding the implementation of ML in this context within for different datasets.

A particular study by (Jarrah et al., 2023) made strides in seismic analysis by developing an algorithm focused on normalizing and analyzing internal signals. This algorithm proved effective in pinpointing the occurrence of primary seismic events by analyzing the temporal discrepancies recorded across various stations. This approach not only improved the precision in detecting the initial seismic waves but also highlighted the capability of ML techniques to decode and synthesize complex interrelationships among various seismic parameters. Such capabilities are crucial in addressing intricate challenges in earthquake engineering, challenges that are often beyond the scope of traditional methodologies (Xie et al., 2020).

1.2. Research Problem

The unpredictability of seismic tremors presents a significant challenge in reducing their catastrophic impact on human lives and economies. Traditional methods for earthquake prediction have proven insufficient in providing precise and timely alerts. Given that over 500,000 earthquakes occur annually, many of which go undetected due to their minor nature, there is a critical need for more accurate prediction models. Enhancing these models requires a comprehensive approach, integrating advanced feature engineering and selection processes. The advancement of ML offers a promising avenue for improving earthquake prediction. However, the application of various ML models, such as Linear Regression

(LR), Random Forest Regressor (RFR), Support Vector Regression (SVR), eXtreme Gradient Boosting Regressor (XGBoost Regressor), and DL models, in seismic prediction - explicitly used in a regression context, using datasets like the United States Geological Survey (USGS) dataset and Los Alamos National Laboratory (LANL) datasets - has not been thoroughly evaluated. This gap in knowledge and practice forms the core of the research problem: assessing the effectiveness of ML models in accurately predicting seismic activities and understanding the challenges and complexities involved in applying these models to seismic data. Additionally, this study will conduct a comparative analysis of different ML models to further understand their performance in seismic prediction. USGS dataset will be used to predict the magnitude of an earthquake, while LANL dataset will be used to predict the time remaining for the next event.

1.3. Research Questions

The current study was conducted to answer the following questions:

1. How effective are various ML models, including LR, RFR, SVR, XGBoost Regressor, and DL networks, in predicting seismic activities?
2. What is the relationship between model selection, dataset characteristics (like USGS and LANL), and predictive accuracy in seismic prediction?
3. How do factors like feature engineering and selection, along with the need for comprehensive data, affect the efficacy of ML models in earthquake prediction?
4. What are the challenges and complexities involved in applying ML techniques to seismic data modelling?

1.4. Research Objectives

The main objectives of this research are as follows:

- To evaluate and compare the efficacy of different ML models in predicting seismic activities using the LANL and USGS datasets.
- To analyze the relationship between the choice of ML models, characteristics of seismic datasets, and the accuracy of predictions.
- To investigate the impact of overfitting and the necessity of comprehensive data in the context of ML models for earthquake prediction.
- To explore the challenges and complexities inherent in applying ML techniques to seismic data, contributing to the advancement of predictive tools in seismology.

1.5. Research Significance

1. Advancing Earthquake Prediction Methods: This study aims to contribute significantly to the field of seismology by advancing the methods of earthquake prediction using ML models, potentially leading to more accurate and timely predictions.

2. Informed Decision-Making: The findings can aid scientists, geophysicists, and disaster management authorities in making informed decisions by utilizing more reliable earthquake prediction models.

3. Risk Mitigation and Safety Improvement: Enhanced earthquake prediction models could play a crucial role in mitigating risks and improving safety protocols for vulnerable populations.

4. Research and Technological Development: The study will add to the existing body of knowledge in the intersection of ML and seismology, potentially leading to further research and technological innovations.

5. Policy Implications: The insights gained from this study could inform policymaking in terms of urban planning, disaster preparedness, and resource allocation for regions prone to seismic activities.

1.6. Research Scope and Limitations

The scope of this research is focused on utilizing ML models for earthquake prediction, specifically targeting the prediction magnitude with the USGS dataset and prediction of timing with the LANL dataset. However, it's crucial to acknowledge certain limitations inherent in this approach. Firstly, earthquake prediction encompasses the comprehensive forecasting of the occurrence, location, and magnitude of seismic events. This research narrows this broad objective by concentrating on predicting the timing of the next event using the LANL dataset, which is derived from laboratory experiments rather than real-world seismic activity. Consequently, while valuable for theoretical and methodological development, the LANL dataset may not fully capture the complexities of natural earthquakes. Secondly, the USGS dataset, employed for predicting earthquake magnitudes, encompasses real-world data but is limited to records from the year 2023 which consists of 17730 instances. Together, these limitations delineate the research's boundaries and highlight areas for potential future expansion.

1.7. Thesis Structure

The rest of thesis report is structured as follows, Chapter 2 presents a background and literature review of previous work on ML and earthquake predictions, chapter 3 describes the methodology conducted in the research and performance metrics used, the results obtained with detailed discussion

in chapter 4. Conclusion and recommendations for future work will be in chapter 5.

Chapter Two

Background and Literature Review

2.1. Background

2.2. Literature Review

Chapter Two

Background and Literature Review

2.1. Background

In this section, researcher outlines the ML models that will be utilized in this research, as well as discussing traditional techniques for earthquake prediction and their limitations.

2.1.1. Implemented Models

Researcher explores and evaluates a comprehensive suite of ML and DL models applied to two distinct datasets: USGS dataset for magnitude prediction and LANL dataset for 'time_to_failure' prediction. These models were selected to address the specific characteristics and challenges presented by each dataset, aiming to provide a thorough analysis and comparison of their performance in predictive tasks.

1) LR

Linear Regression is a foundational technique in statistics and ML, aimed at modeling the linear relationship between a dependent variable and one or more independent variables. It operates by estimating the coefficients of the linear equation, minimizing the difference between observed and predicted values. Simple Linear Regression deals with one independent variable, while Multiple Linear Regression involves multiple. The method's effectiveness is often measured by the R-squared metric, indicating the proportion of variance in the dependent variable explained by the independent variables. This technique is pivotal for prediction and analysis across various fields, including economics and the sciences, due to its simplicity and interpretability (James et al., 2013).

2) RF

RF is an ensemble ML algorithm that builds multiple decision trees to make more accurate and stable predictions. It's effective for both classification and regression tasks. By aggregating the results of individual trees, it minimizes the risk of overfitting, a common issue with single decision trees. The algorithm enhances diversity through bootstrap aggregating, or bagging, by training each tree on a random subset of data and features (Liaw & Wiener, 2001). This approach makes RF versatile, requiring minimal data preprocessing and capable of handling missing values and categorical data. Hyperparameters, like the number of trees, can be optimized for better performance, often using scikit-learn's GridSearchCV.

3) SVR

SVR applies the principles of Support Vector Machines (SVM) to regression problems. It seeks to find a function that deviates from the actual observed targets y by a value no greater than ϵ for each training point, while also being as flat as possible. SVR achieves this by constructing a linear model that fits within a margin of tolerance ϵ but is penalized for data points that fall outside this margin. The flexibility of SVR comes from its ability to use different kernel functions to handle linear and non-linear relationships. The choice of kernel (e.g., linear, polynomial, radial basis function) allows the SVR to adapt to various data structures. SVR is particularly useful for datasets with high dimensionality or when the relationship between the independent and dependent variables is complex and non-linear (Drucker et al., 1996).

4) XGBoost

XGBoost is an advanced implementation of gradient boosting (GB) algorithms, known for its efficiency, flexibility, and portability. It is designed to optimize large-scale tree boosting, providing a highly efficient, scalable, and fast solution for supervised learning tasks. XGBoost improves upon traditional GB methods by introducing several enhancements, such as a novel tree learning algorithm that handles sparse data, parallel and distributed computing for faster execution, and improved regularization to prevent overfitting. These features make XGBoost a powerful tool for a wide range of predictive modeling tasks, including classification, regression, and ranking problems. Its performance and speed have made it a popular choice among data scientists and ML practitioners for winning numerous ML competitions (Chen & Guestrin, 2016).

5) LSTM

LSTM networks are a special kind of RNN capable of learning long-term dependencies. Introduced by Hochreiter & Schmidhuber, (1997), LSTMs were designed to overcome the vanishing gradient problem that affects standard RNNs, enabling the model to retain information for long periods. LSTMs achieve this through their unique architecture, which includes memory cells that store information and gates (input, output, and forget gates) that regulate the flow of information into and out of the cell, as well as the retention of information over time. This makes LSTMs particularly suited for tasks involving sequential data, such as time series analysis, natural language processing, and speech recognition, where the

ability to remember previous information is crucial for understanding context and making predictions.

6) GRU

GRU is a type of RNN introduced by Cho et al., (2014). GRU is designed to solve the vanishing gradient problem encountered by traditional RNNs, like LSTM networks, but with a simpler structure. GRUs streamline the architecture by using two gates: an update gate and a reset gate. These gates determine how much past information needs to be passed along to the future, making GRUs capable of capturing dependencies from long sequences of data without the need for a large memory unit. This efficiency allows GRUs to perform well on tasks requiring the modelling of sequential data, such as language translation, speech recognition, and time series analysis, with potentially fewer parameters and computational simplicity compared to LSTMs.

2.1.2. Review of Conventional Methods for Earthquake Prediction

This section critically explores conventional methods in earthquake prediction, tracing their evolution from ancient observational practices to early scientific assessments.

1) Animal Behaviour

Historically, the study of the seismic cycle has been a priority, leading to the exploration of various hypotheses, including the predictive potential of animal behavior (Tributsch, 1982). It's believed that animals might detect environmental changes signaling impending disasters, such as shifts in electromagnetic fields and air pressure, which remain undetected by humans. Ancient texts from Greece and Rome document instances where animals evacuated areas prior to seismic events. Fusakichi Omori, a Japanese seismologist, in 1912, observed animals exhibiting unusual behaviors before earthquakes (Tributsch, 1982). For instance, cats, dogs, and horses appeared agitated, while snakes and rats fled the scene. More recent research from 1989 and 2011 also supported these observations, noting peculiar behaviors in cats and dogs just before seismic events.

2) Historical Seismic Data Analysis

Analyzing historical earthquake data is a common method for forecasting, as it provides insights into the seismic likelihood of a region (Ogata, 1999). Zones with a history of frequent seismic activity are typically at higher risk of future earthquakes. Moreover, this data helps in understanding the relationship between earthquakes and other geographic markers like tectonic plate edges, faults, and the Earth's crust stress levels.

Therefore, this historical data is invaluable for refining prediction models and methodologies, particularly in areas prone to seismic disturbances.

3) Radon Gas Emission

Radon, an odorless, colorless, and tasteless radioactive gas, forms from the decay of uranium in the Earth's crust and can be detected in air, water, and soil (Cothorn & Smith, 1987). Numerous studies have investigated the link between radon emissions and seismic activity. For example, abnormal radon concentrations were recorded in Bakreswar, West Bengal, India, and were associated with seismic events near the Nicobar and Andaman Islands (Cothorn & Smith, 1987). Similar correlations were observed in NW Slovenia, western Turkey, Kerman Province in Iran, northern Pakistan, and Balakot Bagh, Pakistan (Baskaran, 2016; Burgmann & Dresen, 2008). Studies have shown that radon levels tend to spike before an earthquake and drop immediately afterward, as seen before the 2011 Tohoku earthquake in Japan and other significant quakes in Taiwan, California, and Italy.

4) Seismic Gap Hypothesis

The Seismic Gap Hypothesis (SGH) is another approach for predicting future earthquakes. It posits that areas with a low frequency of significant quakes are likely to experience them, based on the notion that over time, the stress accumulation on faults leads to potential energy build-up, eventually resulting in an earthquake (Kagan & Jackson, 1991). The focus of this hypothesis is on identifying regions with long intervals between earthquakes and assessing the likelihood of substantial seismic events. Historical data, such as the Chi-Chi earthquake in Taiwan in 1999

and the Tōhoku earthquake in Japan in 2011, supports the reliability of SGH in earthquake forecasting.

5) Geodetic Measurements

Geodetic techniques, including GPS, InSAR, and leveling, are employed to study the Earth's surface, shedding light on tectonic movements. These methods enable scientists to detect seismic shifts or pre-earthquake changes (Reinisch et al., 2020). Research using geodetic data enhances understanding of fault movements and stress accumulation leading up to earthquakes. Implementing such measurements can improve early warning systems, providing advance notice of potential earthquakes and enabling preventive measures.

2.1.3. Limitations of Traditional Approaches

One major challenge in analyzing animal behavior for earthquake prediction is the difficulty in observing and recording such behavior on a large scale (Grant & Halliday, 2010). Animal reactions could be influenced by various environmental factors, not necessarily related to seismic activity, complicating the interpretation of these behaviors. More robust and reliable methods are needed to effectively use animal behavior in earthquake forecasting (Grant & Halliday, 2010).

The reliance on historical seismic data for earthquake prediction faces several hurdles. The record of earthquakes, which began with the advent of seismometers in the late 19th century, is incomplete, especially for remote areas. This leads to gaps in understanding a region's seismic history (Stein & Wyssession, 2009). Additionally, the accuracy and detail in older seismic records are often limited, posing challenges in estimating earthquake recurrence intervals. Furthermore, past seismic data may not

accurately represent the current stress and deformation in the Earth's crust, which can change over time due to various geological and anthropogenic factors. Earthquake prediction models based on historical data must also incorporate additional data and observations, such as current seismic activity and geodetic measurements, to enhance accuracy.

Predicting earthquakes based on Radon gas emissions is complicated by various influencing factors, like weather patterns and geological changes, which can affect Radon levels in the environment (King, 1986). This variability makes it difficult to rely solely on Radon gas emissions for accurate earthquake predictions. Despite numerous studies, the connection between Radon gas levels and seismic events remains uncertain due to limited research.

The Seismic Gap Hypothesis (SGH) presupposes that earthquakes follow a pattern and that regions without recent seismic activity are more prone to future earthquakes (Kagan & Jackson, 1991). However, SGH may not apply universally, as each earthquake has unique characteristics and tectonic plates vary in their activity levels. This hypothesis can also lead to false alarms, as a lack of earthquakes in an area does not guarantee an impending event. Other factors, like stress changes or tectonic movements, also influence seismic activity.

The effectiveness of geodetic measurements is subject to various factors, such as atmospheric conditions, satellite positioning, and equipment accuracy (Burgmann & Dresen, 2008). This makes it challenging to discern whether observed surface shifts are precursors to earthquakes or just normal variations. Moreover, geodetic measurements often detect changes over extended periods, limiting their usefulness for immediate earthquake prediction. Their utility is further constrained by the need for specific sites and times for data collection, hindering the creation of a comprehensive picture of subsurface changes across a region.

2.2. Literature Review

In a comprehensive examination of ML applications for predicting various phenomena, several studies demonstrate the broad adaptability and potential of these models across different fields. The researcher reviews studies related to earthquake prediction using different ML methods.

The utilization of the USGS dataset for earthquake prediction has been central to several studies that employ ML and DL techniques to enhance prediction accuracy and reliability. These studies, spanning various methodologies and regions, collectively demonstrate the potential of advanced computational approaches in understanding and forecasting seismic events. Asim et al. (2018) advanced the field with their innovative model that integrates SVR, Hybrid Neural Networks (HNN), and Enhanced Particle Swarm Optimization (EPSO). Applied across seismic regions such as Hindukush, Chile, and Southern California, the model was trained on USGS earthquake catalogs from January 1980 to December 2016. Targeting the prediction of earthquakes with magnitudes of 5.0 and above, it adopted a binary classification framework. Similarly, Hoque et al. (2020) focused on earthquake magnitude prediction using an Artificial Neural Network (ANN) model across seismically active areas including Japan, Turkey, Greece, and the Indian Subcontinent. Their work, which also utilized USGS datasets, aimed at predicting the likelihood of earthquakes exceeding specific magnitude thresholds within a 15-day period, classifying the task as a binary classification problem. The study conducted by Jain et al. (2021) contributed to the field by deploying ML and DL algorithms such as RF Regression, Multilayer Perceptron (MLP)

regression, and SVR to predict earthquake magnitudes based on positional and depth parameters. Utilizing USGS datasets, their evaluation focused on the RMSE metric, with RF Regression achieving an RMSE of 1.731, MLP regression an RMSE of 1.647, and SVR an RMSE of 1.720. Finally, Mondol (2021) explored various ML models for predicting earthquake magnitude and depth, including RF, LR, Polynomial Regression, and LSTM. Training on real-world earthquake data, the study assessed model performance using metrics like R-squared (R^2), Explained Variance (ExVar), and MSE. Polynomial regression was modestly effective for magnitude prediction with R^2 of 0.013 and MSE of 0.1809, showing slight predictive power. The RF model excelled in depth prediction, demonstrating strong accuracy with high R^2 and ExVar scores. LR and LSTM showed limited effectiveness, with the LSTM model notably underperforming, as indicated by negative R^2 scores.

Recent studies have leveraged the LANL dataset to advance methodologies in earthquake prediction. Zaidi et al. (2020) developed three predictive models: RF, a fusion of CNN with LSTM layers, and a hybrid model combining hand-crafted features with CNN-LSTM layers. These models achieved a significant benchmark in earthquake prediction with a MAE of 1.51. Simultaneously, Brykov et al. (2020) employed the XGBoost algorithm, enhanced by extensive feature engineering, to analyze seismic data. By evaluating 48 statistical features and selecting an optimal subset of 10, they attained a noteworthy MAE of 1.913. In their study on leveraging ML for earthquake prediction, (Kumar et al., 2021) found that the Light Gradient Boosting Mechanism (LGBM) outperformed other models, including LR, SVM, RF Regression, Case-Based Reasoning, and XGBoost, in terms of prediction accuracy. By extracting 40 statistical features from this dataset and employing a methodical cross-validation

approach for testing, the study aimed to refine earthquake prediction methods. Remarkably, the LGBM model achieved the most favorable outcome, with a MAE for time to failure predictions of approximately 2.4. Furthering this domain, Yang et al. (2022) introduced Auto-REP, an automated regression model designed for the prediction of laboratory-scale seismic events. Auto-REP streamlined the predictive process through automated feature selection, algorithm choice, and hyperparameter optimization via Bayesian techniques, achieving an MAE of 1.52 and an MSE of 1.59. Finally, Zhang et al. (2023) study marks a significant contribution to earthquake prediction methodologies, employing Mel spectrogram and Mel Frequency Cepstral Coefficient (MFCC) for feature extraction from seismic signals. A DL model integrating LSTM, one-dimensional (1D) and two-dimensional (2D) CNNs, was proposed, alongside a stacking model fusion strategy that combines GB trees with DL models for optimal performance. The best result was achieved with this stacking strategy, yielding a Test-MAE of 2.238.

In their study, (Hason et al., 2023) utilized an advanced computational technique known as Molecular Particle Swarm Optimization (PSO) to refine the creation of predictive models focusing on peak ground acceleration (PGA). This research centered around the seismic region of Iraq, incorporating a comprehensive dataset of 187 historical records spanning from 2004 to 2020, sourced from global seismic stations relative to the nearest epicenter distance. The primary variables integrated into these models included the Richter Earthquake Magnitude (MW), average shear-wave velocity (Vs30), focal depth (FD), and proximity to the epicenter (REPi). However, the study faced limitations due to the scarcity of localized PGA response data and the selection of only four parameters.

Consequently, a more extensive, international dataset was employed as an alternative.

The study conducted by Usman et al., (2022) focuses on predicting ground surface deformation (GSD) resulting from earthquakes by employing ML models such as NN, RF, k-Nearest Neighbors (kNN), and GB. These models utilize data from Synthetic Aperture Radar (SAR) and the USGS earthquake data catalog, showcasing the integration of diverse data sources for enhanced prediction capabilities. The study contributes by offering a comparative analysis of different ML models for GSD prediction, although the performance and applicability of these models may be limited by the diversity and volume of the available datasets.

Proposing an earthquake detector and forecaster employing deep neural networks with LSTM units, focusing on earthquake data from the Himalaya region. The study's novelty lies in the sophisticated application of LSTM for earthquake prediction, offering potential advancements in early warning systems. However, the model's performance in diverse seismic regions and its ability to generalize to different earthquake magnitudes remain areas in their future work. Both ANN and LSTM models have been found to underperform in scenarios involving large-scale events. (Bhargava & Pasari, 2022).

Employing seven different ML classifier algorithms, namely Bayes Net, Random Tree, Simple Logistic, RF, Logistic Model Tree (LMT), ZeroR, and LR. These classifiers are applied to six diverse datasets pertaining to India and its neighboring regions, utilizing the Weka tool for model development. In the study by Debnath et al., (2021), the focus is on leveraging ML classifiers to forecast earthquake magnitudes, categorizing

them into three distinct ranges: fatal, moderate, and mild. The study offers a comprehensive analysis of different algorithms for earthquake forecasting, particularly in the Indian context, with potential limitations related to the scalability and generalizability of the models as the study is confined to India and its nearby regions.

The study by Berhich et al., (2020) leverages LSTM models to predict earthquake magnitudes in Morocco, a region vulnerable to seismic activities. The researchers utilized a dataset from the National Geophysical Institute, covering seismic events from 1900 to 2019. Two approaches were employed: one using a singular model for all datasets and another dividing the data based on magnitude ranges to tailor the prediction models to specific types of seismic events. The LSTM models demonstrated significant potential in earthquake prediction. The approach of decomposing datasets based on magnitude ranges yielded more accurate predictions, particularly for larger earthquakes. This suggests that considering the distinct characteristics of seismic events can enhance prediction models. The study's LSTM models outperformed traditional ANN models, highlighting the effectiveness of LSTM in handling complex, time-series data for earthquake prediction. In contrast, the study's reliance on evaluation metrics from the training phase without extensive validation on unseen data raises concerns about the models' generalization ability and the risk of overfitting. Additionally, the lack of integration of diverse data types beyond seismic activity, such as environmental and geological data, may limit the models' predictive accuracy.

An innovative model for predicting the occurrence and location of earthquakes proposed by (Al Banna et al., 2021), leveraging the capabilities of LSTM networks enhanced with an attention mechanism. By

examining the seismic activity in Bangladesh, a region known for its vulnerability to earthquakes, the research introduces an approach that combines various LSTM architectures and dense layers to forecast seismic events. Utilizing data from the earthquake catalog of Bangladesh, the model employs seismic indicators as features to predict earthquakes in the subsequent month. The incorporation of an attention mechanism aims to bolster the model's accuracy in forecasting earthquake occurrences, achieving a notable accuracy of 74.67%. Additionally, a regression model predicts the earthquake epicenter's distance from a predefined location, demonstrating a RMSE of 1.25. However, the study identifies limitations such as limited geographic applicability, reliance on a predefined set of seismic indicators, and the model's prediction timeframe restricted to monthly occurrences.

The study conducted by (Kavianpour et al., 2023) introduces an innovative earthquake prediction model that combines CNN, BiLSTM and Attention Mechanism (AM) to predict the number of expected earthquakes in a month over the past 50 years and their maximum magnitude for nine different regions. This model addresses the challenge of earthquake prediction by leveraging the strengths of CNN and BiLSTM to capture spatial and temporal dependencies in seismic data, enhanced by AM to prioritize significant features for accurate prediction. Investigating the model's performance over different prediction windows (e.g., weekly, daily) could enhance its utility for short-term earthquake early warning systems.

An innovative artificial intelligence-based model for earthquake prediction in Indonesia, employing the RF Regression algorithm. The model utilizes a comprehensive dataset of Indonesian earthquake records spanning from January 1900 to January 2022. Key attributes include date,

location, latitude, longitude, hypocenter, magnitude, depth of the earthquake, maximum peak ground acceleration, modified Mercalli intensity, and site class. The target prediction label is the modified Mercalli intensity, aiming to enhance earthquake mitigation strategies in Indonesia. With a split of 80% training data and 20% testing data, the model achieved a remarkable accuracy of 98.8%, demonstrating low error rates with mean absolute error and mean square error values of 5.4% and 4.6%, respectively. (Wijaya et al., 2022).

The study conducted by (Abdul Salam et al., 2021) introduces two groundbreaking earthquake prediction models employing hybrid ML techniques tailored for Southern California. The dataset used encompasses historical earthquake data for Southern California, spanning from January 1, 1950, to May 31, 1978. The study meticulously processes historical earthquake data to calculate seven seismic indicators, forming the basis for the models' inputs. By integrating the Flower Pollination Algorithm (FPA) with Extreme Learning Machine (ELM) and Least Square Support Vector Machine (LS-SVM), the research presents a novel approach to forecasting earthquake magnitudes within a fifteen-day timeframe. The models, FPA-ELM and FPA-LS-SVM, undergo comprehensive evaluation through performance metrics such as RMSE, MAE, Symmetric Mean Absolute Percentage Error (SMAPE), and Percent Mean Relative Error (PMRE), demonstrating the FPA-LS-SVM model's showing the lowest error rates, RMSE of 0.540 and MAE of 0.434.

Exploring the application of ML techniques to predict major earthquake events (Mallouhy et al., 2019). They categorize earthquake predictions into short-term predictions (hours or days in advance) and forecasts (months to years in advance), focusing on the latter due to the complex and chaotic nature of earthquake phenomena. The novelty of their

work lies in applying eight different ML algorithms on a real earthquake dataset to classify events as negative or positive major earthquakes. The algorithms include RF, Naive Bayes, LR, MLP, AdaBoost, K-nearest KNN, SVM, and Classification and Regression Trees (CART), each with various hyperparameters and evaluation metrics. The exceptional performance of RF in this study can be credited to its ensemble method, which combines multiple decision trees to improve prediction accuracy. It demonstrated a well-rounded performance regarding MAE and RMSE, posting values of 0.312 and 0.416, respectively.

Table 1 shows the summary of related works to earthquake predictions using ML. The literature review reveals several attempts to address different aspects of earthquake predictions. However, these studies often address only parts of the broader prediction challenge, leaving critical gaps in comprehensive forecasting capabilities. This thesis identifies a crucial research gap: the need for effective ML models to accurately predict the magnitude and timing of seismic activities is critical, as is understanding the challenges and complexities involved in applying these models to seismic data. Additionally, this study will conduct a comparative analysis of different ML models to further understand their performance in seismic prediction.

Table 1 : Summary of Related Works

Reference	Dataset Used	Proposed Model(s)	Novelty	Results
(Asim et al., 2018)	USGS	combines SVR, HNN, and EPSO	Predicting the occurrence of earthquakes with a magnitude of 5.0 and above.	Binary Classification
(Hoque et al., 2020)	USGS NEIC	ANN	Predict the occurrence of earthquakes within specific magnitude thresholds for different regions.	higher accuracy Achieved of 87% for India region.
(Jain et al., 2021)	USGS	RF Regression, MLP Regression, SVR	Earthquake prediction using ML and DL based on position and depth from USGS data	RMSE RF: 1.731 MLP:1.647 SVR: 1.720
(Mondol, 2021)	USGS	LR RF Polynomial Regression LSTM	Predicting the magnitude and depth of an earthquake	Polynomial Regression MSE of 0.180 and R^2 of 0.013 for magnitude
(Brykov et al., 2020)	LANL	XGBoost	ML modeling using LANL data, employing feature engineering	MAE was 1.913

(Zaidi et al., 2020)	LANL	RF, CNN + LSTM, Hand-crafted features + CNN + LSTM	Prediction of time until the next earthquake using diverse model combinations	Hand-crafted MAE of 1.510
((Kumar et al., 2021)	LANL	LGBM, LR, SVR, RF	Extracting 40 features from signal data for predicting time to failure.	LGBM with MAE of 2.4
(Yang et al., 2022)	LANL	Auto-REP	Automated feature extraction and hyperparameter optimization using Bayesian techniques using LANL Dataset	MAE: 1.520 MSE: 1.590
(Zhang et al. 2023)	LANL	- Combines LSTM +1D CNN and with 2D CNN Stacking model strategy (combines GB trees with DL models)	employing Mel spectrogram and Mel Frequency Cepstral Coefficient (MFCC) for feature extraction from seismic signals	Stacking strategy model with MAE of 2.238
(Bhargava & Pasari, 2022)	earthquake data from the Himalaya region	LSTM	Earthquake detector and forecaster using deep neural networks for the Himalaya region	Both ANN and LSTM models have been found to underperform in scenarios involving large-scale events.

(Debnath et al., 2021)	Six different dataset	Bayes Net, Random Tree, Simple Logistic, RF, LMT, ZeroR, LR	leveraging ML classifiers to forecast earthquake magnitudes.	LMT outperforms other with 99.94 Acc
(Berhich et al., 2020)	dataset from the National Geophysical Institute	LSTM	leverages LSTM models to predict earthquake magnitudes in Morocco	Training phase MAE of 0.075 MSE of 0.014
(Al Banna et al., 2021)	data from the earthquake catalog of Bangladesh	LSTM with attention mechanism	Earthquake occurrence prediction using LSTM with an attention mechanism based on seismic indicators	Earthquake occurrence (classification) Accuracy 74.67 % And Distance (regression) RMSE 1.558
(Kavianpour et al., 2023)		combines CNN, BiLSTM and Attention Mechanism (AM)	predict the number of expected earthquakes in a month over the past 50 years and their maximum magnitude for nine different regions.	proposed method achieved RMSE= 0.074 MAE =0.076 R2=0.906
(Wijaya et al., 2022)	dataset of Indonesian earthquake records (1900-2020)	RFR	predict the MMI as a measure of the intensity and damage level caused by an earthquake	accuracy of 98.8%, MAE 5.4% and MSE 4.6%,

Salam et al (2021)	historical earthquake data for Southern California	FPA ELM LS-SVM FPA-ELM FPA-LS-SVM	the research presents a novel approach to forecasting earthquake magnitudes within a fifteen-day timeframe	FPA-LS-SVM achieved RMSE of 0.540 and MAE of 0.434 .
(Mallouhy et al., 2019)	Earthquake Data Center in Northern California	LR, RF, Naive Bayes, MLP, AdaBoost, KNN, SVM, and CART	Predict the major of an earthquake as negative or positive.	RF achieved MAE and RMSE, posting values of 0.312 and 0.416, respectively.

Chapter Three

Methods and Procedures

3.1. Research Design

3.2. Implemented Models

3.3. Performance Metrics

Chapter Three

Methods and Procedures

This chapter provides an overview of the methodology implemented in this research.

3.1. Adopted Methodology

Figure 1 illustrates a comprehensive ML pipeline tailored for earthquake predictive modelling, which is crucial for the development of accurate and reliable predictions. Initially, raw data undergoes a series of preprocessing steps, including cleaning, and exploratory analysis, to ascertain its readiness for further processing. Subsequently, this prepared data is enhanced through feature engineering techniques such as extraction, normalization, and selection, which collectively fine-tune the input variables for the predictive models. Subdividing the data into train and test sets is a pivotal step, and this is where the train-test split method comes into play. With the data duly partitioned, the training phase engages a diverse array of algorithms ranging from conventional ones like LR and SVR to more complex DL architectures including LSTM, and GRU, as well as advanced ensemble methods like XGBoost and RFR. Post-training, the models are ushered into a rigorous testing phase with the intent of accurately predicting a specific outcome. The efficacy of these models is rigorously quantified through various performance metrics such as MSE, MAE, and R-squared, all of which are instrumental in discerning the accuracy of the forecasts and in selecting the superior model for deployment. This meticulous and methodical approach is designed to

enhance the predictive power of the models, ensuring that the projections made are both trustworthy and applicable.

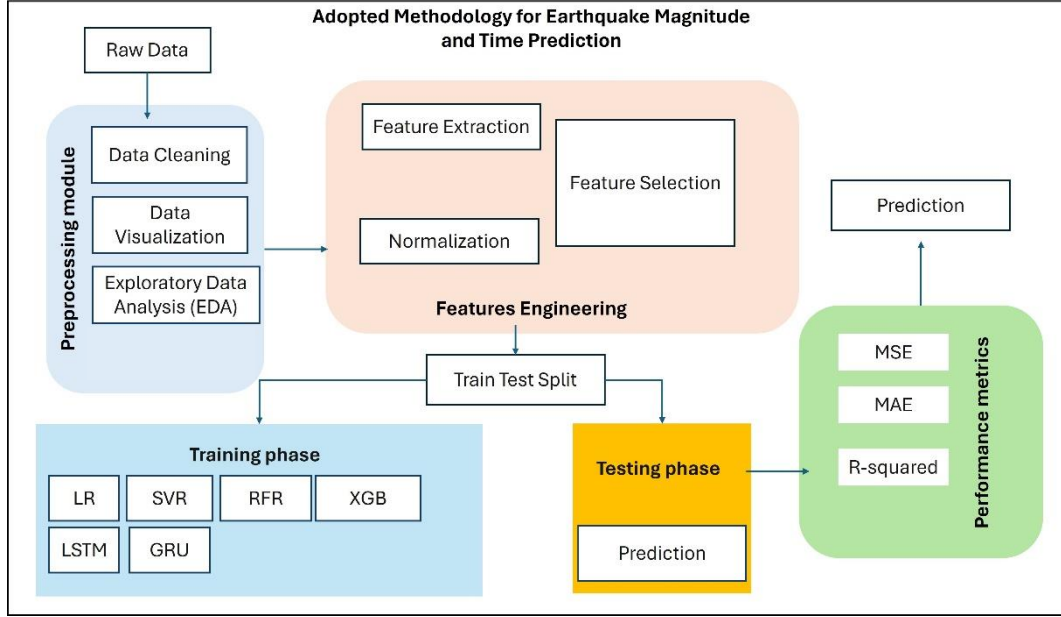


Figure 1: Adopted Methodology Used

3.2. Tested Approaches

The selection of ML models for the current research, specifically, LR, RFR, SVR, XGBoost, and two DL proposed architectures, LSTM and GRU was influenced by their demonstrated efficacy in prior research (Berhich et al., 2020; Jain et al., 2021; Kumar et al., 2021). The goal is to adapt and refine these models through comprehensive hyperparameter tuning and adjustments to their configurations, ensuring their optimal performance for the specific requirements and data of this research. Detailed descriptions of these approaches were provided in section 2.1.1.

3.3. Performance Metrics

To evaluate the effectiveness of different models, a variety of metrics were employed, acknowledging that no single metric provides a definitive

measure of model performance. The evaluation strategy incorporated three commonly used metrics for assessing prediction methods: MAE, MSE and R-squared. The definitions and computations of these performance metrics, utilizing y for actual values, $\hat{y}$ for predicted values and n is the number of observations, are as follows:

- 1) MAE, or Mean Absolute Error, calculates the average of absolute differences between the predicted and actual values. It is given by the formula:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

MAE is sensitive to the scale of the data, making it less suitable for comparing series with different magnitudes. Nevertheless, it is a popular choice for evaluating models on the same dataset due to its resilience to outliers.

- 2) MSE is a statistical metric used to measure the accuracy of a model by quantifying the average of the squared differences between the predicted and actual values. It's an indicator of the quality of an estimator; a MSE of zero indicates perfect prediction accuracy, with no deviation between predicted and observed outcomes. As errors in model predictions increase, the MSE value rises, indicating lower model accuracy.

The formula for MSE is expressed as:

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n} \quad (2)$$

MSE is commonly referred to as mean squared deviation (MSD). It is particularly useful in quantifying the variance of errors, as it squares each difference, thus assigning greater weight to larger errors. This makes MSE sensitive to outliers and can help in identifying models that are particularly poor at predicting extreme values.

- 3) R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (3)$$

Where, SS_{res} is the sum of squares of residuals, and SS_{tot} is the total sum of squares. In essence, R-squared quantifies the degree of correlation between the predicted and actual values, offering a metric to gauge the accuracy of a predictive model.

Chapter Four

Results and Discussion

4.1. Data Sources

4.1. Results

4.2. Discussion

Chapter Four

Results and Discussion

4.1. Data Sources

This section provides an overview of the two datasets utilized in conducting the experiments of this thesis.

4.1.1. USGS Dataset

The USGS Earthquake Dataset provides a detailed and real-time catalog of seismic activities worldwide. Each record encapsulates comprehensive details about an earthquake, including geographic coordinates, occurrence time, depth, magnitude, and other relevant attributes. This dataset's dynamic nature, with continuous updates, makes it an invaluable resource for understanding and analyzing seismic events globally. The USGS Earthquake dataset, sourced from the U.S. Geological Survey (<https://earthquake.usgs.gov/earthquakes/search/>), encompasses comprehensive details of global seismic events. It reflects the USGS's commitment to providing scientific data to understand Earth's dynamics, mitigate natural hazards, and manage natural resources.

The dataset compiles records spanning from the 1st of January to the 31st of December 2023, including only seismic events with a magnitude of 3.5 or higher. It contains 17,730 instances, each with 22 features: time, latitude, longitude, depth, gap, mag, magType, nst, dmin, rms, net, id, updated, place, type, horizontal, depthError, magError, magNst, status,

locationSource, magSource. Table 2 shows the description of the dataset features.

Table 2 :USGS dataset Description

Feature	Description	Datatype
Time	Time when the event occurred	Long Integer
latitude	Decimal degrees latitude. Negative values for southern latitudes.	decimal
longitude	Decimal degrees longitude. Negative values for western longitudes.	decimal
Depth	Depth of the event in kilometers.	decimal
Mag	The magnitude for the event	decimal
magType	The method or algorithm used to calculate the preferred magnitude for the event	string
Nst	The total number of seismic stations used to determine earthquake location.	integer
Gap	The largest azimuthal gap between azimuthally adjacent stations (in degrees)	decimal
dmin	Horizontal distance from the epicenter to the nearest station (in degrees)	decimal
rms	The root-mean-square (RMS) travel time residual, in sec, using all weights.	decimal
net	The ID of a data contributor.	string
Id	A unique identifier for the event	string
updated	Time when the event was most recently updated.	Long integer
place	Textual description of named geographic region near to the event.	string
type	Type of seismic event.	string
horizontalError	Uncertainty of reported location of the event in kilometers.	decimal

depthError	Uncertainty of reported depth of the event in kilometers.	decimal
magError	Uncertainty of reported magnitude of the event	decimal
magNst	The total number of seismic stations used to calculate the magnitude for this earthquake.	integer
status	Indicates whether the event has been reviewed by a human.	string
locationSource	The network that originally authored the reported location of this event.	String
magSource	Network that originally authored the reported magnitude for this event.	string

The dataset is instrumental for:

- Real-time earthquake monitoring and reporting.
- Assessing earthquake impacts and hazards.
- Conducting research on the causes and effects of earthquakes. Figure

2 shows sample of USGS dataset.

```
[ ] # Display the first 5 rows of data
df.head()
```

	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	...	updated	place	type	horizontalError	depthError	magError	magNst	status
0	2023-12-31T23:18:09.731Z	-4.2590	143.5602	133.548	4.2	mb	40.0	117.0	6.237	0.53	...	2024-03-02T21:39:37.040Z	60 km WSW of Angoram, Papua New Guinea	earthquake	12.56	7.300	0.103	26.0	reviewed
1	2023-12-31T22:52:31.316Z	-2.6886	141.4208	10.000	4.2	mb	26.0	78.0	9.164	0.68	...	2024-03-02T21:39:37.040Z	13 km E of Vanimo, Papua New Guinea	earthquake	12.18	1.924	0.121	19.0	reviewed
2	2023-12-31T22:46:19.016Z	-17.5065	-178.3327	540.654	4.4	mb	46.0	49.0	3.455	0.54	...	2024-03-02T21:39:37.040Z	256 km ENE of Levuka, Fiji	earthquake	8.42	8.128	0.098	30.0	reviewed
3	2023-12-31T22:13:49.174Z	13.9526	144.5892	193.928	4.2	mb	38.0	100.0	0.452	0.53	...	2024-03-02T21:39:37.040Z	54 km NNW of Asan-Maina Village, Guam	earthquake	7.53	6.416	0.103	26.0	reviewed
4	2023-12-31T19:23:30.665Z	-28.0143	-177.2923	91.774	4.1	mb	10.0	141.0	1.348	0.41	...	2024-03-02T21:39:37.040Z	Kermadec Islands region	earthquake	11.85	9.357	0.212	6.0	reviewed

5 rows x 22 columns

Figure 2: First 5 rows of USGS dataset.

4.1.1.1 USGS Dataset Preprocessing

Data preprocessing is a technique that converts given primary data into a clean dataset to make it suitable for use. Here the researcher goes through these steps for the data preprocessing.

1) Data Cleaning

After loading the dataset, a meticulous check for missing values is performed. Columns with a significant number of missing values, specifically 'dmin', 'magError', and 'magNst', are entirely removed from the dataframe to ensure data quality. Conversely, for columns with a minimal number of missing entries, such as 'nst', 'gap', and 'horizontalError', the missing values are imputed with the mean of the available data, preserving as much information as possible. Following the initial cleanup, it's crucial to verify the accuracy and plausibility of the remaining dataset. This includes ensuring all data falls within acceptable ranges, such as verifying that latitude values lie between -90 and 90 degrees, and longitude values are within the -180 to 180 degree range. Such checks help identify any outliers or incorrect entries that could skew analysis results. Lastly, it is important to inspect the target variable, in this case, the magnitude of earthquakes 'mag', to confirm that there are no nonsensical values. Specifically, all earthquake magnitudes should be positive, as a magnitude less than 0 does not represent a valid measurement. This comprehensive approach to preprocessing not only cleans the data but also ensures its integrity for subsequent analysis or modelling tasks.

2) Feature Engineering

Feature engineering is a vital process in data science and ML that involves creating new features from existing data to improve the performance of predictive models. One common challenge in feature

engineering is dealing with various data types appropriately to maximize their utility in analysis. A notable example is the 'time' column in datasets, which often comes in an object data type, essentially a string format. This format is not immediately suitable for time-based analysis, which requires numerical or datetime data types for computations and comparisons. To address this, the 'time' column can be transformed into a datetime format using the ``to_datetime()`` function provided by Pandas and then duplicate records dropped. Regarding other columns in the dataset, it's crucial to review and ensure their data types are appropriate for their respective data. In this scenario, the data types of other columns are deemed suitable for analysis as they are. This means there is no need for additional data type transformations outside of the 'time' column.

3) Feature Selection

In the process of preparing the dataset for models training, some columns are removed from the dataset as follow:

- **net, id, updated:** These columns are likely identifiers and metadata related to the earthquake recording. They are usually specific to the dataset's collection process and do not hold predictive power for the magnitude of earthquakes.
- **status, locationSource, magSource:** These columns may provide information about the status of the event record and the sources for location and magnitude data. Like identifiers, this metadata is typically not directly related to the physical properties of earthquakes that influence their magnitude.
- **type:** it refers to the type of seismic event and in our task values are 'earthquake'.
- **place, region:** Describes the location of the earthquake. While geographical location can significantly influence earthquake

characteristics, it's important to note that the dataset already includes **latitude** and **longitude** values, offering precise, numerical representations of each earthquake's location.

Using both a correlation matrix and mutual information to identify relevant features for predicting the target variable (mag), represents a comprehensive approach to feature selection. This method combines linear (correlation) and non-linear (mutual information) assessments to uncover the most informative predictors within the dataset. Figure 3 shows the correlations matrix for the remaining of numerical features, which is a powerful tool for summarizing and visualizing the strength of relationships between variables. For example, 'mag' and 'nst' show a notable positive correlation, which suggests that as the value of one increases, the value of the other tends to increase as well. Conversely, 'mag' and 'gap' exhibit a strong negative correlation, indicating that a higher value in one may correspond to a lower value in the other. Mutual information technique has been done with the whole features after encoding the categorical features, as it shown int Figure 4, the highest scores are observed for features like 'nst', 'depth error', 'longitude', 'latitude', and 'magnitude', suggesting these variables are most predictive of the outcome. Moreover, In the feature selection process, geological and geophysical attributes were considered, reflecting their importance in earthquake prediction models. While 'depth' exhibited relatively lesser relevance among the features studied, it was retained in the feature selection process because geological and geophysical attributes are considered significant in this context.

after training of the models and leveraging feature importance metrics derived from a RF model, the models were evaluated across three different subsets of selected features:

Subset 1: latitude, longitude, depth, mag, nst, gap, rms, horizontalError, depthError.

4) Data Visualization and Analysis

It is generally a good idea to perform exploratory data analysis (EDA) to gain a better understanding of the data and to identify any patterns or relationships that may exist. EDA can involve tasks such as data visualization and summary statistics.

It makes sense to study the basic statistics regarding the target attribute which is 'mag'. As it is shown from Table 3, The dataset comprises 16592 observations. The mean (average) magnitude across all observations is approximately 4.442, indicating a central tendency around this value. The standard deviation, a measure of dispersion or variability among the magnitudes, is relatively low at 0.441, suggesting that most magnitudes are clustered around the mean. The minimum recorded magnitude is 3.500, while the maximum is significantly higher at 7.800, indicating a wide range of magnitudes within the dataset. The quartile values provide additional insights into the distribution: the first quartile (25%) is at 4.200, the median (50%) is at 4.400, and the third quartile (75%) is at 4.600. These quartiles indicate that half of the observations fall within a narrow range of 4.200 to 4.600, demonstrating a concentration of magnitudes within this interval.

Figure 5 depicts the histogram of 'mag' distribution. From this histogram, we can observe that:

- The distribution has a peak around the 4.4 to 4.5 range, which is in line with the median value provided in the summary statistics.
- The shape of the distribution is skewed to the right, indicating that there are a few higher magnitude values that are less common, with the tail extending towards the maximum value of 7.8.

- The concentration of values is mostly between 4.0 and 5.0, which corresponds to the interquartile range from the summary statistics, where 50% of the data lies.
- The spread towards the lower magnitudes is more abrupt compared to the gradual decline toward the higher magnitudes.

Table 3: Magnitude Descriptive Statistic

	mag
count	16592.000
mean	4.442
std	0.441
min	3.500
quartile 25%	4.200
quartile 50%	4.400
quartile 75%	4.600
max	7.800

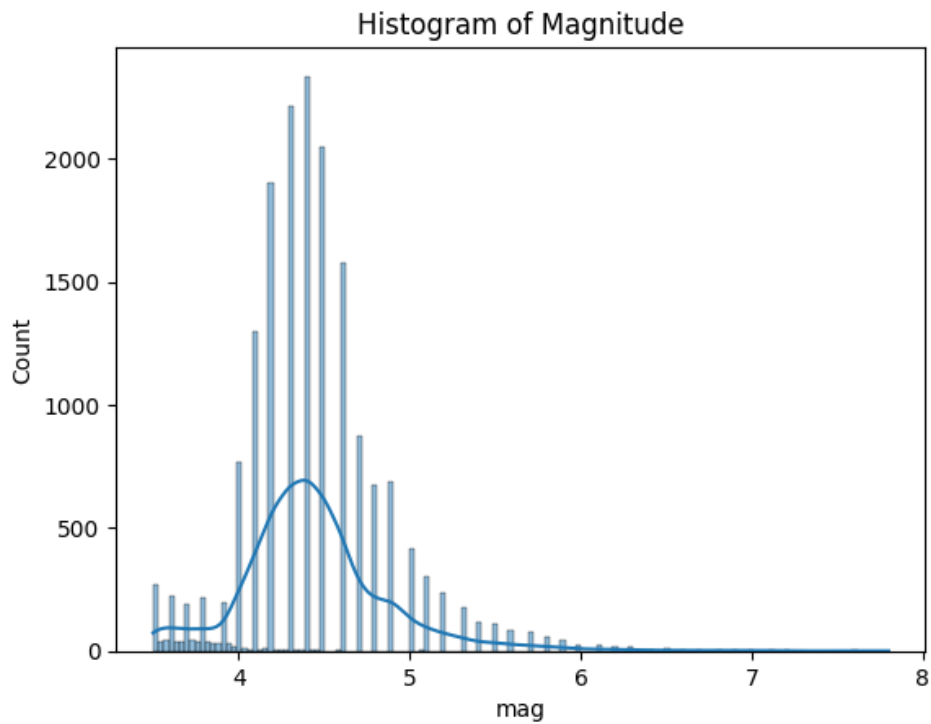


Figure 5:Earthquake Magnitude Histogram

Figure 6 displaying the magnitude of the earthquake over time, and it shows fluctuations in magnitude over the year of 2023.

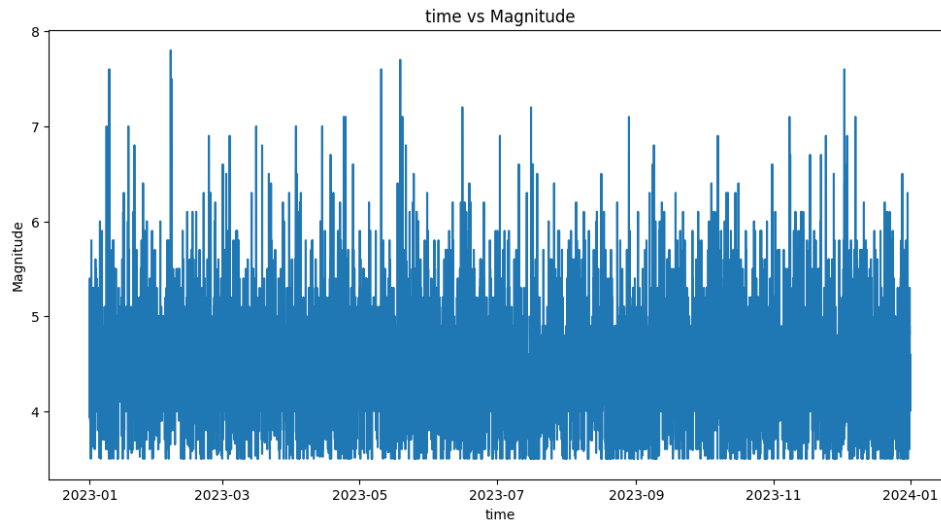


Figure 6: Magnitude over Time

Table 4 shows the descriptive statistics of the selected features. These features include latitude, longitude, depth, gap, and nst (number of stations reporting), each with data for 16,592 observations. The statistics cover measures such as the mean, standard deviation, minimum, and quartiles. Latitude ranges from -65.849 to 86.593, and longitude from -179.998 to 179.999, indicating a global dataset. The depth of the seismic events varies significantly, with a minimum of 0.024 meters and a maximum of 681.238 meters. The gap, which likely measures the angular separation of seismic data collection points, varies from 8 to 347 degrees. Lastly, the number of stations reporting (nst) varies from 3 to 466, highlighting the diverse scale of seismic monitoring efforts captured in the data.

Table 4: Descriptive Statistics for Selected Features

	latitude	longitude	depth	gap	nst
count	16592	16592	16592	16592	16592
mean	6.911	31.781	82.606	108.353	49.029
std	28.740	122.355	133.380	52.787	41.674
min	-65.849	-179.998	0.024	8.000	3.000
quartile 25%	-16.426	-71.456	10.000	70.000	22.00
quartile 50%	7.998	78.0364	29.717	102.000	35.00
quartile 75%	30.638	139.380	88.895	134.000	61.000
max	86.593	179.999	681.238	347.000	466.000

4.1.2. LANL Dataset

The LANL Earthquake Prediction dataset is a specialized collection of seismic signals aimed at forecasting the timing of laboratory earthquakes (Yang et al., 2022). Originating from a renowned experimental setup for studying earthquake physics, this dataset's primary purpose is to predict the impending time-to-failure of laboratory-induced earthquakes. The dataset, provided by Los Alamos National Laboratory, not only contributes to the field of seismology but also supports broader national security objectives.

The dimensions of the data are huge, with over 600 million rows of information. The two columns within the training dataset have the subsequent meaning.

1. Acoustic data: It is the acoustic signal measured within the laboratory experiment, which has an integer datatype.

2. Time to failure: It provides the time of the next failure, which has a float datatype.

The dataset is structured into two primary components: training data and test data.

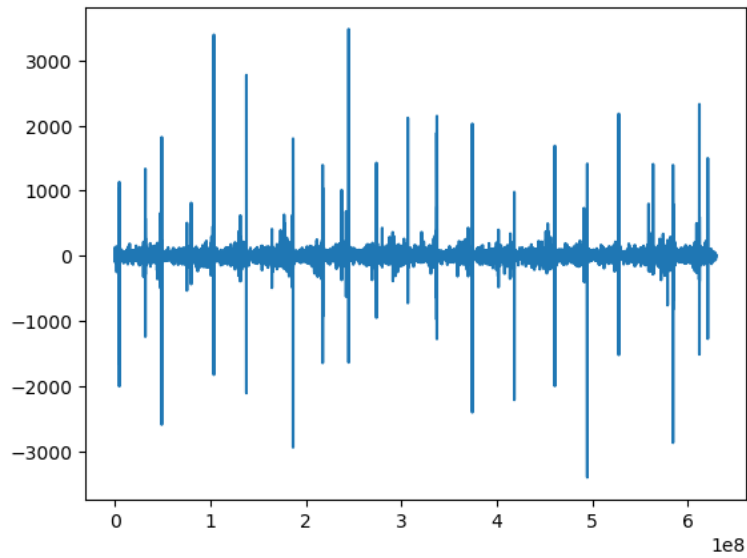
1. Training Data (train.csv): This is a single, continuous segment of experimental data. The file size is 8.89GB.

2. Test Data (Folder test): Comprises numerous small segments of test data. Each file within this folder represents a continuous segment, but collectively, they do not form a continuous series.

In this study, proposed models will use the train dataset using ‘TimeSeriesSplit’ which is a method provided by scikit-learn, used for cross-validation of time series data. Unlike standard cross-validation methods that shuffle data randomly, TimeSeriesSplit keeps the temporal order of observations, making it particularly suitable for time series forecasting where the sequence and timing of data points are crucial.

1) Exploratory Data Analysis (EDA)

Given the impracticality of visualizing every piece of collected data, we opted to conduct and present an EDA on just 1% of the data. This approach allows us to manageably illustrate key insights and patterns within the dataset. Figure 7 depicts 1% Total Acoustic Data, plotted against time and as it is observed the acoustic data shows very complex oscillations with variable amplitude.



When plotting both the Time to Failure and Acoustic Data (sampling 1% of the data) on a single graph, it is observed that there is a noticeable increase in the amplitude of the acoustic data just prior to each failure. We also see that numerous amplitudes have been observed in different moments in time. Figure 8 shows the acoustic data and time to failure: 1% data.

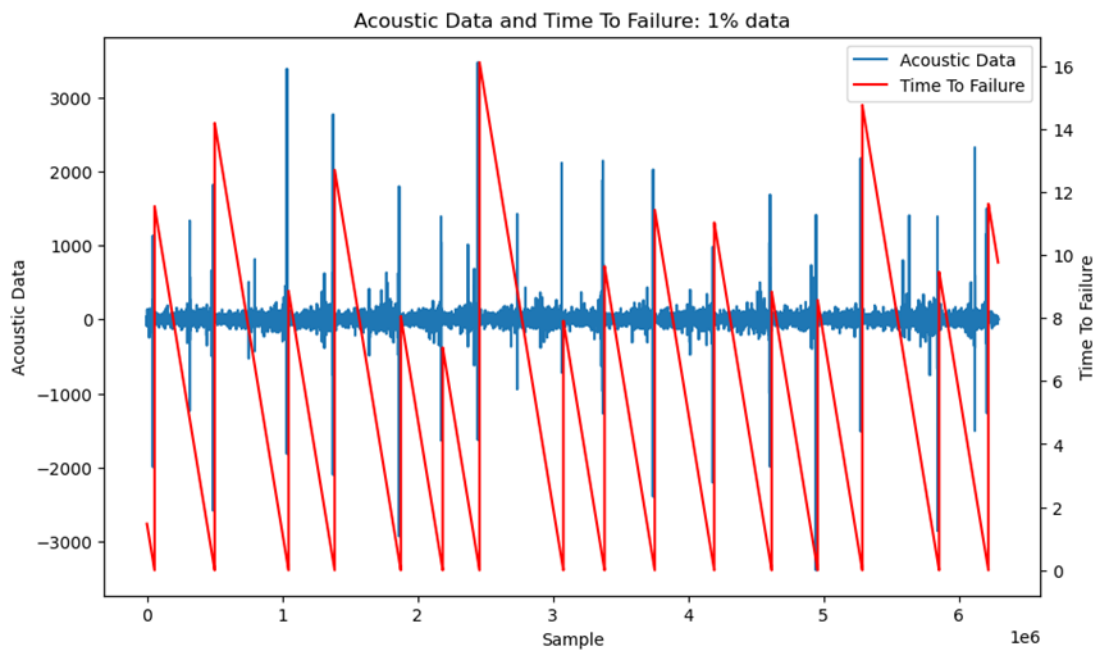


Figure 8: Acoustic Data and Time to Failure (1% data)

Figure 9 presents a density plot for 25,000 samples of acoustic data. It is evident that the bulk of the values fall within the range of approximately -10 to 10. Additionally, the presence of outliers in both the positive and negative directions is clearly visible.

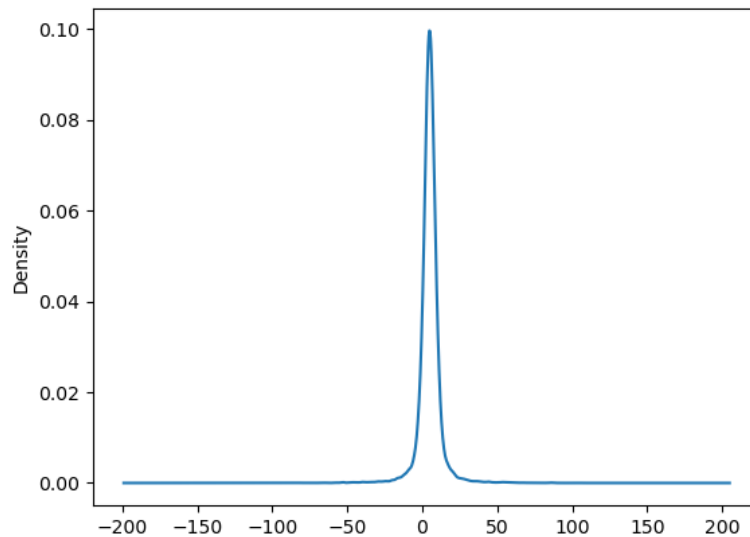


Figure 9: Density for 25k samples of Acoustic Data

Figure 10 displays a density plot for 25,000 instances of "time_to_failure" revealing an uptick in target data density beginning at 1.462 and a decline around 1.470. Furthermore, it is observed that the density levels at 1.464, 1.466, 1.468, and 1.470 remain consistently high and stable.

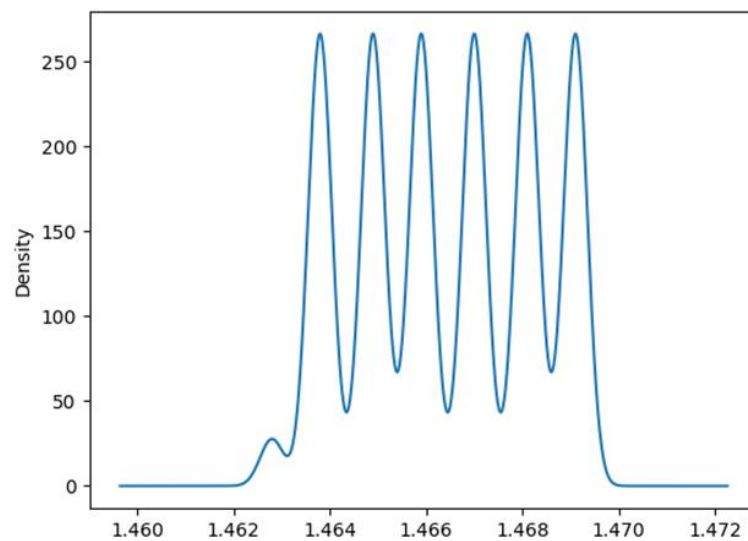


Figure 10: Density plot for 25k instances of "time_to_failure,"

Figure 12 showcases a box plot of 25000 samples of acoustic data, highlighting the presence of outliers within the dataset. It is observed that the 25th, 50th (median), and 75th percentiles are situated between -15 and 15, indicating the central distribution of most of the data points.

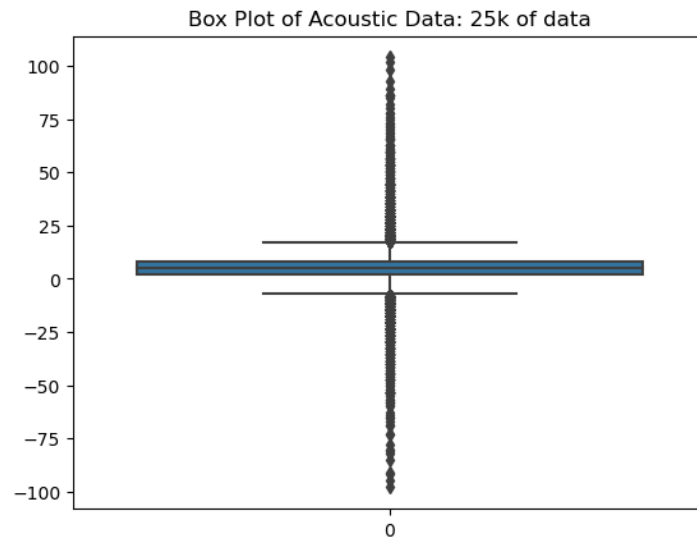


Figure 11: Boxplot of Acoustic Data (25k samples)

Figure 12 illustrates a boxplot of "time to failure" for 25,000 data points. It reveals that there are no outliers within the target feature. Specifically, it indicates that less than 25% of the targets have a "time to failure" of less than 1.465, less than 50% of the targets have a "time to failure" of less than 1.466, and less than 75% of the targets have a "time to failure" of less than 1.468.

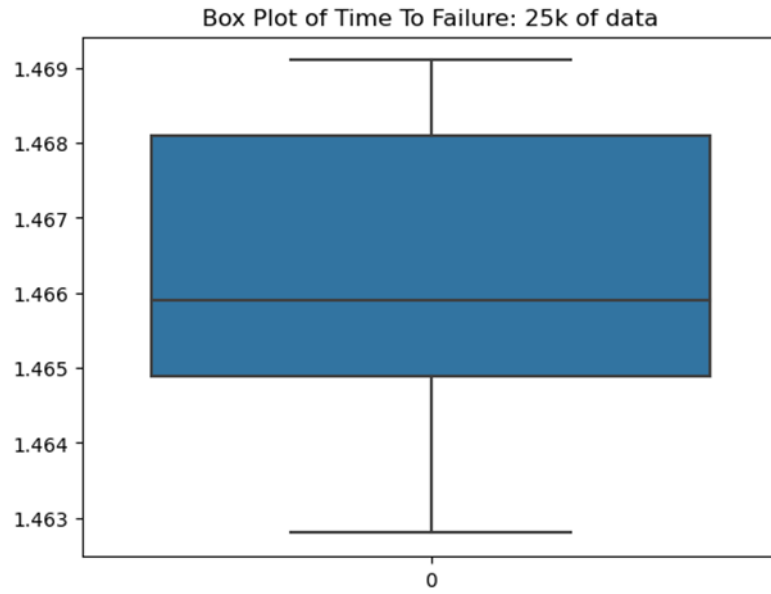


Figure 12: Boxplot of Time to Failure (25k)

2) LANL Data preprocessing and Feature Engineering

The methodology for feature engineering began with the premise that the distribution of Acoustic Data (AD) holds valuable features for analysis. This assumption was supported by practical reasoning, observed changes in AD distribution over time (Figure 8).

With the help of the study by (Kumar et al., 2021), we employed a comprehensive feature engineering strategy to analyze seismic and acoustic data from laboratory micro-earthquake simulations. We extracted a rich set of 58 features from 4194 data segments, each consisting of 150,000 observations. These features encompass basic statistical metrics such as maximum, minimum, median, standard deviation, and interquartile range, along with more complex time series-derived features like Auto-Regressive Model Coefficients, Fast Fourier Transformation Variance, and Spectral Welch Density. Figure 13 shows the data shape and the features used for testing the models.

```

✓ [36] print(Data.shape)
      Data.columns

(4194, 58)
Index(['Min', 'SWD_Min', 'Real_FFT_Min', 'Imag_FFT_Min', 'Max', 'SWD_Max',
      'Real_FFT_Max', 'Imag_FFT_Max', 'Mean', 'SWD_Mean', 'Real_FFT_Mean',
      'Imag_FFT_Mean', 'Median', 'Std', 'SWD_Std', 'Real_FFT_Std',
      'Imag_FFT_Std', 'Kurtosis', 'Skew', 'IQR', 'Abs_Mean', 'Abs_Std', '25%',
      '50%', '75%', '90%', '95%', '99%', 'Var_Roll_100', 'Avg_Roll_100',
      'Min_Roll_100', 'Abs_Mean_Roll_100', '1%_Roll_100', '5%_Roll_100',
      '10%_Roll_100', '90%_Roll_100', '95%_Roll_100', '99%_Roll_100',
      'Var_Roll_1000', 'Avg_Roll_1000', 'Min_Roll_1000', 'Abs_Mean_Roll_1000',
      '1%_Roll_1000', '5%_Roll_1000', '10%_Roll_1000', '90%_Roll_1000',
      '95%_Roll_1000', '99%_Roll_1000', 'Var_Roll_10000', 'Avg_Roll_10000',
      'Min_Roll_10000', 'Abs_Mean_Roll_10000', '1%_Roll_10000',
      '5%_Roll_10000', '10%_Roll_10000', '90%_Roll_10000', '95%_Roll_10000',
      '99%_Roll_10000'],
      dtype='object')

```

Figure 13: LANL Feature set (58).

1. Min, Max, Mean, Median, Std: These are basic statistical measures. 'Min' and 'Max' represent the minimum and maximum values in each segment, respectively. 'Mean' is the average value, 'Median' is the middle value when data is sorted, and 'Std' is the measure of the amount of variation or dispersion of the data points.

2. SWD (Spectral Welch Density): Prefixes like 'SWD_Min', 'SWD_Max', 'SWD_Mean', and 'SWD_Std' indicate measurements derived from the power spectral density estimate using Welch's method, calculated for various statistics like minimum, maximum, mean, and standard deviation.

3. Real_FFT and Imag_FFT: These prefixes refer to the real and imaginary components of the Fast Fourier Transform of the data. 'Real_FFT_Min', 'Real_FFT_Max', etc., represent the minimum, maximum, mean, and standard deviation of the real part of the Fourier transformed data, with similar metrics for the imaginary part.

4. Kurtosis and Skew: 'Kurtosis' measures the tailedness of the probability distribution of a real-valued random variable. 'Skew' measures the asymmetry of the probability distribution of a real-valued random variable about its mean.

5. IQR (Interquartile Range): This is the difference between the 75th percentile and the 25th percentile. It measures the statistical dispersion or spread of the data points.

6. Abs_Mean, Abs_Std: These are the mean and standard deviation calculated on the absolute values of the dataset. 'Abs_Mean' represents the average of the absolute values, and 'Abs_Std' the dispersion of the absolute values.

7. Percentiles (25%, 50%, 75%, 90%, 95%, 99%): These represent the values below which a certain percent of data in the dataset falls. For example, 25% is the value below which 25% of the data lies.

8. Rolling Features: These are statistical features calculated over a rolling window (subsets of data):

- **Var_Roll, Avg_Roll, Min_Roll, Abs_Mean_Roll:** Variance, average, minimum, and mean of absolute values calculated over rolling windows of different sizes (100, 1000, 10000).

- **Percentile_Roll (1%, 5%, 10%, 90%, 95%, 99%):** Percentiles calculated over the rolling windows.

- **The rolling window sizes (100, 1000, 10000)** indicate how many data points are considered for each rolling calculation, providing a moving statistic across the dataset.

In current research, we also tested another set of features alongside those previously extracted. This additional set includes several common statistical metrics and rolling window calculations, with some overlap and some variations compared to the initial set, Figure 14 shows the shape and the features of the second feature set used.

```
print(Data.shape)
Data.columns

(4194, 37)
Index(['meen', 'std', 'min', 'max', 'kurt', 'skew', 'abs_max', 'abs_mean',
      'abs_std', 'q_01', 'q_05', 'q_95', 'q_99', 'avg_roll_10', 'std_roll_10',
      'max_roll_10', 'min_roll_10', 'q_01_roll_10', 'q_05_roll_10',
      'q_95_roll_10', 'q_99_roll_10', 'avg_roll_100', 'std_roll_100',
      'max_roll_100', 'min_roll_100', 'q_01_roll_100', 'q_05_roll_100',
      'q_95_roll_100', 'q_99_roll_100', 'avg_roll_1000', 'std_roll_1000',
      'max_roll_1000', 'min_roll_1000', 'q_01_roll_1000', 'q_05_roll_1000',
      'q_95_roll_1000', 'q_99_roll_1000'],
      dtype='object')
```

Figure 14: LANL Feature set (37)

Common Features:

- Mean ('meen' typo corrected to 'mean')
- Standard Deviation ('std')
- Minimum ('min') and Maximum ('max')
- Kurtosis ('kurt')
- Skewness ('skew')
- Absolute Mean ('abs_mean') and Absolute Standard Deviation ('abs_std')
- Absolute Maximum ('abs_max')

Additional Quantile Features:

- **Quantiles ('q_01', 'q_05', 'q_95', 'q_99'):** Represent the 1st, 5th, 95th, and 99th percentiles of the data, indicating the values below which a certain percentage of observations fall.

Rolling Window Features:

- **Average, Standard Deviation, Maximum, Minimum ('avg_roll', 'std_roll', 'max_roll', 'min_roll'):** These statistics are calculated over rolling windows of different sizes (10, 100, 1000).
- **Quantiles for Rolling Windows ('q_01_roll', 'q_05_roll', 'q_95_roll', 'q_99_roll'):** Percentiles calculated over the same rolling windows.

While both feature sets incorporate basic statistical measures and rolling window features, the new set explicitly includes quantile values and absolute maximum which were not listed in the previous description. Additionally, this set specifies different sizes for rolling windows (10, 100, 1000), providing more granularity and different time scales compared to the previous feature set. The introduction of smaller window sizes like 10 and explicit mention of absolute measures like 'abs_max' can provide more detailed insights into the data's behavior at different scales, which might be crucial for improving model accuracy in predicting seismic events.

The researcher employed these two sets of feature engineering, aiming to identify the optimal model for testing the LANL dataset for earthquake prediction, to enhance the accuracy of predictions.

4.2. Results

This section provides an overview of the implemented models performance of the two datasets USGS and LANL.

4.2.1. USGS Dataset Models Performance

Table 5 illustrates the comparative results for earthquake magnitude prediction using the USGS dataset, with the analysis involving traditional ML algorithms: LR, SVR, RFR, and XGBoost Regressor. The performance of these models is measured using three key metrics: MAE, MSE, and R-squared.

Table 5: Performance Metrics for Earthquake Magnitude Prediction Models on USGS Dataset

Model	MAE	MSE	R-squared
LR	0.203	0.078	0.592
SVR	0.170	0.059	0.686
RFR	0.167	0.055	0.729
XGBoost Regressor	0.188	0.069	0.637

Figure 15 shows the MAE values across the models indicates that RFR had the lowest error at 0.167, suggesting its predictions are closest to the actual magnitudes on average. SVR also performed well with an MAE of 0.170, followed by XGBoost Regressor and LR with 0.188 and 0.203, respectively.

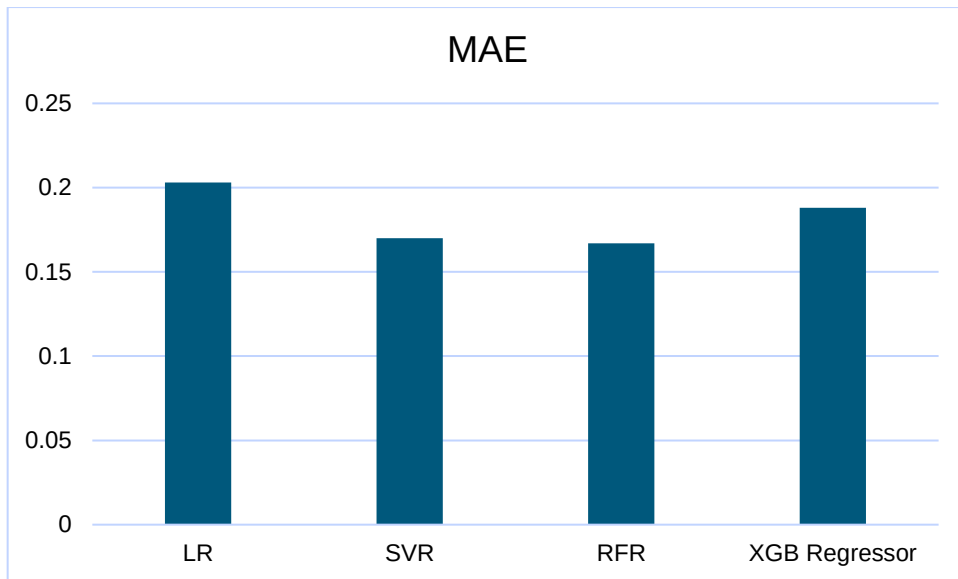


Figure 15: MAE for LR, SVR, RFR and XGBOOST Regressor

Figure 16 depicts the MSE, which penalizes larger errors more severely than smaller ones, RFR once again had the smallest value at 0.055, indicating its predictions were generally more accurate. SVR followed with an MSE of 0.059, XGBoost Regressor had the highest MSE at 0.069, and LR had a notably higher MSE of 0.079.

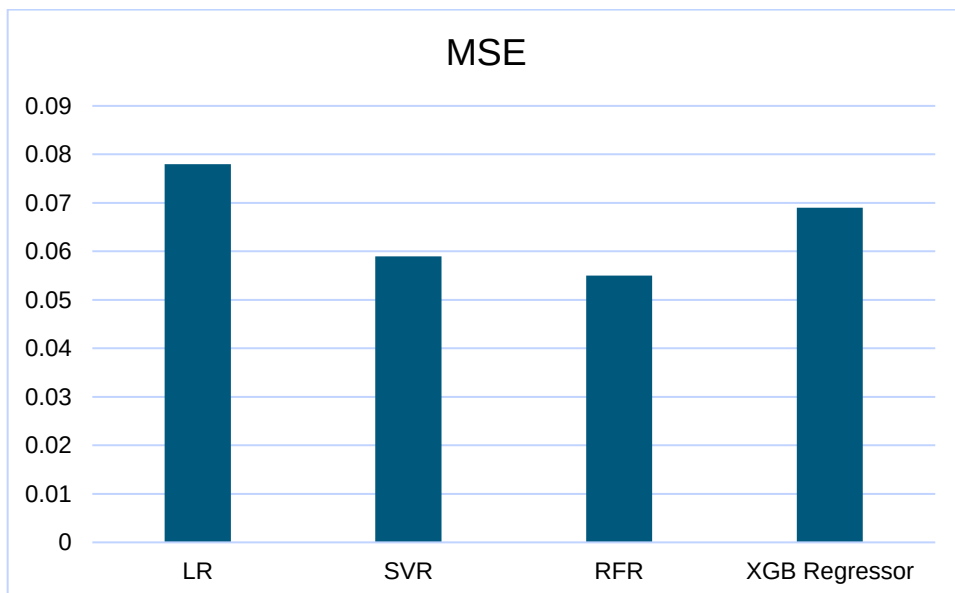


Figure 16: MSE for LR, SVR, RFR and XGBoost Regressor

The R-squared metric, which indicates how much of the variance in the magnitude variable is accounted for by the models, reveals that the RFR is the most effective with an R-squared of 0.709, demonstrating considerable predictive strength. The SVR follows closely with an R-squared of 0.686, signifying strong predictive abilities as well. Meanwhile, the XGBoost Regressor and LR are somewhat less effective in prediction, with R-squared values of 0.637 and 0.592, respectively. These R-squared statistics are depicted in Figure 17.

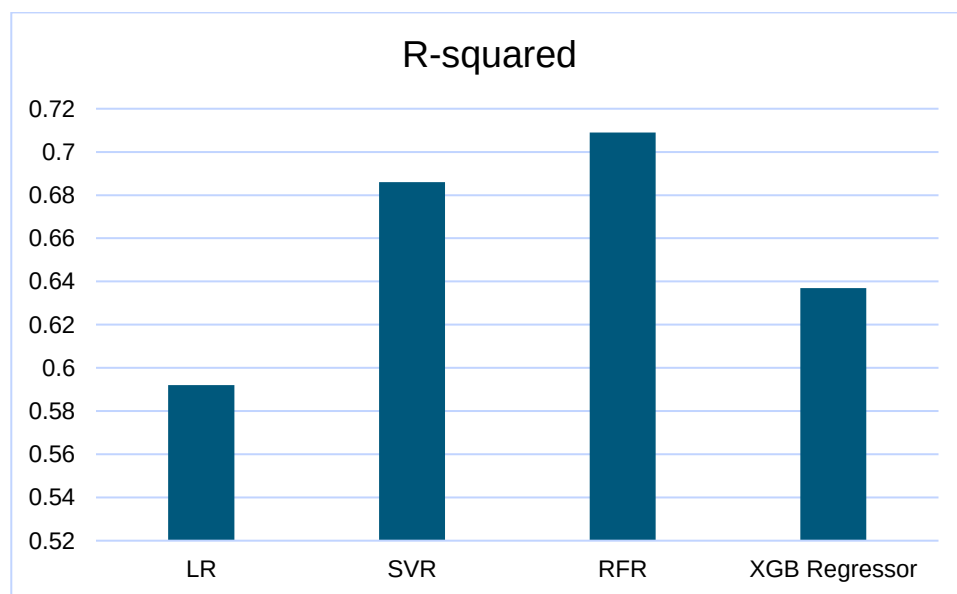


Figure 17:R-squared for LR, SVR, RFR and XGBoost Regressor

Figure 18 shows a strong alignment of predicted values with actual values at lower magnitudes, reflecting the model's accuracy in this range. However, as magnitudes increase, the data points diverge more from the ideal prediction line. One reason for this could be the lower occurrence of higher magnitude earthquakes in the dataset, which results in fewer data points for the model to learn from. This scarcity of high-magnitude examples can lead to less reliable predictions for larger events, as the model has not been as well-trained for these scenarios. Overall, The RFR appears

to be the most promising model for this dataset based on the traditional ML algorithms.

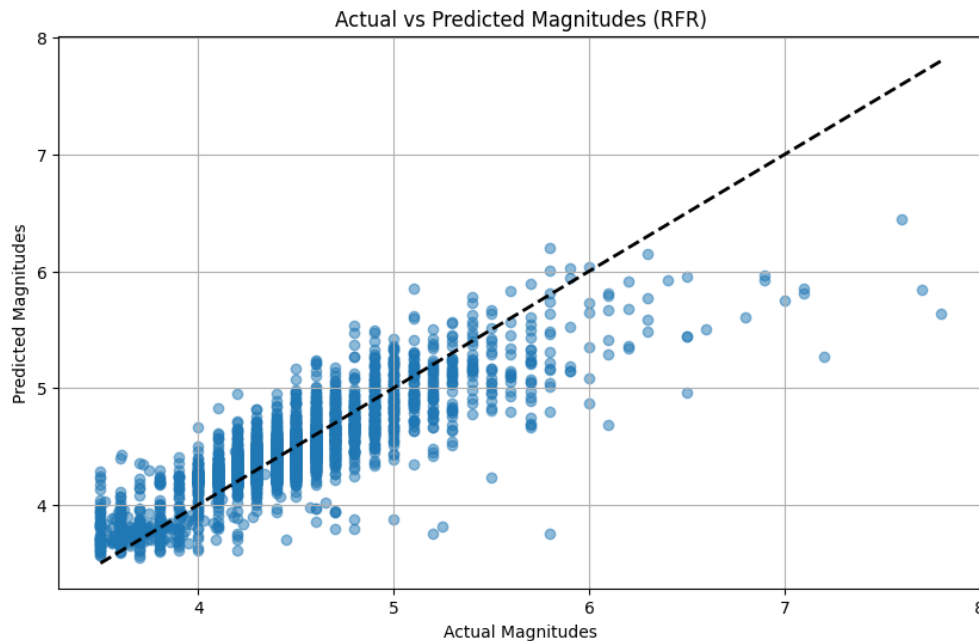


Figure 18: RFR Actual vs Predicted Magnitudes

A comprehensive comparison of DL models applied to the USGS dataset, encompassing a variety of architectures including multiple layers of LSTM, and GRU. The effectiveness of these models is quantitatively analysed using MAE, MSE, and R-squared. Like in the previous models we will try to predict the magnitude of an earthquake using regression analysis. Min-max normalization is used on the dataset. The dataset is split into training and testing sets containing 70% and 30% of the data respectively using sequential split.

1) LSTM

The keras library is used to implement the LSTM model in this research. For an LSTM network we have to set the timestep value, which is the number of previous inputs the model will take into account when predicting the next value. The model is trained with timesteps values 7, 10, 15 and 30 respectively. Two networks used in this model, one with 2 layers

with 100, 50 neurons respectively. And the other network with 3 layers with 100, 50, 50 neurons respectively. And with a 'relu' activation function, and an output layer that predicts values. The 'Adam' activation function is used in the output layer as updating the weights of the network, with a learning rate of 0.001.

Dropout is a regularization technique for neural network models where randomly selected neurons are ignored during training, this makes the model less sensitive to neuron weights and prevents overfitting. Dropout with a 20% is implemented in this model using the built-in dropout function in keras.

Table 6 depicts LSTM model's performance. This performance varied across different timesteps, showing a range of MSE from 0.197 to 0.219, indicating some inconsistency in its predictions. The MAE was also stable at around 0.33. However, the R-squared values were negative, implying that the model may not have captured the predictive structure of the data effectively. A negative value implies that a straight line fits the data better than the tested model (Mondol, 2021).

The second network, MSE results ranged narrowly from 0.189 to 0.207, with the lowest errors observed at shorter timesteps. MAE remained relatively consistent, reflecting a stable average error across the different timesteps tested. R-squared values were mostly negative, similar to the more complex LSTM model, signaling potential overfitting or the model's unsuitability for the dataset's pattern.

Table 6: LSTM Model's Performance

Model Structure	Timestep	MAE	MSE	R-squared
LSTM within 2 layers 100 and 50 neurons	7	0.320	0.197	-0.028
	10	0.332	0.219	-0.129
	15	0.328	0.208	-0.093
	30	0.326	0.208	-0.067
LSTM within 3 layers 100, 50 and 50 neurons	7	0.313	0.192	-0.005
	10	0.379	0.264	-0.387
	15	0.309	0.189	0.022
	30	0.307	0.191	0.017
	48	0.317	0.207	0.014

2) GRU Model

The GRU model architecture is built using Keras, is optimized after several experimental iterations to establish the most effective structure. It starts with an input layer designed to receive sequences of data of a certain length (timestep). The optimized model includes two GRU layers: the first GRU layer has 50 units and returns sequences, making it suitable for chaining with another recurrent layer. The second GRU layer, also with 50 units, does not return sequences, preparing the data for a denser representation. After the GRU layers, there's a dense layer with a single unit, serving as the output layer. This architecture, identified as the best performing setup through comparative testing, is compiled with the mean squared error loss function and the Adam optimizer with learning rate of 0.001. For simplicity, we will focus our discussion on the results of this architecture.

The optimized GRU model's architecture was evaluated at several timesteps values 7, 10, 15, 30 and 48 respectively. Table 7 shows GRU model's performance, demonstrated slightly better consistency in its predictions with MSE values ranging from 0.189 to 0.210. The MAE was consistent, like the LSTM models. R-squared values were low but positive for shorter timesteps, indicating a better fit than the LSTM models but still leaving room for improvement.

Table 7: GRU Model's Performance

Timestep	MAE	MSE	R-squared
7	0.190	0.068	0.308
10	0.191	0.068	0.308
15	0.189	0.069	0.311
30	0.195	0.069	0.312
48	0.210	0.072	0.324

The detailed analysis of DL models on the USGS dataset demonstrates a nuanced view of each model's predictive capabilities. It appears that simpler models, such as the standalone GRU, may be better suited for this dataset, as reflected by the more favourable R-squared values. However, all models exhibited issues with capturing the predictive structure, as indicated by the negative R-squared values in many instances. This suggests that further refinement of the models' architectures, training procedures, or the dataset's feature engineering may be required to enhance predictive performance.

4.2.2. LANL Dataset Models Performance

Table 8 presents the results from an experimental evaluation of several ML models on LANL dataset. The models included LR, RFR, XGBoost Regressor, and SVR, with performance assessed using MAE, MSE and R-squared. The experiments conducted using the two sets of the feature engineering to evaluate the performance of the models.

Table 8: Models performance (LANL)

Models	MAE		MSE		R-squared	
	Feature set (58)	Feature set (37)	Feature set (58)	Feature set (37)	Feature set (58)	Feature set (37)
LR	2.258	2.266	8.094	8.364	0.415	0.396
RFR	2.283	2.295	8.171	8.258	0.410	0.403
XGBoost Regressor	2.386	2.462	8.803	9.330	0.364	0.326
SVR	2.221	2.172	7.851	7.921	0.433	0.428

1) MAE

In evaluating the models using the MAE metric across the two different feature sets, it's evident that each model's performance slightly varies depending on the number of features. For LR model, the MAE is slightly lower with the first feature set (58 features) at 2.258 compared to 2.266 with the second set (37 features). Similarly, the RFR shows a marginal increase in MAE when using fewer features, from 2.283 to 2.295. The XGBoost Regressor demonstrates a more noticeable difference, with MAE increasing from 2.386 to 2.462 as the feature set is reduced, indicating a potential dependency on a larger number of features for optimal performance. Conversely, the SVR model performs better with the smaller feature set, showing a decrease in MAE from 2.221 to 2.172, suggesting it might be more efficient at handling a more concise set of features. Overall, these results highlight the varying sensitivities of different models to the feature set size, emphasizing the importance of feature selection in model performance optimization for earthquake prediction (Figure 19).

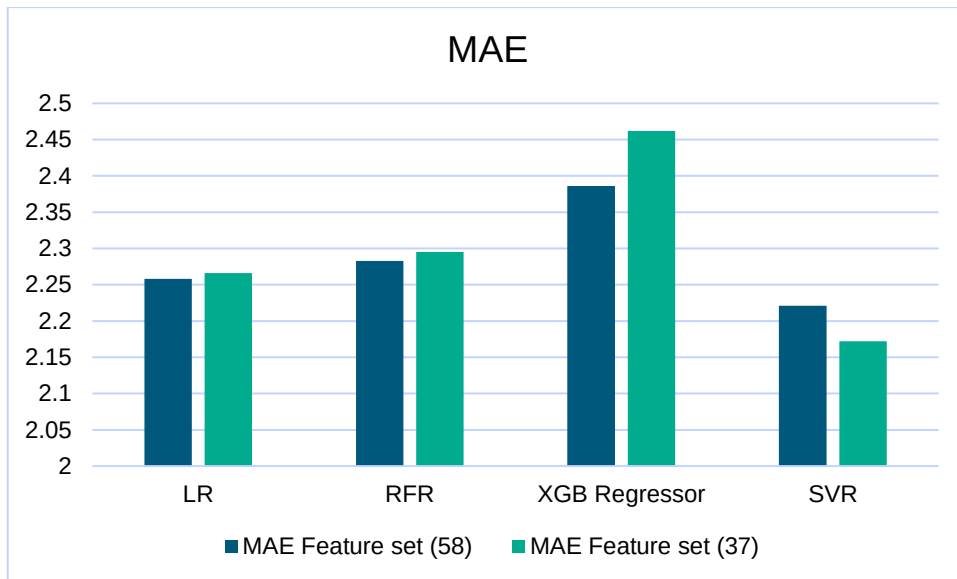


Figure 19: MAE for LR, RFR, XGBoost and SVR (LANL)

2) MSE

Analyzing the performance of the models on the two feature sets using the MSE metric reveals too how model accuracy is influenced by the number of features. Figure 20 shows MSE metric performance among the models, for LR, the MSE increases from 8.094 with the first feature set (58 features) to 8.364 with the second set (37 features), suggesting a slight decline in performance with fewer features. Similarly, RFR experiences a small rise in MSE from 8.171 to 8.258, indicating a minimal impact from reducing the feature count. The XGBoost Regressor shows a more significant increase in MSE, from 8.803 to 9.330, which could imply that XGBoost relies more heavily on a larger set of features for optimal performance. In contrast, the SVR maintains a relatively stable MSE, only slightly increasing from 7.851 to 7.921, suggesting that SVR is less sensitive to changes in the number of features in this dataset. This comparison highlights that while some models like SVR can maintain robustness across different feature sets, others like XGBoost might require more careful feature selection to minimize error for earthquake prediction.

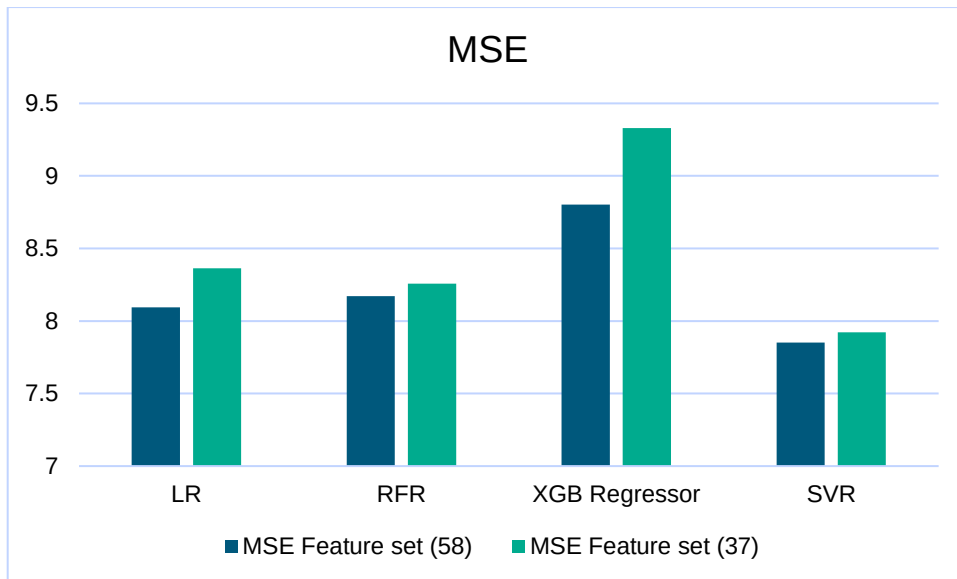


Figure 20: MSE For LR, RFR, XGBoost and SVR (LANL)

3) R-squared

When examining the performance of the models using the R-squared metric across the two different feature sets, we observe distinct impacts on model efficacy related to the number of features utilized. For LR, the R-squared value decreases slightly from 0.415 with the first feature set (58 features) to 0.396 with the second set (37 features), indicating a minor reduction in the model's ability to explain the variability of the response data with fewer features. The RFR also shows a slight decrease in R-squared from 0.410 to 0.403, suggesting a small but noticeable impact on performance due to the reduction in feature count.

In contrast, the XGBoost Regressor exhibits a more significant drop in R-squared from 0.364 to 0.326, which may imply that the reduction of features has a more pronounced effect on its performance, potentially due to losing important variables that help capture more complexity in the data. On the other hand, the SVR maintains a relatively stable R-squared, decreasing slightly from 0.433 to 0.428, indicating that SVR's performance is robust against changes in the number of features (Figure 21). This comparison highlights the varying dependencies of different models on feature sets, with some like XGBoost being more sensitive to feature

reduction, while others like SVR demonstrate greater stability in their explanatory power regardless of feature count.

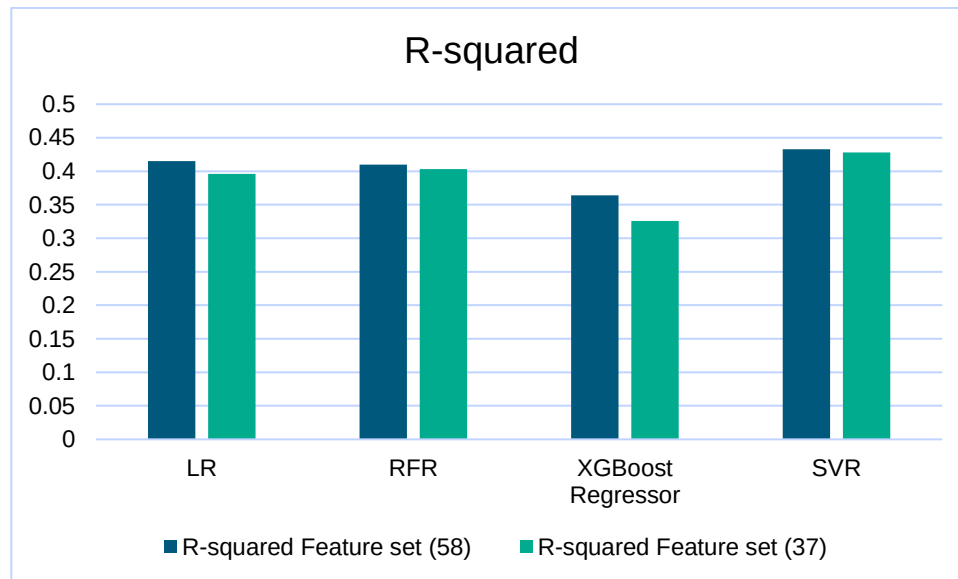


Figure 21: R-squared for LR, RFR, XGBoost and SVR (LANL)

The comparison across models on the LANL dataset reveals that the SVR demonstrates the best performance for the LANL dataset. It shows the lowest error rates and the highest or near-highest explanatory power across the different feature sets, making it the most reliable model among those tested for predicting earthquake parameters in this context.

1) LSTM

Table 9 details five LSTM-based model architectures, each designed for processing sequential data, potentially in tasks like time series prediction. All models use the Keras Sequential API and consist of multiple LSTM layers interspersed with Dropout layers to prevent overfitting.

Model 1 and Model 2 both employ simpler configurations, with Model 1 incorporating two LSTM layers of 50 and 32 units, and a single Dropout layer with a rate of 0.3. Model 2 starts with a larger first LSTM layer of 100 units followed by a second layer of 32 units, also followed by

a single Dropout layer with the same dropout rate. Both models terminate with a dense output layer using ReLU activation for regression outputs.

Model 3 increases complexity by using three LSTM layers with units of 50, 32, and 16, each outputting sequences to the next, accompanied by Dropout layers at rates of 0.2 and 0.3 placed strategically to enhance regularization.

Model 4 is the most complex, with four LSTM layers starting at a substantial 150 units and tapering down to 80, 32, and 16 units. It features three Dropout layers with rates of 0.2 and 0.3. This model, like the others, finishes with a single-unit dense layer with ReLU activation, optimized for regression tasks.

Model 5 mirrors the structure of Model 4 but standardizes the LSTM units more evenly, using two initial layers of 50 units each, followed by layers of 32 and 16 units, also outputting sequences. It incorporates three Dropout layers at a consistent rate of 0.2.

Table 9: LSTM Model Architectures

No	Model architecture
1	<pre> model = Sequential() model.add(LSTM(50, return_sequences=True,input_shape=(X_train.shape[1],X_train.shape[2]))) model.add(LSTM(32)) model.add(Dropout(0.3)) model.add(Dense(1 , activation='relu')) </pre>
2	<pre> model = Sequential() model.add(LSTM(100, return_sequences=True,input_shape=(X_train.shape[1],X_train.shape[2]))) model.add(LSTM(32)) model.add(Dropout(0.3)) model.add(Dense(1 , activation='relu')) </pre>
3	<pre> model = Sequential() model.add(LSTM(50, return_sequences=True,input_shape=(X_train.shape[1],X_train.shape[2]))) model.add(Dropout(0.2)) model.add(LSTM(32, return_sequences=True)) model.add(Dropout(0.3)) model.add(LSTM(16)) model.add(Dense(1 , activation='relu')) </pre>
4	<pre> model = Sequential() model.add(LSTM(150, return_sequences=True,input_shape=(X_train.shape[1],X_train.shape[2]))) </pre>

	<pre> model.add(Dropout(0.2)) model.add(LSTM(80, return_sequences=True)) model.add(Dropout(0.3)) model.add(LSTM(32, return_sequences=True)) model.add(Dropout(0.2)) model.add(LSTM(16)) model.add(Dense(1 , activation='relu')) </pre>
5	<pre> model = Sequential() model.add(LSTM(50, return_sequences=True,input_shape=(X_train.shape[1],X_train.shape[2]))) model.add(Dropout(0.2)) model.add(LSTM(50, return_sequences=True)) model.add(Dropout(0.2)) model.add(LSTM(32, return_sequences=True)) model.add(Dropout(0.2)) model.add(LSTM(16, return_sequences=True)) model.add(Dense(1 , activation='relu')) </pre>

Several experiments have assessed the performance of the different model architectures. Table 10 showcase the best evaluation results for each model architecture across different setups with 37 and 58 features. Each model was evaluated with a varying window size, number of epochs, and batch size as they showed better results.

Table 10: LSTM Performance for LANL dataset

LSTM Configuration	Window Size	Epochs	Batch Size	MAE		MSE		R-squared	
				Features (37)	Features (58)	Features (37)	Features (58)	Features (37)	Features (58)
1	7	100	32	2.391	2.299	9.033	9.341	0.349	0.258
2	7	100	32	2.542	2.299	11.004	9.007	0.207	0.284
3	7	32	32	2.115	1.938	6.346	5.759	0.539	0.542
4	20	32	32	2.140	1.907	6.349	5.351	0.544	0.573
5	7	32	32	1.794	1.789	5.190	5.000	0.602	0.603

1) MAE

Model 1 and Model 2, both configured with a window size of 7, 100 epochs, and a batch size of 32, showed similar trends. Model 1 produced a

MAE of 2.391 with 37 features and improved to 2.299 with 58 features, indicating better accuracy with more data inputs. Model 2 had a higher initial MAE of 2.542 with 37 features, but like Model 1, it improved to 2.299 with 58 features, highlighting consistent benefits from increased feature counts in this setup.

Model 3, with shorter training at 32 epochs, displayed superior efficiency, achieving a MAE of 2.115 with 37 features and further improving to 1.938 with 58 features, the best improvement among all models when using more features. This suggests that Model 3 is particularly robust, benefiting significantly from additional data without extensive training.

Model 4 extended the temporal context with a window size of 20 and maintained the same epoch and batch size as Model 3. It achieved a MAE of 2.140 with 37 features and an even lower MAE of 1.907 with 58 features, the lowest recorded, implying that a larger window size effectively captures more relevant temporal patterns that enhance prediction accuracy with more features.

Lastly, Model 5 demonstrated remarkable effectiveness and versatility, producing the lowest MAE of 1.794 with 37 features and slightly improving to 1.789 with 58 features. This model's performance indicates a near-optimal use of its configuration, showing minimal gains from additional features, which suggests that it's already operating at high efficiency with the provided data.

Overall, these results indicate that increasing the number of features generally enhances model accuracy. However, the impact varies by specific model configurations, with some benefiting more from additional data or altered training durations and window sizes. Model 5 stands out for

its exceptional efficiency and robustness across different feature setups (Figure 22).

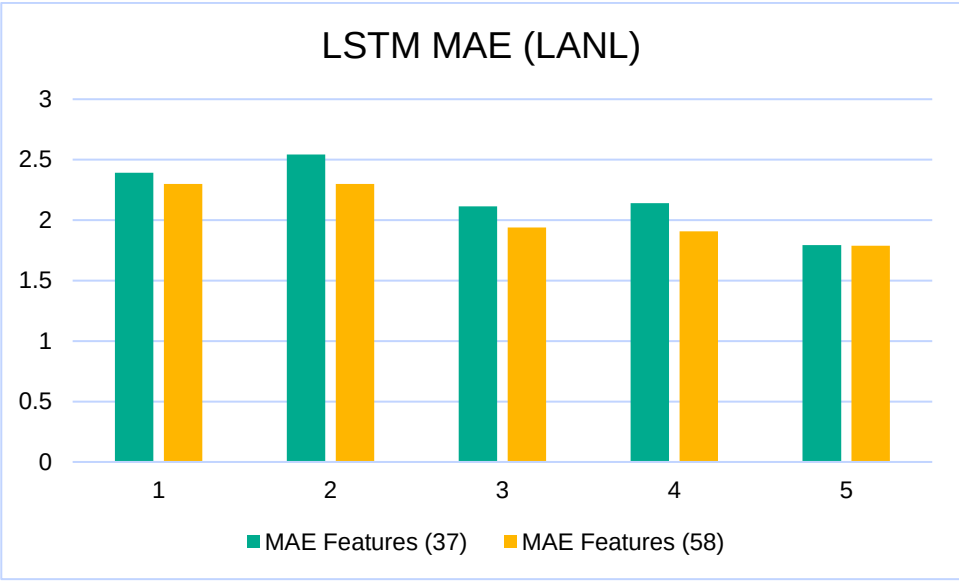


Figure 22: LSTM MAE Metric (LANL)

2) MSE

Figure 23 provides MSE values for the five LSTM models architecture under conditions of 37 and 58 features, revealing how each model adapts to varying input complexities. Model 1, with a window size of 7, 100 epochs, and a batch size of 32, showed an MSE of 9.033 with 37 features, which slightly worsened to 9.341 with 58 features, indicating a decline in performance with more complex data. In contrast, Model 2 improved significantly when handling more features, decreasing from an MSE of 11.004 with 37 features to 9.007 with 58 features, suggesting it processes larger datasets more efficiently. Model 3 also showed enhanced performance with an increase in features, with MSE dropping from 6.346 to 5.759, demonstrating its ability to leverage additional data for better accuracy. Model 4, using a larger window size of 20, reduced its MSE from 6.349 to 5.351 when the number of features increased, which points to its effective use of a broader temporal context to improve prediction accuracy. Finally, Model 5, which had the lowest MSE of 5.190 with 37 features, further improved to 5.000 with 58 features, underscoring its robustness and

optimal performance regardless of feature set size. Overall, these results illustrate varying degrees of adaptability across models, with most showing improved accuracy as they processed more complex datasets.

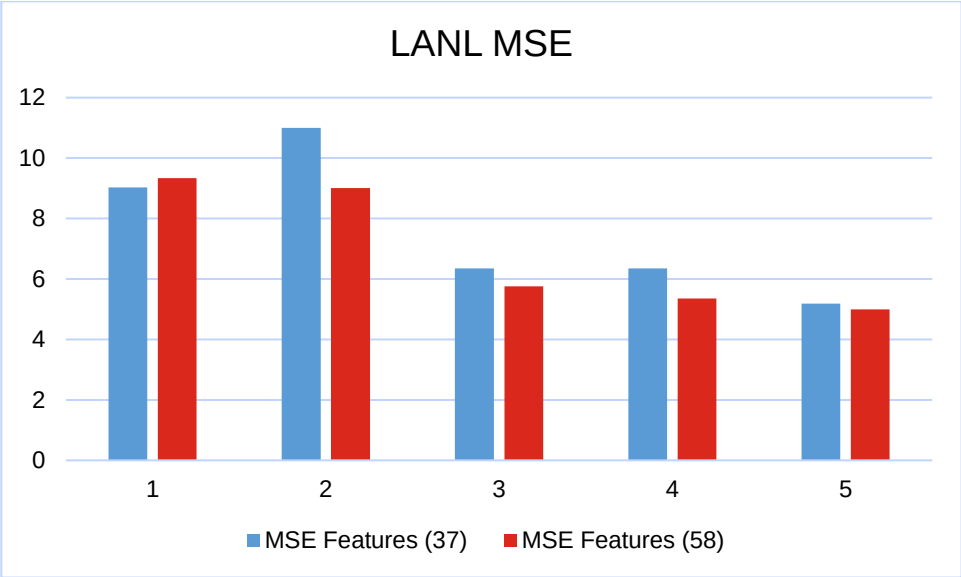


Figure 23: LSTM MSE Metric (LANL)

3) R-squared

Figure 24 provides insights into the R-squared values for the five LSTM models architecture, reflecting their performance in terms of explained variance when trained with 37 and 58 features. Model 1, operating with a window size of 7, 100 epochs, and a batch size of 32, showed a decrease in R-squared from 0.349 with 37 features to 0.258 with 58 features, indicating a reduction in model performance with increased feature complexity. Conversely, Model 2 experienced a slight improvement in R-squared from 0.207 to 0.284 when the feature count was increased, suggesting it better handles more complex datasets. Model 3 maintained stable performance with an R-squared of 0.539 increasing marginally to 0.542, demonstrating its robustness across different feature sets. Model 4, which utilized a larger window size of 20, showed an improvement from an R-squared of 0.544 with 37 features to 0.573 with 58 features, reflecting its effective use of the broader temporal context to

capture more variance with more features. Lastly, Model 5, which already had the highest R-squared of 0.602 with 37 features, marginally improved to 0.603 with 58 features, indicating its superior and consistent predictive accuracy regardless of the feature set size. Overall, these results highlight how different LSTM configurations respond to an increase in the number of features, with most models showing stable or improved performance, particularly when handling more complex datasets.

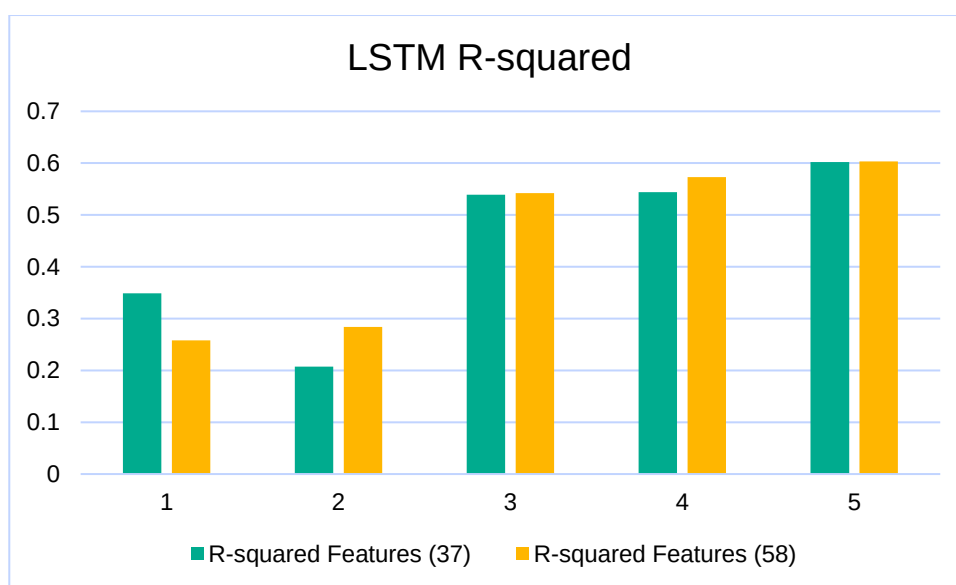


Figure 24: LSTM R-squared Metric (LANL)

Among the five evaluated LSTM model architectures, Model 5 consistently excels across various evaluation metrics MAE, MSE, and R-squared values, indicating its superior performance. It achieved the lowest MAE and MSE scores in both the 37 and 58 feature sets, highlighting its precision and effectiveness in minimizing prediction errors. Additionally, Model 5 demonstrated the highest R-squared values, which reflects its exceptional ability to capture and explain variance within the dataset. This consistent top performance across different metrics and feature sets underscores Model 5's robustness and reliability, making it an optimal choice for tasks requiring precise and accurate sequence predictions.

Figure 25 depicts the performance of Model 5 architecture in predicting the time to failure for earthquakes. The model seems to capture the overall downward trend in the time to failure quite well, suggesting it has learned the general pattern of the sequence. However, it struggles with the exact timing and magnitude of the sharp drops, which are the critical earthquake events. These discrepancies are particularly noticeable at the points where the blue line has steep vertical drops, the model's predictions either precede or follow the actual event, and the predicted values don't match the steepness of the actual failure events.

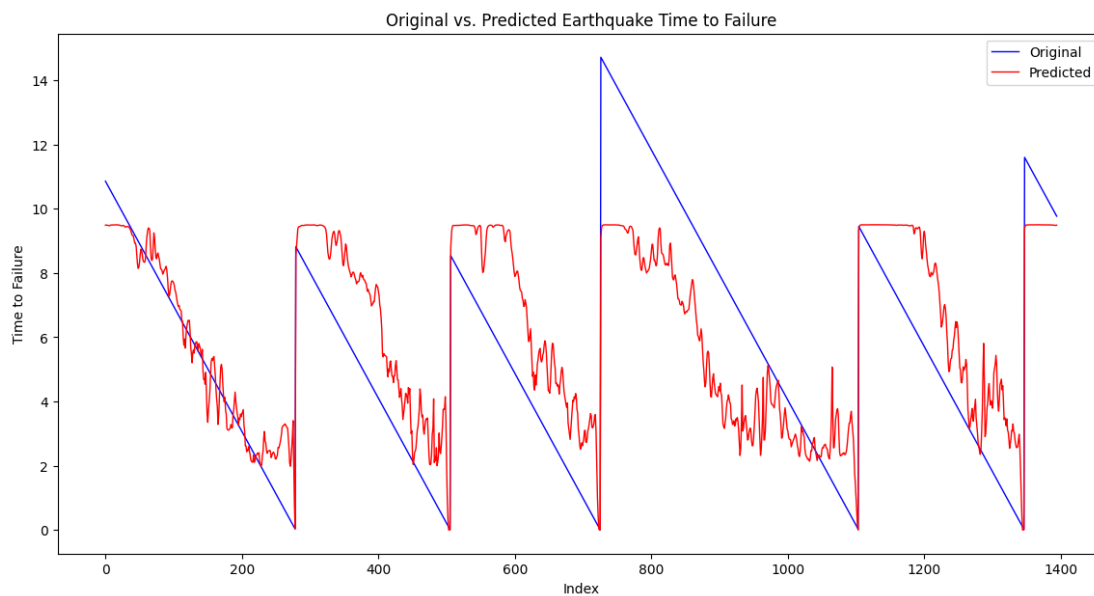


Figure 25: LSTM Model 5 Architecture Prediction Performance (LANL)

2) GRU

Table 11 outlines four distinct GRU neural network architectures. Each model utilizes an input layer shaped to the training data dimensions, followed by GRU layers interspersed with dropout layers for regularization. The first architecture features a 50-unit GRU layer that passes sequences to a subsequent 32-unit GRU layer, culminating in a single-unit dense output layer. The second architecture expands on this by

introducing an additional 50-unit GRU layer, each followed by a dropout layer with a 0.2 dropout rate, before proceeding to the 32-unit GRU layer. The third model increases the initial GRU layer to 100 units followed by a 48-unit layer, both returning sequences and interspaced with 0.2 dropout layers, before transitioning to a 32-unit GRU layer. This allows for a deeper and potentially more nuanced understanding of the data. Lastly, the fourth model incorporates a sequential configuration of GRU layers with 50, 32, and finally 16 units, each followed by a dropout layer set at a 0.2 rate, indicating a design choice that might strike a balance between learning capacity and complexity.

All models are compiled with mean squared error as the loss function and optimized with Adam, suggesting that they are calibrated to minimize prediction errors effectively.

Table 11: GRU Architecture Used for LANL dataset

No	Model Architecture
1	<pre> input_layer = Input(shape=(X_train.shape[1], X_train.shape[2])) gru_layer = GRU(50, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) gru_layer = GRU(32)(dropout_layer) output_layer = Dense(1)(gru_layer) model = Model(inputs=input_layer, outputs=output_layer) model.compile(loss='mean_squared_error', optimizer='adam') </pre>
2	<pre> input_layer = Input(shape=(X_train.shape[1], X_train.shape[2])) gru_layer = GRU(50, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) gru_layer = GRU(50, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) gru_layer = GRU(32)(dropout_layer) output_layer = Dense(1)(gru_layer) model = Model(inputs=input_layer, outputs=output_layer) model.compile(loss='mean_squared_error', optimizer='adam') </pre>
3	<pre> input_layer = Input(shape=(X_train.shape[1], X_train.shape[2])) gru_layer = GRU(100, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) gru_layer = GRU(48, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) </pre>

	<pre> gru_layer = GRU(32)(dropout_layer) output_layer = Dense(1)(gru_layer) model = Model(inputs=input_layer, outputs=output_layer) model.compile(loss='mean_squared_error', optimizer='adam') </pre>
4	<pre> input_layer = Input(shape=(X_train.shape[1], X_train.shape[2])) gru_layer = GRU(50, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) gru_layer = GRU(32, return_sequences=True)(input_layer) dropout_layer = Dropout(0.2)(gru_layer) gru_layer = GRU(16)(dropout_layer) output_layer = Dense(1)(gru_layer) model = Model(inputs=input_layer, outputs=output_layer) model.compile(loss='mean_squared_error', optimizer='adam') </pre>

Several experiments assessed the performance of the four GRU model architectures. Table 12 shows the best evaluation results for each model architecture with setups using 37 and 58 features. Each model was evaluated with varying window sizes, numbers of epochs, and batch sizes, as these configurations yielded better results.

Table 12:GRU Performance (LANL)

Configuration	Window size	Epochs	Batch Size	MAE		MSE		R-squared	
				Features (37)	Features (58)	Features (37)	Features (58)	Features (37)	Features (58)
1	7	32	32	1.907	1.990	5.179	6.075	0.588	0.517
2	10	32	16	1.931	2.179	6.440	7.544	0.511	0.400
3	10	32	16	1.935	1.993	6.603	6.187	0.475	0.507
4	7	32	32	1.777	1.870	4.192	5.877	0.610	0.531

1) MAE

In an assessment of GRU model architectures using the LANL dataset for earthquake prediction, Model 4 outperformed the others, achieving the lowest MAE across both feature sets, with scores of 1.777 for 37 features and 1.87 for 58 features. This suggests an optimal synergy between its

window size of 7 and the given epochs and batch size for capturing seismic patterns. Model 1 followed, showing strong accuracy, particularly with 37 features, where it scored an MAE of 1.907, slightly increasing to 1.990 with the larger feature set. Models 2 and 3, with a window size of 10 and trained for 32 epochs with a batch size of 16, demonstrated comparable performance, but with marginally higher errors, indicating a slightly less effective configuration for the task. Model 2 had MAEs of 1.931 and 2.179, while Model 3 scored 1.935 and 1.993 for the 37 and 58 feature sets respectively. The performance gradient across these models underscores the nuanced impact of model configuration on predictive accuracy in earthquake timing, with Model 4 leading the way in effectiveness. Figure 26 shows these outcomes.

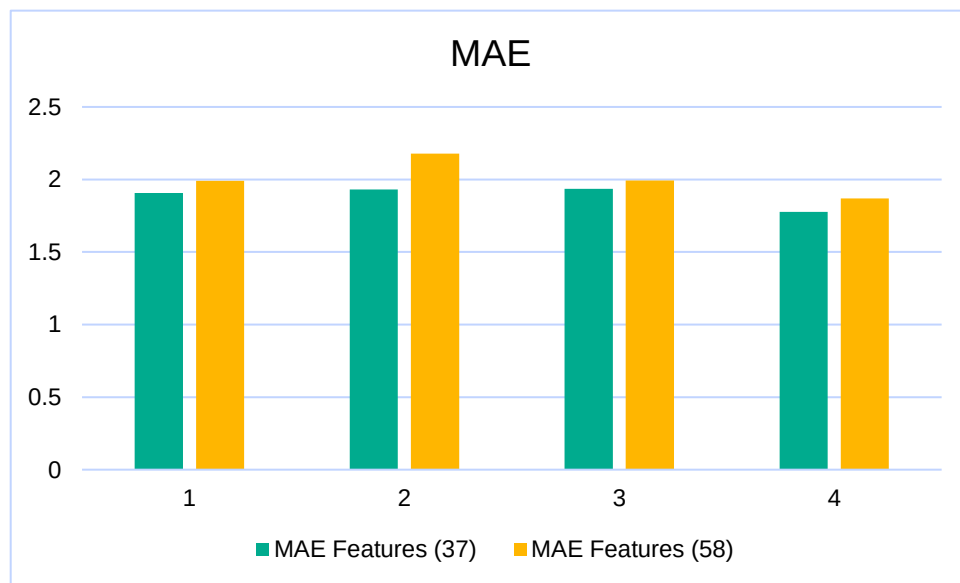


Figure 26: GRU MAE Metric (LANL)

2) MSE

Figure 27 provides insights into the performance of the four GRU models architecture. Model 4 stands out with the most favorable MSE scores, registering 4.192 for 37 features and 5.877 for 58 features, indicating its superior predictive capabilities likely due to an optimal

window size and training setup. Model 1 also shows promise, with its MSE increasing from 5.179 with 37 features to 6.075 with 58 features, suggesting a moderate impact from the increase in feature set size. Meanwhile, Models 2 and 3, with larger window sizes of 10, had higher MSE values (6.44 and 7.544 for Model 2, 6.603 and 6.187 for Model 3), reflecting less precision in their predictions with both feature sets.

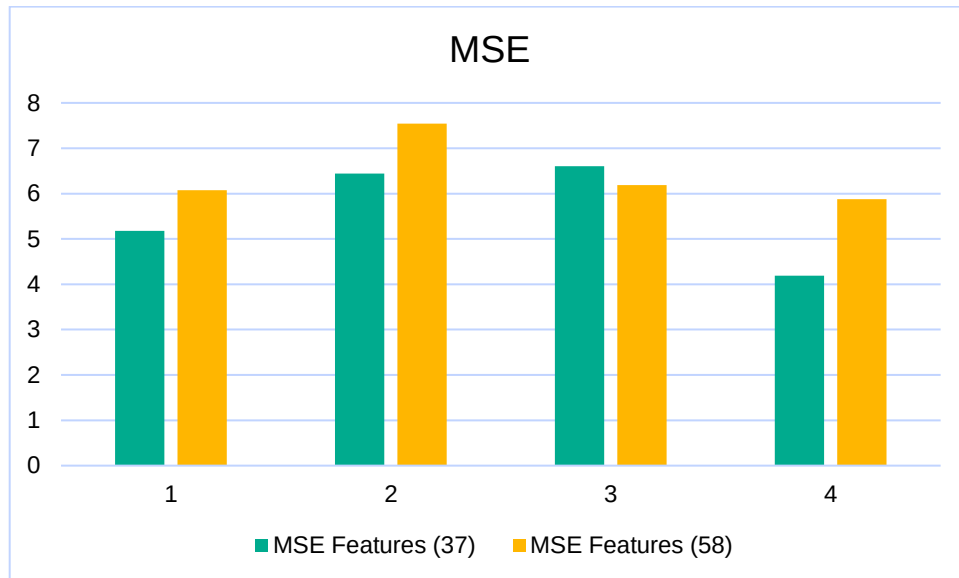


Figure 27: MSE metrics For GRU (LANL)

3) R-squared

The Figure 28 outlines the R-squared performance metrics for the four GRU models architecture. Model 4 showcases the highest effectiveness, registering R-squared values of 0.61 for 37 features and 0.531 for 58 features, indicating its superior ability to explain the variance in the seismic activity data. Model 1 also performs robustly, with R-squared scores of 0.588 for 37 features and 0.517 for 58 features, reflecting its competence in modelling the data patterns, albeit slightly less effectively than Model 4. Conversely, Models 2 and 3, with a window size of 10, demonstrate lower adaptability; Model 2's performance drops from 0.511 with 37 features to 0.4 with 58 features, while Model 3 shows R-squared values of 0.475 and 0.507 for 37 and 58 features respectively, indicating a marginal

improvement with increased feature complexity. These findings highlight the varying abilities of each model to capture and predict the complex temporal patterns of earthquake data, with Model 4 consistently outperforming the others.

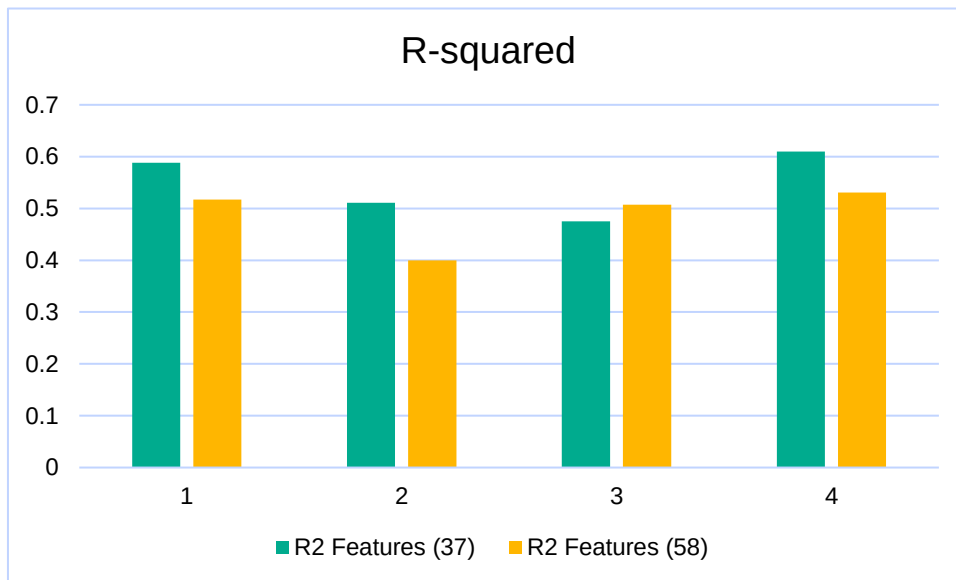


Figure 28: GRU R-squared Metric (LANL)

Based on the assessments from the three key metrics, MAE, MSE, and R-squared values, Model 4 consistently emerges as the top performer across all these key metrics for earthquake timing prediction using the GRU architecture. This model demonstrated the lowest MAE and MSE scores, indicating a high precision in forecasting and effectiveness in minimizing prediction errors. Additionally, it achieved the highest R-squared values, which signifies its superior ability to explain the variance observed in the dataset. Figure 29 illustrates the performance of Model 4 in predicting when earthquakes will occur. As we can see that the prediction often closely follows the pattern of the original, suggesting the model generally does well in estimating when the earthquakes might happen. However, it's not perfect; the prediction sometimes fails to align precisely with the sharp

drops in the original, which are the critical moments when earthquakes actually took place. Despite these discrepancies, the model appears to capture the overall trend quite effectively, indicating its usefulness in estimating the countdown to seismic events.

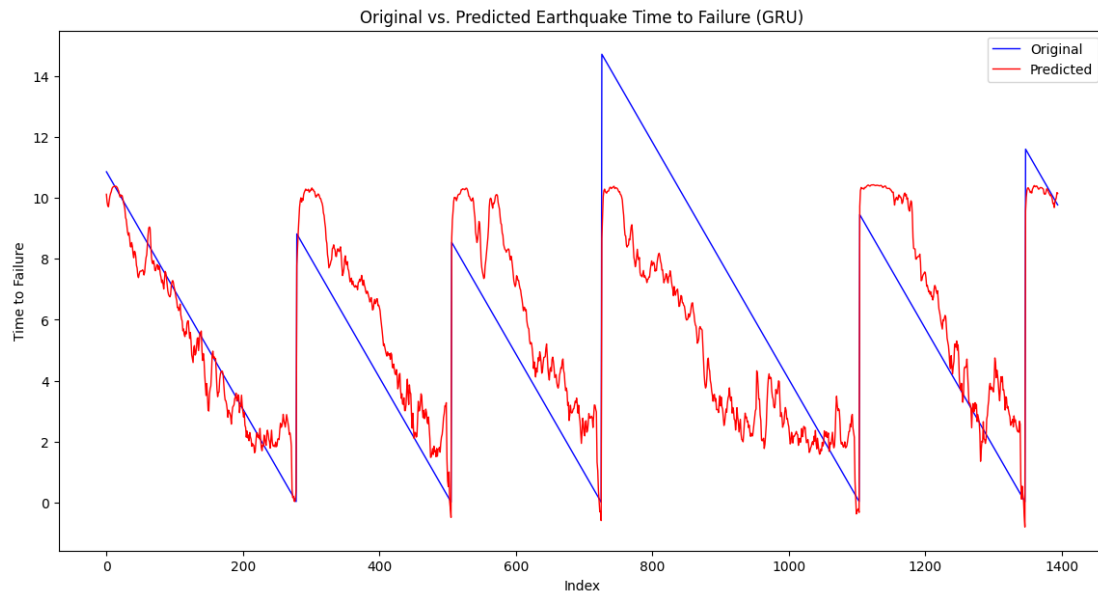


Figure 29: GRU Model Prediction

4.3. Comparison with other studies

In current study, utilizing USGS dataset for predictive modeling, the researcher significantly advanced the modeling techniques compared to previous efforts by Jain et al. (2021) and Mondol (2021). Specifically, the application of the RF model demonstrated substantial improvements, achieving an MAE of 0.167, which is markedly lower than the 2.996 reported by Jain et al. and 0.20550 by Mondol. Furthermore, the MSE of 0.055 and an R^2 of 0.729 in the RF model indicate a superior model fit and predictive accuracy. Similarly, the implementation of SVR by the current study recorded an MAE of 0.170 and an MSE of 0.059, significantly refining the predictive accuracy over the 2.958 MAE achieved by Jain et al. (Jain et al., 2021). Although Mondol's LR model showed slightly better

error metrics (MAE of 0.1831 and MSE of 0.0013) compared to the LR model in this study (MAE of 0.203, MSE of 0.078), the researcher's work broadens the analytical scope by including additional models such as XGBoost and GRU. These models exhibited competitive performance, with MAEs of 0.188 and 0.189 respectively, indicating robust predictive capabilities and enhancing the analytical framework (Mondol, 2021). Contrasting with Mondol's application of LSTM, which resulted in an MAE of 0.2131 and a negative MSE, the LSTM model in this study, while showing higher MAE and MSE values, indicates potential areas for further refinement. Overall, the introduction of advanced techniques like XGBoost and GRU in this study not only refines the accuracy of traditional models but also paves the way for innovative approaches in future research within this domain. Table 13 summarises these comparisons.

Table 13: Other studies Comparison (USGS)

Reference	Dataset	Models	MAE	MSE	R ²
(Jain et al., 2021)	USGS	RF	-	2.996	-
		MLP	-	2.713	-
		SVR	-	2.958	-
(Mondol, 2021)	USGS	Polynomial Regression	-	0.180	0.0132
		LR	-	0.1831	0.0013
		RF	-	0.20550	-0.1207
		LSTM	-	0.2131	-0.1553
Salam et al (2021)	historical earthquake data for Southern California	FPA-LS-SVM	0.434	-	-
Current study	USGS	RF	0.167	0.055	0.729
		LR	0.203	0.078	0.592
		SVR	0.170	0.059	0.686
		XGBoost regressor	0.188	0.069	0.637
		LSTM	0.309	0.189	0.022
		GRU	0.189	0.069	0.311

In current study on the LANL dataset, the researcher's approach to predictive modeling is juxtaposed with prior studies by Yang et al. (2022), Zaidi et al. (2020), Brykov et al. (2020), Zhang et al. (2023), and Kumar et al. (2021). The researcher's deployment of diverse models showcases a breadth of analysis aimed at improving prediction accuracy and model robustness. The application of LSTM and GRU models by the researcher yielded notably superior results in terms of MAE and MSE. The LSTM model produced an MAE of 1.789 and MSE of 5.000, showing a substantial

increase in predictive accuracy with an R^2 of 0.603, which is considerably higher than that of previous studies. Similarly, the GRU model recorded an MAE of 1.777 and an MSE of 4.192, alongside an impressive R^2 of 0.610, suggesting a strong model fit and reliability in predictions. Comparatively, the other models used in the current study, such as LR, SVR, RF, and XGB, demonstrated less competitive performance metrics. For instance, LR reported an MAE of 2.258, MSE of 8.094, and an R^2 of 0.415; SVR showed slight improvements with an MAE of 2.221 and an MSE of 7.851, achieving an R^2 of 0.428. RF and XGBoost also showed less favorable outcomes with RF posting an MAE of 2.283, MSE of 8.171, and R^2 of 0.410, and XGBoost with an MAE of 2.386, MSE of 8.803, and an R^2 of 0.364 (Table 14)

Table 14: Other studies Comparison (LANL)

Reference	Models	MAE	MSE	R^2
(Yang et al., 2022)	Auto-REP	1.520	1.590	-
(Zaidi et al., 2020)	Hand-crafted model	1.510	-	-
(Brykov et al., 2020)	XGBoost	1.913	-	-
(Zhang et al. 2023)	Stacking strategy model	2.238	-	-
((Kumar et al., 2021)	LGBM	2.4	-	-
Current study	LR	2.258	8.094	0.415
	SVR	2.221	7.851	0.428
	RF	2.283	8.171	0.410
	XGB	2.386	8.803	0.364
	LSTM	1.789	5.000	0.603
	GRU	1.777	4.192	0.610

In contrast to the researcher's current study, previous studies such as Yang et al. (2022), who used an Auto-REP model with an MAE of 1.520 and MSE of 1.590, and Zaidi et al. (2020), who employed a hand-crafted model with an MAE of 1.510, showed better predictive performance compared to most of the models employed in the current study. However, the LSTM and GRU models used by the researcher still stand out with their superior performance. The LSTM model achieved an MAE of 1.789 and MSE of 5.000, coupled with an R^2 of 0.603, while the GRU model reported an MAE of 1.777, MSE of 4.192, and an R^2 of 0.610. These results highlight the effectiveness of advanced recurrent neural network models like LSTM and GRU in achieving more accurate and reliable earthquake predictions on the LANL dataset (). This demonstrates potential for further improvements and innovations in traditional machine learning models, including LR, SVR, RF, and XGB, suggesting that further research and development could enhance their performance to meet or exceed the benchmarks set by earlier studies.

Overall, the study demonstrates the efficacy of using advanced RNN models over traditional ML approaches for complex datasets like LANL, highlighting a significant step forward in enhancing predictive accuracy and model reliability in the field of seismology (Yang et al., 2022; Zaidi et al., 2020; Brykov et al., 2020; Zhang et al., 2023; Kumar et al., 2021).

4.4. Results Discussion

The experimental results obtained from applying various ML and DL models to the USGS and LANL datasets provide valuable insights into the challenges and potentials of predicting seismic activities. This section

offers a detailed discussion of the findings, highlighting the performance of the different models, the influence of model architecture and training parameters.

4.4.1. USGS Dataset

For the USGS dataset, traditional ML models, including LR, SVR, RFR, and XGBoost Regressor, showed varied performances with RFR and SVR outperforming others in terms of MAE, MSE, and R-squared. These results underscore the effectiveness of ensemble and support vector models in handling complex, non-linear relationships inherent in geological data.

DL models, particularly those based on LSTM and GRU networks, demonstrated promising but inconsistent results. The nuanced differences in performance across models with varying architectures (e.g., number of layers, units, and dropout rates) and training parameters (e.g., epochs and batch sizes) suggest that there is no one-size-fits-all approach.

4.4.2. LANL Dataset

Among the models evaluated, the DL approaches, specifically LSTM and GRU, demonstrated superior predictive abilities compared to traditional ML models. The LSTM model achieved an MAE of 1.789, an MSE of 5.000, and the highest R-squared value of 0.603. Similarly, the GRU model recorded an MAE of 1.777, an MSE of 4.192, and an R-squared of 0.610, indicating slightly better performance than LSTM in handling the dataset's complexities.

Traditional ML models, LR, SVR, RF, and XGBoost regressor exhibited lower performance metrics. LR, for instance, had an R-squared

value of 0.415, indicating a moderate fit, while XGBoost had the lowest R-squared of 0.364 among the models, reflecting its lesser suitability for this dataset despite its general robustness in many other predictive tasks.

Answers to the research questions:

1. Effectiveness of ML Models in Predicting Seismic Activities

Among the models tested, the RFR exhibited exemplary performance across both datasets, particularly with the USGS data, achieving an MAE of 0.167, an MSE 0.055 of, and an R^2 of 0.729. This indicates a strong predictive capability with minimal error, suggesting its robustness in handling the nonlinearities and complexities inherent in seismic data. On the other hand, the DL models, specifically LSTM and GRU, showed superior performance on the LANL dataset, highlighting their efficacy in capturing temporal dependencies and complex patterns within seismic activity data.

2. Relationship Between Model Selection, Dataset Characteristics, and Predictive Accuracy

The relationship between model selection and dataset characteristics is pivotal in achieving high predictive accuracy. The success of RFR in the USGS dataset and LSTM/GRU in the LANL dataset underscores the importance of selecting models that are well-suited to the specific features and challenges presented by different datasets. The USGS dataset, which focuses on magnitude prediction, benefits from the robust feature handling and ensemble learning approach of RFR, while the temporal sequence

forecasting in the LANL dataset is effectively captured by the sequential nature of LSTM and GRU models.

3. Impact of Feature Engineering and Comprehensive Data on ML Efficacy

Feature engineering and the comprehensiveness of the dataset play critical roles in enhancing the performance of ML models. Effective feature selection and engineering help in highlighting the essential characteristics that contribute to the predictive accuracy, thereby optimizing the model's performance. Moreover, the need for comprehensive, high-quality data is crucial in training models that can generalize well to real-world conditions and provide reliable predictions.

4. Challenges and Complexities in Applying ML to Seismic Data Modelling

The application of ML techniques to seismic data modelling introduces several challenges and complexities, including the handling of large, noisy datasets, the need for substantial computational resources, and the requirement for sophisticated data preprocessing techniques. Moreover, the inherent unpredictability and variability of seismic events make it difficult to achieve high accuracy, necessitating ongoing research and adaptation of ML models to improve their predictive power.

Chapter Five

Conclusion and Future Work

5.1. Conclusion

5.2. Future Work

Chapter Five

Conclusion and Future Work

5.1. Conclusion

This research has embarked on an ambitious journey to harness ML technologies for the prediction of earthquake magnitude and timing, utilizing data from the USGS and the LANL websites. The USGS dataset was analyzed to predict earthquake magnitudes through regression analysis, while the LANL dataset was employed to forecast the timing of seismic events. A comprehensive evaluation of various ML algorithms including LR, SVR, RFR, XGBoost Regressor, and advanced DL models LSTM and GRU were conducted to determine their predictive capabilities using metrics such as MAE, MSE, and R-squared.

The results indicate that the choice of model significantly impacts the predictive accuracy. For instance, the RFR showcased robust predictive capabilities when applied to the USGS dataset, primarily due to its ability to handle complex feature interactions, which is crucial for modeling the nonlinear relationships typical in earthquake magnitude data. In contrast, traditional models did not perform as well on the LANL dataset, which required capturing temporal patterns and sequences.

On the LANL dataset, advanced DL models LSTM and GRU outperformed traditional models. This superior performance can be attributed to their architecture, which is particularly suited for analyzing sequential and time-series data, allowing them to effectively capture the temporal dynamics and dependencies present in the timing data of seismic events. This highlights the necessity of choosing a model that aligns with

the specific characteristics and requirements of the dataset to optimize predictive performance.

Moreover, the disparity in model performance across the two datasets underscores the importance of understanding dataset characteristics, such as data structure and underlying patterns, feature engineering and selection, which dictate the suitability of different ML approaches. For earthquake prediction tasks, where accuracy is critical, selecting the right model based on a thorough analysis of the data can significantly enhance the reliability and effectiveness of the predictions.

The research presented in this thesis contributes to the field of seismology by providing a detailed analysis of the predictive performance of various ML models. It sheds light on the critical factors that influence model performance, including the choice of model, the nature of the dataset, and the preprocessing and feature engineering and selection of dataset attributes. The insights gained from this study underscore the potential of ML in enhancing our understanding of seismic activities and point towards the need for continued research, focusing on model improvement, data enrichment, and the exploration of advanced algorithms for better prediction accuracy. In doing so, this work takes a significant step forward in the pursuit of more reliable and robust seismic activity prediction tools, aiming to contribute to the broader goal of earthquake preparedness and risk mitigation.

5.2. Future Work

Building on the comprehensive analysis and insights gained from this thesis, which focused primarily on predicting earthquake magnitude and timing, several promising directions for future research emerge to advance the field of seismic activity prediction using machine learning models.

Future efforts should also aim to predict the location of earthquakes, integrating a broader set of indicators such as seismic activity, historical patterns, epicenter locations, groundwater and radon gas levels, animal behavior, and the risks of foreshocks and aftershocks. Additionally, enhancing model performance through exploring data enrichment techniques and incorporating advanced computational methodologies will be crucial. These initiatives will help develop more sophisticated and accurate tools for earthquake preparedness and risk mitigation.

5.3. References

- Abdul Salam, M., Abdelminaam, & Salama, D.-D. (2021). Earthquake Prediction using Hybrid Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*, 12, 654–665. <https://doi.org/10.14569/IJACSA.2021.0120578>
- Al Banna, M., Ghosh, T., Nahian, M. J., Taher, K., Kaiser, M. S., Mahmud, M., Hossain, M., & Andersson, K. (2021). Attention-Based Bi-Directional Long-Short Term Memory Network for Earthquake Prediction. *IEEE Access*, PP, 1–1. <https://doi.org/10.1109/ACCESS.2021.3071400>
- Asim, K., Idris, A., Iqbal, T., & Martínez-Álvarez, F. (2018). Earthquake prediction model using support vector regressor and hybrid neural networks. *PLOS ONE*, 13, e0199004. <https://doi.org/10.1371/journal.pone.0199004>
- Baskaran, M. (2016). Radon as a Tracer for Earthquake Studies. In M. Baskaran (Ed.), *Radon: A Tracer for Geological, Geophysical and Geochemical Studies* (pp. 205–228). Springer International Publishing. https://doi.org/10.1007/978-3-319-21329-3_10
- Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth

- geoscience. *Science (New York, N.Y.)*, 363(6433), eaau0323.
<https://doi.org/10.1126/science.aau0323>
- Berhich, A., Belouadha, F.-Z., & Kabbaj, M. I. (2020). *LSTM-based Models for Earthquake Prediction* (p. 7).
<https://doi.org/10.1145/3386723.3387865>
- Bhargava, B., & Pasari, S. (2022). Earthquake Prediction Using Deep Neural Networks. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1, 476–479.
<https://doi.org/10.1109/ICACCS54159.2022.9785011>
- Brykov, M. N., Petryshynets, I., Pruncu, C. I., Efremenko, V. G., Pimenov, D. Y., Giasin, K., Sylenko, S. A., & Wojciechowski, S. (2020). Machine Learning Modelling and Feature Engineering in Seismology Experiment. *Sensors*, 20(15), Article 15.
<https://doi.org/10.3390/s20154228>
- Burgmann, R., & Dresen, G. (2008). Rheology of the Lower Crust and Upper Mantle: Evidence from Rock Mechanics, Geodesy, and Field Observations. *Annu. Rev. Earth Planet. Sci*, 36, 531–567.
<https://doi.org/10.1146/annurev.earth.36.031207.124326>
- Cothorn, C. R., & Smith, J. E. (Eds.). (1987). *Environmental Radon*. Springer US. <https://doi.org/10.1007/978-1-4899-0473-7>
- Debnath, P., Chittora, P., Chakrabarti, T., Chakrabarti, P., Leonowicz, Z., Jasiński, M., Gono, R., & Jasińska, E. (2021). Analysis of

- Earthquake Forecasting in India Using Supervised Machine Learning Classifiers. *Sustainability*, 13. <https://doi.org/10.3390/su13020971>
- Elnashai, A. S., Cleveland, L. J., Jefferson, T., & Harrauld, J. (2008). Impact of Earthquakes on the Central USA. *MAE Center Report 08-02*. <https://hdl.handle.net/2142/8971>
- Gao, Y., Zhang, K., Wu, Z., Tian, T., Gai, H., Zuo, J., & Li, B. (2022). An Early-Mid Holocene process of regional desertification recorded in aeolian sediments from the northern slope of the middle Himalayan Mountains. *Climate of the Past Discussions*, 1–26. <https://doi.org/10.5194/cp-2022-17>
- Grant, R., & Halliday, T. (2010). Predicting the unpredictable; evidence of pre-seismic anticipatory behaviour in the common toad. *Journal of Zoology*, 281, 263–271. <https://doi.org/10.1111/j.1469-7998.2010.00700.x>
- Hason, M. M., Hanoon, A. N., & Abdulhameed, A. A. (2023). Particle swarm optimization technique-based prediction of peak ground acceleration of Iraq's tectonic regions. *Journal of King Saud University - Engineering Sciences*, 35(7), 463–473. <https://doi.org/10.1016/j.jksues.2021.06.004>
- Hassin, A., & Abbood, D. (2021). Machine Learning System for Human – Ear Recognition Using Scale Invariant Feature Transform. *Artificial*

Intelligence & Robotics Development Journal, 1–12.

<https://doi.org/10.52098/airdj.20217>

Hoque, A., Raj, J., Saha, A., & Bhattacharya, P. (2020). Earthquake Magnitude Prediction Using Machine Learning Technique. In N. Kar, A. Saha, & S. Deb (Eds.), *Trends in Computational Intelligence, Security and Internet of Things* (pp. 37–53). Springer International Publishing. https://doi.org/10.1007/978-3-030-66763-4_4

Jain, R., Nayyar, A., Arora, S., & Gupta, A. (2021). A comprehensive analysis and prediction of earthquake magnitude based on position and depth parameters using machine and deep learning models. *Multimedia Tools and Applications*, 80(18), 28419–28438. <https://doi.org/10.1007/s11042-021-11001-z>

Jarah, N. B., Hashim, K. M., & Hassin, A. H. (2023). Earthquake prediction in Iraq using machine learning techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(1), Article 1. <https://doi.org/10.11591/ijeecs.v29.i1.pp322-329>

Jordan, T. H. (2006). Earthquake Predictability, Brick by Brick. *Seismological Research Letters*, 77(1), 3–6. <https://doi.org/10.1785/gssrl.77.1.3>

- Kagan, Y. Y., & Jackson, D. D. (1991). Seismic Gap Hypothesis: Ten years after. *Journal of Geophysical Research: Solid Earth*, 96(B13), 21419–21431. <https://doi.org/10.1029/91JB02210>
- Kavianpour, P., Kavianpour, M., Jahani, E., & Ramezani, A. (2023). A CNN-BiLSTM model with attention mechanism for earthquake prediction. *The Journal of Supercomputing*, 79(17), 19194–19226. <https://doi.org/10.1007/s11227-023-05369-y>
- King, C.-Y. (1986). Gas geochemistry applied to earthquake prediction: An overview. *Journal of Geophysical Research: Solid Earth*, 91(B12), 12269–12281. <https://doi.org/10.1029/JB091iB12p12269>
- Kulkarni, R. (2012). A Review Of Application Of Data Mining In Earthquake Prediction. *International Journal of Computer Science & Information Technology (IJCSIT)*.
- Kumar, D. S. A., Kumar, A., Dhanraj, A., & Thakur, A. (2021). EARTHQUAKE PREDICTION USING MACHINE LEARNING. *International Research Journal of Modernization in Engineering Technology and Science*. <https://api.semanticscholar.org/CorpusID:258721909>
- Mallouhy, R., Jaoude, C. A., Guyeux, C., & Makhoul, A. (2019). Major earthquake event prediction using various machine learning algorithms. *2019 International Conference on Information and*

- Communication Technologies for Disaster Management (ICT-DM)*, 1–7. <https://doi.org/10.1109/ICT-DM47966.2019.9032983>
- Meenal, R., Kailash, K., Michael, P. A., Joseph, J. J., Josh, F. T., & Rajasekaran, E. (2022). Machine learning based smart weather prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, 28(1), Article 1. <https://doi.org/10.11591/ijeecs.v28.i1.pp508-515>
- Mondol, M. (2021). *Analysis and Prediction of Earthquakes using different Machine Learning techniques*. <https://doi.org/10.13140/RG.2.2.15085.10727>
- Ogata, Y. (1999). Seismicity Analysis through Point-process Modeling: A Review. In M. Wyss, K. Shimazaki, & A. Ito (Eds.), *Seismicity Patterns, their Statistical Significance and Physical Meaning* (pp. 471–507). Birkhäuser. https://doi.org/10.1007/978-3-0348-8677-2_14
- Reinisch, E. C., Ali, S. T., Cardiff, M., Kaven, J. O., & Feigl, K. L. (2020). Geodetic Measurements and Numerical Models of Deformation at Coso Geothermal Field, California, USA, 2004–2016. *Remote Sensing*, 12(2), Article 2. <https://doi.org/10.3390/rs12020225>
- Stein, S., & Wysession, M. (2009). *An Introduction to Seismology, Earthquakes, and Earth Structure*. John Wiley & Sons.

- Tributsch, H. (1982). *When the Snakes Awake: Animals and Earthquake Prediction*. <https://www.semanticscholar.org/paper/When-the-Snakes-Awake%3A-Animals-and-Earthquake-Tributsch/6dab63ba531ed82a95e9242c88bbc18a60feb90d>
- Wijaya, U., Kusrini, & Muhammad, A. H. (2022). Indonesian Seismic Mitigation using Earthquake Predicted Artificial Intelligence Model. *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, 349–354. <https://doi.org/10.1109/ICOIACT55506.2022.9972091>
- Xie, Y., Ebad Sichani, M., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 36(4), 1769–1801. <https://doi.org/10.1177/8755293020919419>
- Yang, F., Kefalas, M., Koch, M., Kononova, A. V., Qiao, Y., & Bäck, T. (2022). Auto-REP: An Automated Regression Pipeline Approach for High-efficiency Earthquake Prediction Using LANL Data. *2022 14th International Conference on Computer and Automation Engineering (ICCAE)*, 127–134. <https://doi.org/10.1109/ICCAE55086.2022.9762437>
- Zaidi, T., Samy, A., Kocatürk, M., & Ateş, H. F. (2020). Learned vs. Hand-Crafted Features for Deep Learning Based Aperiodic Laboratory Earthquake Time-Prediction. *2020 28th Signal Processing and*

Communications Applications Conference (SIU), 1–4.

<https://doi.org/10.1109/SIU49456.2020.9302474>

الملخص

يمكن أن تساهم تقنية تنبؤ دقيقة وفعالة بالزلازل في إنقاذ العديد من الأرواح والاقتصاد. ومن منطلق هذا الهدف، تستكشف هذه الأطروحة تطبيق نهج التعلم الآلي في تنبؤ خصائص الزلازل، باستخدام بيانات المسح الجيولوجي الأمريكي (USGS) للزلازل بقوة 3.5 وما فوق للعام الماضي (2023) وبيانات مختبر لوس ألاموس الوطني (LANL). تم استخدام مجموعة بيانات USGS لتوقع قوة الزلازل من خلال التحليل الانحداري، بينما تركز مجموعة بيانات LANL على التنبؤ بتوقيت الأحداث الزلزالية القادمة. تم تقييم مجموعة متنوعة من خوارزميات التعلم الآلي بدقة، بما في ذلك الانحدار الخطي (LR)، والانحدار الداعم الناقل (SVR)، ومنظم الغابات العشوائية (RFR)، ومنظم (XGB) XGBoost، ونماذج التعلم العميق المتقدمة مثل الذاكرة قصيرة المدى طويلة الأمد (LSTM)، ووحدات البوابة المتكررة (GRU) تم تقييم أداء النماذج باستخدام عدة مقاييس تقييم، وهي الخطأ المطلق الوسطي (MAE)، ومتوسط الخطأ التربيعي (MSE)، ومربع الارتباط (R^2). من بين النماذج التي تم تقييمها، أظهر RFR أداءً متميزًا على مجموعة بيانات USGS، حيث حقق خطأ مطلق وسطي قدره 0.167، وخطأ تربيعي وسطي قدره 0.055، ومربع ارتباط بلغ 0.729. يبرز هذا قدرته القوية في نمذجة العلاقات المعقدة ضمن بيانات قوة الزلازل. على الجانب الآخر، تفوق نموذج GRU في مجموعة بيانات LANL، حيث حقق أفضل أداء بخطأ مطلق وسطي قدره 1.777، وخطأ تربيعي وسطي قدره 4.192، ومربع ارتباط بلغ 0.610. تم تحقيق هذه النتيجة بعد هندسة وخلق 37 ميزة محددة بناءً على عدد عمليات رياضية، مما يدل على أهمية هندسة الميزات في تعزيز دقة التنبؤ للنموذج لتوقيت الأحداث الزلزالية. تشير النتائج إلى أن الجهود المستقبلية يجب أن تركز على تحسين عمليات هندسة واختيار الميزات ودمج مؤشرات متنوعة أخرى إضافية مثل النشاط الزلزالي، والأنماط التاريخية، والبيانات الجيولوجية، لتحسين قدرات التنبؤ لهذه النماذج بشكل أكبر. سيساعد هذا النهج في تطوير أدوات أكثر تطوراً ضرورية للتأهب الفعال للزلازل وتخفيف المخاطر.

جامعة البحرين

برامج الدراسات العليا

كلية العلوم



التنبؤ بالزلازل باستخدام نهج تعلم الآلة

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمة

إعداد

فتحان طائل صالح شاجره

20082472

إشراف

أ.د. نبيل حويحي
أستاذ

جامعة البحرين

مملكة البحرين
مارس 2024