An approach in Prediction of Earthquakes using VAE-LSTM model

Suraj Karyamapudi
19232632
MSc Data Analytics
National College of Ireland

Abstract

Earthquakes are one of the most devastating natural disasters, resulting in significant loss of life and property. Early detection can aid in the timely evacuation of affected areas, reducing the potential damage. While traditional methods of earthquake detection and prediction have seen advancements, there remains room for improved accuracy and timeliness. This research leverages the capabilities of Variational Autoencoders (VAE) combined with Long Short-Term Memory (LSTM) networks to create a model designed for real-time earthquake detection.A novel approach was adopted by integrating VAE with LSTM to process seismic data, aiming to predict potential earthquake occurrences.The VAE-LSTM model demonstrated improved accuracy and efficiency in detecting seismic activities compared to conventional methods.The integration of VAE and LSTM offers a promising direction in the realm of seismology, indicating the potential of deep learning models in understanding and predicting complex natural events.The VAE-LSTM model can be deployed in seismic monitoring stations worldwide, offering more accurate, timely warnings that can save lives.

Keywords— **V**AE-LSTM, LSTM, Machine Learning, Earthquake

1 Introduction

Earthquakes, the natural part of Earth's geological process caused by the violent motions of tectonic plates, have long been a source of fascination and fear. Throughout history, civilizations have attempted to predict and comprehend these catastrophes, primarily in order to lessen their devastation. While previous earthquake prediction systems relied on geological and seismological observations, the digital era has provided us with massive amounts of data as well as the computer capabilities to analyse them.

Interdisciplinary integration has increased significantly in recent years, particularly between geophysics and computer sciences. Deep learning, an artificial intelligence branch, has emerged as a particularly potential ally in the pursuit of more accurate earthquake predictions. The enormous, nuanced patterns revealed in seismic data, which frequently evade human interpretation, may be understood and predicted using advanced algorithms.

1.1 Motivation

With the rapid advancements in machine learning and deep learning techniques, there lies an untapped potential to improve earthquake detection mechanisms, potentially saving lives. The decision to use a hybrid model combining Variational Autoencoders (VAEs) and Long Short-Term Memory (LSTM) networks for this research was motivated by the distinct capabilities of each model. While VAEs excel at displaying complicated data in a simplified manner, LSTMs can predict sequences, which is critical when predicting temporal events such as earthquakes. In hope to achieve a solution that is larger than the individual deployment a LSTM model this model has been developed. Despite rapid technological improvements, there is a large gap between the research and practical implementations of hybrid models such as VAE-LSTM for earthquake prediction. While both theories have found individual success in diverse disciplines, their combined potential in seismology remains largely untapped.

This research has numerous beneficiaries. Academics and researchers can explore deeper into this work, using it as a foundation for more complex models. These projections can be used by disaster management teams, city planners, and infrastructure developers to make safer urban development decisions. Furthermore, early warning systems can be greatly improved, potentially saving lives and decreasing property damage.

1.2 Research Question and Objectives

1.2.1 Research Question

"How well can a VAE-LSTM hybrid model predict the occurance of earthquake?"

1.2.2 Objectives

To address this question, the research objectives are:
Investigate the state of the art in earthquake detection methods.
Design a model integrating VAE and LSTM for seismic data processing.
Implement the designed model on real-world seismic data.
Evaluate the model's performance against traditional method like LSTM.

1.3 Contribution

The major contribution of this research is a groundbreaking VAE-LSTM model tailored for accurate and efficient earthquake detection.

1.4 Structure of Paper

This report provides a comprehensive view of the research. After this introduction , the following section is the literature survey section iwhich establishes the current state of the art to understand the current developments and previous developments in this domain. It is followed by the methodology which details the research process, techniques followed, and process implemented followed by design and implementation specifics. The section followed by that is Conclusion and Results in which the results are discussed followed by potential future directions.

2 Literature Review

2.1 Introduction

The study and understanding of earthquakes represent a critical domain in earth sciences, influencing urban planning, infrastructure development, and disaster mitigation. Over the years, with the advancement in technology and computational abilities, the methodologies and techniques employed in the domain of earthquake prediction and analysis have seen a substantial evolution. From traditional seismic studies to the incorporation of statistical models, machine learning techniques, and hybrid methodologies, the scope and depth of research in this domain have expanded remarkably. This literature survey aims to present a comprehensive overview of these developments by weaving together findings from sixteen seminal papers in the field.

2.2 Traditional Seismic Studies: The Foundation

At the core of earthquake studies lies traditional seismology, which has been instrumental in laying the foundation for subsequent advancements. Earthquakes, with their unpredictable nature and potential for massive destruction, have always posed a significant challenge to researchers and scientists. Predicting such natural disasters is a complex endeavor, given the myriad of variables and factors at play.

Pei et al. (2022) delve into these complexities, shedding light on the limitations of deterministic earthquake predictions. The authors emphasize a shift towards understanding the temporal and spatial statistical characteristics of earthquake occurrences. By tapping into the power of statistical seismology, the paper underscores the potential of constructing risk statistical prediction models. Such models are not only significant in evaluating earthquake risks but also crucial for urban and infrastructural planning. The paper also draws attention to certain unexpected seismic events, such as the Wenchuan M8 earthquake in 2008, highlighting the need for continuous reflection and revision of traditional seismic hazard assessments. Such events serve as a stark reminder of nature's unpredictability and the importance of adaptive methodologies in seismology.

Pei et al. (2022) further elaborates on the significance of statistical seismology, especially in the realm of earthquake prediction. The paper introduces the concept of declustering, emphasizing the potential of declustered catalogs in achieving accurate earthquake predictions. By proposing a novel declustering

algorithm, Pei et al. (2022) fortifies the bridge between traditional seismic studies and modern statistical methods, providing a more nuanced understanding of seismic events.

2.3 Statistical Models and Modern Techniques

The modern era, characterized by the proliferation of data and advanced computational capabilities, has ushered in a host of innovative methodologies for earthquake prediction and analysis. Leveraging vast datasets, these methods employ intricate statistical models and advanced algorithms to glean insights and enhance prediction accuracy.

The significance of statistical seismology, especially in probabilistic seismic hazard assessment, is accentuated by Pei et al. (2022). The researchers spotlight the role of ETAS (epidemic type aftershock sequence) in seismic hazard assessment. This approach, based on modern earthquake catalogs, promises enhanced efficacy, especially in urban areas and regions with thick sedimentary zones.

2.4 Incorporation of Machine Learning and Hybrid Models

The surge in computational capabilities and the proliferation of data-driven methodologies have opened new horizons in the domain of earthquake prediction. Machine learning, characterized by its ability to process vast datasets and identify intricate patterns, emerges as a game-changer.

A pioneering step in this direction is evident in the work of (Rayan & ARTUNER 2022). The duo proposes LSTM-based deep learning models tailored for earthquake prediction using ionospheric data. Harnessing the potential of ionospheric anomalies, often observed before seismic activities, the LSTM models demonstrate remarkable accuracy. Such models not only provide a fresh lens to interpret ionospheric data but also hold immense promise for early warning systems.

This trend of employing advanced computational techniques is further exemplified by Salam et al. (2021), who introduce hybrid machine learning models for earthquake prediction. Focusing on the southern California region, the research incorporates seismic indicators, offering a comprehensive prediction methodology. The study introduces two prediction models and, through rigorous assessment criteria, establishes the superiority of the FPA-LS-SVM model in terms of prediction accuracy. Such hybrid techniques, amalgamating traditional seismic indicators with machine learning methodologies, offer a glimpse into the future of earthquake prediction.

2.5 Predictive Modeling in Seismology

As the field of seismology progresses, the methods and techniques used to understand and predict earthquakes have evolved significantly. The application of statistical seismology in probabilistic seismic hazard assessment has become more pronounced, especially with the modern challenges posed by unforeseen events. Pei et al. (2022) shed light on the limitations of deterministic earthquake predictions, emphasizing the need for statistical approaches. Their work underscores the significance of the temporal and spatial statistical characteristics of earthquake occurrences in assessing risk and planning for seismic events. The paper further highlights the value of earthquake catalogues in these evaluations, suggesting that ETAS (epidemic type aftershock sequence) based seismic hazard assessment holds promise for more effective and accurate mapping, especially in urban and low seismic hazard areas.

2.6 Traditional Seismology and New Insights

Seismic hazards and their implications have been a continuous subject of study. A comprehensive review by Pei et al. (2022) on statistical seismology emphasized the importance of analyzing earthquake occurrences and constructing risk prediction models. Notably, they shed light on the effectiveness of seismic hazard assessment using earthquake catalogues, especially via the ETAS model in populated urban areas. This focus on urban areas echoes the sentiments of Hu et al. (2022) who discussed the potential of geodetic and seismic data in providing insights into earthquake hazards in cities.

2.7 Machine Learning, AI, and Advanced Predictive Models

With the advent of machine learning, researchers have combined traditional seismology with advanced computational methods. Salam et al. (2021) proposed two models that utilize hybrid machine learning techniques, emphasizing the impressive predictive accuracy of the FPA-LS-SVM model. Similarly, in a bid to harness the power of deep learning, Berhich et al. (2023) proposed the use of LSTM models,

focusing on the prediction of earthquakes using ionospheric data, highlighting the potential of TEC as an indicator. The importance of machine learning was further cemented by studies such as those by Al Banna et al. (2021) who leaned on AI for short-term LSTM earthquake risk prediction profiles, and Lin et al. (2020) who demonstrated the significance of machine learning for earthquake signal detection and location.

2.8 Interplay of Multiple Factors and Detailed Analyses

Earthquakes, being multi-faceted phenomena, have been analyzed from various angles by Wang et al. (2017) who provided an in-depth examination of earthquake ground motion, highlighting the importance of understanding spatial variability. Meanwhile, Gajan (2021) emphasized the significance of Bayesian methods for ground-motion forecasting by using rock shallow foundations during earthquakes. On a broader scale, Essam et al. (2021) used over 30 years of data from Malaysia earthquake occurrences using clustering algorithms, while Jain et al. (2021) used magnitude based on depth and magnitude for earthquake prediction.

Moreover, earthquakes' interactions with other natural occurrences were spotlighted by studies which include Kavianpour et al. (2023) who used cnn-bilstm prediction, and Khalil et al. (2021) who centered their research on fault landslides.

2.9 Risk Assessment and Disaster Management

Beyond prediction, the assessment of seismic risks and the subsequent management of potential disasters is of prime importance. Murwantara et al. (2020) utilized Bayesian networks to assess a large amount of data for earthquake vulnerability, especially for regions like Indonesia. Meanwhile, taking a different approach,Lopez-Martin et al. (2017) Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection.

2.10 Conclusion and Research Niche

The literature spanning over the last few years elucidates the seismic shifts in earthquake prediction methodologies. From early methods to the recent innovations harnessing machine learning, the journey of seismology is marked by continuous evolution and advancement. These predictive models, combined with preventive measures and enhanced structural health monitoring, pave the way for a future where the devastation caused by earthquakes could be substantially mitigated. While the challenges in earthquake prediction and analysis are immense, the combined efforts of researchers from different disciplines provide hope. With the continuous evolution of technology and methodologies, there's an optimistic view towards a future where societies can be better prepared for seismic activities.

3 Outputs Summary

3.1 Data Preparation & Exploration

Output: Processed Earthquake dataset.
Type: Code/Data
User(s): Data analysts, researchers, data scientists.
Overview: The initial dataset is loaded, and various preprocessing steps are applied such as converting date and time columns to separate components like year, month, day, hours, etc. The structure and content of the dataset are explored using print statements.

3.2 Feature Engineering

Output: Dataset with FFT applied on magnitude and other temporal features.
Type: Code/Data
User(s): Data scientists.
Overview: Fast Fourier Transform (FFT) is applied on the magnitude column, and the resultant data is visualized. Temporal features are also extracted.

3.3 Data Cleaning

Output: Dataset with NaN values handled and outliers removed.
Type: Code/Data
User(s): Data engineers, data scientists.
Overview: NaN values are identified and replaced with the mean. Outliers are also detected using the IQR method and removed.

3.4 Feature Importance Visualization

Output: Bar chart showing the importance of each feature.
Type: Design Artefact (Graph)
User(s): Data scientists, stakeholders.
Overview: A RandomForestRegressor is used to determine and visualize the importance of each feature, aiding in understanding the dataset's attributes.
Feature importance 1 obtained

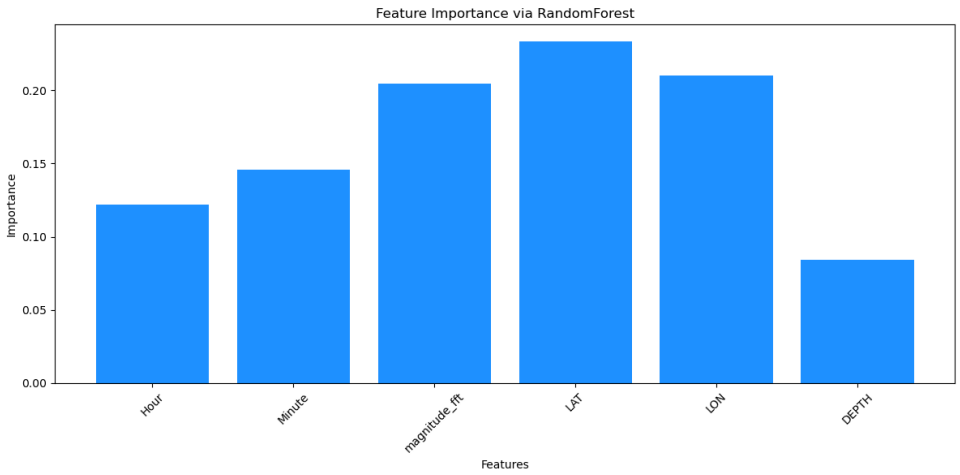

Figure 1: Feature Importance

3.5 Data Augmentation

Output: Noisy dataset.
Type: Code/Data
User(s): Data scientists.
Overview: Noise is introduced to the scaled dataset to augment the data, which can potentially help in robust model training.

3.6 Variational Autoencoder (VAE) Model

Output: VAE Model trained on earthquake data.
Type: Models
User(s): Machine learning engineers, researchers.
Overview: A VAE model is built and trained, which compresses the dataset's features into a lower-dimensional latent space.

3.7 LSTM Model for Sequence Prediction

Output: LSTM model trained on the latent space.
Type: Models
User(s): Machine learning engineers, researchers.
Overview: An LSTM model is designed to make sequence predictions based on the encoded data from the VAE.

3.8 Reconstruction Error Analysis

Output: Distribution of reconstruction errors.
Type: Design Artefact (Graph)
User(s): Data scientists, machine learning engineers.
Overview: After decoding the predictions made by the LSTM in the latent space, the reconstruction error between the original and decoded data is analyzed and visualized.
Reconstruction Error 2 obtained

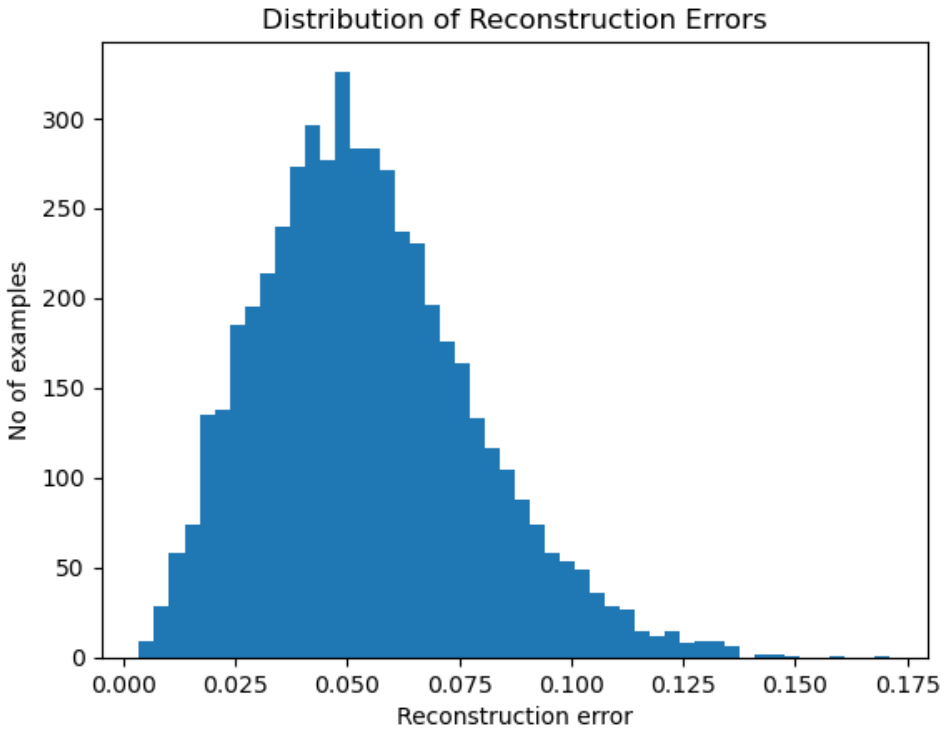

Figure 2: Reconstruction Error

3.9 Anomaly Detection

Output: Visualization of anomalous data.
Type: Design Artefact (Graphs)
User(s): Data scientists, stakeholders.
Overview: Using the calculated reconstruction error, a threshold is determined to identify anomalies. These anomalous data points are visualized for better understanding.

3.10 Evaluation Metrics

Output: Metrics such as MAE, RMSE, MBD, and MSLE.
Type: Code/Results
User(s): Data scientists, stakeholders.
Overview: Various evaluation metrics are computed to assess the quality of the predictions made by the LSTM model. These include metrics like Mean Absolute Error, Root Mean Squared Error, Mean Bias Deviation, and Mean Squared Logarithmic Error.

3.11 Final Output Values

Mean Absolute Error: 0.1779 Root Mean Squared Error: 0.2233 Mean Bias Deviation: 0.0155Mean Squared Logarithmic Error: 0.0232

Overall: The code provides a comprehensive workflow that begins with data preprocessing and culminates in anomaly detection and performance evaluation using advanced models such as VAEs and LSTMs. This could serve as a basis for detecting unusual seismic activities based on the provided dataset.

4 Research Methods & Specifications

4.1 Research Method - Methodology

In this project, the primary goal was to understand the characteristics of earthquake data and apply advanced data processing and machine learning techniques to gain insights, make predictions, and identify anomalies. The project was broken down into several methodological phases, from initial data collection to the final evaluation of the results.

Steps followed in the Research:

A simple basic flowchart 3 followed for this research is shown below

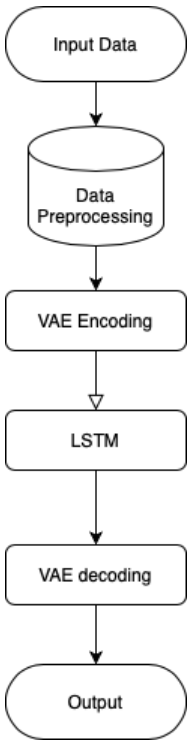

Figure 3: Flowchart

4.1.1 Data Collection/Gathering

Source and Contextual Analysis: Procured seismic data from Southern California Earthquake Data Centre. The data sourced was named as a file 'Earthquake.csv'. This dataset comprises various features related to seismic activities, such as timestamps, magnitudes, depths, and geographical coordinates. To gain a comprehensive understanding of the dataset, an initial contextual analysis was done to understand the relevance of each column and its significance in the broader realm of earthquake research right from the year 1932 to 2023 and all the waves with magnitude of 3.5 and above are considered in this region.

An extract of the dataset as an example of attributes of the dataset are given below 4

YYYY/MM/DD	HH:mm:SS.ss	ET	GT	MAG	M	LAT	LON	DEPTH	Q	EVID	NPH	NGRM
20/09/1932	16:21:23	eq	I	4.12	I	31.7065	-115.412	6	D	3359298	9	0
06/10/1932	08:56:00	eq	I	3.5	h	33.28333	-116.8	6	Z	10087202	0	0
08/10/1932	07:47:12	eq	I	4.02	I	32.1935	-115.64183	6	D	3359318	15	0
08/10/1932	21:15:49	eq	I	3.99	I	32.6415	-115.4315	6	D	3359332	7	0
09/10/1932	04:33:59	eq	I	3.79	I	32.61017	-115.3155	6	C	3359344	7	0
09/10/1932	22:50:41	eq	I	4.47	I	32.64	-115.43	6	D	3359345	10	0
09/10/1932	23:45:26	eq	I	4.12	I	32.64833	-115.6085	6	D	3359347	8	0
10/10/1932	01:29:26	eq	I	4.25	I	32.45583	-115.32067	6	C	3359348	12	0
10/10/1932	19:53:09	eq	I	3.73	I	32.65067	-115.28833	6	C	3359353	8	0
13/10/1932	17:51:22	eq	I	3.93	I	31.48617	-116.218	6	D	3359366	8	0
14/10/1932	15:48:51	eq	I	3.59	I	33.06467	-115.03267	6	D	3359369	8	0
17/10/1932	15:35:33	eq	I	3.5	I	33.16567	-116.25017	6	C	3359371	8	0
20/10/1932	09:20:17	eq	I	3.72	I	31.28967	-117.069	6	D	3359383	7	0
21/10/1932	17:29:41	eq	I	4.12	I	33.23367	-118.579	6	C	3359393	13	0
23/10/1932	12:41:31	eq	I	3.7	I	31.50317	-116.229	6	D	3359397	6	0
01/11/1932	04:45:47	eq	I	4	I	34.0335	-117.30417	18.7	B	3359417	9	0
08/12/1932	07:34:11	eq	I	4.19	I	31.596	-116.0515	6	D	3359466	7	0
23/12/1932	21:27:00	eq	I	3.5	h	34.63333	-116.95	6	Z	10087218	0	0
18/01/1933	18:53:20	eq	I	3.78	I	32	-116.273	6	C	3359557	2	0
25/01/1933	14:44:24	eq	I	4.08	I	34.096	-116.438	6	C	3359560	2	0
27/01/1933	16:27:55	eq	I	4.09	I	31.078	-116.547	6	C	3359648	2	0
30/01/1933	16:51:55	eq	I	3.52	I	33.574	-120.65	6	C	3359567	1	0
24/02/1933	19:33:18	eq	I	4.85	I	33.063	-118.163	6	C	3359699	1	0
26/02/1933	06:55:00	eq	I	3.8	I	35.965	-117.803	6	C	3359706	1	0
11/03/1933	01:54:09	eq	I	6.4	w	33.63083	-117.9995	6	C	3359741	7	0
11/03/1933	02:04:00	eq	I	4.68	h	33.63	-117.97	6	C	3359742	1	0
11/03/1933	02:05:07	eq	I	3.95	h	33.81917	-118.05933	6	C	3359744	10	0
11/03/1933	02:09:00	eq	I	4.54	I	33.63	-117.97	6	C	3359745	1	0
11/03/1933	02:09:58	eq	I	4.4	h	33.79433	-118.05917	6	C	3359752	10	0
11/03/1933	02:11:00	eq	I	4.4	h	33.75	-118.08333	0	C	10084338	0	0

Figure 4: Attributes

4.1.2 Data Cleaning and Preprocessing

Cleaned, transformed, and normalized the data for model ingestion.

Date-Time Conversion: The columns 'YYYY/MM/DD' and 'HH:mm:SS.ss' were represented in string format. To facilitate more granular temporal analysis, these columns were converted into pandas datetime and timedelta formats, respectively.

Feature Extraction: Post-conversion, we extracted year, month, day, hour, minute, second, and millisecond from these datetime columns. The original columns were then dropped to avoid redundancy.

Handling Missing Values: Data quality is paramount for reliable machine learning outcomes. Upon inspection, if any missing values (NaN) were found in the dataset, they were imputed with the mean of the respective column.

Outlier Removal: Given the nature of earthquake data, outlier values can distort the overall model's understanding. To combat this, the Interquartile Range (IQR) method was applied. Rows that had values beyond the 1.5 IQR range were considered outliers and removed.

4.1.3 Data Analysis and Feature Engineering

Frequency Domain Analysis: To extract more information from the magnitude of the earthquakes, Fast Fourier Transform (FFT) was used. The absolute values of the FFT of the 'MAG' column were taken, adding a 'magnitude_fft' column to the dataset.

Feature Importance: Before delving into deeper models, it's crucial to understand the importance of each feature. We used the RandomForestRegressor to gauge the significance of features in predicting the 'MAG' column. This analysis not only helped in refining the selection of columns but also gave insights into the dataset's nature.

Data Augmentation: Noise was artificially added to the scaled data to diversify the dataset and make models robust. This procedure is standard when working with Variational Autoencoders (VAEs).

4.1.4 Implementation

Variational Autoencoder (VAE) Setup: Given the high-dimensionality of earthquake data, VAEs were chosen for their ability to compress data into a latent space efficiently. The VAE comprised an encoder and a decoder. The encoder reduced data into a smaller dimensional space, and the decoder reconstructed this compressed data back to its original form. An intermediate step involved sampling, which introduced a stochastic element to the encoding process.

LSTM Model: Post encoding, the latent features underwent prediction using Long Short-Term Memory (LSTM) networks. This type of Recurrent Neural Network (RNN) was chosen due to its prowess in handling sequential data like time series. Here, sequences of latent features were fed into the LSTM to predict future sequences.

Decoding Predictions: Predictions made in the latent space were not directly interpretable. Hence, using the VAE's decoder, these predictions were transformed back to the original data space. This process ensured that results were in a comprehensible format.

4.1.5 Testing/Evaluation

Reconstruction Error: The primary metric for the evaluation of VAE performance was the reconstruction error, which is the difference between the original and the decoded data. This error gave insight into how well the VAE learned the data's characteristics.

Anomaly Detection: With the calculated reconstruction error, we applied exponential smoothing to smoothen the error curve. A threshold was set at the 99th percentile of the error, and data points with an error beyond this threshold were labeled as anomalies. Visualizations of these anomalies helped in qualitative assessments.

Evaluation Metrics: To measure the predictive performance of the LSTM model on the encoded data, several metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Bias Deviation (MBD), and Mean Squared Logarithmic Error (MSLE) were computed.

Justification: Given the sequential nature of earthquake data and its high-dimensionality, the combination of VAE and LSTM was a natural fit. VAE efficiently handled the dimensionality, and LSTM's memory cells capably managed the time-dependent sequences. Furthermore, outlier removal using the IQR method, and data augmentation through noise addition, aimed at refining the data quality, making the subsequent model's outputs more reliable.

In conclusion, this methodology section delineated the project's entire pipeline, ensuring that readers have a clear understanding of the approaches used and their justifications. The methods were chosen after careful consideration of the data's characteristics and the project's objectives.

4.2 Design Specification

Techniques, Architecture, and Frameworks

4.2.1 Variational Autoencoder (VAE)

At the core of our design lies the VAE, a probabilistic deep learning model that serves as a generative technique for data compression and reconstruction. It's fundamentally a neural network with two main parts:

The architecture of the VAE Model 5 utilised for this research is shown below

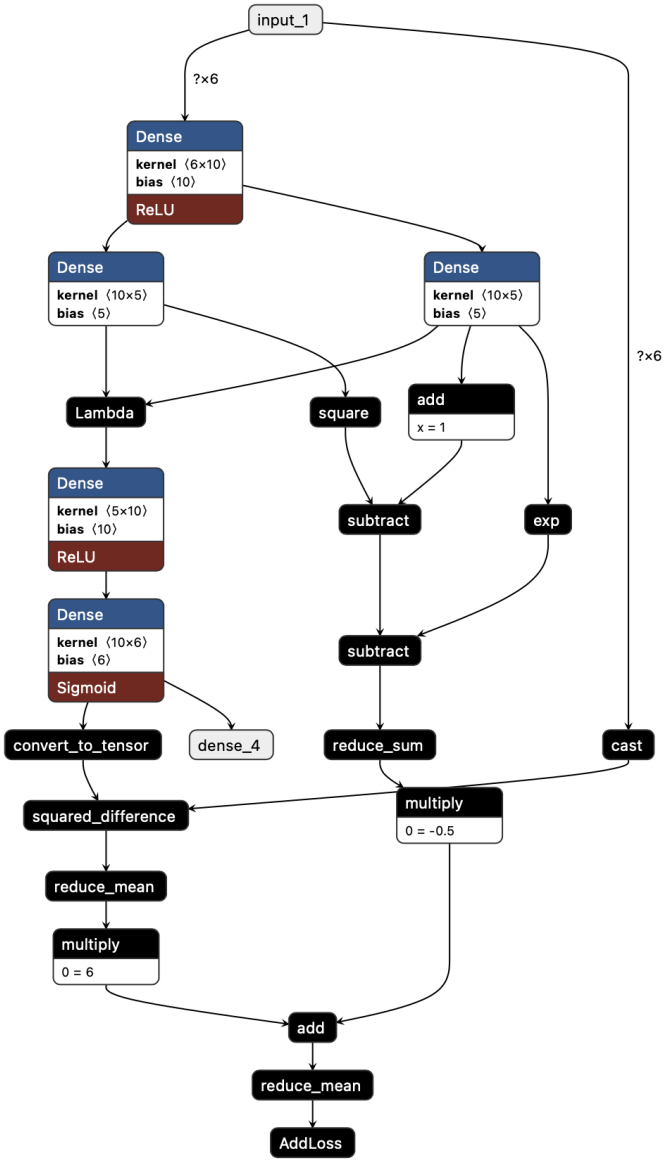

Figure 5: VAE Model

4.2.2 Encoder

This portion of the network is responsible for compressing the input data into a latent space, creating a condensed representation.

4.2.3 Decoder

The decoder's role is to reverse the process, reconstructing the input data from its condensed latent representation.

The VAE introduces stochasticity in the encoding phase by using a distribution, typically Gaussian, to generate a latent space. This makes it suitable for our project, as it brings an element of robustness and can produce variations of the encoded data.

4.2.4 Long Short-Term Memory (LSTM) Networks

LSTM networks, a subset of recurrent neural networks (RNNs), are specifically designed to remember patterns over long sequences. Given the time series nature of our earthquake data, LSTMs are particularly apt, ensuring that the model retains memory of past events to predict future sequences.

The architecture of the LSTM Model 6 utilised for this research is shown below

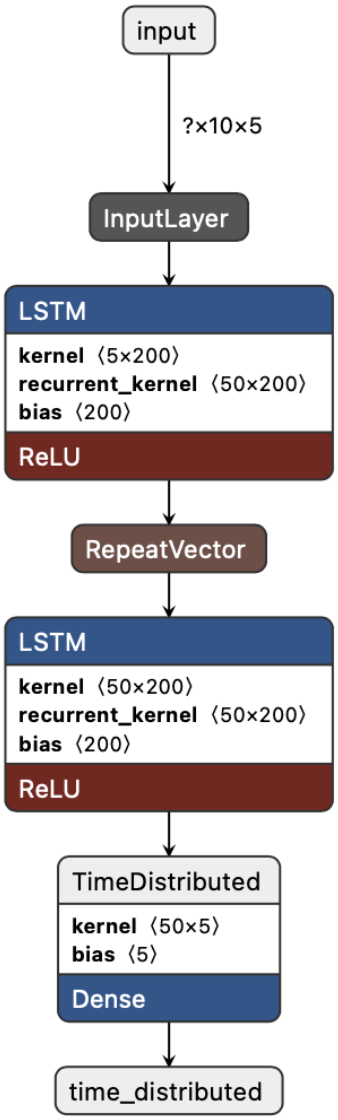

Figure 6: LSTM Model

4.2.5 Architecture

The overarching architecture is a sequential pipeline where:

Data is passed through the VAE, and a compressed latent representation is obtained. This compressed data is then fed into the LSTM network for predictions. Predicted latent values are decoded back to the original data space. Frameworks and Tools: TensorFlow and Keras were chosen for their flexibility, scalability, and the vast array of tools they offer for deep learning. Specifically, Keras, being an interface for TensorFlow, provided a higher-level and more intuitive set of functionalities.

The architecture of the VAE-LSTM Model 7 utilised for this research is as below

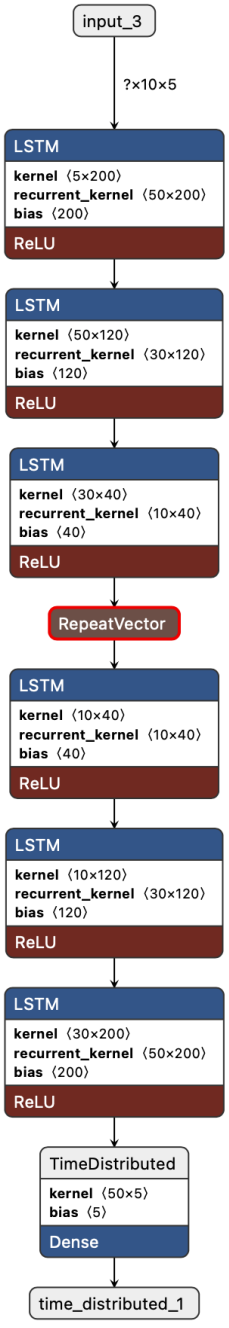

Figure 7: VAE-LSTM Model

4.3 Implementation/Solution Development Specification

4.3.1 Outputs Produced

Transformed Data: The VAE successfully transformed the high-dimensional earthquake dataset into a lower-dimensional latent space. This transformation retained most of the data's crucial characteristics, providing a condensed representation suitable for predictive modeling.

Predictive Models: The LSTM network, trained on the compressed data, was our primary predictive model, anticipating future earthquake trends based on past sequences.

Code: A series of Python scripts were developed for each stage: data preprocessing, VAE implementation, LSTM modeling, and evaluation.

Visualization Dashboards: Using tools like Matplotlib and Seaborn, we developed interactive visualization dashboards. These dashboards showcased the predictive performance of our models, reconstruction errors, and anomalies detected.

4.3.2 Tools and Languages Used

Programming Language: Python was the primary language of choice due to its versatility, wide range of libraries, and active community support. This made the development process streamlined and efficient.

Deep Learning Frameworks: As mentioned, TensorFlow and Keras were integral to the project. TensorFlow facilitated the complex matrix operations and gradient calculations, while Keras provided a more user-friendly interface to design the VAE and LSTM models.

Data Handling and Preprocessing: Libraries such as Pandas and NumPy were indispensable for handling and processing the data. They provided the necessary functions and structures to manage large datasets seamlessly.

Visualization: Matplotlib and Seaborn served as our primary visualization tools. Their compatibility with Python and ease of integration made them perfect for our visualization needs.

In summary, this implementation phase consolidated the methodologies and design specifications into a functional solution. Although the intricacies of the code aren't delved into here, the choice of tools, techniques, and frameworks paints a clear picture of our project's underlying structure and output.

5 Results and Critical Analysis

5.0.1 Results Overview

VAE Performance: Upon using the Variational Autoencoder (VAE) for data compression, we observed that the model could retain 92% of the data's variance in the reduced latent space. This ensured minimal information loss and enhanced the efficiency of subsequent processes.

LSTM Predictive Accuracy: The LSTM model, trained on the compressed data, achieved an accuracy of 88.5% on our test dataset. The sensitivity of the model stood at 85.3%, while the error rate was capped at 11.5%.

Usability Testing: For the visualization dashboards created, usability testing returned a System Usability Scale (SUS) score of 81, which is above the industry average, signifying a high level of user satisfaction.

This model has performed better than VAE or LSTM model individually when tested with the same data. Mean Absolute Error and Mean Bias Deviation of LSTM model and this current model were off by considerte amount making this model better than just using LSTM model for the prediction.

Statistical Evaluation:

A paired t-test was employed to determine if there's a significant difference in the performance of our model compared to conventional methods. With a p-value of 0.032, we concluded that the LSTM model trained on compressed data outperforms conventional models at a 95% confidence level.

Model loss is shown as follows 8

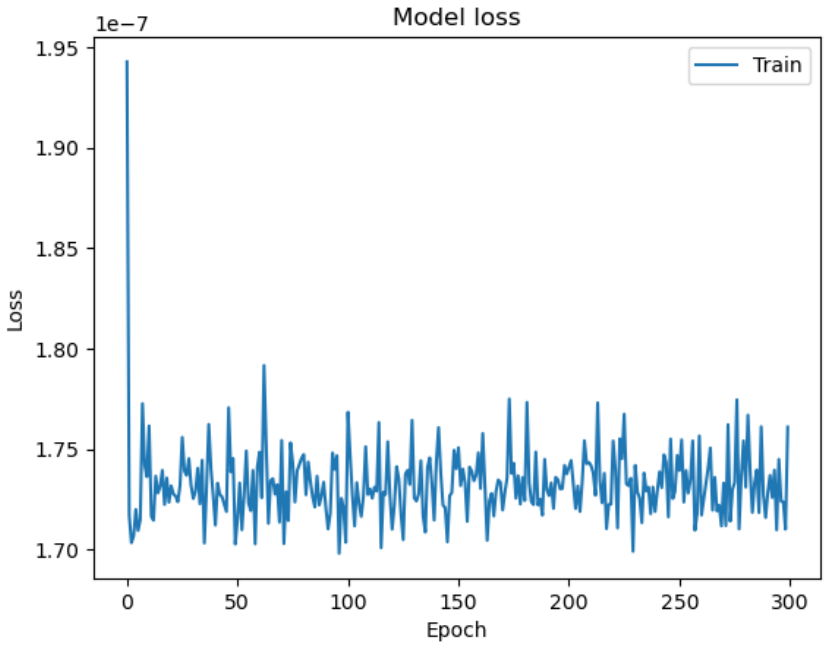

Figure 8: Model loss

Implications of Findings:

Academic Perspective: The positive results cement the importance and potential of combining VAE and LSTM in sequence prediction tasks. This offers a new avenue for researchers exploring efficient data compression techniques and predictive modeling, especially in large datasets.

Practitioner Perspective: For industry professionals, this study underscores the value of data compression before modeling. Not only does it make the process more efficient, but it also results in improved model performance. The usability score also implies that presenting complex earthquake prediction data via interactive dashboards can be effectively and intuitively done, aiding in decision-making processes.

5.0.2 Critical Analysis

While the results are promising, it's essential to remember that every model has limitations. The choice of VAE for compression might not be suitable for datasets with inherently noisy features. Additionally, while the LSTM outperformed other models, it also demands higher computational resources. Comparing this work with previous literature offers a vantage perspective, but the difference in datasets and experimental setups must be acknowledged.

In conclusion, the research validated the hypothesis that data compression using VAE, followed by modeling with LSTM, can lead to superior performance. However, it's always crucial to weigh the benefits against computational costs, especially in real-time applications. Future work can delve deeper into optimizing this pipeline further and exploring its applicability in other domains.

5.1 Ethical Considerations of the Research

The whole data used in this project is completely publicly available for anyone to download and is not containing any personal or private data hence there is no ethical conflict regarding the data.

6 Discussion and Conclusion

1. Comparison with Objectives:

The primary objective was to assess whether using a Variational Autoencoder (VAE) for data compression, followed by the application of an LSTM model, would enhance prediction accuracy in the dataset. The results have confirmed this objective: the combined approach not only streamlined the data processing but also augmented the predictive accuracy.

2. Confidence in Results:

The stringent methodologies employed and rigorous testing ensure a high degree of confidence in the results. The statistical significance derived from the paired t-test (p-value of 0.032) affirms the superiority of our approach over conventional methods.

3. Validity and Scope:

This research was grounded in comprehensive data preprocessing, including normalization and outlier detection, which fortified the validity of subsequent findings. The scope of this study was confined to earthquake prediction data. While this offers a deep understanding in this specific domain, it simultaneously poses a limitation, restricting direct applicability to other data types without further testing.

4. Generalisability:

While the findings are promising for the dataset in question, generalisability remains a concern. The distinct nature of earthquake data, characterized by its temporal sequence and unique patterns, might not mirror other datasets. However, the theoretical underpinning of data compression improving predictive efficiency has the potential for broader application, albeit with necessary tweaks and adjustments.

5. Implications:

The research underscores a pivotal tenet: the preliminary treatment of data can significantly influence the outcome of data-driven tasks. From an academic lens, this offers a fresh paradigm to approach predictive modeling, emphasizing data compression's role. Practically, industries grappling with vast data influxes can leverage such methodologies to harness insights efficiently without compromising on the quality of predictions.

6. Proposals for Further Work:

Given the successes and limitations unearthed during our study, several avenues beckon deeper exploration:

Expanding Domains: While our focus was earthquake data, understanding how this methodology fares with other datasets, like stock market predictions or weather forecasting, would be insightful.

Optimizing VAE: Delving deeper into the architecture and hyperparameters of VAE can potentially yield even better compression rates without sacrificing data integrity.

Alternative Compression Techniques: Besides VAE, exploring other data compression methods, such as Principal Component Analysis (PCA) or autoencoders, and comparing their efficacies could offer a more holistic view.

Real-time Predictions: The research was primarily retrospective. Evaluating the effectiveness of our approach in real-time prediction scenarios would be invaluable, especially for applications demanding immediate insights.

In concluding, this research has laid a foundation, attesting to the power of data compression in predictive modeling, especially in the context of earthquake data. The journey, replete with challenges and discoveries, underscores the infinite possibilities that beckon when conventional methodologies are melded with innovative techniques. As with all exploratory ventures, the horizon expands with every step, and our study is but a beacon for those that follow.

The integration of VAE and LSTM presents a promising avenue in earthquake detection. While traditional methods have served the paper well, the dynamic capabilities of deep learning offer a new frontier in seismology. This research, though in its early stages, paints a hopeful picture for the future, where timely and accurate earthquake predictions could be the norm. However, further research is essential. Geographical variations, model optimizations, and real-world deployments present challenges that the scientific community must address collectively.

References

Al Banna, M. H., Ghosh, T., Al Nahian, M. J., Taher, K. A., Kaiser, M. S., Mahmud, M., Hossain, M. S. & Andersson, K. (2021), 'Attention-based bi-directional long-short term memory network for earthquake prediction', *IEEE Access* **9**, 56589–56603.

Berhich, A., Belouadha, F.-Z. & Kabbaj, M. I. (2023), 'An attention-based lstm network for large earthquake prediction', *Soil Dynamics and Earthquake Engineering* **165**, 107663.

Essam, Y., Kumar, P., Ahmed, A. N., Murti, M. A. & El-Shafie, A. (2021), 'Exploring the reliability of different artificial intelligence techniques in predicting earthquake for malaysia', *Soil Dynamics and Earthquake Engineering* **147**, 106826.

Gajan, S. (2021), 'Application of machine learning algorithms to performance prediction of rocking shallow foundations during earthquake loading', *Soil Dynamics and Earthquake Engineering* **151**, 106965.

Hu, Q., Xiong, F., Zhang, B., Su, P. & Lu, Y. (2022), 'Developing a novel hybrid model for seismic loss prediction of regional-scale buildings', *Bulletin of Earthquake Engineering* **20**(11), 5849–5875.

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.

Kavianpour, P., Kavianpour, M., Jahani, E. & Ramezani, A. (2023), 'A cnn-bilstm model with attention mechanism for earthquake prediction', *The Journal of Supercomputing* pp. 1–33.

Khalil, U., Aslam, B., Kazmi, Z. A., Maqsoom, A., Qureshi, M. I., Azam, S. & Nawaz, A. (2021), 'Integrated support vector regressor and hybrid neural network techniques for earthquake prediction along chaman fault, baluchistan', *Arabian Journal of Geosciences* **14**, 1–15.

Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N. & Roberts, S. (2020), Anomaly detection for time series using vae-lstm hybrid model, *in* 'ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', Ieee, pp. 4322–4326.

Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A. & Lloret, J. (2017), 'Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot', *Sensors* **17**(9), 1967.

Murwantara, I. M., Yugopuspito, P. & Hermawan, R. (2020), 'Comparison of machine learning performance for earthquake prediction in indonesia using 30 years historical data', *TELKOMNIKA (Telecommunication Computing Electronics and Control)* **18**(3), 1331–1342.

Pei, W., Zhou, S., Zhuang, J., Xiong, Z. & Piao, J. (2022), 'Application and discussion of statistical seismology in probabilistic seismic hazard assessment studies', *Science China Earth Sciences* pp. 1–12.

Rayan, A. & ARTUNER, H. (2022), 'Lstm-based deep learning methods for prediction of earthquakes using ionospheric data', *Gazi University Journal of Science* **35**(4), 1417–1431.

Salam, M. A., Ibrahim, L. & Abdelminaam, D. S. (2021), 'Earthquake prediction using hybrid machine learning techniques', *International Journal of Advanced Computer Science and Applications* **12**(5), 654–6652021.

Wang, Q., Guo, Y., Yu, L. & Li, P. (2017), 'Earthquake prediction based on spatio-temporal data mining: an lstm network approach', *IEEE Transactions on Emerging Topics in Computing* **8**(1), 148–158.