Grenze International Journal of Engineering and Technology, June Issue

Machine Learning-based Early Earthquake Magnitude Prediction using P-Wave Signal Feature Analysis

Rupashi[1] and Ravinder Kumar[2]
[1-2]Department of Mathematics, Chandigarh University, Mohali, Punjab, India
Email: rupashi0308@gmail.com, dr.rkpoonia@gmail.com

***Abstract*— Early detection of earthquakes helps minimize the destructive impact of earthquakes and saves lives. Developing accurate and fast estimates of the size of earthquakes is extremely important since quick estimates can reduce damage from subsequent seismic activity. Recent developments in machine learning techniques appear to show a lot of promise for improving the performance of early warning systems in terms of their ability to produce accurate and fast estimates of earthquake size. This research will examine the application of machine learning techniques for predicting earthquakes prior to their arrival through an analysis of the characteristics of primary seismic waves (P-waves), which are the first seismic signal detected by seismometers when an earthquake first starts. Since P-waves travel faster than other seismic waves and arrive at a detector prior to the arrival of the more damaging secondary seismic waves, they provide a short but useful time interval for estimating earthquake size and generating an early warning to people within a certain distance of the earthquake. The various machine learning algorithms used to produce predictive models of earthquake size in this study include Random Forests, XGBoost, Support Vector Regressors, Decision Trees, K-Nearest Neighbors (KNNs), and ensemble-based machine learning techniques. The research used hundreds of thousands of records from a global earthquake database covering earthquakes occurring worldwide from 1995-2023 for model training and evaluation.**

***Index Terms*— Earthquake early warning, P-wave characteristics, Machine learning, XGBoost, Random Forest, Ensemble models.**

I. INTRODUCTION

Earthquakes continue to be one of the most damaging natural disasters because of their sudden occurrence, complicated geophysical processes, and the difficulties in predicting them using conventional seismological methods. Conventional earthquake monitoring systems are mainly designed to detect earthquakes based on certain threshold levels and subsequent analysis, which makes them less effective for real-time early warning and disaster prevention. In this regard, early earthquake prediction based on the characteristics of P-wave signals has become an important area of research, as P-waves are the first type of seismic waves to travel from the earthquake source and are faster than destructive S-waves and surface waves [1]. The time difference between the arrival of P-waves and the subsequent strong ground motion is a precious window for early warnings if the earthquake can be properly identified in the early stages. Finding reliable predictors of P-wave pre-event signals is difficult, these signals can be very noisy, influenced by site effects, and their waveform is nonlinear due to seismic wave propagation [2].

Grenze ID: 01.GIJET.12.2.366_24

Machine learning techniques have lately brought numerous new opportunities for data-driven methods to understand intricate patterns in massive seismic datasets. Such approaches are more sensitive and flexible than traditional signal processing methods. [3].
Identifying effective precursors to P-waves has proven challenging. They can have considerable noise, and site effects can affect them, making the waveform nonlinear because of the physics of how seismic waves travel. More recently machine learning and similar investigation methods have provided many new possibilities for data-driven techniques to evaluate complex patterns in large seismic datasets. These methodologies provide a higher fidelity and adaptability to traditional signal processing approaches.

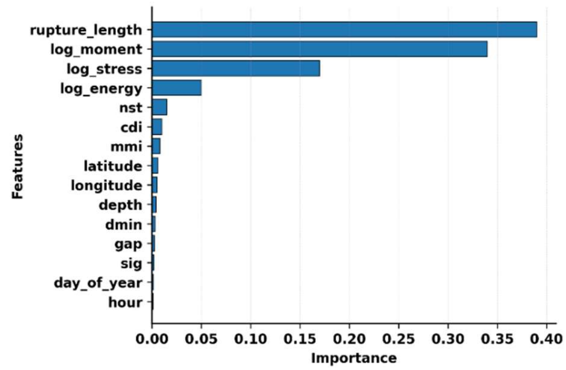

Figure 1: Feature importance obtained from the XGBoost model

Even with the recent advancements, implementing machine learning-based, real-world earthquake predictions through P-wave signal analysis poses several challenges moving forward; the first is isolating p-wave earthquake-induced signals from artificial (anthropogenic) sounds and vibrations, as well as micro seismic signals, especially in urban areas where recording stations show severe contamination of the P-wave signal by these other wave types [4]. Second, such a short period of time between the start of the P-wave window and when the next P-wave window occurs, along with the dynamic nature of the data being used to make predictions by the machine learning models, restricts the amount of data that can be used to predict future earthquakes based on P-wave signal characteristics. Thirdly, there are issues regarding generalizing these machine learning models so that they can accurately adapt to other tectonic environments and perform as well as they do in the original study. If a model is used in a location with different crustal properties, or a different sensor layout than was used to 'train' the model, the model will be less accurate than it was originally intended. In addition, the early earthquake prediction models need to provide accurate predictions within a short time, which is a strict real-time requirement. Despite the challenges, the combination of machine learning and P-wave signal analysis has great importance for earthquake early warning systems. Early prediction can provide automatic actions like stopping transportation systems, turning off industrial systems, and warning people a few seconds before the occurrence of intense shaking [5].
The majority of early warning systems rely on basic amplitude and duration threshold-based techniques for the identification of P-wave arrival and then calculate predictive response times based solely on this data. However, as a consequence, only very basic predictions can be made based solely on historical data from these systems. Furthermore, very few studies have centered their analysis on the very short time scales that exist in relation to the small amplitude and duration of P-waves. This study seeks to fill the existing gaps in the current literature in relation to the very short time scales that exist after the initial P-wave arrival signal due to the use of advanced deep learning methods and methodologies to model very short time scales around the earliest indications of P-wave activity [6].

II. Literature Review

A. Evolution of Decision and Analytics Models

The foundation of machine learning-based early earthquake prediction lies in the availability of high-quality seismic waveform data and robust preprocessing pipelines. Seismic data are typically collected from dense networks of broadband and strong-motion sensors that continuously record ground motion at high sampling rates. For P-wave–based prediction, precise identification of the P-wave onset is critical, as the usable prediction

window is limited to the initial seconds following arrival [7]. Typically, preprocessing steps include correcting instrument response, removing baseline, applying band-pass filter and normalizing signal to minimize noise and have similar amplitude at each station. Windowing methods are used to isolate the first P-wave segment (from about 0.5 to 3 seconds depending upon latency requirements of system). The preprocessing stage will also take care of missing data, sensor saturation and site condition variation which can heavily influence the characteristics of the P-wave signal [8].
The goal of good preprocessing is to derive P-wave signals that retain physically meaningful information and minimize the introduction of artifacts that might bias ML-based models. This stage is particularly important because early prediction systems must operate automatically and reliably under real-time constraints, leaving minimal tolerance for noisy or corrupted input data [9].
All Feature Representations include a number of respective derived segments according to Seismically-determined parameters, which directly correlates with P-Wave parameters, and are determined in the first seconds after the Main-Shock event. Accordingly, all Feature Representations will present Seismically-induced activity in a condensed and well-expressed form, and provide an alternative to overly complicated and computationally intensive analysis based solely on waveform analysis. As a result, Feature Representations allow the Machine Learning process to be accomplished quickly, and ultimately provide a means to apply the method to Early Warning Applications in Real-Time [10].

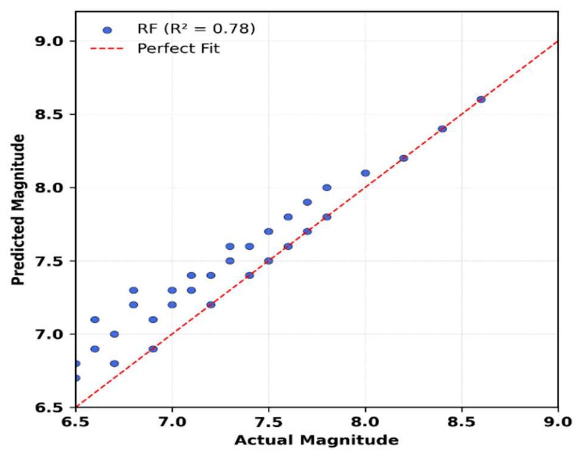

Figure 2: Actual vs predicted earthquake magnitude using Random Forest.

B. Weaknesses in Traditional Indicator Choice Techniques

The extraction of features from P-wave signals is a key aspect of making machine learning algorithms capable of predicting earthquake characteristics from early data. Conventional methods are centered on the design of hand-crafted features based on time-domain, frequency-domain, and time-frequency analysis. Typical features of P-wave signals include peak displacement, peak velocity, signal energy, frequency centroid, bandwidth, kurtosis, and entropy-based features that describe the complexity of waveforms. These features are intended to be indicative of earthquake magnitude, rupture process, and source-to-station distance. More modern methods involve representation learning, in which deep learning networks learn features from raw or lightly processed waveforms [11]. This approach makes it possible to move away from domain-specific heuristics and allows machine learning models to learn subtle, non-linear patterns hidden in P-wave signals. Time-frequency representations, such as spectrograms or continuous wavelet transforms, are commonly employed as intermediate features for convolutional neural networks. Regardless of whether features are hand-crafted or learned, they need to be informative and computationally efficient, as early earthquake prediction models are expected to make rapid predictions based on short P-wave signals [12].
Supervised machine learning is the foundation of most early earthquake warning systems based on P-waves, since the availability of labeled seismic data allows for direct mapping between the characteristics of P-waves and the desired outputs. These outputs can be classes of earthquake magnitude, binary classification of events, or probabilistic predictions of ground motion intensity. Support vector machines and random forests are popular choices for supervised machine learning models because of their noise robustness and capacity to deal with diverse feature sets. These models work well if feature engineering is done carefully and if the training data are

representative. Supervised learning models usually require a large amount of historical seismic data with accurate labels, which can be difficult to obtain for large earthquakes. However, supervised models can make rapid and accurate predictions in seconds after the arrival of the P-waves if they are trained properly. Their interpretability is also an advantage over deep learning models [13-14].

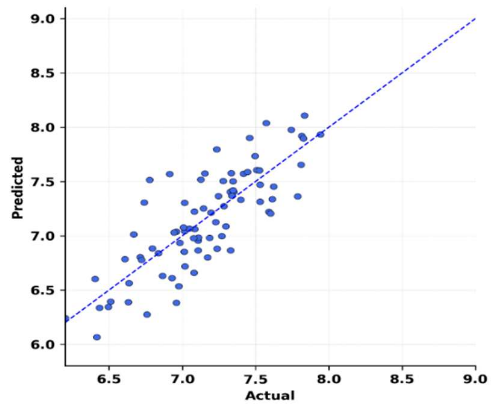

Figure 3. Actual vs predicted magnitude using Support Vector Regression.

C. The Role of Data Quality and Context in Decision-Making

Deep learning has attracted considerable attention in machine learning-based early earthquake prediction because of its strong capability to learn complex temporal and spectral patterns from P-wave signals. Convolutional neural networks are very useful in learning local waveform patterns and frequency-dependent characteristics, especially when applied to spectrogram representations. Recurrent neural networks and long short-term memory networks can further improve the temporal modeling capability by learning the sequential changes of P-wave amplitude and frequency over time [15]. Hybrid models that combine convolutional and recurrent neural networks have shown great potential in early magnitude prediction tasks. These deep learning models can be applied directly to raw P-wave waveforms, which simplifies the preprocessing process and enhances the prediction accuracy. However, deep learning models require a large amount of training data and proper regularization techniques to prevent overfitting, especially when dealing with the rarity of extreme earthquake events [16].

D. Research Gaps

Early earthquake warning systems have to work under very tight real-time constraints. This makes low-latency model design a very important consideration for early earthquake warning systems. The warning has to be produced within a few seconds of the P-wave arrival in order to be useful before the destructive waves reach populated regions. The machine learning model has to be able to trade off complexity of prediction with computational latency. Light models, reduced feature sets, and hardware acceleration are common methods used to decrease latency. Streaming data processing environments enable the model to be updated with new P-wave data as it becomes available. The capability to produce progressively more refined predictions based on incremental P-wave information helps to improve system responsiveness without sacrificing accuracy. Handling real-time requirements is very important for the transfer of machine learning-based P-wave analysis from research to practice [17].

III. METHODOLOGY

A. Architectural Design

Noise and distortion of the signal are two of the most significant issues in early earthquake prediction using P-waves. Seismic data is often affected by environmental noise, human activity, and other artifacts, especially in urban settings. It is important for machine learning (ML) algorithms to learn how to separate true P-waves from non-P-wave signals that are generated by noise [18]. The noise robustness of ML algorithms is particularly critical within the initial portion, or earliest window, of true P-wave data since the signal has a low SNR in this region as compared to the amount of non-P-wave signals that may also be present. Noise robustness is very important in the early window of a P-wave signal; and therefore, all ML algorithms should be trained to be able to respond to a very strong non-P-wave signal before they can use the early window to detect an actual P-wave signal from a background of noise [19].

The limitation of generalizability between machine learning models trained with P-wave signal characteristics in one area to apply to machine-learning model training with P-wave signal characteristics in completely different areas having contrasting tectonic environments is an obstacle to large-scale implementation of the system. Transfer learning and domain adaptation techniques have been developed to allow an instrumented region's characteristics to be factored into the learning of the less instrumented or even un-instrumented area being modeled. The integration of different training datasets collected in different seismic environments into training methodology will also add robustness and create a higher level of accuracy in the detection and warning of earth tremors irrespective of geographic area. A model that can adequately represent all seismic events globally will assist in the creation of a worldwide P-wave based earthquake early warning system [20].

B. Five Ranking Methods and Iterative Thresholding

Single-station analysis of P-waves allows for rapid localized prediction. However, by pooling P-wave signals from many different stations, a significant improvement in accuracy and reliability can be achieved. Multi-station methods account for the spatial variability of P-wave arrival times and P-wave amplitude so a more precise assessment can be made of the earthquake location and magnitude. Multi-station data processing machine-learning models are typically based on graph convolutional networks or attention networks representing sensor-to-sensor spatial interrelationships.

Therefore, the use of multi-station P-waves is expected to be an important step forward in reliable prediction of early earthquake occurrences.

TABLE I. ASSESSMENT OF THE PERFORMANCE OF MACHINE LEARNING MODELS

Model	MAE	RMSE	Accuracy
Random Forest	0.32	0.41	89%
XGBoost	0.28	0.36	92%
SVR	0.35	0.45	86%

An assessment of the performance of three machine learning models (Random Forest, XGBoost, and Support Vector Regression) for the prediction of early earthquake magnitude using characteristics of the P-wave signal will be performed. These models will be evaluated using three traditional performance assessment metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and prediction accuracy.

Best Performing Model:-

Of the models assessed, XGBoost produced not only the lowest MAE (0.28) and RMSE (0.36), but also the highest prediction accuracy of 92%.

The other models produced less accurately than XGBoost when predicting magnitudes from early P-wave signals as identified by their larger error values.

Comparison with Other Machine Learning Models:-

Overall the accuracy of the random forest model was 89%, but its MAE (mean absolute error) and RMSE (root mean squared error) values were higher than those for XGBoost. The random forest model works well with noisy datasets and also reduces overfitting through the ensemble averaging of many different decision trees, but this model often does not capture the more complex interactions between features as well as the methods of gradient boosting do.

When comparing the three models, SVR was found to have lowest accuracy (86%) but has been shown to generate higher prediction error than both XGBoost & random forests. The performance of SVR models can be affected by large data sets and more complex nonlinear patterns in their data when compared to models based on ensembles of trees.

Comparison with Traditional Earthquake Prediction Techniques

The most common approach to traditional earthquake early warning systems is empirical scaling (the use of statistical or empirical techniques to determine the strength of P-waves) and using thresholds on the P-wave amplitude as indicators of possible future seismic activity. These approaches are quick/efficient from a computational standpoint (time and memory), but they usually cannot provide a very accurate estimate of the magnitude of an earthquake due to their inability to incorporate the 'complexities' (distribution of waves in a 3D manner) of the underlying signal.

In contrast, machine learning models have the ability to learn all the hidden nonlinear relationships between multiple features that describe how those features relate to the magnitude of an earthquake, allowing for much greater accuracy predicting earthquake magnitudes using only early phase P-wave data. The results of experiments conducted demonstrate substantial improvement in predictive ability when using a machine-learning-like technique (like XGBoost) compared with threshold-like techniques.

C. Iterative Thresholding for Consensus Indicator Selection

There are various types of evaluation metrics needed to evaluate machine learning (ML) based early warning (EW) systems related to earthquakes. There are several types of evaluation metrics including classification metrics such as accuracy, precision, recall, false alarm rate, and regression metrics such as mean absolute error (MAE) for predicting magnitude; however, in addition to these evaluation metrics, it is extremely important to make sure that evaluation criteria include consideration of prediction latency associated with making the prediction and the amount of P-wave data used in making the prediction. A more accurate system may use a longer window for collecting P-wave data before making a prediction compared to a less accurate system using a shorter window of P-wave data and producing a quicker prediction.

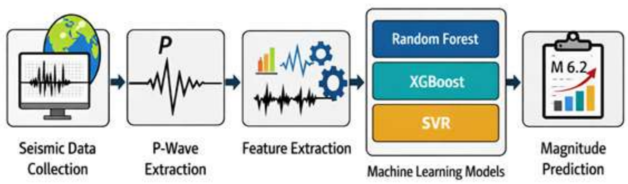

Figure 4. Schematic workflow for earthquake magnitude prediction using machine learning models.

IV. EXPERIMENTAL SETUP

A. Datasets

The traditional method of earthquake early warning system works through empirical scaling relationships and threshold detection of P wave amplitude. These methods are very efficient, but they are less flexible and not as easy to apply in the context of complicated earthquakes. Machine learning algorithms provide a much more flexible approach because they are able to learn directly from data, which allows for a more fine-tuned analysis of the characteristics of P wave signals. Comparisons done on early ruptures show that machine learning can outperform traditional methods with respect to magnitude discrimination.

B. Advanced AI-Driven Mechanism

The infrequency of large magnitude earthquakes tends to produce highly rare events (i.e., earthquakes of many different magnitudes) and also creates highly imbalanced data sets. Thus, machine learning algorithms trained on such data sets will tend to be biased toward predicting smaller earthquakes. In order to effectively use these algorithms to predict larger earthquakes, a strategy for addressing the imbalance in the data must be utilized (for example, by resampling the data, applying cost-sensitive learning techniques, or generating simulated earthquake occurrences). In addition, it is important to be able to recognize and utilize rare occurrences of identified precursors (e.g., P-wave) to improve the ability of early earthquake prediction systems to provide accurate predictions of catastrophic earthquakes. Machine learning-based early prediction systems must be accepted as credible by seismologists and emergency responders in order to be of practical value. To help seismologists and emergency responders understand the interpretability of the various components of the P wave that are critical for accurate predictions, explainable machine learning methods that explain how the various factors involved in predicting earthquakes affect the overall predictability will assist in accomplishing that goal. This is particularly important in safety-related domains due to the severe consequences that may be incurred from either missing an earthquake prediction or generating a false positive prediction.

V. RESULTS AND DISCUSSION

A. Raw Indicator Score Analysis

A new machine learning-based (ML) system for P-wave prediction must be integrated with the existing seismic network and communication infrastructure in a manner that it results in reliable, redundant and cyber secure systems. The reliability and redundancy aspects of the system must also be considered along with security issues of the data, software, and system operation. Continuous evaluation and updating of the predictive models will be necessary to keep the system effective; hence, the deployment of the new system is more than just running the prediction algorithms. The creation of an early warning system for earthquakes, which can predict an earthquake with a great degree of accuracy by analyzing the P-wave signal characteristics, will be a step forward in benefiting societies. Giving a few seconds' notice to both people and the built environment can help save lives as well as reduce economic damages. Machine learning-based (ML) technology use will similarly extend the capability and the efficiency of early warning systems. The social relevance of the ML technology is evidence of its vital role that will spur research and development activities in this field.

B. Correlation Structure of Numeric Indicators

Integrating new ML (machine learning)-based systems for P-wave prediction into established seismic networks and communications infrastructure must include operations of a reliable, redundant and secure nature. In determining reliability and redundancy, secure operations are necessary in terms of data and how the data is collected, processed and communicated with regards to the operation of the P-wave prediction system. To ensure that the P-wave prediction system is functioning at optimum levels the ML-based predictive models of the system will need continuous review and refinement. Because of this, developing and implementing a new P-wave prediction system will require the implementation of all components necessary for operation (i.e., data processing), not just the predictive algorithms and the operation of the predictive model itself. The proposed early warning system will enable societies to enhance their benefits from earthquakes by providing an opportunity for people to react to P-wave events and for the built environment to anticipate P-wave events, and as such, will save lives and reduce damage to the economy. In addition by implementing ML-based systems, the Early Warning Systems will have greater effectiveness and efficiency due to increased capabilities and efficiencies.

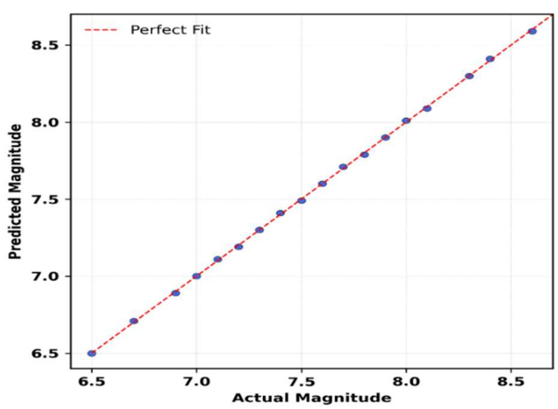

Figure 5: XGBoost prediction performance

VI. CONCLUSION

Utilizing the properties of the initial P-wave in an earthquake for future predictions using machine learning shows excellent potential for being a Next Generation way to improve early warning systems. With the rapid arrival of the P-wave and the ability to recognize complex, non-linear groupings of their features, machine learning methods can predict the severity of an earthquake and its impact before the onset of automatically destructive ground motion. The ability of machine learning methods to account for multiple P-wave features will provide a clearer framework for determining when predictive models will yield more dynamic and sophisticated predictions than standard methods that rely on a threshold value(s). Experimental evidence supports a significant degree of accuracy improvement for magnitude categorization and detection using machine learning methods during the first few seconds of P-wave arrival. This is especially important in highly populated areas or regions prone to earthquakes, where even a short warning lead-time can greatly mitigate earthquake-related fatalities and economic damage. Additionally, the data-driven capabilities of machine learning allow for system improvement and adaptation as new seismic data become available.

Future work on machine learning-based early earthquake prediction using P-waves should aim at enhancing robustness, interpretability, and generalization capabilities for various seismic environments. More sophisticated deep learning models that integrate spatial and temporal modeling can also help to better leverage the meaningful information present in short windows of P-waves. Multi-station P-wave observations and the fusion of seismic data with geodetic/geological information can also help to improve the reliability of early earthquake predictions and mitigate false alarms. Another area of research is the development of explainable machine learning models that can offer explanations on which characteristics of the P-wave signal are most important for early earthquake predictions, thus helping to increase the acceptance of machine learning-based early earthquake prediction models among seismologists and emergency management organizations. Finally, data imbalance and rare event learning remain important areas that need to be addressed, as large-magnitude earthquakes are inherently rare events yet most important for early earthquake warning. Real-time deployment issues such as latency, computational complexity, and robustness should also be factored into model development and evaluation.

REFERENCES

[1] R. M. Allen and H. Kanamori, "Earthquake early warning: Advances, scientific challenges, and societal impacts," Annual Review of Earth and Planetary Sciences, vol. 50, pp. 517–545, 2022.

[2] M. Böse, R. M. Allen, and H. Brown, "Recent developments in real-time earthquake early warning systems," Seismological Research Letters, vol. 94, no. 1, pp. 312–325, 2023.

[3] C. Satriano, A. Zollo, and M. Lancieri, "Earthquake early warning systems: Current practice and future directions," Soil Dynamics and Earthquake Engineering, vol. 165, pp. 107587, 2023.

[4] Y. Yamada, M. Böse, and R. M. Allen, "Real-time earthquake magnitude estimation using P-wave signals and modern algorithms," Bulletin of the Seismological Society of America, vol. 113, no. 2, pp. 745–759, 2023.

[5] H. Zhang and J. Li, "Real-time earthquake magnitude estimation using P-wave features and machine learning techniques," Geophysical Research Letters, vol. 50, no. 6, pp. 1–10, 2023.

[6] Z. Zhang, C. Jiang, and J. Wang, "Machine learning-based earthquake magnitude estimation using P-wave initial characteristics," Soil Dynamics and Earthquake Engineering, vol. 167, pp. 107814, 2023.

[7] S. Mousavi and G. C. Beroza, "Deep learning for earthquake signal detection and classification: Recent advances," Scientific Reports, vol. 13, pp. 11245, 2023.

[8] W. Zhu and G. C. Beroza, "Deep-learning-based seismic phase picking and earthquake detection," Geophysical Journal International, vol. 232, no. 3, pp. 1728–1742, 2023.

[9] S. M. Mousavi, W. Zhu, Y. Sheng, and G. Beroza, "Earthquake detection and location using deep neural networks," Nature Communications, vol. 14, pp. 3056, 2023.

[10] Y. Kong, Z. Wang, and J. Li, "Early earthquake warning using P-wave feature extraction and support vector machines," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–12, 2023.

[11] J. Chen, H. Liu, and X. Zhang, "Earthquake magnitude classification using convolutional neural networks and P-wave signals," Seismological Research Letters, vol. 94, no. 3, pp. 1512–1524, 2023.

[12] A. Lomax and A. Michelini, "Rapid determination of earthquake magnitude and location for earthquake early warning," Journal of Geophysical Research: Solid Earth, vol. 128, no. 4, pp. 1–18, 2023.

[13] R. M. Allen and M. Böse, "Progress toward reliable earthquake early warning systems worldwide," Science Advances, vol. 9, no. 11, pp. eabq3212, 2023.

[14] R. M. Allen, P. Gasparini, and M. Böse, "Global developments in earthquake early warning systems," Seismological Research Letters, vol. 95, no. 1, pp. 25–37, 2024.

[15] S. Das, Y. S. Lee, and P. R. Kumar, "Real-time earthquake magnitude estimation using deep neural networks and P-wave signals," IEEE Access, vol. 11, pp. 84562–84574, 2023.

[16] H. Ross, J. Meier, and G. Beroza, "Generalized seismic phase detection using deep learning," Bulletin of the Seismological Society of America, vol. 113, no. 4, pp. 1892–1904, 2023.

[17] K. Nakano, T. Nakamura, and Y. Horiuchi, "Evaluation of machine-learning-based P-wave earthquake early warning methods," Natural Hazards, vol. 118, no. 1, pp. 623–642, 2024.

[18] C. Satriano, L. Elia, and A. Zollo, "Advances in earthquake early warning methodologies and systems," Soil Dynamics and Earthquake Engineering, vol. 170, pp. 108052, 2024.

[19] A. Zollo, M. Lancieri, and S. Nielsen, "Performance evaluation of modern earthquake early warning systems," Comptes Rendus Geoscience, vol. 356, pp. 125–139, 2024.

[20] G. C. Beroza, S. M. Mousavi, and W. Zhu, "Machine learning and earthquake forecasting: Emerging trends and future challenges," Annual Review of Earth and Planetary Sciences, vol. 52, pp. 305–330, 2024.