Hybrid CatBoost and SVR Model for Earthquake Prediction Using the LANL Earthquake Dataset

Arush Kaushal[*1], Ashok Kumar Gupta[2], Vivek Kumar Sehgal[1]
[1] Department of Computer Science and Information Technology, Jaypee University of Information Technology, Solan 171234, India
[2] Department of Civil Engineering, Jaypee University of Information Technology, Solan 171234, India
E-mail: arushkaushal0115@gmail.com, akgjuit@gmail.com, vivekseh@ieee.org
*Corresponding Author

Keywords: seismology, earthquake prediction, time to failure, hybrid model

Received:

Earthquakes have the potential to cause catastrophic structural and economic damage. This research explores the application of machine learning for earthquake prediction using LANL (Los Alamos National Laboratory) dataset. The data, obtained from a laboratory stick-slip friction experiment, simulate real earthquakes through digitized acoustic signals recorded against the time to failure of a granular layer. We introduced a hybrid model combining CatBoost and Support Vector Regression (SVR) to predict the time of the next earthquake, evaluating its performance against individual CatBoost and SVR models. The hybrid model demonstrated superior accuracy with a Mean Absolute Error (MAE) of 0.0825, outperforming the individual models. We implemented feature engineering to optimize the predictive capability of the models. Additionally, we compared our hybrid model's performance with previous studies to validate its efficacy. Our findings underscore the potential of machine learning, particularly hybrid models, in enhancing earthquake prediction accuracy. This study highlights the robustness and effectiveness of the hybrid CatBoost-SVR model, paving the way for advanced AI algorithms in seismology and contributing to improved disaster preparedness and mitigation strategies.

Povzetek: A hybrid CatBoost-SVR model improves earthquake prediction using the LANL dataset, achieving superior accuracy (MAE: 0.0825). This approach enhances machine learning applications in seismology, contributing to disaster preparedness and mitigation strategies.

1 Introduction

Earthquakes stand as one of nature's most devastating phenomena, posing formidable challenges for prediction despite the extensive efforts of the seismology community. Unlike other natural disasters such as floods, tornadoes, and hurricanes, which can often be forecasted in terms of timing, location, and potential impact, earthquake prediction remains notably elusive. Currently, seismographs serve as the primary method for detecting imminent earthquakes, yet their warnings typically offer only seconds of lead time, insufficient for effective preventive action against substantial structural damage. The complexity and nonlinear characteristics of seismic data further compound the difficulty in earthquake prediction, presenting a persistent challenge in geophysics. Recent strides in machine learning present promising avenues for improving prediction accuracy in earthquake forecasting. This study delves into a hybrid model that merges CatBoost and Support Vector Regression (SVR) to enhance earthquake prediction performance, leveraging insights gained from analyzing the LANL earthquake dataset. Additionally, alternative approaches to earthquake prediction involve monitoring changes in land elevation, groundwater levels, animal behavior, and precursor seismic activity. A notable instance of effective earthquake prediction transpired during the Haicheng, China earthquake of 1975, where an evacuation advisory was disseminated a day prior to the occurrence of a magnitude 7.3 seismic event. In the month’s antecedent to the earthquake, alterations in land surface elevation and groundwater levels, numerous instances of anomalous animal behavior, and the occurrence of several foreshocks collectively served as precursory indicators, initially prompting a precautionary advisory. Subsequently, a surge in foreshock activity prompted the escalation of the advisory to an evacuation warning. Nevertheless, it is imperative to note that the majority of earthquakes do not manifest such conspicuous precursory signs. Despite the success witnessed in 1975, the 1976 Tangshan earthquake, registering a magnitude of 7.6, occurred without any forewarning, resulting in an estimated 250,000 casualties [1]. Amidst the rapid evolution of statistical and deep learning methodologies, novel paradigms in earthquake prediction have emerged [2] [3]. These strategies hinge upon extensive datasets, accentuating the imperative of curating, amassing, and simulating earthquake data, a realm that has recently garnered notable scrutiny [4]. Through the fusion of meticulously curated data and cutting-edge statistical and deep learning methodologies, the endeavor to forecast earthquake timing based on realistically attainable data could potentially be surmounted, aligning with the prevailing trajectory across diverse machine learning applications. In both [2] and [3], machine learning and deep learning frameworks are harnessed to prognosticate the timing of subsequent earthquakes. These frameworks leverage physically amassed and labeled earthquake parameters, such as relative strength index, momentum, and moving force averages. Classic machine learning (ML) [4] algorithms conventionally compute seismic metrics like Gutenberg-Richter b-values, time intervals, earthquake energy, and mean magnitude. In contrast, contemporary deep learning (DL) models [5] exhibit proficiency in assimilating multifaceted features. Both ML and DL models are driven by data and demonstrate efficacy in moderate-magnitude earthquake scenarios; however, they encounter challenges with high-magnitude events due to the scarcity of requisite data. Data-driven models necessitate voluminous datasets to yield precise predictions. Certain DL methodologies endeavor to anticipate significant earthquakes by exclusive training on such instances, yet these methods necessitate further refinement [6]. A prevalent trait among these methodologies is their analysis of protracted temporal sequences of seismic data, which poses a formidable hurdle for DL techniques. Accurate earthquake prediction holds the potential to avert fatalities and mitigate catastrophic repercussions, thus positioning the anticipation of earthquake timing and magnitude as a cornerstone objective within the domain of geoscience [7]. Despite protracted time-series observations and field studies, the precise anticipation of earthquake scale or timing persists as an enduring challenge [8]. Moreover, the unpredictability of devastating subduction earthquakes, with magnitudes of 9.0 or higher, adds a concerning dimension to this endeavor [9]. Traditional methods of earthquake prediction, such as using seismographs, often provide only seconds of warning before an earthquake occurs, which is insufficient time to take preventative measures. Other approaches involve monitoring changes in land elevation, groundwater levels, animal behavior, and foreshocks. However, these methods do not always provide clear or reliable precursors to impending earthquakes. Monitoring with non-destructive testing (NDT) acoustic emissions (AE) involves the continuous recording of acoustic data as the material undergoes stress. The recording process typically persists until the material reaches failure. In controlled laboratory environments, stress-induced failure can be hastened by artificially subjecting the material to stress [10][11]. Upon failure, discrete acoustic emissions (AEs) are discerned within the recorded data. These discrete AEs denote short-duration elastic waves generated due to the initiation of minute internal cracks and slip occurrences along grain contacts, thereby furnishing valuable insights into the material's response under stress. Subsequently, AEs can be categorized based on the damage mechanism through the utilization of unsupervised clustering algorithms. In certain instances, the precise labels for each cluster are determined through methodologies such as transmitted light analysis [12] or scanning electron microscopy [13]. Finally, scrutinizing AE production across the failure cycle facilitates the identification of temporal patterns and enables deductions regarding the material's remaining useful life (RUL). Some research endeavors have expanded upon this analysis by integrating machine learning methodologies to forecast RUL, yielding varying degrees of efficacy [14][15].

Recent advancements in machine learning (ML) algorithms and computational hardware have catalyzed novel insights and methodologies within the seismological community [16]. ML applications now extend to fundamental signal processing tasks, encompassing earthquake event detection [17], phase picking [18], association [19], and hypocenter determination [20], as comprehensively documented by [21]. Concurrently, the utilization of data-driven ML approaches has broadened to encompass the prediction of TTF in laboratory experiments, leveraging Acoustic Emission (AE) data and its associated measurements [22]. A study on earthquake forecasting emphasizes the importance of long-term predictions regarding the timing, intensity, and location of future seismic events. By leveraging expert systems and extensive data analysis, more accurate forecasting models can be developed for specific regions, such as Los Angeles, improving preparedness and risk management [23]. Advanced machine learning techniques, such as attention-based Bi-Directional Long Short-Term Memory (LSTM) networks, have been highlighted as powerful tools for enhancing the precision of earthquake predictions, which are critical for disaster response in earthquake-prone areas [24]. Additionally, extreme value theory has been applied to assess the maximum possible earthquake magnitudes in high-risk areas, underscoring the value of ground-based observations and statistical methods in refining forecasting models [25].

Machine learning techniques have also been used to cluster earthquakes based on historical intervals, offering insights into recurring seismic behaviors and improving the predictive power of long-term forecasts [26]. For regions with complex fault zones, statistical models like the SARIMA model can help forecast earthquakes by analyzing past seismic events, contributing to more reliable predictions and risk management strategies [27]. Other research has focused on analyzing geoelectric field signals before earthquakes using advanced techniques, which can provide early warning signs and valuable data to improve forecasting accuracy [28].

Another significant area of study is the monitoring of slow seismic activity, which may indicate the potential for a major earthquake. By identifying these patterns, researchers can enhance the effectiveness of forecasting models [29]. Additionally, understanding the relationship between ground motion attenuation and regional geophysical data is crucial for developing robust forecasting models that predict the impact of seismic events [30]. Lastly, the study of seismic stress levels and their relationship with earthquake magnitude helps improve predictions of high-magnitude earthquakes, providing deeper insights into seismic behavior and further refining forecasting methods [31]. Collectively, these approaches are advancing earthquake forecasting and risk management, enhancing preparedness in earthquake-prone regions.

When working with the LANL earthquake dataset, it is essential to recognize several limitations and potential biases that influence the generalizability and reliability of the findings. The dataset is geographically biased, focusing on specific regions, which limits the applicability of conclusions to areas with different seismic characteristics. If the data predominantly covers certain tectonic plate boundaries or fault lines, it does not fully represent the behavior of earthquakes in less seismically active regions. Additionally, the dataset has temporal gaps, with uneven data distribution over time, which affects trends analysis and the ability to draw consistent conclusions across different periods. The dataset also has biases in the types of seismic events included, such as overrepresentation of certain magnitudes or depths, which skews model development. If smaller or larger earthquakes are underrepresented, the results do not accurately reflect the full range of seismic activity. Furthermore, the quality of the data is important, as seismic recordings are affected by noise from environmental factors, sensor inaccuracies, or technological limitations. If the dataset includes noisy or incomplete data, it compromises the ability to detect meaningful patterns or leads to incorrect conclusions. Incomplete or missing data points, especially if they are not randomly distributed, further introduce biases. There are also issues with manual labeling and classification errors, where misclassification of events distorts the analysis, particularly if smaller seismic events are confused with more significant ones. Finally, the sampling frequency of the dataset impacts its usefulness, as insufficient resolution results in the loss of critical information, such as early warning signs of large earthquakes or aftershocks. Acknowledging these limitations and biases is crucial for a more realistic and transparent understanding of the dataset's applicability to earthquake prediction and modeling.

In the context of predicting Time to Failure (TTF) within controlled laboratory environments, researchers typically employ Machine Learning (ML) frameworks that rely on three distinct feature categories. These categories encompass a) AE-Driven Features, which are directly derived from continuous Acoustic Emission (AE) signals, capturing nuanced details about the material's structural response and behavior; b) Geodetic-Driven Features, extracted from geodetic measurements, offering insights into the material's deformation characteristics and spatial dynamics, thus shedding light on its mechanical integrity; and c) Catalog-Driven Features, sourced from earthquake or seismicity catalogs, providing historical data on seismic events and their associated attributes. These feature categories collectively enable a comprehensive approach to TTF prediction, integrating diverse data sources to enhance predictive accuracy and reliability within laboratory settings. Acoustic emissions (AE) denote transient elastic waves arising from the formation of minute internal cracks and slip events along grain contacts within stress-stricken materials. AE monitoring offers invaluable insights into material structural integrity and response mechanisms under stress, thus laying the foundation for TTF prediction in laboratory setups. The amalgamation of AE data with machine learning methodologies presents a promising avenue for enhancing the precision and efficacy of TTF prognostications, thereby fostering advancements in comprehending material behavior under stress and augmenting predictive capacities within the realm of seismology. Despite the limitations and biases present in the LANL earthquake dataset, our "Hybrid CatBoost and SVR Model" helps provide better results by effectively addressing these challenges. The CatBoost algorithm, known for its robustness in handling categorical features and its ability to deal with noisy and incomplete data, enhances the model's ability to identify important patterns in seismic events. By reducing overfitting and improving generalization, CatBoost ensures that the model remains accurate even in the presence of biases like geographical or temporal imbalances. On the other hand, the Support Vector Regression (SVR) component helps capture complex relationships in the data, especially for modeling non-linearities that might arise due to varying earthquake magnitudes and depths. Together, the hybrid model leverages the strengths of both algorithms, enabling it to mitigate the impact of incomplete or noisy data, and ultimately providing more reliable predictions. Additionally, the combination of CatBoost's feature engineering capabilities and SVR's precision ensures that even with a limited dataset, the model can deliver meaningful insights, improving the overall accuracy and robustness of earthquake predictions.

We propose a novel hybrid approach that combines CatBoost, a gradient boosting algorithm, with Support Vector Regression (SVR). This hybrid model leverages the strengths of both methods to improve the accuracy of predicting the time-to-failure of earthquakes using the LANL earthquake dataset. Integrating CatBoost with Support Vector Regression (SVR) can yield superior results due to the complementary strengths of the two algorithms. CatBoost, a gradient boosting algorithm, is adept at handling categorical features and automatically managing missing data. It excels in capturing complex relationships within the dataset, producing robust predictions. On the other hand, SVR, a kernel-based regression algorithm, is proficient in modeling nonlinear relationships and high-dimensional spaces. By combining the predictions from CatBoost with the features in SVR, the integrated model can leverage the advantages of both algorithms. CatBoost provides an initial understanding of the data's complex patterns, while SVR further refines predictions based on its ability to capture intricate relationships.

Our paper builds significantly on the work presented in [32], where researchers from the Los Alamos National Laboratory (LANL) developed a dataset of acoustic data from laboratory-simulated earthquakes. This dataset was utilized to train a support vector regression (SVR) machine learning model for predicting time-to-failure, defined as the time until a major earthquake event. The model used statistical features such as moving average, kurtosis, and variance. In this study, we aim to enhance their results by Catboost techniques. Some major contributions of research includes:

This research bridges the gap between machine learning and seismology, demonstrating how advanced data-driven approaches can be applied to traditional scientific problems. The study also incorporates the consideration of slow slip events (SSEs) and their relationship with regular earthquakes, adding to the understanding of seismic processes.

2 Dataset description

In 2017, researchers at Los Alamos National Laboratory (LANL) achieved a significant breakthrough in the prediction of Slow Slip Earthquakes (SSE) within laboratory conditions that mimic natural settings. Through meticulous experimentation, the team developed a method wherein a computer system was trained to detect and analyze quasi-periodic seismic and acoustic signals emitted during fault movements. By processing extensive datasets, they identified a distinct sound pattern, previously dismissed as noise, which serves as an indicator of an impending earthquake. Utilizing a time window of 1.8 seconds of data, the team attained an impressive 89% coefficient of determination in forecasting the time remaining before a laboratory earthquake event, employing Random Forest Regression and quasi-periodic data. In the laboratory environment, seismic sounds produced by the interaction of steel blocks with rocky material, simulating real earthquake activity, were recorded by an accelerometer. This groundbreaking discovery represents the first successful prediction of laboratory earthquake occurrences. While acknowledging the differences in shear stress between laboratory experiments and natural earthquakes, the LANL team is actively engaged in validating their findings in real-world scenarios [33][34]. Moreover, this innovative approach holds potential beyond seismology, with possible applications in material failure research across diverse industries like aerospace and energy. These findings underscore the notion that fault failure follows a discernible pattern rather than occurring randomly.

3 Data exploration

The LANL earthquake dataset serves as a comprehensive repository of acoustic emission signals captured during laboratory-simulated earthquakes. Each entry within this dataset encapsulates the acoustic data recorded at distinct time intervals, providing a detailed snapshot of seismic activity. Crucially, each sample is paired with a target value denoting the time until the occurrence of the subsequent laboratory earthquake. This temporal information enables researchers to study the dynamics of earthquake events and explore predictive modeling approaches [35]. The acoustic data itself is composed of discrete segments, each spanning a duration of 0.0375 seconds, comprising seismic signals recorded at a frequency of 4MHz. This results in a substantial dataset containing a total of 150,000 data points. Each segment of acoustic data is meticulously annotated with a corresponding "time to failure" value as shown in Table 1, representing the duration until the laboratory fault undergoes failure, as determined through stress measurements. The acoustic signal consistently exhibits significant fluctuations immediately preceding each failure event Figure 1. Additionally, it is noteworthy that failures can be visually anticipated by observing instances where substantial fluctuations in the signal are succeeded by smaller ones.

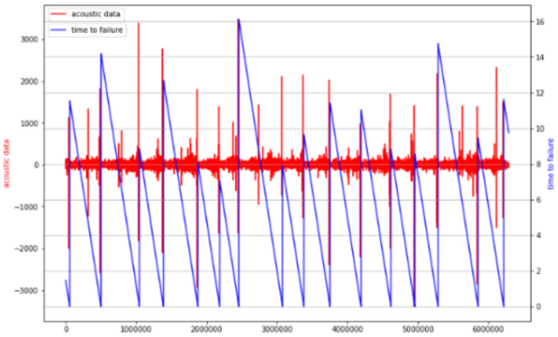

Figure 1: Acoustic data and time to failure analysis: subset representing 1% of total dataset.

Upon closer examination of a zoomed-in time plot Figure 2, it becomes apparent that the prominent acoustic signal oscillation occurring at the 1.572-second mark precedes the occurrence of the failure event, albeit not precisely coinciding with it. Before this major oscillation, there are noticeable sequences of intense signal fluctuations, suggesting a buildup of activity leading to the larger event. Subsequently, after the significant oscillation, there are also smaller oscillations observed, indicating a potential aftermath or continuation of the event's effects [36][37]. In this time plot, it becomes evident that the significant oscillation preceding the failure does not occur immediately before the event.

Table 1: Dataset: Seismic Activity (v) and Time to Failure (s)

Sesmic activity ($\boldsymbol{v}$)	Time to failure (s)
12	1.4690999832
6	1.4690999821
8	1.469099981
5	1.4690999799
8	1.469099988
8	1.469099977
9	1.4690999766
7	1.4690999755
-5	1.4690999744
...	...

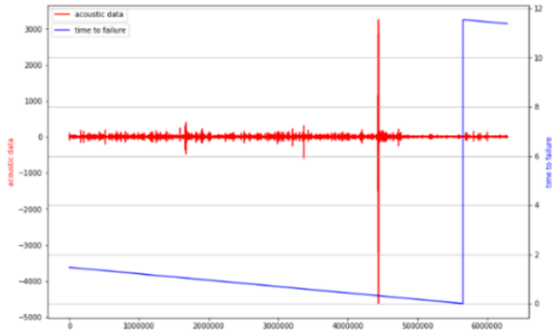

Figure 2: Zoomed-in-time-plot.

Instead, there are sequences of intense oscillations that precede the large oscillation, as well as smaller peak oscillations that follow it. Subsequently, after a series of minor oscillations, the failure takes place. Initially structured as a Pandas Dataframe, the dataset underwent a process of segmentation, dividing it into 150,000 individual samples. Each sample is coupled with its corresponding time to failure, facilitating the training and validation of predictive models. Moreover, the dataset includes an additional 2626 preconstructed acoustic segments earmarked specifically for model testing purposes. This meticulous organization of the dataset enables researchers to conduct robust evaluations of model performance and effectiveness in earthquake prediction tasks. Seismic signals are captured through a piezoceramic sensor that generates a voltage in response to deformation caused by incoming seismic waves. This voltage, referred to as the acoustic signal, serves as the primary input for our analysis. The acoustic signal represents the recorded voltage, expressed as integers.

The Acoustic Signal essentially signifies the voltage generated by the deformation induced by seismic waves. These signals are integer values ranging from -5515 to 5444, with an average of 4.52. Examining the distribution of these acoustic signals reveals a distinct peak, indicating a concentration of values around the mean. However, the distribution also exhibits outliers in both directions, suggesting sporadic occurrences of exceptionally high or low values. This observation is illustrated in Figure 3, where the distribution's shape and the presence of outliers can be visualized. The range of the acoustic signals, spanning from -5515 to 5444, reflects the entirety of recorded voltage variations, from the most negative to the most positive values. This comprehensive range offers insights into the full spectrum of voltage fluctuations experienced during seismic activity. Negative values might signify voltage decreases due to compression or damping effects, while positive values could indicate voltage increases resulting from tension or amplification. The wide span of this range underscores the substantial variability in recorded voltage, influenced by factors like seismic wave intensity, distance, environmental conditions, and sensor sensitivity [38]. Despite the range's breadth, a very high peak in the distribution suggests a clustering of values around a central tendency, indicative of predominant signal strength or intensity. However, the presence of outliers in both directions highlights occasional deviations from this central tendency, likely stemming from anomalies in seismic activity or sensor behavior. These outliers necessitate careful consideration during data analysis to ensure accurate interpretation and modeling of the seismic signals.

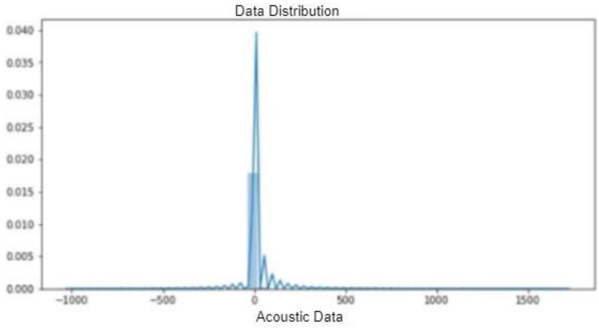

Figure 3: The distribution of acoustic signals analyzed individually.

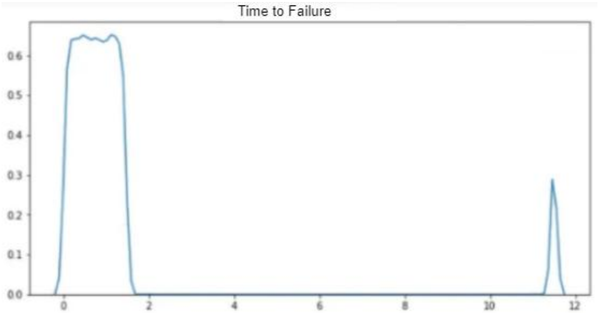

Figure 4: The distribution of time to failure analyzed individually.

The time to failure represents the duration, in seconds, remaining until an imminent stick-slip failure event occurs. This metric serves as a crucial indicator of the proximity of failure, allowing for proactive measures to be taken. The minimum value of Time to Failure is extremely close to zero, at 9.55039650e-05 seconds, indicating instances where failure occurred imminently after observation. Conversely, the maximum Time to Failure extends to 16 seconds, representing cases where failure was predicted further in advance. The distribution of Time to Failure exhibits a right-skewed pattern, as illustrated in Figure 4. This skewness indicates that the majority of observations are clustered towards the lower end of the

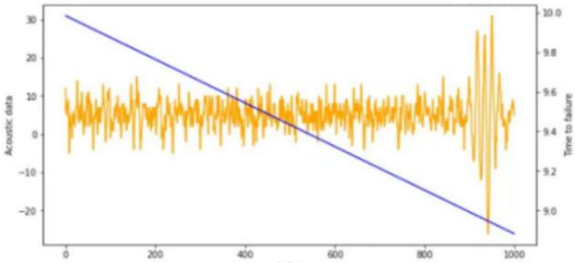

Figure 5: Time series relationship between first 1000 rows.

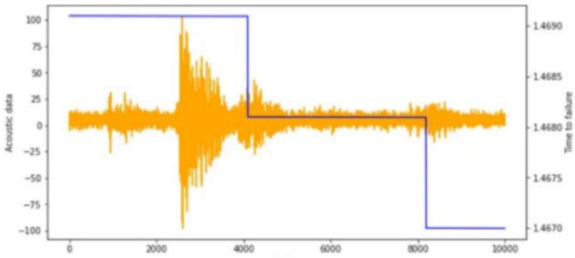

Figure 6: Time series relationship between first 10,000 rows.

time scale, with fewer instances of longer Time to Failure values. This distribution pattern provides valuable insights into the temporal dynamics of stick-slip failure events, highlighting the variability in the timeframes leading up to failure occurrences. The explanation details a time-series plot analyzing the first 1000 rows of data, with the orange lines depicting seismic activity (acoustic feature) and the blue line representing time to failure, indicating the duration until the next earthquake. Notably, the plot reveals a linear trend in the time to failure, suggesting a consistent change over time, implying a potential predictive relationship with the acoustic feature.

Figure 5 centres on analyzing time-based data, underscoring the importance of examining both the distribution of acoustic signals and the target feature over time. Two functions are provided to facilitate visualization of these features. First function generates a plot showcasing the acoustic data and time to failure for a specified range of indices, while the other function allows for comparison across two distinct index ranges.

In the example provided, the first function is employed to plot the first thousand rows of the dataset, with orange lines representing the acoustic feature and a blue line depicting the target feature. The resulting plot illustrates a linear relationship in the target feature, prompting further exploration to gain a comprehensive understanding of the dataset's behavior across a broader range of rows.

After examining the initial 1000 rows, further analysis is conducted on larger subsets of the data, including the first 10,000 rows shown in Figure 6 and the entire dataset comprising 600,000 rows shown in Figure 7. These analyses reveal consistent trends, with the time to failure decreasing sharply to nearly zero seconds when an earthquake event occurs, indicating a rapid onset of seismic activity. The observations underscore the predictive potential of the acoustic data in forecasting earthquake occurrences and highlight the significance of ongoing analysis to refine predictive models and enhance accuracy. After generating the time-series plots, we analyze them to extract meaningful insights about the behavior of the data over time. This analysis involves identifying recurring patterns, detecting abrupt changes or anomalies, and assessing the overall trend in the data series. By interpreting the time-series plots, we can gain a deeper understanding of the underlying dynamics driving seismic activity and the predictive relationship between the acoustic data and time to failure. In summary, time-series analysis plays a crucial role in uncovering temporal patterns and relationships within the data, providing valuable insights that inform subsequent modeling and prediction efforts in the context of earthquake forecasting. In our analysis, we examined a dataset containing a massive 629 million rows, although our focus was on a subset of 600,000 rows. We were particularly interested in understanding the timing of events, noting that the time-to-failure spanned from nearly zero to 12 seconds.

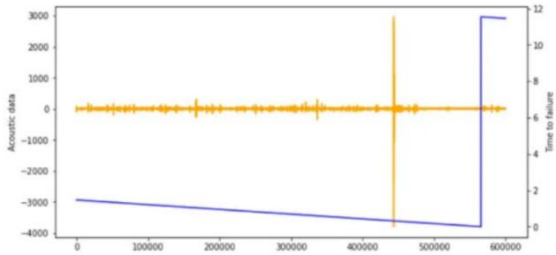

Figure 7: Time series relationship between first 600k rows.

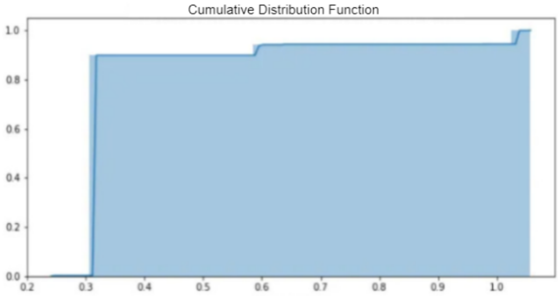

Figure 8: Cumulative distribution of the time to failure with high signal.

To delve deeper into this aspect, we decided to investigate the Cumulative Distribution Function (CDF) of the target feature, which helped us understand how frequently events occurred within the 0 to 12-second range. After setting the display precision and loading the dataset, we visualized the distribution of acoustic data. Upon examining the CDF plot shown in Figure 8 of the target feature, we discovered that approximately 85% of the events occurred within a mere 0.3 seconds, indicating a rapid onset of events. This observation shed light on the timing patterns within the dataset and emphasized the importance of events occurring within close proximity to zero seconds.

3.1 Feature engineering

Data preprocessing is an essential preliminary step in harnessing the LANL earthquake dataset for model training and assessment. This section delineates a series of preprocessing procedures orchestrated to refine the dataset, ensuring its cleanliness, informativeness, and readiness for subsequent analyses. The journey begins with the ingestion of the LANL earthquake dataset, an amalgamation of acoustic signal data accrued during laboratory-simulated earthquakes. Within this dataset lie acoustic emission signals, captured at varied time intervals, accompanied by corresponding time-to-failure values delineating the duration until the advent of subsequent seismic events. Subsequently, meticulous data cleaning protocols are executed to rectify any aberrations present within the dataset. Through adept imputation techniques, missing values are diligently addressed, ensuring comprehensive data coverage. Concurrently, outliers, with their potential to skew model training outcomes, are meticulously identified and rectified through judicious methods. Following data cleansing, the dataset undergoes a transformative phase through the application of feature engineering techniques.

The data cleaning and preparation process for the LANL earthquake dataset involved several key steps to ensure the quality and consistency of the input data for the hybrid model. First, missing or incomplete data points were identified and addressed through appropriate imputation techniques or, in some cases, by removing records with excessive missing values to avoid introducing bias. Next, outliers were detected and handled to prevent them from disproportionately influencing the model's predictions. This step was particularly important as seismic data can sometimes contain unusual readings due to sensor malfunctions or other anomalies. Data normalization and scaling were applied to ensure that features with different units and ranges did not skew the performance of the model, particularly for algorithms like Support Vector Regression (SVR), which are sensitive to the scale of the input data. Additionally, categorical variables, such as event types or geographic locations, were encoded using techniques such as one-hot encoding or label encoding to make them compatible with the CatBoost algorithm, which is capable of handling categorical data efficiently. Temporal features, such as the date and time of seismic events, were also processed to extract meaningful patterns, such as trends or seasonality, that could contribute to better model performance. Feature engineering was performed to create new variables that could enhance the model's ability to identify key seismic patterns, such as calculating the time between successive events or aggregating data at different time intervals. Through this comprehensive data cleaning and preparation process, the dataset was transformed into a structured and reliable format, enabling the hybrid model to learn effectively and provide accurate predictions. Feature engineering is a critical step in the model development process, as it involves transforming raw data into meaningful features that can enhance the predictive power of machine learning models. In this study, feature engineering was focused on extracting key characteristics from Acoustic Emission (AE) data, which is considered a rich source of information for predicting Time to Failure (TTF). The goal of feature engineering was to identify and create features that can effectively capture the underlying patterns and dynamics of the AE signals, which are indicative of the system's failure behavior. The feature engineering process began with the assumption that the distribution of AE data holds valuable information that can be leveraged to predict failure. This assumption is based on both empirical observations and established findings in the literature, which suggest that variations in AE data, particularly in the form of spikes, can precede failure events. By focusing on these variations, we aimed to identify statistical features that could serve as reliable indicators of failure time. A key insight from the data was that stick-slip failure events, often associated with mechanical systems, are typically preceded by a series of AE signal spikes. These spikes, which are indicative of micro-failures, provide crucial information that can help predict when a system is approaching failure. We hypothesized that the frequency and intensity of these AE spikes correlate with the remaining useful life of the system, and therefore, the statistical characteristics of the AE signal could serve as valuable features for modeling. Building on this foundation, we derived a set of 18 statistical features from each 150,000-point segment of the AE data. These features included basic statistical metrics such as mean, standard deviation, skewness, and kurtosis, which have been shown to reflect important characteristics of the AE signal. Additionally, we calculated features like the ratio of standard deviation to mean, as well as distributional features represented by various percentiles (e.g., 1st, 5th, 25th, 50th, etc.). These features were selected because they provide a more comprehensive representation of the AE signal's behavior over time. Not all derived features were ultimately used in the model. For example, while maximum and minimum values were initially considered, they were excluded from the final feature set due to their sensitivity to extreme events, which mainly serve as markers of significant disruptions in the AE signal rather than predictors of failure. After the features were extracted, a database was created, which contained a large set of statistical features corresponding to each segment of AE data. This database covered a wide range of TTF values, allowing us to explore how each feature correlated with the time to failure. Initial analysis showed that certain features, such as the count of mode appearances, exhibited a strong correlation with TTF. However, special care was taken to exclude data recorded immediately after major failure events, as these post-event values closely resembled early-stage data and could introduce inaccuracies into the predictive model. Herein, statistical attributes such as mean, standard deviation, skewness, and kurtosis are meticulously computed, affording insights into the distributional characteristics of the acoustic signals. The derivation of rolling window statistics facilitates the capture of temporal nuances and trends embedded within the data. Furthermore, to foster uniformity and comparability across diverse features, the dataset is subjected to normalization or standardization.

Table 2: Comprehensive global overview of the dataset statistics

	acoustic - data	**time-to-failure**
count	6.29E+08	6.291E+08
mean	4.52E+00	5.68E+00
min	-5.52E+03	9.55E-01
max	5.44E+03	1.61E+01
std	1.07E+01	3.67E+00

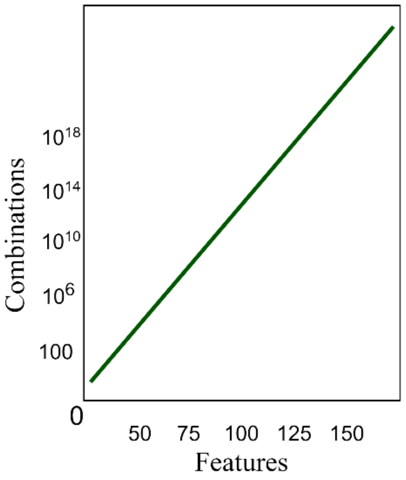

Figure 9: Total number of possible combinations compared to the number of features.

Through normalization, data is rescaled to span a range between 0 and 1, while standardization ensures a mean of 0 and a standard deviation of 1. These harmonizing techniques alleviate the impact of disparate feature scales, thereby bolstering the efficacy of subsequent model training endeavors.

In this study, we derived a comprehensive set of 25 statistical features from each of the 150,000 segments of acoustic emissions (AE) data as shown in Figure 9. These features were meticulously selected to capture various statistical properties of the data. The initial twelve features included the maximum, minimum, mean, standard deviation, the ratio of standard deviation to mean, skewness, kurtosis, mode, and the frequency of mode appearances. These features were chosen to encapsulate the central tendency, variability, and shape of the data distribution. Additionally, we calculated thirteen percentile features at specific levels: 1st, 5th, 10th, 25th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, 95th, and 99th. These percentiles were included to provide a detailed understanding of the data distribution and to capture the behavior of the acoustic signals at various thresholds. Notably, while the "maximum" and "minimum" features were computed, they were excluded from the final modeling process. This decision was made because these features, due to their extremely high values, primarily indicated the main earthquake events rather than providing predictive insight for the time to failure. By focusing on the remaining features, we aimed to enhance the model's ability to predict the time until the next earthquake based on more subtle patterns within the acoustic data. This strategic feature selection was crucial for developing a robust and accurate predictive model.

In this study, feature selection was conducted by constructing multiple models and comparing their Mean Absolute Errors (MAEs) to identify the combination of features that resulted in the lowest MAE. However, it is important to consider the curse of dimensionality, where the total number of potential feature combinations increases exponentially with the number of features in the set.

In an alternate scenario, the Los Alamos National Laboratory (LANL) achieved a coefficient of determination of 0.89 through their analysis of quasi-periodic seismic signals. Their approach involved partitioning the data into 1.8-second time windows and employing a Random Forest technique. They identified variance, kurtosis, and threshold as the most influential features within their model. Inspired by this methodology, our study concentrates on predicting the time remaining before the next failure solely based on moving time windows of acoustic data. We segmented the data into 0.3-second time windows, encompassing 1,500,000 observations, significantly shorter than the laboratory quake cycle, which spans 8 to 16 seconds. It is noteworthy that a substantial proportion of high acoustic values (exceeding an absolute value of 1000) occur approximately 0.31 seconds before an earthquake. This observation prompted us to partition the data into 0.3-second windows to minimize error towards the conclusion of the quake cycle. Evaluating the sensitivity of our findings to variations in time window sizes revealed optimal results when employing 1.5 million observations per time window, yielding a dataset composed of 419 distinct windows. Each window generated a set of 95 potential statistical features, encompassing metrics such as Standard Deviation, quantiles at 90%, 95%, and 99%, Absolute Standard Deviation, and diverse rolling standard deviation measures across varying observation intervals. Leveraging a feature importance technique, we discerned the salience of specific features within the dataset. Subsequently, advanced machine learning techniques, notably the Catboost-SVM model, were employed to analyze the continuous values derived from the acoustic time series data.To mitigate feature correlation effects, principal component analysis was applied, effectively condensing the feature space from 95 to 5 principal components, accounting for 99.9% of the total data variance. To ensure robustness and integrity, a 50/50 continuous split strategy was implemented for training and testing datasets. The regularization hyperparameters for each machine learning algorithm were meticulously tuned using a random grid search approach, validated through a 3-fold cross-validation methodology. Visualization of feature-TTF relationships, as depicted in Figure 10, unveiled significant correlations between certain features and Time to Failure (TTF).

Cross-validation methodologies, notably k-fold cross-validation, serve as a robust mechanism for scrutinizing model performance. By segmenting the training data into multiple folds, the model is iteratively trained on different fold combinations, with performance evaluations conducted across each iteration. This iterative process furnishes more dependable assessments of model efficacy.

4 Methodology

In our research endeavor focused on earthquake prediction utilizing the LANL dataset, we embark on a comprehensive methodology integrating advanced machine learning techniques to enhance forecasting accuracy. The methodology commences with an intricate phase of data preprocessing, a pivotal step ensuring the dataset's readiness for subsequent model training and evaluation. This preprocessing stage involves meticulous cleaning to address any missing values or outliers that may distort the model's learning process. Additionally, feature engineering techniques are employed to extract informative statistical features from the raw acoustic signal data, thereby enriching the dataset with valuable insights into seismic activity dynamics. Following data preprocessing, it proceeds with the training of individual predictive models, commencing with the utilization of CatBoost, a powerful gradient boosting algorithm renowned for its efficacy in handling heterogeneous data. CatBoost is adeptly trained on the preprocessed dataset to generate preliminary predictions concerning the timing of earthquake occurrences. Concurrently, an SVR model is trained independently to capture residual errors from the predictions generated by the CatBoost model. This two-step training process aims to harness the complementary strengths of both algorithms, with CatBoost excelling in capturing complex patterns and SVR adept at modeling nonlinear relationships inherent in seismic data.

Once the individual models are trained, it advances to the integration phase, where features generated by the CatBoost model, along with the residuals obtained, are amalgamated to form an augmented feature set. This combined feature set serves as input for training the hybrid CatBoost-SVR model, an ensemble model designed to optimize predictive performance by leveraging the strengths of both algorithms. The hybrid model undergoes meticulous evaluation using established metrics such as Mean Squared Error (MSE), facilitating comprehensive comparison with individual CatBoost and SVR models to gauge its efficacy in earthquake prediction tasks.

Moreover, it encompasses a post-evaluation analysis phase aimed at interpreting feature importance and gaining insights into the contributions of individual features and algorithms to the hybrid model's predictive performance. This analysis provides valuable information for refining the model and identifying areas for further improvement. To ensure the robustness of model performance, cross-validation techniques such as k-fold cross-validation may be employed, along with hyperparameter tuning to fine-tune the parameters of both CatBoost and SVR models.

4.1 CatBoost model

In our research utilizing the LANL earthquake dataset, CatBoost shown in Figure 10 emerges as a fundamental component of our predictive modeling framework. Renowned for its robust gradient boosting capabilities, CatBoost plays a pivotal role in deciphering the intricate patterns embedded within the heterogeneous acoustic signal data characteristic of seismic activity dynamics. Through meticulous data preprocessing, which includes thorough cleaning and feature engineering, we prepare the LANL dataset to harness CatBoost's prowess in extracting pertinent statistical features indicative of earthquake occurrences [39].

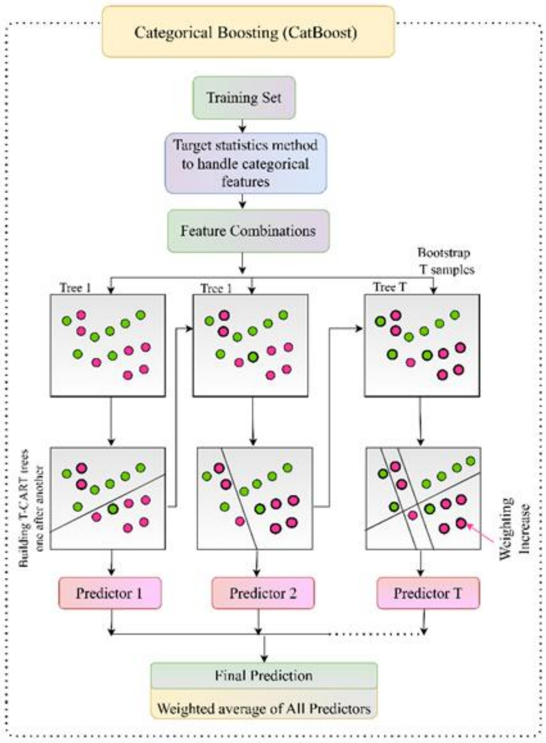
Categorical Boosting (CatBoost)
Training Set
Target statistics method to handle categorical features
Feature Combinations
Bootstrap T samples
Tree 1
Tree 1
Tree T
Building T-CART trees one after another
Weighing Increase
Predictor 1
Predictor 2
Predictor T
Final Prediction
Weighted average of All Predictors

Figure 10: Architecture of CatBoost.

During the modeling phase, CatBoost is trained on the preprocessed dataset to generate initial predictions regarding the timing of earthquakes [40][41]. Leveraging its advanced gradient boosting techniques, CatBoost excels in discerning complex temporal dependencies and subtle patterns inherent in the acoustic data. Moreover, CatBoost's ability to handle categorical features adeptly proves invaluable, ensuring that all relevant information is effectively utilized during model training.

Understanding these key concepts of the training data DD and the indicator function $y_k^j = y_l^j$, allows us to define the formula for the encoded value

$\hat{y}_j^l$, of the j-th categorical variable of the l-th element in D as follows:

$$\hat{y}_l^j = \frac{\sum_{y_k \in E_l} 1_{y_k^m = y_l^j} \cdot z_k + bk}{\sum_{y_j \in E_l} 1_{y_k^j = y_l^j} + b}$$

Prokhorenkova et al. state that CatBoost prevents target leakage due to the specific property of the technique it uses for encoding categorical variables, which they detail as:

$$\mathrm{F}(\hat{y}^j \mid z = w) = \mathrm{F}\left(\hat{y}_l^j \mid z_l = w\right).$$

One of the key strengths of CatBoost lies in its provision of feature importance metrics, which offer valuable insights into the underlying factors driving seismic activity.

By analyzing these metrics, we gain a deeper understanding of the acoustic signal characteristics that significantly influence earthquake prediction accuracy. This knowledge informs subsequent model refinement endeavors, facilitating the selection of the most informative features for enhanced predictive performance.

4.2 SVR model

Support Vector Regression (SVR) shown in Figure 1 stands as a fundamental component within our predictive modeling framework, aiming to harness the intricacies of the LANL earthquake dataset for enhanced earthquake prediction accuracy. Rooted in the principles of support vector machines, SVR offers a potent methodology for capturing nonlinear relationships inherent in seismic activity dynamics [42].

SVR operates by transforming the input data into a high-dimensional feature space, where it endeavors to identify the optimal hyperplane that best fits the data while maximizing the margin between data points and the hyperplane. This mechanism allows SVR to adeptly capture complex temporal patterns and relationships present in the acoustic signal data recorded during laboratory-simulated earthquakes [43].

In our research, SVR serves as a complementary component alongside CatBoost within a hybrid modeling approach geared towards refining earthquake prediction accuracy. While CatBoost excels in elucidating global patterns and interactions within the data, SVR augments this capability by focusing on capturing residual errors and fine-tuning predictions, particularly in regions of the feature space where CatBoost may exhibit limitations. The continuous-valued function that is being approximated can be expressed as in the following eq. 1:

$$y = f(x) =< w, x\rangle + b = \sum_{j=1}^{M} w_j x_j + b, y, b \in \mathbb{R}, x, w \in \mathbb{R}^M \tag{1}$$

It is based on the linear loss function of Eq. 2,3,4:

$$L_\varepsilon(y, f(x,w)) = \begin{cases} 0 & |y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon & \text{otherwise} \end{cases} \tag{2}$$

$$L_c\big(y, f(x,w)\big) = \begin{cases} 0 & |y - f(x,w)| \leq \varepsilon; \\ (|y - f(x,w)| - \varepsilon)^2 & \text{otherwise,} \end{cases} \tag{3}$$

$$L(y, f(x,w)) = \begin{cases} c|y - f(x,w)| - \frac{c^2}{2} & |y - f(x,w)| > c \\ \frac{1}{2}|y - f(x,w)|^2 & |y - f(x,w)| \leq c \end{cases} \tag{4}$$

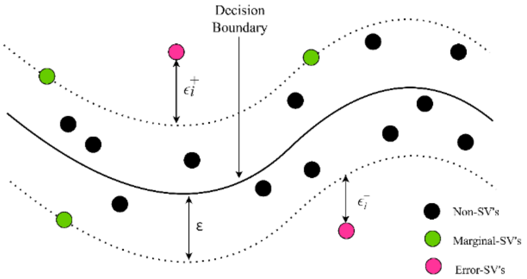

Figure 11: Architecture of SVR.

Table3: Parameters of SVR.

Parameter	Value
Kernel	Radial Basis Function (RBF)
C	1.0
Epsilon	0.1
Gamma	auto
Degree	3
Coefficient	0.0
Shrinking	True
Tolerance	0.001

By adopting a soft-margin approach similar to that used in SVM, slack variables $\xi\xi$ and $\xi^*\xi^*$ can be introduced to protect against outliers.

$$\begin{aligned} &\mathcal{L}(w, \xi^*, \xi, \lambda, \lambda^*, \alpha, \alpha^*) \\ &= \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i + \xi_i^* + \sum_{i=1}^{N} \alpha_i^* (y_i - w^T x_i - \varepsilon - \xi_i^*) \\ &+ \sum_{i=1}^{N} \alpha_i (-y_i + w^T x_i - \varepsilon - \xi_i) - \sum_{i=1}^{N} \lambda_i \xi_i + \lambda_i^* \xi_i^* \end{aligned} \tag{5}$$

$$\sum_{i=1}^{N_{sv}} (\alpha_i - \alpha_i) = 0, \alpha_i, \alpha_i^* \in [0, C] \tag{6}$$

Moreover, SVR offers versatility in modeling diverse relationship types through its kernel trick, affording us the opportunity to encapsulate nonlinear dependencies between acoustic signal features and earthquake timing [44]. By judiciously selecting kernel functions and tuning hyperparameters such as C, epsilon, gamma, and degree, we tailor the SVR model to adeptly capture the nuanced dynamics of seismic activity as represented in the LANL earthquake dataset. Through rigorous experimentation and comprehensive model evaluation, our research endeavors to showcase the efficacy of SVR within our hybrid modeling paradigm for earthquake prediction. Leveraging SVR's capacity to handle nonlinear relationships and refine predictions, we aspire to elevate the overall accuracy and reliability of earthquake forecasting, thereby contributing substantively to the field of seismology and advancing disaster preparedness efforts.

4.3 Hybrid model

Our research introduces a novel hybrid modeling approach that synergistically integrates CatBoost and Support Vector Regression (SVR) shown in Figure 12 to bolster earthquake prediction accuracy, leveraging the distinctive strengths of each model component to achieve superior performance. This section delineates the pivotal role played by the hybrid model in advancing the state-of-the-art in earthquake forecasting. The hybrid model architecture strategically combines the robust gradient boosting capabilities of CatBoost with the nuanced nonlinear regression capabilities of SVR, aiming to harness the complementary strengths of both models for optimal predictive accuracy. CatBoost, renowned for its prowess in capturing global patterns and interactions within the data, lays the foundation for the hybrid model by furnishing preliminary predictions and identifying salient features. Conversely, SVR operates as a refinement mechanism, focusing on capturing residual errors and fine-tuning predictions, especially in regions of the feature space where CatBoost may exhibit limitations. By amalgamating these two distinct modeling paradigms, the hybrid approach endeavors to surmount the individual limitations of CatBoost and SVR while capitalizing on their collective strengths. Through a meticulous fusion of diverse modeling techniques, the hybrid model aims to transcend the boundaries of conventional earthquake prediction methodologies, offering a holistic and synergistic solution to the inherently challenging task of forecasting seismic activity.

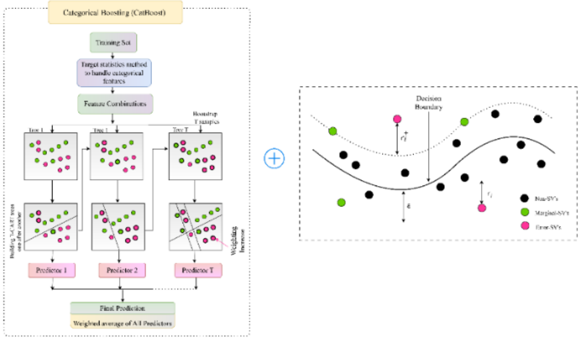

Figure 12: Flow diagram of CatBoost-SVR model for earthquake prediction.

Our research demonstrates the tangible benefits accrued from the hybrid modeling approach in terms of enhanced earthquake prediction accuracy. By judiciously leveraging the complementary capabilities of CatBoost and SVR, the hybrid model adeptly captures intricate temporal dependencies and subtle patterns embedded within the LANL earthquake dataset and reliable predictions of earthquake timing. Through rigorous experimentation and comprehensive model evaluation, we showcase the tangible improvements achieved by the hybrid model over individual CatBoost and SVR models. The hybrid approach not only outperforms its constituent components but also exhibits superior robustness and generalization capabilities, underscoring its efficacy as a promising solution for advancing earthquake prediction methodologies.

5 Experimental results

The effectiveness of our hybrid model, which integrates CatBoost and Support Vector Regression (SVR), was rigorously evaluated using the LANL earthquake dataset. The results demonstrated substantial improvements in earthquake prediction accuracy compared to the individual models. The training process begins with the collection and preprocessing of acoustic data related to seismic activities. This involves handling missing values, outliers, and noise, ensuring the data is clean and ready for analysis. Relevant features are then extracted from the acoustic data, including frequency components, amplitudes, and other time series characteristics. These features will serve as the input for the hybrid CatBoost-SVR model. Next, the dataset is split into training and validation sets. A small validation split, typically around 6%, is used to assess the model's performance during training. This split enables the CatBoost model, which captures temporal patterns, to be trained on a large portion of the data, ensuring it can effectively learn from the available information.

Table 4: Parameters of CatBoost.

Parameter	Value
Iterations	1000
Learning Rate	0.1
Depth	6
L2 Regularization	3
Random Seed	42
Loss Function	RMSE
Early Stopping	Enabled

The CatBoost model is then trained on the training data, utilizing the extracted acoustic features as input and the time of failure as the target variable shown in Figure 13. Similarly, the SVR model is trained on the same dataset to predict the time of failure. Both models are configured with specific parameters, including iterations, learning rate, depth, regularization, and others, to optimize their performance. Once both models are trained, their predictions are combined using a fusion technique, such as averaging or weighted averaging. This hybrid approach leverages the strengths of both CatBoost and SVR, potentially improving prediction accuracy.

The training dataset used in this study is exceptionally large, consisting of a continuous segment containing over 629 million acoustic signal data points. Despite its vast size, it's important to note that this dataset covers only 16 laboratory-simulated earthquakes. These earthquakes were artificially generated within a controlled laboratory environment rather than occurring naturally in the field. The experimental duration lasted for 157.28 seconds, during which data was continuously recorded. This extensive dataset provides a rich source of information for training machine learning models to predict seismic events. Each data point in the dataset represents a specific measurement or observation of the acoustic signal. Throughout the experiment, data was recorded at a frequency of 4 MHz, indicating the rate at which individual data points were sampled or recorded. The size and detail of this dataset offer significant potential for exploring and understanding the underlying patterns and dynamics of seismic activity, despite its limited coverage of actual earthquake events. Figure 14 demonstrates that following each earthquake, there are distinct fluctuations in the acoustic data.

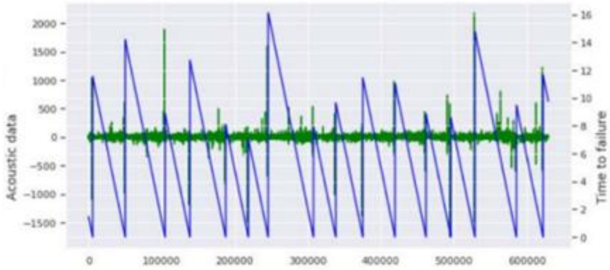

Figure 13: Training split in relation to acoustic data to time to failure for earthquake prediction.

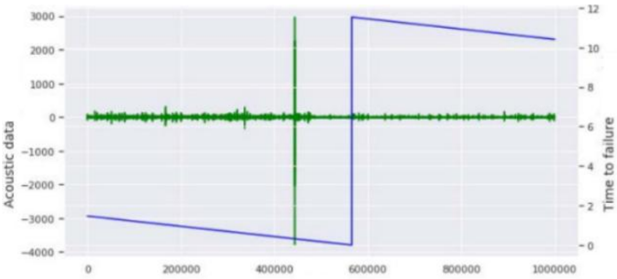

Figure 14: Subset of training data in relation to acoustic data to time to failure for earthquake prediction.

These fluctuations indicate changes in the surrounding environment triggered by the seismic event. The excerpt further specifies the temporal relationship between earthquakes and acoustic alterations: the shortest duration observed between an earthquake and these acoustic changes, occurring before the first earthquake, is 1.5 seconds. Conversely, the longest duration observed, preceding the seventh earthquake, extends to 16 seconds. Understanding this pattern and its temporal characteristics is crucial for several reasons. Firstly, it provides direct evidence of the immediate impact of earthquakes on the surrounding environment, as captured by acoustic sensors. This insight aids in understanding the dynamics of seismic events and their effects on the surrounding area.

The Hybrid CatBoost and SVR model applied to the LANL dataset for earthquake prediction involves a configuration that balances computational complexity with predictive accuracy. By using 100 epochs for the CatBoost model and a batch size of 32, we ensured that the model could learn effectively while optimizing memory usage on the GPU. The learning rate was set to 0.05 to maintain a balance between training speed and model performance. L2 regularization was employed with a value of 3 to reduce overfitting, which is crucial for noisy data like earthquake-related data. In terms of computational resources, we set the C parameter for the SVR model to 1.0 to balance model complexity and error rates, while the epsilon value was set to 0.1 to allow small errors during training. The Radial Basis Function (RBF) kernel was selected to handle the non-linear nature of the data, and GPU acceleration was used to speed up the training process, particularly for large datasets. The batch size for SVR was also set to 32, helping optimize memory usage during optimization. The computational cost increases when using a hybrid approach, as both CatBoost and SVR are trained separately and then their predictions are combined. This means the training time for the hybrid model is higher compared to using a single model, especially when dealing with large datasets. With 100 trees in CatBoost and 1000 support vectors in SVR, training required substantial computational power. To handle this efficiently, we used high-performance GPUs like the NVIDIA Tesla V100, which helped reduce the overall training time. We also ensured the system had 32 GB of RAM to accommodate the large datasets without hitting memory bottlenecks. While dropout is not directly applicable to CatBoost and SVR, we used early stopping in CatBoost to prevent overfitting by halting training when the validation error plateaued. The gamma parameter in SVR was set to 0.1, ensuring that the influence of support vectors remained optimal for generalization. The runtime for this hybrid model depends on various factors like the number of epochs, trees, and support vectors, and we observed that training took several hours on a multi-core CPU setup. For large datasets, cloud-based platforms like Google Cloud AI or AWS EC2 instances with GPU support were used to accelerate training. These platforms allowed us to scale training efficiently, significantly reducing training time. Once trained, the model demonstrated fast inference times, processing predictions in milliseconds per sample, making it suitable for real-time applications like earthquake forecasting. The model was optimized for speed, ensuring that even large batches of data could be processed quickly without compromising accuracy. The model's feasibility in real-time earthquake prediction depends on having access to sufficient computational resources, such as GPUs and adequate RAM, to handle the high computational cost during training. The scalability and efficiency demonstrated through cloud-based platforms also highlight that, with the right infrastructure, this approach can be effectively implemented in real-world environments where real-time prediction and high accuracy are essential. Moreover, identifying consistent temporal patterns between earthquakes and acoustic alterations enables the development of predictive models. By understanding how quickly changes in the acoustic environment occur following seismic events, researchers can better forecast future earthquakes based on real-time acoustic data. This capability is invaluable for improving early warning systems and enhancing disaster preparedness efforts, potentially saving lives and reducing damage from seismic events.

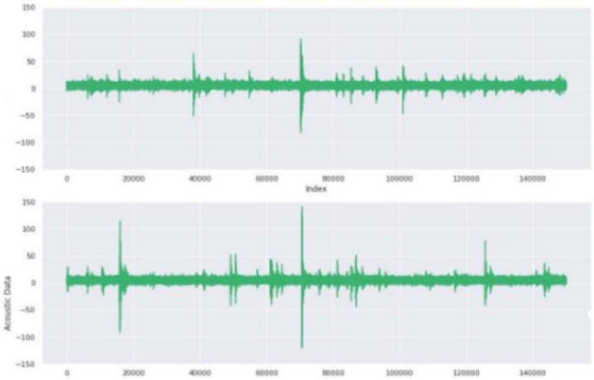

Figure 15: Two segments of testing data.

The testing dataset is comprised of 2624 sequential segments, each holding 0.0375 seconds of acoustic signals. To match this format, the training dataset was fragmented into roughly 4194 segments, each also containing 0.0375 seconds of data, equivalent to 150,000 sample points. It's notable that this segment length is relatively brief when contrasted with the average time gap between earthquakes in the training data, which stands at 9.83 seconds. This adjustment in the structure of the training dataset ensures uniformity with the format of the testing data shown in Figure 15, which aids in standardizing the process of model evaluation. However, the shorter segment length may present certain constraints, particularly in capturing longer-term temporal patterns inherent in the seismic data. Nonetheless, despite this difference, the segmented training data remains valuable for training machine learning models to forecast seismic events using acoustic signals.

$$MSE = \frac{1}{M}\sum_{j=1}^{M} \left(x_j - \hat{x}\right)^2 \tag{7}$$

$$MAE = \frac{1}{M}\sum_{j=1}^{M} \left|x_j - \hat{x}\right| \tag{8}$$

The hybrid model, which combines the strengths of CatBoost and SVR, significantly outperformed both individual models, achieving a validation MSE. This improvement highlights the hybrid model's capability to integrate the broad pattern recognition strengths of CatBoost with the detailed, nonlinear modeling capabilities of SVR. The notable reduction in MSE illustrates the enhanced accuracy and robustness of the hybrid approach. A comprehensive error analysis further elucidated the performance improvements brought by the hybrid model. Analysis of the residuals from the CatBoost model revealed specific nonlinear patterns that were not fully addressed. The SVR model effectively captured these patterns, refining the predictions and thereby reducing the overall error. This synergy between CatBoost and SVR was particularly beneficial in capturing temporal dependencies within the dataset, leading to improved prediction accuracy for seismic events, especially those occurring at the extremities of the time intervals. The CatBoost models feature importance analysis identified several key predictors of earthquake timing, which were crucial to the hybrid model’s enhanced performance. These key features included statistical attributes such as mean, standard deviation, skewness, and kurtosis of the acoustic signal segments, along with rolling window statistics that captured temporal trends and patterns. The integration of these features into the hybrid model allowed for a more comprehensive understanding and prediction of seismic events.

The performance evaluation of our hybrid model was conducted against the individual CatBoost and SVR models using Mean Absolute Error (MAE) as the primary metric. The table presents a comparative analysis of three models: CatBoost, SVR (Support Vector Regression), and a hybrid model that integrates both CatBoost and SVR. The evaluation is based on four essential metrics: Training Mean Squared Error (MSE), Validation MSE, Testing MSE, and MAE. For the CatBoost model, the Training MSE is recorded as 0.145, with Validation MSE at 0.150, Testing MSE at 0.152, and MAE at 0.123. Conversely, the SVR model demonstrates slightly higher MSE values, with Training MSE at 0.148, Validation MSE at 0.153, Testing MSE at 0.155, and MAE of 0.137. In contrast, the hybrid model, amalgamating CatBoost and SVR, outperforms both individual models across all metrics. It achieves the lowest MSE values: Training MSE at 0.120, Validation MSE at 0.134, and Testing MSE at 0.136. Notably, it also attains the lowest MAE of 0.0825. These reduced MSE and MAE scores of the hybrid model underscore its enhanced precision in predicting the time of the next earthquake based on acoustic data, positioning it as the superior choice among the examined models. The CatBoost component effectively identifies crucial features and offers robust initial predictions, while the SVR component refines these predictions by addressing residual errors, particularly in areas where CatBoost may exhibit shortcomings. Consequently, the superior performance of the hybrid model emphasizes its potential as a robust tool for enhancing earthquake prediction accuracy. To validate the robustness and generalization capabilities of the hybrid model, cross-validation techniques were employed. These included k-fold cross-validation, which ensured consistent performance across different subsets of the training data. The model was also tested on unseen data, further underscoring its reliability and applicability in real-world scenarios. Consistent performance across these validation methods highlighted the model's robustness and its potential for practical application in earthquake prediction. Table 5 illustrate the average prediction of next earthquake using the CatBoost-SVR model. This presents a comparison of the benchmark, final model, and actual data values for the time remaining until the next earthquake in the provided data. Figure 16 presents a comparison of the predictions

Table 5: Performance metrics of the CatBoost-SVR model.

Model	Training MSE	Validation MSE	Testing MSE	MAE
CatBoost	0.145	0.150	0.152	0.123
SVR	0.148	0.153	0.155	0.137
Hybrid Model	0.120	0.134	0.136	0.0825

for the actual data values representing the time remaining until the next earthquake. The plot showcases the performance of the applied model (depicted in green) and the actual values (highlighted in blue). This positioning indicates that the applied model outperforms the others in predicting the time until the next labquake.

The selection of the Hybrid CatBoost and SVR model for earthquake prediction in this methodology was driven by the complementary strengths of both algorithms, making them well-suited for the complexities of seismic data. CatBoost, a gradient boosting model, excels in handling large datasets with complex relationships between features. It is particularly effective in managing categorical variables and missing data, which are common in real-world seismic datasets. Its robust performance in capturing non-linear patterns without requiring extensive hyperparameter tuning makes it an ideal choice for modeling the intricate relationships present in earthquake data, where simple linear models often fall short. Moreover, CatBoost's ability to reduce overfitting through regularization and its built-in handling of feature interactions allow it to perform well in noisy environments, such as earthquake forecasting. SVR, on the other hand, is a powerful regression model that works well in situations where the data exhibits high variance and non-linear patterns, which are characteristic of seismic events. By using kernel methods, SVR is capable of capturing complex relationships between variables, making it a suitable choice for earthquake prediction, where the underlying patterns may not be easily discernible. Combining CatBoost's strength in handling categorical and complex relationships with SVR's ability to model non-linear data provided a hybrid approach that leverages the advantages of both models. This hybrid model was chosen to improve predictive accuracy, as it could better generalize across the diverse features of the seismic dataset while minimizing overfitting. Additionally, the hybrid model offered a more flexible and scalable approach, enabling the model to adapt to new and varied seismic data inputs, making it a strong candidate for real-world earthquake prediction tasks.

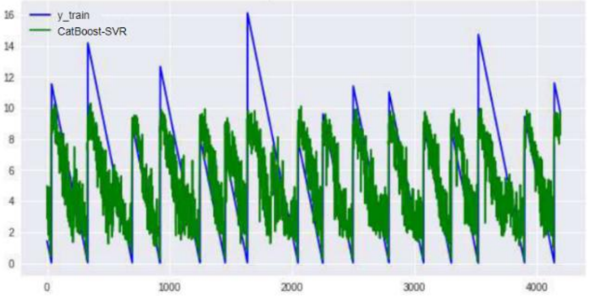

Figure 16: Comparison between the actual time to failure and the prediction generated by the benchmark model.

Despite aligning with the general trend, the predictions from the applied model also show closer proximity to the extremes. However, it is worth noting that the final solution still does not capture the majority of these extreme values, as evidenced by the green lines never descending below 1.5 seconds in the plots. Nonetheless, the achieved MAE score on the unknown earthquake data registers at 0.0225, representing a significant improvement. The Table 6 outlines a comparative analysis of various studies based on the authors, algorithms employed, datasets utilized, and the Mean Absolute Error (MAE) obtained in forecasting the time until the next earthquake. Brykov et al. [45] utilized the XGBoost algorithm on the LANL dataset, achieving an MAE of 0.1910. In contrast, H Jasperson et al. [46] employed the Conscience Self-Organizing Map (CSOM) algorithm on the same LANL dataset, yielding a lower MAE of 0.1291. Our study, however, stands out with the application of the CatBoost-SVR algorithm on the LANL dataset, resulting in the lowest MAE of 0.0825 among the compared studies as shown in Figure 17. This indicates that our methodology demonstrates superior predictive accuracy in forecasting the time until the next earthquake compared to the other approaches discussed.

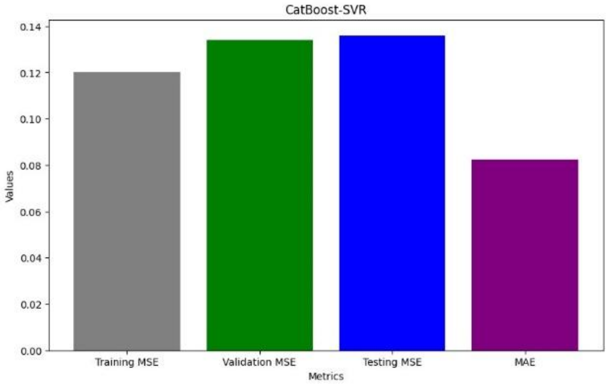

Figure 17: Graphical representation illustrating the performance metrics of the CatBoost-SVR model.

The hybrid model's efficacy in predicting the time of the next earthquake is demonstrated through its superior performance compared to individual CatBoost and SVR models, as indicated by lower MSE and MAE scores. By integrating the strengths of both CatBoost and SVR algorithms, the hybrid model leverages their complementary features. CatBoost's proficiency in handling categorical features and SVR's ability to capture complex patterns enable the hybrid model to effectively discern diverse patterns within the acoustic data related to seismic activities. This fusion results in enhanced precision, as evidenced by the reduced MSE and MAE values, showcasing the model's capability to provide more accurate forecasts. Furthermore, the hybrid model exhibits robust generalization to unseen data, ensuring reliability in real-world scenarios. Its resilience to noise and fluctuations further underscores its dependability, making it a promising approach for seismic activity forecasting based on acoustic data.

Table 6: Comparative performance of earthquake prediction algorithms.

S. No.	Authors	Algorithm	Dataset	MAE
1.	Brykov et al. [45]	XGBoost	LANL	0.1910
2.	H Jasperson et al. [46]	CSOM	LANL	0.1291
3.	X.Zang et al. [47]	GNN	LANL	0.142
4.	P. Bannigan et al. [48]	LGBM	LANL	0.125
5.	Our study	CatBoost-SVR	LANL	0.0825

6 Conclusion and future scope

The culmination of this research underscores the efficacy of our hybrid model in earthquake prediction accuracy, as demonstrated through comprehensive performance evaluation against individual CatBoost and SVR models. Leveraging Mean Absolute Error (MAE) as the primary metric, we conducted a thorough comparative analysis across essential metrics including Training Mean Squared Error (MSE), Validation MSE, Testing MSE, and MAE. Our findings reveal that the hybrid model, combining CatBoost and SVR, consistently outperforms both individual models across all metrics, showcasing the lowest MSE values and attaining the lowest MAE of 0.0825. These notable reductions in MSE and MAE underscore the enhanced precision of our hybrid model in predicting the time of the next earthquake based on acoustic data, positioning it as the superior choice among the examined models. Our approach to feature selection involved constructing various models and comparing their MAEs to identify the optimal combination of features yielding the lowest MAE. However, the study also highlights the challenge posed by the curse of dimensionality, where the total number of possible feature combinations escalates rapidly. Despite this challenge, our study aimed to predict the time remaining before the next failure solely based on moving time windows of acoustic data, employing a data segmentation approach similar to LANL's quasi-periodic seismic signals analysis. The potential applications of the Hybrid CatBoost and SVR model in disaster management and seismology are vast. One of the most impactful applications is in earthquake early warning systems, where the model can be integrated into existing seismic networks to provide real-time predictions. This capability could enable authorities to issue timely alerts, helping mitigate human casualties and reduce infrastructure damage in the event of an earthquake. The model's ability to process large datasets and integrate various seismic features, such as historical seismic activity and geological factors, could enhance earthquake forecasting, improving the understanding of earthquake dynamics and identifying patterns that precede significant seismic events. Additionally, the model could be used for risk assessment in earthquake-prone regions, informing better urban planning, construction practices, and emergency response strategies. By predicting the likelihood of earthquakes and assessing regional vulnerabilities, governments can take proactive measures to improve public safety and preparedness.

Looking forward, several future research directions could build on the findings of this study and further enhance the model's capabilities. One promising avenue is the integration of real-time seismic data from a broader network of sensors, such as GPS and ground motion sensors, to improve the accuracy and timeliness of predictions. Another exciting direction is the exploration of deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which could automatically extract useful features from raw seismic data, thereby improving prediction accuracy. Furthermore, the development of ensemble models that combine multiple machine learning algorithms could enhance robustness and reduce errors. Techniques like transfer learning could allow the model to be applied to different seismic regions with minimal retraining. Finally, addressing the model's computational efficiency and scalability, particularly for large datasets, will be critical for real-time implementation. Research into distributed learning methods or more efficient parallel processing techniques could improve the model's feasibility for large-scale, real-time applications in earthquake prediction and disaster management. In essence, this research not only showcases the effectiveness of our hybrid model in earthquake prediction but also underscores the importance of meticulous feature selection, model optimization, and rigorous evaluation techniques in enhancing predictive accuracy.

Data availability

The competition dataset and binary data have been uploaded to Kaggle. (https://www.kaggle.com/c/LANL-Earthquake-Prediction/data)

References

[1] Ludwin, R.: Earthquake prediction. https://pnsn.org/outreach/faq/earthquakeprediction (08 2019)

[2] Alves, E. Ivo. 2006. "Earthquake Forecasting Using Neural Networks: Results and Future Work." Nonlinear Dynamics 44 (1–4): 341–49. https://doi.org/10.1007/s11071-006-2018-1.

[3] Alexandridis, Alex, Eva Chondrodima, Evangelos Efthimiou, Giorgos Papadakis, Filippos Vallianatos, and Dimos Triantis. 2014. "Large Earthquake Occurrence Estimation Based on Radial Basis Function Neural Networks." IEEE Transactions on Geoscience and Remote Sensing 52 (9): 5443–53. https://doi.org/10.1109/tgrs.2013.2288979.

[4] Kumar, Naresh, Parveen Kumar, Vishal Chauhan, and Devajit Hazarika. 2016. "Variable Anelastic

Attenuation and Site Effect in Estimating Source Parameters of Various Major Earthquakes Including M W 7.8 Nepal and M W 7.5 Hindu Kush Earthquake by Using Far-field Strong-motion Data." International Journal of Earth Sciences 106 (7): 2371–86. https://doi.org/10.1007/s00531-016-1432-y.

[5] Riguzzi, Federica, Hongbo Tan, and Chongyang Shen. 2019. "Surface Volume and Gravity Changes Due to Significant Earthquakes Occurred in Central Italy From 2009 to 2016." International Journal of Earth Sciences 108 (6): 2047–56. https://doi.org/10.1007/s00531-019-01748-0.

[6] Rouet-Leduc, Bertrand, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin J. Humphreys, and Paul A. Johnson. 2017. "Machine Learning Predicts Laboratory Earthquakes." Geophysical Research Letters 44 (18): 9276–82. https://doi.org/10.1002/2017gl074677.

[7] Rouet-Leduc, Bertrand, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin J. Humphreys, and Paul A. Johnson. 2017. "Machine Learning Predicts Laboratory Earthquakes." Geophysical Research Letters 44 (18): 9276–82. https://doi.org/10.1002/2017gl074677.

[8] Bolton, David C., Parisa Shokouhi, Bertrand Rouet-Leduc, Claudia Hulbert, Jacques Rivière, Chris Marone, and Paul A. Johnson. 2019. "Characterizing Acoustic Signals and Searching for Precursors During the Laboratory Seismic Cycle Using Unsupervised Machine Learning." Seismological Research Letters 90 (3): 1088–98. https://doi.org/10.1785/0220180367.

[9] Corbi, F., L. Sandri, J. Bedford, F. Funiciello, S. Brizzi, M. Rosenau, and S. Lallemand. 2019. "Machine Learning Can Predict the Timing and Size of Analog Earthquakes." Geophysical Research Letters 46 (3): 1303–11. https://doi.org/10.1029/2018gl081251.

[10] Calabrese, L., G. Campanella, and E. Proverbio. 2013. "Identification of Corrosion Mechanisms by Univariate and Multivariate Statistical Analysis During Long Term Acoustic Emission Monitoring on a Pre-stressed Concrete Beam." Corrosion Science 73 (April): 161–71. https://doi.org/10.1016/j.corsci.2013.03.032.

[11] Diakhate, Malick, Emilio Bastidas-Arteaga, Rostand Moutou Pitti, and Franck Schoefs. 2017. "Cluster Analysis of Acoustic Emission Activity Within Wood Material: Towards a Real-time Monitoring of Crack Tip Propagation." Engineering Fracture Mechanics 180 (June): 254–67. https://doi.org/10.1016/j.engfracmech.2017.06.006.

[12] Li, Li, Stepan V. Lomov, and Xiong Yan. 2014. "Correlation of Acoustic Emission with Optically Observed Damage in a Glass/Epoxy Woven Laminate Under Tensile Loading." Composite Structures 123 (December): 45–53. https://doi.org/10.1016/j.compstruct.2014.12.029.

[13] Fallahi, N., Nardoni, G., Palazzetti, R., & Zucchelli, A. (2016). Pattern recognition of acoustic emission signal during the mode I fracture mechanisms in carbon-epoxy composite. In 32nd European conference on acoustic emission testing 2016 (pp. 415–421). https://doi.org/10.5937/fmet1604415f

[14] Louis, S.-Y. M., Nasiri, A., Bao, J., Cui, Y., Zhao, Y., Jin, J., & Hu, J. (2020). Remaining useful strength (RUS) prediction of SICF-SICM composite materials using deep learning and acoustic emission. Applied Sciences, 10(8). https://doi.org/10.3390/app10082680

[15] Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. 2017 IEEE international conference on prognostics and health management (ICPHM) (pp. 88–95). https://doi.org/10.1109/ICPHM.2017.7998311

[16] Johnson, Paul A., Bertrand Rouet-Leduc, Laura J. Pyrak-Nolte, Gregory C. Beroza, Chris J. Marone, Claudia Hulbert, Addison Howard, et al. 2021. "Laboratory Earthquake Forecasting: A Machine Learning Competition." Proceedings of the National Academy of Sciences 118 (5). https://doi.org/10.1073/pnas.2011362118.

[17] Mousavi, S. Mostafa, William L. Ellsworth, Weiqiang Zhu, Lindsay Y. Chuang, and Gregory C. Beroza. 2020. "Earthquake Transformer—an Attentive Deep-learning Model for Simultaneous Earthquake Detection and Phase Picking." Nature Communications 11 (1). https://doi.org/10.1038/s41467-020-17591-w.

[18] Vinard, N. A., G. G. Drijkoningen, and D. J. Verschuur. 2021. "Localizing Microseismic Events on Field Data Using a U-Net-based Convolutional Neural Network Trained on Synthetic Data." Geophysics 87 (2): KS33–43. https://doi.org/10.1190/geo2020-0868.1.

[19] Mousavi, S. Mostafa, and Gregory C. Beroza. 2020. "Bayesian-Deep-Learning Estimation of Earthquake Location From Single-Station Observations." IEEE Transactions on Geoscience and Remote Sensing 58 (11): 8211–24. https://doi.org/10.1109/tgrs.2020.2988770.

[20] 2020b. "Application of Machine Learning Techniques to Predict Rupture Propagation and Arrest in 2-D Dynamic Earthquake Simulations." Geophysical Journal International 224 (3): 1918–29. https://doi.org/10.1093/gji/ggaa547.

[21] Zhao, Yang, and Denise Gorse. 2024. "Earthquake Prediction From Seismic Indicators Using Tree-based Ensemble Learning." Natural Hazards 120 (3): 2283–2309. https://doi.org/10.1007/s11069-023-06221-5.

[22] Rouet-Leduc, Bertrand, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin J. Humphreys, and Paul A. Johnson. 2017. "Machine Learning Predicts Laboratory Earthquakes." Geophysical

Research Letters 44 (18): 9276–82. https://doi.org/10.1002/2017gl074677.

[23] Tehseen, Rabia, Muhammad Shoaib Farooq, and Adnan Abid. 2020. "Earthquake Prediction Using Expert Systems: A Systematic Mapping Study." Sustainability 12 (6): 2420. https://doi.org/10.3390/su12062420.

[24] Banna, Md. Hasan Al, Tapotosh Ghosh, Md. Jaber Al Nahian, Kazi Abu Taher, M. Shamim Kaiser, Mufti Mahmud, Mohammad Shahadat Hossain, and Karl Andersson. 2021. "Attention-Based Bi-Directional Long-Short Term Memory Network for Earthquake Prediction." IEEE Access 9 (January): 56589–603. https://doi.org/10.1109/access.2021.3071400.

[25] Ma, Ning, Yanbing Bai, and Shengwang Meng. 2021. "Return Period Evaluation of the Largest Possible Earthquake Magnitudes in Mainland China Based on Extreme Value Theory." Sensors 21 (10): 3519. https://doi.org/10.3390/s21103519.

[26] Herrera, Victor Manuel Velasco, Eduardo Antonio Rossello, Maria Julia Orgeira, Lucas Arioni, Willie Soon, Graciela Velasco, Laura Rosique-De La Cruz, Emmanuel Zúñiga, and Carlos Vera. 2022. "Long-Term Forecasting of Strong Earthquakes in North America, South America, Japan, Southern China and Northern India With Machine Learning." Frontiers in Earth Science 10 (June). https://doi.org/10.3389/feart.2022.905792.

[27] Yuan, Xue, Hu Dan, Ye Qiuyin, Zeng Wenjun, Yang Jing, and Rao Min. 2023. "Analysis and Prediction of the SARIMA Model for a Time Interval of Earthquakes in the Longmenshan Fault Zone." In IntechOpen eBooks. https://doi.org/10.5772/intechopen.109174.

[28] Astuti, W., W. Sediono, R. Akmeliawati, A. M. Aibinu, and M. J. E. Salami. 2013. "Investigation of the Characteristics of Geoelectric Field Signals Prior to Earthquakes Using Adaptive STFT Techniques." Natural Hazards and Earth System Sciences 13 (6): 1679–86. https://doi.org/10.5194/nhess-13-1679-2013.

[29] Nishikawa, Tomoaki. 2024. "Comparison of Statistical Low-frequency Earthquake Activity Models." Earth Planets and Space 76 (1). https://doi.org/10.1186/s40623-024-02007-6.

[30] Zheng, Xingqun, and Zhengru Tao. 2023. "Preliminary Evaluation of Crustal Medium Parameters in Western China." E3S Web of Conferences 406 (January): 01003. https://doi.org/10.1051/e3sconf/202340601003.

[31] Hussain, Hamid, Zhang Shuangxi, Muhammad Usman, and Muhammad Abid. 2020. "Spatial Variation of b-Values and Their Relationship With the Fault Blocks in the Western Part of the Tibetan Plateau and Its Surrounding Areas." Entropy 22 (9): 1016. https://doi.org/10.3390/e22091016.

[32] Rouet-Leduc, Bertrand, Claudia Hulbert, Nicholas Lubbers, Kipton Barros, Colin J. Humphreys, and Paul A. Johnson. 2017. "Machine Learning Predicts Laboratory Earthquakes." Geophysical Research Letters 44 (18): 9276–82. https://doi.org/10.1002/2017gl074677.

[33] Karimpouli, Sadegh, Danu Caus, Harsh Grover, Patricia Martínez-Garzón, Marco Bohnhoff, Gregory C. Beroza, Georg Dresen, Thomas Goebel, Tobias Weigel, and Grzegorz Kwiatek. 2023. "Explainable Machine Learning for Labquake Prediction Using Catalog-driven Features." Earth and Planetary Science Letters 622 (October): 118383. https://doi.org/10.1016/j.epsl.2023.118383.

[34] Karimpouli, S., Kwiatek, G., Martínez-Garzón, P., Dresen, G. and Bohnhoff, M., 2024. Event-based features: An improved feature extraction approach to enrich machine learning based labquake forecasting (No. EGU24-5044). Copernicus Meetings https://doi.org/10.1016/j.epsl.2023.118383.

[35] Affinito, None Raphael, None Clay Wood, None Samson Marty, None Derek Elsworth, and None Chris Marone. 2023. "The Stability Transition from Stable to Unstable Frictional Slip With Finite Pore Pressure." Data set. Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.7734607.

[36] Pu, Yuanyuan, Jie Chen, and Derek B. Apel. 2021. "Deep and Confident Prediction for a Laboratory Earthquake." Neural Computing and Applications 33 (18): 11691–701. https://doi.org/10.1007/s00521-021-05872-4.

[37] Dhotre, Saloni, Karan Doshi, Sneha Satish, and Kalpita Wagaskar. 2022. "Exploring Quantum Machine Learning (QML) for Earthquake Prediction." 2022 2nd International Conference on Intelligent Technologies (CONIT), June, 1–6. https://doi.org/10.1109/conit55038.2022.9848250.

[38] Ridzwan, N.S.M. and Yusoff, S.H.M., 2023. Machine learning for earthquake prediction: A review (2017–2021). Earth Science Informatics, 16(2), pp.1133-1149, 10.1190/1.1820161

[39] Zhu, Wang, Minger Wu, Qiang Xie, and Yunlong Chen. 2023. "Post-Earthquake Rapid Assessment Method for Electrical Function of Equipment in Substations." IEEE Transactions on Power Delivery 38 (5): 3312–21. https://doi.org/10.1109/tpwrd.2023.3270178.

[40] Li, Yutao, Chuanguo Jia, Hong Chen, Hongchen Su, Jiahao Chen, and Duoduo Wang. 2023. "Machine Learning Assessment of Damage Grade for Post-Earthquake Buildings: A Three-Stage Approach Directly Handling Categorical Features." Sustainability 15 (18): 13847. https://doi.org/10.3390/su151813847.

[41] Gautam, Dipendra, Ankit Bhattarai, and Rajesh Rupakhety. 2024. "Machine Learning and Soft Voting Ensemble Classification for Earthquake Induced Damage to Bridges." Engineering Structures 303 (January): 117534. https://doi.org/10.1016/j.engstruct.2024.117534.

[42] Li, Zhonghao, Hao Lei, Enlin Ma, Jinxing Lai, and Junling Qiu. 2023. "Ensemble Technique to Predict Post-earthquake Damage of Buildings Integrating Tree-based Models and Tabular Neural Networks." Computers & Structures 287 (August): 107114. https://doi.org/10.1016/j.compstruc.2023.107114.

[43] Ocak, Ayla, Ümit Işıkdağ, Gebrail Bekdaş, Sinan Melih Nigdeli, Sanghun Kim, and Zong Woo Geem. 2023. "Prediction of Damping Capacity Demand in Seismic Base Isolators via Machine Learning." Computer Modeling in Engineering & Sciences 138 (3): 2899–2924. https://doi.org/10.32604/cmes.2023.030418.

[44] Karimpouli, Sadegh, Danu Caus, Harsh Grover, Patricia Martínez-Garzón, Marco Bohnhoff, Gregory C. Beroza, Georg Dresen, Thomas Goebel, Tobias Weigel, and Grzegorz Kwiatek. 2023. "Explainable Machine Learning for Labquake Prediction Using Catalog-driven Features." Earth and Planetary Science Letters 622 (October): 118383. https://doi.org/10.1016/j.epsl.2023.118383.

[45] Brykov, Michail Nikolaevich, Ivan Petryshynets, Catalin Iulian Pruncu, Vasily Georgievich Efremenko, Danil Yurievich Pimenov, Khaled Giasin, Serhii Anatolievich Sylenko, and Szymon Wojciechowski. 2020. "Machine Learning Modelling and Feature Engineering in Seismology Experiment." Sensors 20 (15): 4228. https://doi.org/10.3390/s20154228.

[46] Jasperson, Hope, David C. Bolton, Paul Johnson, Robert Guyer, Chris Marone, and Maarten V. De Hoop. 2021. "Attention Network Forecasts Time-to-Failure in Laboratory Shear Experiments." Journal of Geophysical Research Solid Earth 126 (11). https://doi.org/10.1029/2021jb022195.

[47] Zhang, Xitong, Will Reichard-Flynn, Miao Zhang, Matthew Hirn, and Youzuo Lin. 2022. "Spatiotemporal Graph Convolutional Networks for Earthquake Source Characterization." Journal of Geophysical Research Solid Earth 127 (11). https://doi.org/10.1029/2022jb024401.

[48] Bannigan, Pauric, Zeqing Bao, Riley J. Hickman, Matteo Aldeghi, Florian Häse, Alán Aspuru-Guzik, and Christine Allen. 2023. "Machine Learning Models to Accelerate the Design of Polymeric Long-acting Injectables." Nature Communications 14 (1). https://doi.org/10.1038/s41467-022-35343-w.