Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.080044

ARTICLE

A Metaheuristic Football Optimization Algorithm Integrated with Large Language Models for Automated Seismic Time-Series Modeling

Amal H. Alharbi[1], Marwa M. Eid[2,*], Nima Khodadadi[3], Ebrahim A. Mattar[4] and Sayed Elkenawy[5,6]

[1]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[2]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt
[3]Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA
[4]College of Engineering, University of Bahrain, Sakhir, Bahrain
[5]School of ICT, Faculty of Engineering, Design and Information & Communications Technology (EDICT), Bahrain Polytechnic, Isa Town, Bahrain
[6]Jadara Research Center, Jadara University, Irbid, Jordan

*Corresponding Author: Marwa M. Eid. Email: mmm@ieee.org

Received: 02 February 2026; Accepted: 15 April 2026; Published: 27 May 2026

ABSTRACT: Seismic time series forecasting remains challenging due to the nonlinearity, non-stationarity, and noise of earthquake data, and because deep learning models are sensitive to preprocessing and hyperparameter settings. Although recent studies have improved neural architectures and optimization techniques, preprocessing is often treated as a fixed or manually designed stage, with limited integration into model optimization. To address this, this paper proposes an integrated, data-driven modelling framework that combines guided preprocessing with systematic hyperparameter optimization for seismic prediction, specifically forecasting earthquake magnitude from seismic catalog time-series data, with experiments conducted on Canadian seismic records. The method uses a Large Language Model to guide data preparation and feature engineering, rather than fully automate them, and applies deep learning-based forecasting with the N-HITS architecture, optimized via metaheuristic-assisted feature selection and hyperparameter tuning. The Football Optimization Algorithm (FbOA), employed as a metaheuristic optimization strategy in this study, is evaluated and compared with several well-known optimizers under identical conditions. The results show significant performance gains, with FbOA achieving superior accuracy, robustness, and convergence compared to baseline and competing methods. Notably, error metrics are reduced (MSE 3.10×10^{-7}, RMSE 5.57×10^{-3}), with high performance indicators ($r = 0.982$, $R^2 = 0.979$, NSE = 0.981, WI = 0.985). These results highlight the value of integrating guided preprocessing with optimization and demonstrate a scalable framework for high-precision time-series prediction in geophysical and related domains.

KEYWORDS: Seismic time-series forecasting; large language models; metaheuristic algorithms; football optimization algorithm; earthquake modeling

1 Introduction

The rapid increase in seismic monitoring systems, coupled with the ongoing growth of earthquake catalogs over many decades, has provided unparalleled data power for earthquake analysis and prediction [1–3]. Today, seismological repositories can record seismic events at high temporal resolution over extensive geographic areas, enabling large-scale research into patterns in seismicity behavior and the construction of computational models for seismic hazard evaluation [4–6]. Simultaneously, machine learning (ML) has

become a standard paradigm of predictive modelling, with flexible nonlinear function approximation and the ability to learn large-scale datasets, and with possibilities to capture complex interactions between heterogeneous variables in a scientific and engineering domain, declining to a standard paradigm [7–11]. These advances have prompted research aimed at applying ML methods to seismological computations, such as spatiotemporal pattern analysis, magnitude estimation, event characterization, and predictive time-series modeling [12–14]. In this study, the prediction task is specifically defined as forecasting earthquake magnitude from seismic catalog time-series data, rather than modeling full ground-motion waveforms or the complete physical earthquake process.

However, the growing volume and complexity of seismic datasets do not necessarily translate into better model performance. Earthquake catalogs are essentially observational datasets [15–17] compiled from measurements taken in distinct tectonic environments and under varying conditions of station coverage, sensor technologies, and institutional processing standards, which can introduce substantial heterogeneity in data completeness, consistency, and reliability across regions and time periods [18,19]. This heterogeneity presents problems unique to those encountered in curated benchmark datasets commonly used in machine learning research [18,19]. In practice, the accuracy of ML-based seismic modeling heavily relies on the likelihood of raw observational records, in terms of their representation and transformation into a feature space that can be learned effectively [18,20]. By extension, data quality and preprocessing cannot be regarded as habitual decision-support procedures; instead, they determine the statistical properties of the dataset and directly impact the representational learning ability of ML models [19,20].

Seismic catalogs often contain incomplete values, inconsistent attribute reporting, and mixed data types, including numerical, categorical, and text. These properties can be explained by variations in the reporting period, the availability of instruments, regional processing bilaterals, and a lack of knowledge about the localization of events and parameter estimates [21–23]. These attributes pose significant challenges for ML systems, which tend to be sensitive to missingness mechanisms, scale distortions across features, and time-varying distributional shifts [18,19]. Although the modeling architecture is well-designed, unstable optimization, inaccurate parameter estimation, and poor generalization performance can be encountered due to the presence of missing values poorly treated or the existence of under-scaled (or over-scaled) features [18,20]. Thus, preprocessing is an important process in which observational seismic data are converted into mathematically coherent forms that can be learned effectively [19,20].

The significance of data quality is particularly strong in seismic applications since the process of an earthquake is nonlinear, non-stationary and spatially correlated [18,19]. The seismic time series can be characterized by sudden bursts, protracted silence, regime transitions, and intricate clustering patterns, in contrast to most engineering datasets, where the noise is often assumed to be close to homoscedastic and the time series patterns are stationary [18]. Moreover, seismic variables are usually used to represent physical processes and observation constraints. Some of the features, including mmi, cdi, gap, and nst, reflect this duality: they represent the information that is connected to the intensity of shaking, the station coverage in the azimuth, and station density: physical phenomena of events and the environment of measurement. These variables have domain significance; their associations may provide important insights into the reliability of the catalog and the event descriptions [19]. Preprocessing options that do not account for these semantics, e.g., indiscriminate normalization, naive deletion, simplistic imputation, etc., may distort physical relationships and compromise the interpretability and validity of the learned representations [18,20].

More generally, preprocessing decisions have inductive biases because they specify which parts of the data are amplified, suppressed, or restructured prior to learning [19]. The feature distributions and interdependencies can be fundamentally transformed by options for scaling, encoding, aggregation, discretization, and transformation. This is the most important issue in spatiotemporal seismic modeling, where learning

performance is not only a matter of the quality of individual features but also of preserving temporal continuity and spatial coherence [18]. Earlier studies have demonstrated that preprocessing schemes can have a strong impact on optimization stability, convergence behavior, sensitivity to noise, and sensitivity to rare but high-impact events [18,20]. In extreme events and highly variable datasets, although relatively rare, such as those related to earthquakes, preprocessing effects can dictate the success of learning algorithms in capturing rich and meaningful patterns, or, rather, they contribute to long-term outages and sampling projections [18,19].

Irrespective of these considerations, preprocessing in seismic ML pipelines is largely a human exercise, driven by heuristics [19]. Common techniques tend to assume consistent rules for handling missing values, generic scaling techniques for homogeneous variables, and traditional representations of categorical data, such as alert, tsunami, and magtype. Although these types of heuristics provide ease, they may incorporate subjective assumptions and may not extrapolate across data sets with varying regional characteristics, reporting conventions, or station-network geometry [18,19]. Additionally, the manual design of preprocessing pipelines is impractical and only gets worse with increased dimensionality and dataset size, as well as the complexity of modeling tasks, which now demand more advanced deep learning architectures sensitive to data structure [18,19]. Such restrictions encourage the creation of automated preprocessing systems that can reason about the semantics of datasets, the processes that generate data, and data modeling goals [19,24,25].

Large Language Models (LLMs) have offered a promising path in this direction, enabling semantically and context-aware reasoning over descriptions of datasets [26–28]. In contrast to rule-based systems, LLMs can learn natural-language metadata and attribute specifications, infer probabilistic relationships between variables, and even suggest the logic behind transformations in a manner that is both statistically powerful and semantically meaningful. This, in principle, enables the generation of preprocessing pipelines dynamically based on a dataset's structure rather than templates. Such capability is especially appealing for seismic modeling, which, in most cases, can demand domain-specific interpretation of physical properties and measurement signs. Incorporating this interpretive layer into the preprocessing step can minimize human bias, achieve greater consistency, and improve transferability between datasets and across seismic regions when generating LLM-based pipelines.

Nevertheless, the use of LLMs for code generation, the automation of exploratory data analysis, and overall machine learning rationales remain underresearched for guided preprocessing in seismic data analysis. Recent studies are not systematic in assessing how LLMs produce preprocessing pipelines, the degree of coherence and context-sensitivity of those pipelines, and their effects on subsequent model performance. Simultaneously, AutoML systems have advanced in model selection and hyperparameter optimization, but generally treat preprocessing as a fixed initial task or a restricted operator-selection task. These systems are constrained by fixed transformation spaces and may not semantically interpret seismic variables in a domain-specific manner. This reveals a research gap in controlled studies that evaluate reasoning-based preprocessing design for geophysical datasets and its downstream effects.

To fill this gap, the current paper explores the use of an LLM as a guided preprocessing assistant for earthquake magnitude forecasting from seismic catalog time-series data, together with deep learning forecasting and metaheuristic optimization. The study addresses three questions: (i) whether an LLM-guided approach can construct coherent and context-aware preprocessing pipelines for seismic data, (ii) the degree to which preprocessing guided by the architectural factors affects the performance of the baseline model, and (iii) whether feature selection and metaheuristic optimization can further help to improve the results under the best preprocessing pipeline. Overall, the paper promotes preprocessing as an intelligent, dataset-specific design task within data-centric predictive modeling.

The contributions of the research are as follows:

- In this work, we propose an integrated seismic time-series modeling framework for earthquake magnitude forecasting that combines LLM-guided preprocessing, deep learning forecasting, and metaheuristic optimization, thereby lessening manual heuristics and trial-and-error design decisions.
- The LLM-guided context-aware preprocessing pipeline interprets seismic attributes semantically and systematically prepares time-series data for learning based on temporal, spatial, and physical features, serving as a guided support tool rather than a fully autonomous preprocessing mechanism.
- The research thoroughly evaluates the proposed framework by comparing it with state-of-the-art deep learning models and popular metaheuristic optimization algorithms, using a common experimental protocol to ensure fairness and reproducibility.
- The combined task of feature selection and hyperparameter optimization is addressed using population-based metaheuristic algorithms, with special regard to the Football Optimization Algorithm, which demonstrates the efficiency of concurrent optimization policies in strengthening models and improving generalization.
- A comprehensive comparative analysis is conducted between baseline and optimized settings to highlight the value added of guided preprocessing and systematic optimization for non-stationary, very complex seismic data.
- The proposed methodology creates a scalable, modular design paradigm that can be easily extrapolated onto other geophysical and data-intensive time-series applications and serves as a basis for future studies on intelligent, end-to-end machine learning systems.

The rest of this paper is intended to discuss these goals systematically. After this introduction, the theoretical background and associated literature in the area are surveyed, including seismic data modeling, LLM-guided preprocessing, and optimization methods. Then, the system configuration and experimental setup are presented in detail, including the dataset and model features. The following sections discuss data-centric seismic machine learning in detail, including its analysis and implications, limitations, and future research directions.

2 Literature Review

Deep learning (DL) has become one of the dominant approaches in earthquake science, previously transforming the way seismic hazards are tracked, interpreted, and addressed. In disaster assessment, early warning, forecasting, structural response analysis, and infrastructure resilience, DL-based methods are evidently superior to traditional, rule-based, and empirically parameterized methods. The review below summarizes the existing research solely on the basis of the provided abstract, outlining trends in method use and areas of application, as well as the unresolved problems.

It has been thoroughly demonstrated that deep learning for Earthquake Disaster Assessment (EDA) has been successfully applied in large-scale systematic reviews. A comprehensive survey of 204 peer-reviewed articles indicates that DL techniques have been used at all temporal stages of the earthquake disaster, including pre-earthquake preparedness, real-time response, and post-earthquake recovery [29]. In this literature, the assessment targets can generally be categorized into disaster-related objects (e.g., earthquakes and secondary hazards) and physical objects (e.g., buildings, infrastructure, and geographical areas). CNN-based image classification prevails after earthquake damage, especially with remote sensing imagery, whereas seismic and social media information are more useful for providing quick situational awareness. Despite significant performance improvements, the review highlights persistent problems, including the lack of labeled data, interregional model generalization, and the need for multimodal learning approaches that leverage heterogeneous data sources.

Broadly speaking, the advantages of using DL techniques in earthquake prediction and forecasting have been reported as their ability to handle large-scale data and learn nonlinear patterns that are often too complex to discover using more traditional machine learning methods [30]. Its applications include the determination of earthquake magnitude, the identification of seismic signals, the identification of ionospheric electron density, and the detection of radon gas anomalies. Experimental research conducted on laboratory-generated earthquake analogs also reveals that DL architectures such as CNNs and LSTMs are capable of forecasting fault stress development, time-to-fail, and rupture behavior across various seismic schemata, including aperiodic and slow-slip phenomena [31]. Based on those progresses, the RECAST framework proposes neural temporal point processes to identify earthquake catalogs directly, without being limited by the theoretical studies of traditional statistical models, providing better forecast accuracy once adequate amounts of training data exist [32]. All these findings indicate that DL provides a scalable and adaptable approach to reliable earthquake forecasting, though no one can predict the long term.

Members of the Deep learning field have played a particularly disruptive role in Earthquake Early Warning (EEW) systems, where Earthquake parameters need to be estimated quickly and with high accuracy. The hybrid autoencoder-CNN system accurately predicts the magnitude and location using just 3 seconds of P-wave onset time, enabling almost immediate transmission of warnings via IoT-based warning systems [33]. Likewise, EEWNet shows that without manually engineered features, raw P-wave waveforms can yield faster, more accurate magnitude predictions than empirically determined peak-displacement techniques [34]. CNN predicted PGA models are also built upon on-site EEW with improved performance, namely, higher accuracy, less uncertainty, as well as better cross-regional generalization, through the automated extraction of discriminative features in short seismic windows of the data sets [35]. End-to-end fully convolutional systems build on this by simultaneously determining earthquakes and computing source parameters from continuous streams of waveforms, enabling solutions to be dynamically updated as more station data arrive [36]. Beyond demonstrating the applicability of DL-based EEW to critical infrastructure systems such as high-speed railways, transformer-enhanced CNN architectures can achieve high alert accuracy and meaningful lead times while operating within operational constraints.

In addition to early warning, DL has made significant contributions to seismic signal processing and event detection. Installations that consist of dense neural forms that combine phase picking, association, and detection phases can deliver superior quality over conventional discrete pipelines by maintaining information across tasks and by imposing a physical constraint on the pipeline through back-projection styles that fracture sensors into physical measurements [37]. DL-based denoising models are useful in urban settings with high levels of anthropogenic noise to remove noise and denoise low-SNR seismic signals, contributing to earthquake localizations and understanding fault behavior in the subsurface, and, furthermore, resulting in greater fault understanding by limiting sensor noise [38]. Reliability of popular detection systems, like EQTransformer, has also been methodically tested, as in doubt about coming up with consistent and replicable list of earthquakes, then the use of uncertainties in the inferences need to be implemented successfully to stabilize outcomes of these queries and unitary frameworks of earthquake classification and forecasting approaches are necessary to gain improved results and outcomes of inquiries [39].

The use of deep learning has also enabled the successful characterization of earthquake sources and the estimation of earthquake magnitude. FMNet demonstrates that it is possible to estimate focal mechanisms in both real time and on the fly by synthetically training datasets as quickly as possible, allowing characterization of fault geometries in relief areas with no seismic history (as of today) [40]. Relative analysis of DL architectures to estimate magnitudes in real-time also shows that when the S-wave information is taken into consideration, prediction results are much more accurate, and the additional seismic phase representations are also important to be taken into account simultaneously with them, enhancing accuracy and preventing

the underreporting of classified seismic phenomena [41]. Other sensing modalities, such as electromagnetic and acoustic emission signals, have also been used by CNN-based models to categorize earthquake magnitudes and identify precursory signals to address data imbalance and noise challenges [42,43].

In the field of structural response and damage assessment, DL can swiftly assess the impacts of earthquakes on the built environment at a scalable level. CNN-based multi-input neural networks are developed to evaluate the cumulative damages in reinforced concrete buildings and achieve almost perfect accuracy of damage state classification with just a regular version of the mainshock-aftershock sequences [44]. Time-frequency analysis assisted by 1D-CNNs and Bayesian optimization also optimizes post-earthquake damage diagnosis, especially when limited training data are available [45]. Lightweight CNN-based transfer learning methods achieve high accuracy on damage images, particularly for distinguishing structural damage, and model interpretability techniques such as Grad-CAM improve model transparency [46]. These and other related studies demonstrate the growing application of DL models in infrastructure, highlighting their potential for supporting diverse infrastructure-related analysis, monitoring, and decision-making tasks, such as real-time prediction of earthquake-induced roadway damage using automated segmentation of that damage [47], prediction of seismic reactions of high-rise buildings under seismic excitation by stochastic vibration analysis of supersystems of vehicles and bridges, and real-time prediction. Complex dynamic effects and uncertainty propagation, e.g., earthquake-fire interactions, are also addressed by hybrid DL frameworks that span complex dynamics and the population and propagation of uncertainty [48].

Lastly, DL has already begun shaping the effects on earthquake risk mitigation at the societal and educational levels. YOLO-based object detection systems enable real-time detection of hazardous interior objects during earthquakes and can be used to interactively learn disaster preparedness and risk awareness on a disaster preparedness and risk awareness training platform [49]. The applications demonstrate the growing role of DL in areas beyond scientific analysis, with direct contributions to public safety and education.

All in all, the reviewed studies show that deep learning has turned into an inseparable component of the entire lifecycle of earthquake research. DL-based technology greatly improves the efficiency, reliability, and scalability of an earthquake response system and disaster management infrastructure by facilitating end-to-end education, overcoming limitations in its environment/handcrafted features, and motivating it to infer timely information. However, issues pertaining to data availability, the measure of uncertainty, interpretability, and operational implementation are critical research areas to be pursued in the future. The studies reviewed cover the entire continuum of earthquake management, from hazard knowledge and forecasting to real-time early warning, catalog construction, and, afterward, structural and infrastructure evaluation. To explain the way that each contribution can be alternatively incorporated into this ecosystem, Table 1 summarizes the research focus, core methodological design, and the major findings and contributions found in each abstract. This hierarchical comparison brings to the fore methodological trends and allows defining the clear distinction between (i) end-to-end waveform learning to earthquake early warning, (ii) network- and catalog-scale monitoring models, (iii) source characterization and precursor-based learning, and (iv) engineering-oriented damage-response-safety applications.

Table 1: Condensed synthesis of deep learning applications across earthquake monitoring, early warning, forecasting, and damage assessment.

Ref.	Focus	Methodology	Key Contributions
[29]	EDA overview	Systematic review (204 studies)	Defines DL landscape for EDA; CNN dominance in damage detection; highlights data scarcity and multimodal opportunities.

(Continued)

Table 1 (continued)

Ref.	Focus	Methodology	Key Contributions
[30]	Prediction survey	Review of ML/DL forecasting studies	Shows increasing reliance on DL for magnitude and precursor analysis due to scalability and representation learning.
[33]	EEW (M, loc)	AE–CNN on 3 s waveforms	Enables near-instant magnitude/location estimation; suitable for IoT-based EEW deployment.
[34]	EEW (M)	End-to-end DL on raw P-wave	Faster and more accurate magnitude estimation than Pd without feature engineering.
[35]	EEW (PGA)	CNN on early P-wave	Improves PGA prediction accuracy, timeliness, and cross-region generalization.
[40]	Source mechanism	FMNet (synthetic training)	Real-time focal mechanism estimation; transferable to data-sparse regions.
[36]	EEW streams	Fully convolutional network	Joint detection and parameter estimation from continuous waveforms.
[37]	Detection pipeline	End-to-end multi-station DL	Improves detection via joint optimization and kinematic constraints.
[49]	Public safety	YOLO-based indoor detection	Real-time identification of hazardous indoor objects for earthquake education.
[38]	Urban denoising	DL waveform denoiser	Recovers low-SNR signals and improves urban earthquake localization.
[31]	Lab forecasting	DL + AR sequence models	Predicts fault stress and failure timing beyond single seismic cycles.
[32]	Catalog forecasting	Neural temporal point process	Outperforms ETAS with long catalogs; scalable seismicity modeling.
[42]	EM precursors	CNN + SMOTE	Accurate magnitude classification under noisy, imbalanced EM data.
[43]	AE precursors	DL on AE time series	High-accuracy real-time seismic precursor detection.
[50]	InSAR review	DL architectures survey	Reviews CNN/RNN/GAN/Transformer use in deformation detection.
[51]	Building EEW	DNN surrogate model	Real-time prediction of high-rise seismic responses.
[52]	Damage vision	Transfer learning + BO	Accurate and explainable post-earthquake damage classification.
[53]	Multi-hazard	Wavelet–LSTM	Captures earthquake–fire interaction effects on stability.
[46]	Cumulative damage	CNN multi-input model	Reliable damage indexing under mainshock–aftershock sequences.

(Continued)

Table 1 (continued)

Ref.	Focus	Methodology	Key Contributions
[48]	RC damage	Time–freq + 1D-CNN	Robust post-event damage diagnosis with data augmentation.
[39]	Detection reliability	Uncertainty-aware DL	Improves reproducibility and reduces false positives.
[44]	Magnitude models	DL comparison study	Shows S-wave inclusion improves real-time magnitude accuracy.
[45]	Road damage	Segmentation DL	Pixel-level road damage mapping using rare-event datasets.
[41]	VBS dynamics	BiGRU + attention	Predicts stochastic vehicle–bridge responses with uncertainty.
[47]	Railway EEW	CNN–Transformer	High-accuracy alerting for rail systems within 10 s of P-wave.

3 Experimental Framework

The proposed framework is built on five methodological principles. First, preprocessing is treated as a primary design component rather than a fixed preliminary step, since seismic forecasting performance depends strongly on how temporal, spatial, and physical attributes are represented. Second, the LLM is used as a guided preprocessing assistant to support context-aware transformation design, rather than as a fully autonomous decision maker. Third, model learning and optimization are evaluated under controlled and identical experimental conditions so that performance differences can be attributed to preprocessing and optimization quality rather than procedural variation. Fourth, feature selection and hyperparameter optimization are addressed jointly through metaheuristic search in order to improve both representational efficiency and predictive performance. Finally, reproducibility is maintained through standardized prompts, fixed decoding settings, shared data splits, and common evaluation metrics across all experiments. This part outlines the experimental design used in the current work, including details on the earthquake dataset properties, the architecture of the Large Language Model (LLM) system, the protocol used to generate the preprocessing pipeline, the architecture of the baseline model and model training scheme, and the evaluation process. The experimental design was designed to be fair, reproducible, and robust against preprocessing effects on modeling and optimization. The general methodological procedure followed in this study is schematically shown in Fig. 1. The figure provides a detailed overview of the designed end-to-end modeling pipeline, which starts with raw earthquake data and proceeds through automated preprocessing, model training, model optimization, and performance evaluation. The diagram in Fig. 1 indicates that the first step is to input an earthquake dataset, which must undergo an initial intelligent data preprocessing stage mediated by Large Language Models (LLMs). The aim of this stage is to optimize data quality, ensure semantic consistency, and transform temporally structured sources into a format usable by advanced forecasting models. After preprocessing, the data is divided into training and test sets to ensure objectivity in model assessment. A set of baseline deep learning models is subsequently trained on the ready data, with numerous metaheuristic algorithms being used to optimize feature selection and hyperparameters. The key in this optimization layer is the proposed Football Optimization Algorithm (FbOA), which interacts with the baseline models and competing optimizers to optimize model parameters. Lastly, the refined models can

be tested using end-to-end performance measurements, completing the modeling loop. Fig. 1 simplifies the process of presenting the logical dependencies and interaction among preprocessing, learning, optimization and evaluation phases in the proposed system by visually integrating all the key elements and data flows.

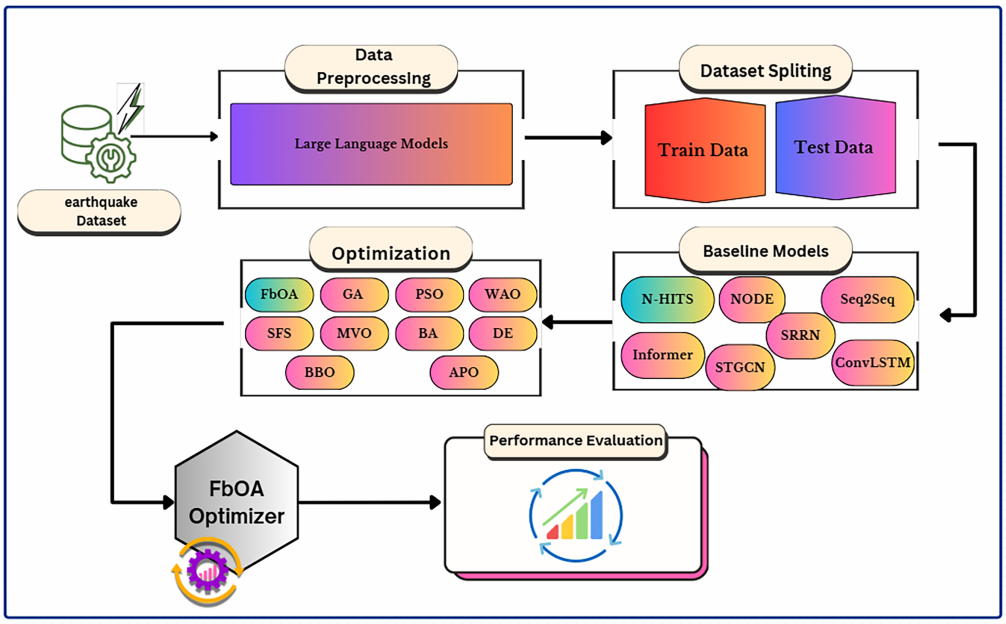

Figure 1: Overview of the proposed LLM-guided preprocessing and FbOA-driven optimization framework for seismic time-series modeling.

3.1 Dataset Overview

The dataset used in this paper is geographically limited to Canada and contains 101,365 earthquake records from 1 January 1985 to recent years, with each record corresponding to a single seismic event. The dataset incorporates time, place, space, and physical characteristics of earthquakes, which provides a basis for comprehensive seismic analysis, straightforward machine learning, and straightforward modeling. The time coverage allows studying long-term trends and assessing seasonality, whereas the spatial and physical features support the study of seismicity in the region and time-series forecasting.

Each earthquake record includes a timestamp in ISO-8601 format, the earthquake's date and time. It is this quality that serves as the basis for time ordering, seasonal breakdown, and time series modeling. Geographic data is in the form of latitudes and longitudes, expressed in decimal degrees, which together determine the exact location (surface) of the epicenter of the earthquake and can be used to map it spatially and analyze distribution patterns in the area.

An additional datum would be depth, measured in kilometers, which indicates the depth of the underground source of the seismic tremor. The depth of an earthquake is one of the vital physical parameters, which are usually correlated with the intensity of shaking on the ground and the probability of surface damage. The magnitude attribute provides a logarithmic representation of earthquake magnitude and energy release, which is a major aim variable in seismic modelling. Also, the magnitude type attribute specifies the scale or method used to compute the reported magnitude, either ML or MN, based on regional jurisdiction and available data.

To improve association, the dataset includes a textual place description that identifies the approximate position of each earthquake relative to familiar cities or features. Although it is not as readable to machines, this feature can provide contextual geographic data to aid exploratory analysis. An abbreviated geographic classification is derived from the column labeled "Unnamed: 7", which corresponds to a province or regional identifier (e.g., NB, BC, QC) extracted from the dataset and used to support spatial grouping. In contrast, the column labeled "Unnamed: 8" contains inconsistent or sparsely populated entries (e.g., missing values or non-informative labels such as "AFTERSHOCK"), resulting in limited analytical value; therefore, it is excluded during preprocessing.

In general, the dataset meets the standards for earthquake catalog equipment used by reputable seismic monitoring institutions, such as the United States Geological Survey and national seismological services. This is because it integrates temporal, spatial, and physical variables, which are highly suited to seismic hazard assessment, spatiotemporal pattern analysis, and earthquake modeling using machine learning.

3.2 LLM System Configuration

The experimental design was intended to test the ability of Large Language Models (LLMs) to serve as guided preprocessing designers under highly constrained conditions. All experiments were carried out with the same prompting templates, consistent decoding settings, and the same execution environments to ensure methodological rigor and eliminate confounding factors. This design will ensure that any perceived variations in preprocessing reasoning and downstream modeling behaviour are caused solely by the LLM's reasoning and generative capacity, and not by prompts, other randomness, or external system configuration variations.

The LLM used in this study is Qwen/Qwen3-Coder-480B-A35B-Instruct, which is accessed via its publicly available implementation. Qwen3-Coder-480B-A35B-Instruct is a large-scale causal language model trained over a number of months in pre-training and post-training processes, with special focus on code generation, guided reasoning support, and long context comprehension. This model has a total of 480 billion parameters, of which 35 billion are activated during inference, enabling it to achieve expressive representational power and computational efficiency through selective parameter use.

The model is structurally divided into 62 transformer layers and uses Grouped Query Attention (GQA), with 96 attention heads for query projections and 8 for key-value projections. This structure is more efficient in terms of attention and exhibits contextual faithfulness in long sequences. Besides, the model includes a mixture-of-experts (MoE) configuration with 160 specialists, of which 8 are dynamic and activated at each inference step. This specialist-directed strategy enables the model to tailor its decision-making processes to task context, which can be especially beneficial for complex, multi-step processes such as dataset deciphering and assisting in structuring a preprocessing pipeline.

The major advantage of Qwen3-Coder-480B-A35B-Instruct is that it can support very long context length, up to 262,144 tokens, by default. This ability is essential to data-centric machine learning: it enables data description, attribute definition, statistical summaries, unstructured procedural instructions, and more to be presented to the model as part of one rationality. This long-context ability in the current paper enabled the LLM to think structurally about the seismic dataset format and to assist in formulating preprocessing logic that accounts for temporal relationships, physical semantics, and modeling goals, without chopping off or losing its contextual integrity.

The LLM was presented as a guided calculation assistant that takes the dataset metadata and assists in producing preprocessing logic in natural language and in a formatted procedural language. Prompting templates were optimally standardized across all experimental runs to ensure they were par. The inference

parameter models, such as temperature, top-k, and decoding strategy, were set to deterministic values to remove stochastic variability and improve reproducibility. With this managed arrangement, the paper provides a principled analysis of the LLM's capability to deduce the semantics of seismic data and to assist in producing coherent, context-driven preprocessing pipelines to downstream time-series modeling.

3.3 Pipeline Generation Protocol

This paper used only one Large Language Model to assist in producing an integrated preprocessing and time-series pipeline for seismic data. The pipeline was based solely on the interpretation of dataset structure, attribute semantics, and temporal modeling objectives generated by the model, under controlled prompting and guidance, and did not use predefined preprocessing templates, rule systems, or manually developed heuristics. It is a style of preprocessing that relies on reasoning, encouraging transformations driven by semantic and statistical factors rather than procedural orthodoxy.

The resulting preprocessing protocol reflects a sensible, internally consistent plan to prepare the earthquake data for teaching with advanced time-series learning. All stages of the pipeline play specific roles in the analysis, ensuring that the temporal structure, physical significance, and statistical characteristics of the data are maintained and represented appropriately.

Date–Time Feature Extraction: The first step in the preprocessing pipeline was the systematic extraction of features from the original date attribute, which was in ISO-8601 format. As the seismic analysis was based on temporal dynamics, the LLM extracted several temporal features, such as year, month, day, hour, and day of the week. These derived variables enable multi-scale time aggregation and the discovery of long-run, seasonal, diurnal, and weekday-weekend patterns. This action enriches the model's temporal features by explicitly encoding the periodicity of all temporal and calendar structures, thereby increasing temporal dependence and the recurrence of seismic patterns.

Magnitude Categorization: To enhance interpretability and facilitate the categorical analysis of earthquake intensity, the pipeline added a derived categorical variable, *mag_category*. Earthquake events were categorized into ordinal classes that included micro, small, big, and great earthquakes that were determined through the known magnitude thresholds. This transformation not only retains the natural hierarchy of magnitude values but also allows statistical analysis within groups; the visualization can be done as a group of classes and is insensitive to extreme values. The classification enables analysis of the distribution of seismic activity across intensity measures and models that leverage both numerical and categorical mapping of the event's strength.

Stationarity Assessment (ADF Test): Since temporal modeling is the key component of the study, an official stationarity test, the Augmented Dickey-Fuller (ADF) test, was implemented in the pipeline. An ADF test was also used to analyse the presence of unit-root behaviour in the seismic time series, indicating whether the series exhibits stochastic trends or mean-reverting behaviour. Some significant outputs, such as the ADF statistic, p-value, and critical values at common levels of significance, were examined. The criterion for deciding whether to differentiate or further transform the data before model training was a p-value threshold of 0.05. This is a step that ensures temporal dependencies are appropriately defined and that modeling assumptions are grounded in statistical evidence.

Seasonality Strength Evaluation: To further characterize temporal structure, the pipeline evaluated the strength of seasonal components through time-series decomposition. Seasonality strength was quantified by comparing the residual variance to the observed variance, yielding a measure of seasonality strength. A value of 0.69 was obtained, indicating a strong, consistent seasonal pattern in the seismic time series.

This quantitative assessment provides clear evidence of recurring temporal behavior and informs modeling decisions related to seasonal representation, trend separation, and forecasting horizon design.

Collectively, integrating temporal feature extraction, magnitude categorization, stationarity testing, and seasonality analysis yielded a statistically grounded, semantically coherent preprocessing pipeline. By aligning preprocessing steps with both domain knowledge and formal statistical diagnostics, the LLM-assisted preprocessing pipeline ensures that the seismic dataset is suitably structured for advanced time-series learning. Importantly, this approach preserves the physical and temporal characteristics of earthquake activity while minimizing arbitrary or heuristically driven transformations, thereby establishing a robust foundation for downstream deep learning and optimization.

In order to examine the underlying temporal structure of the variability of earthquake magnitude, a seasonal decomposition of the time series of monthly earthquake magnitude was carried out as presented in Fig. 2. This decomposition breaks the actual series into four additive components: the original observed signal, the long-term trend, the repetitive seasonal pattern, and the residual fluctuations. The observed component is the raw time-varying magnitude, whereas the trend component captures the slow change in the background seismic activity. Seasonality is used to represent recurring patterns of intra-annual processes that may be due to cataloging or reporting processes rather than actual physical seismic processes, and the residuals are irregular short-term changes that it fails to explain by trends or seasonality. It is clear from the figure that long-term variation is smooth, the seasonal component has not changed over the years, and the residuals are concentrated around zero, indicating that most of the systematic temporal structure has been effectively captured by the decomposition.

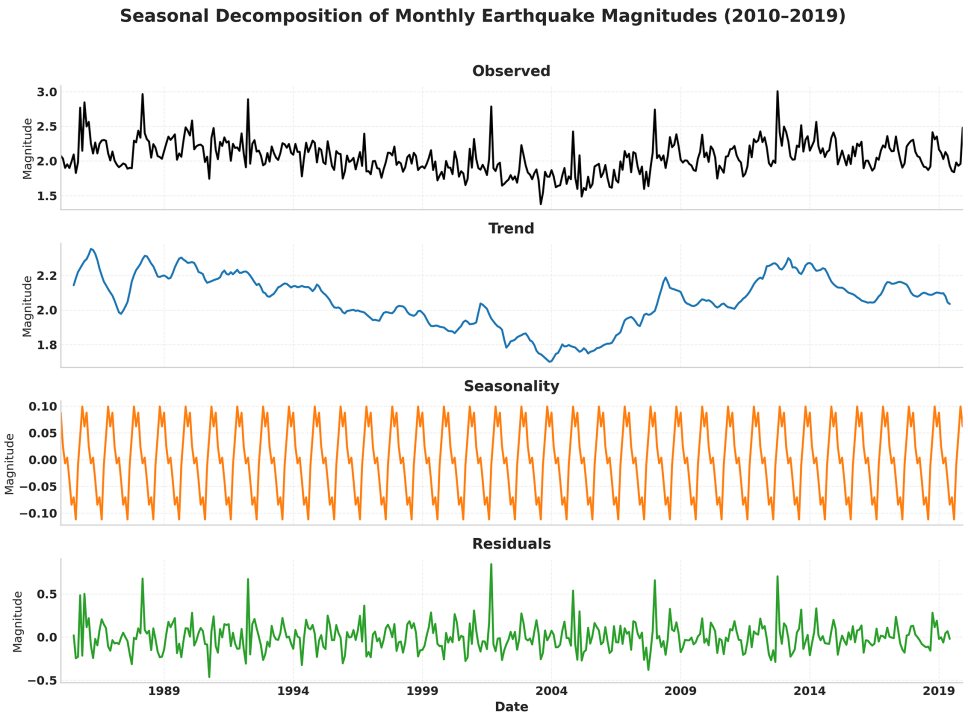

Figure 2: Seasonal decomposition of monthly earthquake magnitudes into observed, trend, seasonal, and residual components.

To study how seismic activity varies across Canada, a comparative distribution analysis of seismic magnitudes for the top five seismically active regions was performed, as shown in Fig. 3. The violin plots

combine kernel density estimation and boxplot statistics, allowing the magnitude and central tendency, as well as outlier behaviour, to be visualised simultaneously in each region. The width of every violin indicates the proportionate frequency of earthquake magnitudes, whereas the enclosed boxplots point out the median and interquartile range. As shown in the figure, a wider magnitude range and larger upper extremes are observed in the northern areas of the continent, such as British Columbia and Alaska, whereas Quebec shows a more concentrated distribution with a shorter median magnitude, highlighting the intense spatial heterogeneity of regional seismicity.

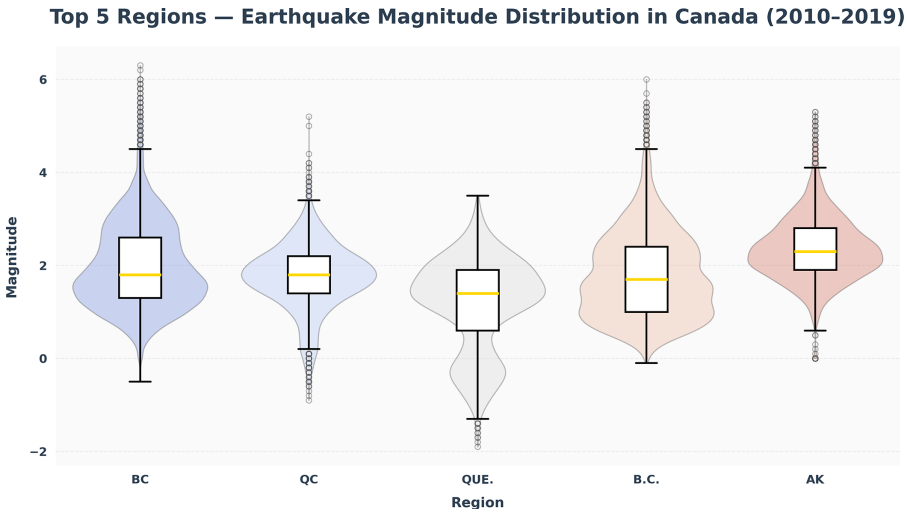

Figure 3: Violin plot of earthquake magnitude distributions for the top five seismically active regions in Canada (2010–2019).

Autocorrelation and partial autocorrelation analysis were conducted in order to determine the structure of temporal dependence and lagged relationships in the earthquake time series, as shown in Fig. 4. The autocorrelation function (ACF) measures the correlation between observations at different time lags and provides insight into the persistence and memory effects of a given series. Complementarily, the partial autocorrelation function (PACF) of a time-series (autoregressive component) correlation of a time-series at each lag by removing the cross-effect of any intervening lag to establish the correct sequence of autoregressive elements in a time-series model. As shown in the figure, the ACF shows slow decay at various lags, suggesting a high degree of temporal persistence, whereas the PACF shows a series of spikes at low lags with rapid decay, suggesting a low-order seismic process.

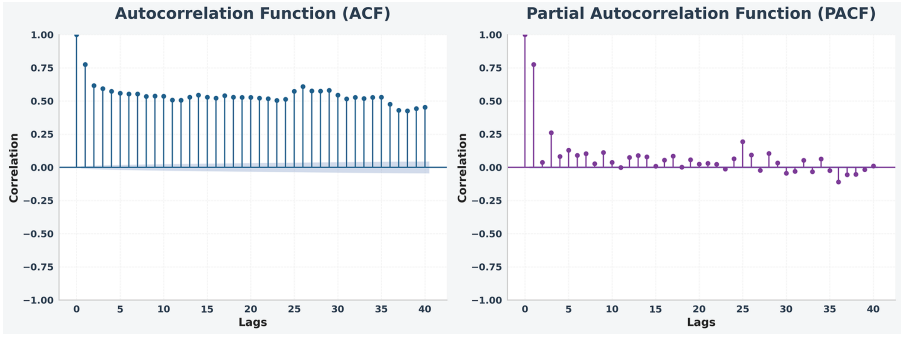

Figure 4: Autocorrelation (ACF) and partial autocorrelation (PACF) functions of the earthquake time series.

To examine temporal trends in seismic activity in Canada over a long period, an annual frequency analysis of registered earthquakes was adopted, as shown in Fig. 5. The figure illustrates the development in the number of earthquakes per year during the period, allowing us to outline interannual variability, long-term trends, and extreme activity years. Annotated reference lines indicate the year with the strongest seismic activity and the least active year, providing context for understanding changes in seismic earthquake occurrence. As shown in the figure, the trend of seismic activity has been marked by a strong upward trend after the mid-1990s, with high values in the late 2000s and early 2010s, after which the trend shifted steadily, though cyclically, to high levels.

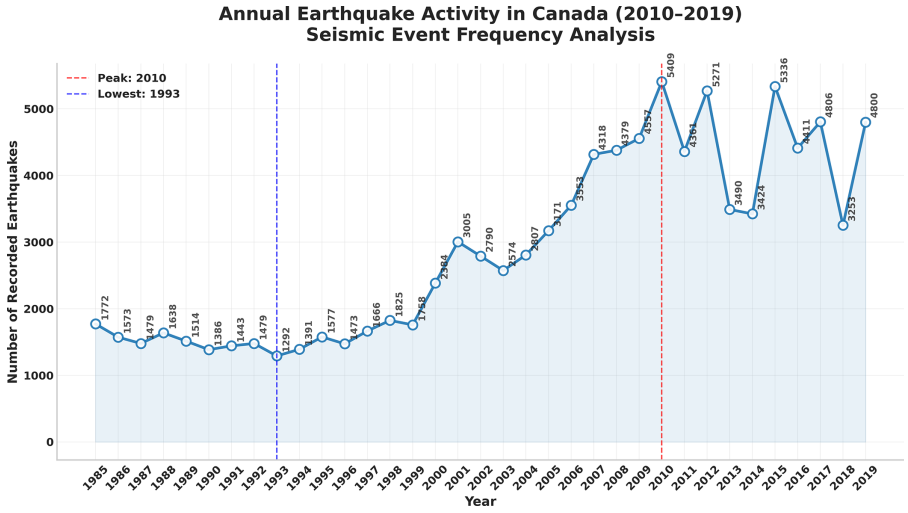

Figure 5: Annual earthquake activity in Canada showing the number of recorded seismic events per year.

In order to analyse how the central tendency and variability of seismic activity would change over time, a rolling statistical analysis of the number of monthly earthquakes was conducted, as shown in Fig. 6. The figure presents both the original monthly earthquake series and the 12-month rolling means and standard deviation, allowing assessment of long-term trends and temporal variability. The rolling standard deviation captures variations in dispersion and episodic instability, whereas the rolling mean averages out short-term fluctuations to focus on long-term changes in seismic activity. From the figure, it is clear that the mean number of earthquakes and their variance rise significantly after the beginning of the 2000s, indicating a shift towards a more dynamic and active seismic regime in the second half of the study period.

3.4 Model Architecture and Training

Time-series modeling of earthquakes is a complex learning problem due to the nature of nonlinear dynamics, irregular time variations, and the simultaneous presence of temporal, spatial and physical characteristics. To ensure the evaluation of preprocessing guided by LLMs is not biased by a particular modeling bias, this study uses a wide range of baseline architectures across various paradigms of sequential and spatiotemporal learning. The chosen models are: **Sequence-to-Sequence (seq2seq) Encoder–Decoder Neural Networks**, **N-HiTS: Neural Hierarchical Interpolation Time Series Forecasting**, **Informer**, **Neural Oblivious Decision Ensembles (NODE)**, **Spatio-Temporal Graph Convolutional Networks (STGCN), Self-Regulating Recurrent Networks (SRRN)**, and **convolutional LSTM (convLSTM)**. This heterogeneous selection enables a strict evaluation of the stability of preprocessing strategies generated under LLM-guided reasoning across architectural hypotheses at the first level.

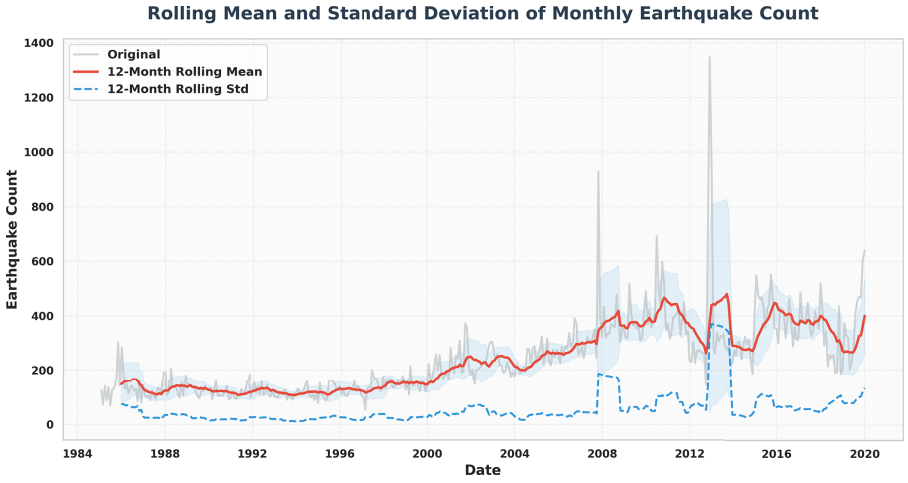

Figure 6: Rolling mean and standard deviation of monthly earthquake counts using a 12-month window.

N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting: N-HiTS addresses long-horizon forecasting through a hierarchical residual learning strategy that decomposes temporal signals across multiple resolution levels [54]. At each hierarchical stack k, the model produces a backcast component $\mathbf{b}^{(k)}$ and a forecast component $\mathbf{f}^{(k)}$ according to

$$\mathbf{b}^{(k)}, \mathbf{f}^{(k)} = g^{(k)}(\mathbf{r}^{(k-1)}), \tag{1}$$

where $\mathbf{r}^{(k-1)}$ denotes the residual input propagated from the previous stack. This design enables N-HiTS to capture both long-term trends and short-term seismic fluctuations within a unified forecasting framework.
Sequence-to-Sequence (seq2seq) Encoder–Decoder Neural Network: The seq2seq architecture formulates forecasting as a conditional sequence learning problem, in which an encoder transforms an input sequence into a compact latent representation and a decoder generates the output sequence conditioned on this representation. Let $\mathbf{X}_{1:T} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ denote a sequence of preprocessed seismic feature vectors over a temporal window of length T. The encoder updates its hidden state according to

$$\mathbf{h}_t = f_{\text{enc}}(\mathbf{x}_t, \mathbf{h}_{t-1}), \tag{2}$$

while the decoder produces the prediction $\hat{\mathbf{y}}_t$ as

$$\hat{\mathbf{y}}_t = f_{\text{dec}}(\mathbf{h}_T, \mathbf{s}_{t-1}), \tag{3}$$

where $\mathbf{s}_{t-1}$ denotes the decoder state. This formulation is well suited for capturing delayed temporal dependencies frequently observed in seismic activity.

Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting: Informer extends transformer-based architectures to long sequence modeling by employing an efficient attention mechanism that prioritizes the most informative temporal dependencies [55]. Given query, key, and value matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, the attention operation is expressed as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V}, \tag{4}$$

where d denotes the feature dimension. This mechanism enables effective modeling of long-range temporal dependencies while maintaining computational efficiency.

Neural Oblivious Decision Ensembles (NODE): NODE integrates differentiable decision tree ensembles within a neural learning framework, enabling the modeling of complex nonlinear feature interactions [56]. The ensemble prediction is defined as

$$\hat{y} = \sum_{m=1}^{M} w_m \, T_m(\mathbf{x}), \tag{5}$$

where $T_m(\mathbf{x})$ represents the output of the m-th oblivious decision tree and w_m denotes its learnable weight. This structure is particularly effective for heterogeneous feature spaces arising from seismic preprocessing pipelines.

Spatio-Temporal Graph Convolutional Network (STGCN): STGCN explicitly models spatial and temporal dependencies by operating on graph-structured seismic representations. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a graph of seismic locations, with node features $\mathbf{X}_t$ at time t. A spatial graph convolution is defined as

$$\mathbf{H}_t = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}_t\mathbf{W}\right), \tag{6}$$

where $\tilde{\mathbf{A}}$ is the adjacency matrix with self-loops, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, and $\mathbf{W}$ is a learnable weight matrix.

Self-Regulating Recurrent Network (SRRN): SRRN introduces an adaptive regulation mechanism to stabilize recurrent learning over long sequences. The hidden state update is defined as

$$\mathbf{h}_t = \alpha_t \odot f(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1}), \tag{7}$$

where α_t denotes a self-regulation factor and $\odot$ represents element-wise multiplication. This mechanism mitigates gradient instability in long seismic time series.

Convolutional LSTM (ConvLSTM): ConvLSTM extends conventional LSTM by replacing matrix multiplications with convolutional operators, thereby preserving spatial locality while modeling temporal evolution. The recurrent update is expressed as

$$\mathbf{H}_t = \mathrm{LSTM}_{\mathrm{conv}}(\mathbf{X}_t, \mathbf{H}_{t-1}), \tag{8}$$

where $\mathbf{X}_t$ and $\mathbf{H}_t$ are tensor-valued representations encoding spatiotemporal structure. All models are trained under the harmonized conditions of an experiment to isolate the effects of LLM-guided preprocessing. The ranges of hyperparameters, learning rates, batch sizes, and the number of epochs are synchronized where possible, and homogeneous data divisions, stopping criteria, and random seeds are used. This controlled training procedure can be used to ensure that any downstream behavior differences are due primarily to preprocessing mechanisms rather than to model training or optimization discrepancies.

3.5 Metaheuristic Algorithms

Metaheuristic algorithms have been part of recent machine learning pipelines, especially when they require complex models, high-dimensional search spaces, or non-convex optimization. In contrast to deterministic or gradient-based optimization procedures, metaheuristics are population-based and stochastic, and do not require clear gradient information. This is what particularly makes them a good choice for guiding the optimization of deep learning models, where objective functions tend to be highly nonlinear, multimodal, and highly sensitive to initialization. Metaheuristic algorithms use natural, physical, or evolutionary mechanisms to achieve strong global search and minimise the risk of local minima. In the current research, metaheuristic algorithms can be used as an optimization layer that supports both the preprocessing and the

deep learning architecture of the model. Their functions are strictly distinct from data preparation and model design so that the effects of optimization may be studied separately. This division is in line with international standards for experimental design, and it is also necessary to ensure that advances in predictive accuracy can be traced to systematic parameter optimization rather than ad hoc manual modifications.

3.5.1 Role in Hyperparameter Optimization

Hyperparameter optimization is an important aspect of fine-tuning deep learning models because hyperparameters directly affect learning dynamics, convergence behavior, and generalization. Some of the parameters that cannot be estimated by the data using backpropagation are learning rates, batch sizes, network depth, regularization coefficients, and architectural control variables and thus have to be selected manually. Poor choices of hyperparameters can result in underfitting, overfitting, nonlinear training or even high computational cost despite the underlying model architecture being well-designed. The approach taken by metaheuristic algorithms resolves the problem by modeling hyperparameter optimization as a global maximization problem. A candidate solution is a given set of hyperparameters, and the quality of its set is measured by an objective function based on model performance on some validation data. During population update, metaheuristics efficiently explore the hyperparameter search space by balancing exploration and exploitation. By doing so, high-quality hyperparameter settings can be automatically discovered without exhaustive grid search or trial-and-error. Optimizing hyperparameters with metaheuristics offers several benefits in deep learning applications. The algorithms first operate in mixed, continuous-discrete search spaces, which are commonly observed when architecture and training-related parameters are optimized simultaneously. Second, they are stochastic, thereby robust to noisy objective functions, which is a common problem of deep learning because weight initialization is random, and mini-batch training is noisy. Lastly, optimization guided by metaheuristics reduces human bias and subjectivity, resulting in a more reproducible, systematically optimized learning pipeline. In this experiment, metaheuristic algorithms are used to automate the selection of critical hyperparameters for the deep learning model, while maintaining a constant preprocessing protocol and model architecture. This structure allows detecting changes in predictive behavior that can be explained by optimizing sources to pinpoint the success of the metaheuristic, as in principled, rather than manual, hyperparameter selection.

3.5.2 Proposed Optimizer: Football Optimization Algorithm (FbOA)

In this paper, the Football Optimization Algorithm (FbOA) is applied as a population-based metaheuristic optimizer for both feature selection and hyperparameter optimization. FbOA is based on the dynamics of football team strategies, in which the (*players*) (their positions) are determined by the coordinated passing and tactical movement to the position of the goal, i.e., the location of high-quality solutions in a complex search space. This analogy has the algorithm functionalizing various passing styles as follows: *short passing*, *lob passing*, and *through-ball passing*. Expected quasi-complementary search behavior is induced by each combination of these passing styles, and collectively, there is a balance between exploitation and exploration. Such a design is justified expressly by the need to address high-dimensional optimization problems, nonlinearity, and multiple local minima, while managing an exploration-exploitation trade-off during the search process.

Population representation and objective.

Let d denote the dimension of the decision space. FbOA maintains a population of N players

$$P = \left\{ \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)} \right\}, \qquad \boldsymbol{\theta}_i^{(t)} \in \mathbb{R}^d,$$

where $\boldsymbol{\theta}_i^{(t)}$ is the position (candidate solution) of player i at iteration t. The quality of each player is evaluated by a fitness function $F(\boldsymbol{\theta})$ (in this work, the validation objective associated with forecasting accuracy and/or the feature-selection criterion). The algorithm iteratively updates players using football-inspired rules that emulate (i) local refinement (short passing), (ii) intermediate transitions (lob passing), and (iii) global diversification (through-ball passing).

Exploration mechanism via football velocity.

FbOA models exploration by assuming that each pass is associated with a *speed/velocity* component that can vary across iterations to enhance space coverage and reduce the likelihood of premature convergence. In the FbOA formulation, the player velocity at iteration n is defined as:

$$V_n = F_{\max}\left(b_x\, a_i\,[F_{\text{ext}} - F_{\min}] + r\, b_y\, a_j\,[F_{\text{best}} - F_{\min}] \cdot \cos\left(\frac{\pi}{\text{Iteration}}\right)\right), \tag{9}$$

where $F_{\max}$ and $F_{\min}$ denote upper and lower force (or search intensity) bounds, b_x and b_y are direction-control coefficients, a_i and a_j are acceleration-like factors, F_{ext} represents an external influence term in the search, F_{best} is the best force/position discovered so far, and r is a random factor to inject stochasticity. The term $\cos\left(\pi/\text{Iteration}\right)$ acts as an iteration-dependent modulator that helps transition the search from broad exploration toward more focused behavior as iterations progress.

Update of the best force term.

To continuously refine the guiding best force/position, FbOA updates F_{best} using the following

$$F_{\text{best}} = \frac{1}{K}\sum_{n=0}^{K}\left(\frac{F_{\max}^{n^2}}{(2n+1)^2}\right), \tag{10}$$

where K is an iteration-dependent factor (described in the FbOA paper as increasing from 0 to 1 to balance exploration and exploitation) and $(2n+1)^2$ serves as a normalization term controlling the update rate. This mechanism strengthens exploitation pressure around promising regions while retaining controlled smoothing through aggregation.

Exploitation (short-passing) update rule.

In the exploitation phase, FbOA focuses on intensive search around promising areas to achieve optimal or near-optimal solutions. The rule of position/state update of the rule of exploitation is specified as follows:

$$F_b\big(S(t+1)\big) = F_i + z_3 F_b\big(S(t)\big) + K\,\sin\left(\frac{\pi}{\text{Iteration}}\right), \tag{11}$$

where $F_b(S(t))$ denotes the current state/solution representation at iteration t, F_i is a current guiding force/position term, z_3 is a control parameter regulating the influence of the current state, and the sinusoidal factor $\sin\left(\pi/\text{Iteration}\right)$ promotes a smooth convergence tendency that accelerates exploitation as the search proceeds.

Mutation operator to escape local optima.

To elevate resiliency against stagnation and local minima, FbOA adds a mutation strategy that introduces structural variation. The formulation of the mutation step is:

$$S(t) = \left(K\, a_q\left(\frac{2n+1}{x}\right) + K\,\cos\left(\frac{\pi}{\text{Iteration}}\right)\right), \tag{12}$$

where a_q scales the mutation magnitude, $\frac{2n+1}{x}$ is a normalization component controlling mutation effect, and the cosine modulation introduces iteration-shaped randomness to maintain diversity. This operator is particularly important in multimodal landscapes, where repeated refinements may otherwise trap the population in suboptimal basins.

Integration in this work: feature selection and hyperparameter optimization.

FbOA plays two complementary roles in this study. For feature selection, each player scores a candidate set of features (or a binary code) based on its predictive performance, preferring smaller subsets. In hyperparameter optimization, each participant simply encodes a candidate hyperparameter vector (continuous/discrete), and the objective is to evaluate its resulting safety and forecasting performance under a controlled training regimen. Pass strategy can be mapped to the following tasks: locally refining high-quality configurations stationary point While passing is a natural extension of high-quality configurations, local jumps to nearby promising basins while passing Lobbing is continuous stationary point While passing is a natural extension of high-quality configurations, local jumps are made to neighboring promising basins, through-ball passing promotes exploration of distant regions of high-dimensional search space (Algorithm 1).

Pseudocode of the Football Optimization Algorithm (FbOA):

Algorithm 1: Football optimization algorithm (FbOA)

Initialize a population of players $P = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N\}$
Set maximum iterations $t_{\max}$ and FbOA parameters (e.g., $F_{\max}$, $F_{\min}$, b_x, b_y, a_i, a_j, z_3, K, a_q)
$t \leftarrow 0$
while $t < t_{\max}$ **do**
    for each player $\boldsymbol{\theta}_i \in P$ **do**
        Evaluate fitness $F(\boldsymbol{\theta}_i)$
    end for
    Update guiding quantities (e.g., F_{best} using Eq. (10))
    for each player $\boldsymbol{\theta}_i \in P$ **do**
        Draw $\gamma \sim \mathcal{U}(0,1)$
        if $\gamma < 0.33$ **then**
            // Short passing: exploitation
            Update state using Eq. (11)
        else if $0.33 \leq \gamma < 0.66$ **then**
            // Lob passing: balanced move
            Compute velocity using Eq. (9) and update $\boldsymbol{\theta}_i$
        else
            // Through ball: global exploration
            Apply a long-range move guided by F_{best} and stochasticity
        end if
        Apply mutation using Eq. (12) when stagnation is detected
        Enforce boundary constraints on $\boldsymbol{\theta}_i$
    end for
    $t \leftarrow t + 1$
end while
return $\boldsymbol{\theta}_{\text{best}}$

3.5.3 Benchmark Optimizers for Comparison

To critically test the proposed Football Optimization Algorithm (FbOA), its optimization power was also evaluated against a collection of widely used metaheuristics that represent differing search philosophies. The benchmark optimizers were selected from known families of evolutionary computation and swarm intelligence and used under identical experimental conditions as FbOA, both for feature selection and hyperparameter optimization. The methods compared are GA, PSO, WAO, SFS, MVO, BA, DE, BBO, and APO, with the abbreviations defined in accordance with the glossary of optimizers used throughout the paper.

- **GA: Genetic Algorithm**, is an evolutionary optimization algorithm which uses selection/crossover mutations in order to sweep through possible solutions, upgrading the fittest ones by their survival philosophy.
- **PSO: Particle Swarm Optimizer**, a swarm-intelligence approach in which particles revise their positions and velocities based on individual experience and collaborative advice, thus offering rapid convergence in terms of cost while preserving population diversity.
- **WAO: Whale Optimization Algorithm**, is a bio-inspired metaheuristic that mimics the foraging behavior of whales, using exploration and exploitation strategies to balance global search with the narrowing of the gap around good candidate solutions.
- **SFS: Stochastic Fractal Search**, a stochastic optimization approach that simulates fractal-inspired diffusion processes to explore the search space and progressively refine solutions through probabilistic neighborhood sampling and adaptive search steps.
- **MVO: Multiverse Optimization**, a physics-inspired metaheuristic that conceptualizes candidate solutions as universes and uses mechanisms analogous to multiverse interactions to guide the search toward high-quality regions of the solution space.
- **BA: Bat Algorithm**, a nature-inspired optimizer based on echolocation behavior, where candidate solutions adjust their positions using loudness and pulse emission concepts to regulate exploration and exploitation.
- **DE: Differential Evolution**, a population-based evolutionary algorithm that generates new candidate solutions by combining scaled differences between individuals and applying selection to retain improved solutions, offering strong robustness in continuous optimization landscapes.
- **BBO: Biogeography-Based Optimizer**, an evolutionary metaheuristic that models the migration of species among habitats, where high-quality solutions share features with others through migration operators while maintaining diversity via mutation-like mechanisms.
- **APO: Puma Optimizer**, a behavior-inspired metaheuristic that models adaptive pursuit strategies to improve search efficiency through dynamic movement patterns that support both exploration of new regions and exploitation around strong candidate solutions.

All benchmark optimizers were implemented with the same data splits, objective function, and stopping conditions, so that they executed all benchmark optimizer methods methodologically in parallel. This managed benchmarking structure is suitable for a meaningful review of FbOA against available metaheuristic-based solutions for open feature selection and hyperparameter optimization.

3.6 Evaluation Metrics

To achieve a holistic, objective evaluation of the model's performance, regression-based evaluation metrics were adopted. These measures were chosen to represent complementary issues: accuracy, magnitudes of error, bias, strength of correlation, and prediction efficiency. The current values are denoted by y_i and $\hat{y}_i$, the values being observed at time step i, respectively, and forecasted, respectively, and bar $\bar{y}$ represents the

average of the observed values, and N is the total number of samples. All the evaluation metrics applied in this study are summarized in Table 2 in mathematical terms. In this study, the target variable is the catalog-reported earthquake magnitude, which represents the size of each seismic event. Magnitude is treated as a dimensionless seismic scale quantity, while the predictor variables are derived from the available catalog attributes, including temporal, spatial, and physical features such as date, latitude, longitude, depth, place, and magnitude type where applicable. Accordingly, error-based metrics such as MAE and RMSE are reported in magnitude units, whereas correlation-based metrics such as R^2 are dimensionless. All these metrics can give a solid evaluation framework. Error measures like MSE, RMSE, MAE, MBE, and RRMSE, and correlation measures (r and R2) all measure the shock, the degree of error in prediction, the degree of bias, and the degree of linear agreement between prediction and observation, respectively. Efficiency-based indicators, such as NSE and WI, are especially informative for geophysical and environmental models, since they compare predictive ability with reference benchmark models.

Table 2: Evaluation metrics and their mathematical formulations used for model performance assessment.

Metric	Mathematical Definition
Mean Squared Error (MSE)	$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
Mean Absolute Error (MAE)	$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$
Mean Bias Error (MBE)	$\mathrm{MBE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$
Pearson Correlation Coefficient (r)	$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}$
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Relative Root Mean Squared Error (RRMSE)	$\mathrm{RRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\bar{y}}$
Nash–Sutcliffe Efficiency (NSE)	$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Willmott Index of Agreement (WI)	$\mathrm{WI} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\lvert\hat{y}_i - \bar{y}\rvert + \lvert y_i - \bar{y}\rvert)^2}$

3.7 Experimental Setup

The experimental setup of the given research ensured methodological rigor, computational stability, and perfect reproducibility of each step of preprocessing, model training, and optimization. Given the high computational cost of deep learning-based time-series forecasting and population-based metaheuristic optimization, the focus was on developing a controlled, well-documented experimental infrastructure. All tests were run on a powerful computer workstation with a multi-core central processing unit and a CUDA-capable graphics processing unit to facilitate parallel computing and faster model training. The hardware setup was chosen to be well-suited to the iterative nature of hyperparameter optimization and the

iterative training procedures required by the experimental setup. Data preprocessing, feature construction, statistical testing, and optimization programs were executed using multi-core CPU resources, and deep learning architecture training was also accelerated with a GPU, especially for hierarchical interpolation, attention schemes, and recurrent structures. This arrangement avoided biases in the comparative assessment of preprocessing strategies or optimization algorithms arising from constraints on computational capacity. The software environment was standardized across all experimental runs to eliminate inconsistencies caused by heterogeneous implementations. Everything was done in Python (version 3.10), leveraging a single scientific computing stack. Implemented models of deep learning, such as N-HITS, Sequence-to-Sequence encoder-decoder networks, Informer, Neural Oblivious Decision Ensembles, Spatio-Temporal Graph Convolutional Networks, convolutional LSTM and Self-Regulating Recurrent Networks, were also executed in a single architecture of deep learning, which supports automatic differentiation, accelerated using the GPU, and optimized utilizing the tensors. It used auxiliary libraries for numerical calculations, Data manipulation, statistical analysis, and visualization. There were also uniform library versions and code conventions used throughout to ensure consistent behavior across experiments. One of the major components of the experimental facility is the implementation of the Football Optimization Algorithm (FbOA) as the primary optimization engine. The FbOA was directly incorporated into the modeling pipeline and was used to select features and optimize a set of hyperparameters, with the same objective function. The optimizer optimized the various solutions tested in the learning models and evaluated them during the same training and validation processes. All population sizes, all FbOA control parameters, and all stopping criteria were clearly defined and tracked, making the optimization process easy to monitor and enabling reproducibility 2. All layers of experimental control were imposed on the forces of reproducibility. Direction-fixed random seeds were implemented at the Python runtime layer, in the deep learning stack, and across all numerical libraries used for stochastic operations. New, deterministic execution modes were provided where possible to minimize the effects of nondeterminism introduced by parallelism and by kernel selection on GPUs. Also, log files containing all configuration details, such as preprocessing specifications, feature selection masks, model hyperparameters, optimizer settings, and training schedules, were systematically logged and stored per experimental run. Such records guarantee that any outcomes may be easily recreated and tested again. The general experimental process was a systematic, progressive pipeline aimed at separating the influence of preprocessing and optimization. During the process, data are preprocessed using IS/ITs, including automatically generated feature extraction based on the LLM, time transformation, consistency measures, and statistical validation. Through a series of non-optimized configurations, baseline models are trained and evaluated to get reference performance levels. The next step involves the use of metaheuristic optimization, where FbOA and hyperparameter settings are improved by FbOA and benchmark optimizers, filtering for possible feature subsets and adjusting hyperparameters. Lastly, the candidate models are reread and tested on a held-out test dataset. This is a staged design that guarantees the observed differences in performance can be unambiguously attributed to the quality of preprocessing and the effectiveness of optimization, rather than to uncontrollable experimental variations.

4 Empirical Results

In this section, the detailed empirical analysis of the modeling framework is conducted to systematically evaluate predictive performance in both non-optimized and optimized settings. The analysis approach ensures that a distinction can be made between baseline behavior when models are trained with the same preprocessing and default hyperparameters, and improved performance with hyperparameter optimization via metaheuristics. The adoption of such a stratified evaluation approach allows the section to create a well-defined performance basis and support an objective evaluation of the input of optimization methods.

The empirical study is based on a broad range of quantitative evaluation metrics that collectively reflect the accuracy of the prediction, systematic bias, strength of correlation, relative error, and overall efficiency. To augment tabular comparisons, several advanced interactive visualization tools are applied to investigate model behavior across multiple facets of analysis. The intended use of these visual analyses is to show trends in performance, variability, robustness, and trade-offs between models and optimization strategies, providing a more in-depth contrast to single-metric comparisons. The baseline results, which are not obtained through hyperparameter optimization, are the subject of the first section of this section. These outcomes serve as a crucial benchmark, illustrating the intrinsic dissimilarity in model decoding when trained under the same preprocessing conditions. The second section studies the effects of hyperparameter optimization using the proposed Football Optimization Algorithm and a collection of benchmark metaheuristics. The comparative analyses are made using error reduction, correlation improvement, convergence behavior, and robustness. Combined, these findings provide a broad empirical basis for assessing the effectiveness of the suggested data-centric, optimization-driven model of seismic segmentation.

4.1 Section Baseline Results without Hyperparameter Optimization

The section provides a detailed discussion of the baseline predictive performance of the considered models before hyperparameter optimization (HPO) or any improvement via a metaheuristic. The aim of such baseline measurement is to establish a strict reference level against which the value of optimization strategies can be quantitatively assessed. To ensure methodological rigor, identical preprocessing procedures, data partitions, and non-optimized (default) hyperparameter settings were applied across all models. As a result, the reported outcomes reflect the *intrinsic modeling capacity* of each architecture rather than the influence of targeted tuning. This baseline analysis is particularly critical in seismic time-series forecasting, where earthquake records are typically sparse, skewed, and dominated by infrequent extreme events. Under such conditions, default configurations often fail to capture domain-specific temporal dynamics, making a controlled baseline indispensable for understanding how effectively each model can extrapolate raw temporal patterns.

The entire range of metrics of regression performance (MSE, RMSE, MAE, MBE, Pearson correlation coefficient), coefficient of determination, R2, RRMSE, NashSutcliffe efficiency, and Willmott index of agreement are summarized in Table 3. These measures all describe predictive accuracy, systematic bias, correlation strength, relative error magnitude and overall efficiency. The presence of absolute and relative error measures, along with indicators of correlation and efficiency, allows for a comprehensive and balanced evaluation of the results. In seismic use, it is quite common to be fooled by a single measure; models might have low average error but low temporal coherence or low regime-shift efficiency, so it can be important to have a richer metric space. In an error-based view, we get a similar result: a clear stratification of predictive accuracy among the considered architectures. N-HITS is able to produce the lowest rate of error, which is 0.0065 MSE, RMSE of 0.0806 and MAE of 0.0462, implying effective bottom predicting ability in its capacity to capture the temporal structure. NODE then yields slightly larger errors (MSE = 0.0093, RMSE = 0.0965, MAE = 0.0581), while Seq2Seq yields an even larger error (MSE = 0.0148, RMSE = 0.1216). The deviations are significantly greater in Informer and STGCN, with RMSE of 0.1473 and 0.1700, respectively, indicating lower performance when set in default environments. ConvLSTM and SRRN have the lowest baseline performance, with RMSE of 0.1961 and 0.2396, and MAEs over 0.10, indicating significant challenges in modeling the seismic time series without optimization. The mean bias error (MBE) values also emphasize the external deviations in baseline predictions. As a case in point, N-HITS has an MBE of 0.0285, compared to SRRN, which has a significantly larger bias of 0.1021. This upward linear trend in MBE across models indicates that architectures with worse performance are not only less accurate but also more generally biased. However,

it is of great concern in seismic forecasting, where such bias may skew long-term trend interpretation and increase error for large-magnitude or eccentric seismic events. Correlation-oriented measures can be used to provide balanced information about the regularity between the expected and observed values. The Pearson correlation coefficient (r = 0.897) and coefficient of determination (R2 = 0.907) for N-HITS are at their peak, indicating that the model aligns well with temporal variability. NODE and Seq2Seq show strong correlations in the values of the variables (0.887 and 0.872, respectively), whereas Informer and STGCN show a gradual decrease in correlation strength. The minimal values of correlation include SRRN, with correlation coefficients of 0.819 and R2 of 0.837, indicating that greater numerical error reduces the likelihood of tracking temporal co-movement. These findings highlight a critical modeling observation: minimizing pointwise error does not necessarily imply good reproduction of seismic variability, and correlation-based measures are necessary to distinguish between localized minimization of error and actual learning of temporal variability.

Table 3: Baseline performance of evaluated models without hyperparameter optimization.

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
N-HITS	0.0065	0.0806	0.0462	0.0285	0.897	0.907	1.40	0.918	0.909
NODE	0.0093	0.0965	0.0581	0.0402	0.887	0.896	1.78	0.905	0.893
Seq2Seq	0.0148	0.1216	0.0690	0.0524	0.872	0.884	2.10	0.886	0.874
Informer	0.0217	0.1473	0.0815	0.0631	0.861	0.874	2.42	0.868	0.857
STGCN	0.0289	0.1700	0.0922	0.0714	0.850	0.863	2.73	0.849	0.843
ConvLSTM	0.0385	0.1961	0.1068	0.0839	0.837	0.851	3.05	0.829	0.828
SRRN	0.0574	0.2396	0.1270	0.1021	0.819	0.837	3.60	0.804	0.812

Baseline performance is further put in perspective with relative and efficiency-based measures. It is observed that the RRMSE also increases steadily from N-HITS (1.40) to SRRN (3.60); thus, the relative error increases significantly as the model's capability declines. On the same note, NSE reduces by 0.918 (N-HITS) to 0.804 (SRRN), and WI reduces by 0.909 to 0.812 in the same range. The above trends imply that weaker architectures result in higher absolute error and less efficient, less effective predictions compared to the reference behaviour. In geophysical and environmental modeling, these efficiency metrics are vital, as they directly document whether a model provides a genuine improvement over fixes and basic frameworks. In general, the numerical findings in Table 3 prove that deep learning models are characterized by significantly different predictive behavior when trained without hyperparameter optimization. The performance differences observed, including error magnitude, bias, correlation, and efficiency, indicate that pure architectural design is insufficient for reliable seismic forecasting in the default setting. These baseline measurements can thus be thought of as a rigid quantitative point of reference for the later sections where feature selection and metaheuristic-based hyperparameter optimization are proposed. Notably, the achieved values in the baseline are not to be understood as the highest possible capacity of the measured architectures, but rather serve to establish controlled initial parameters with respect to which the subsequent enhancements have to be explained by systematic optimization rather than by architectural superiority as such. To provide a highly detailed overview of predictive performance across a variety of assessment criteria, a faceted visualization of model performance metrics was established, as shown in Fig. 7. The different metrics, such as MSE, RMSE, MAE, MBE, Pearson correlation coefficient (r), coefficient of determination (R2), RRMSE, NashSutcliffe efficiency (NSE), and Willmott index of agreement (WI), are indexes that are associated with each panel in the facet grid, thus allowing to have a clear and organised analysis of behaviour of the model between error metrics and correlation metrics. The bar heights denote the metric values for each model, allowing direct

comparison of models within a particular metric and maintaining the same interpretability of the individual evaluation space. As shown in the figure, error-based measures tend to increase with the models, while correlation-based metrics tend to decline steadily, demonstrating the trade-off between a model's accuracy and its explanatory power. Such a multidimensional representation further indicates whether the models exhibit balanced performance profiles or significant asymmetry across the different evaluation dimensions, and thus serves as an informative complement to pure numerical comparisons.

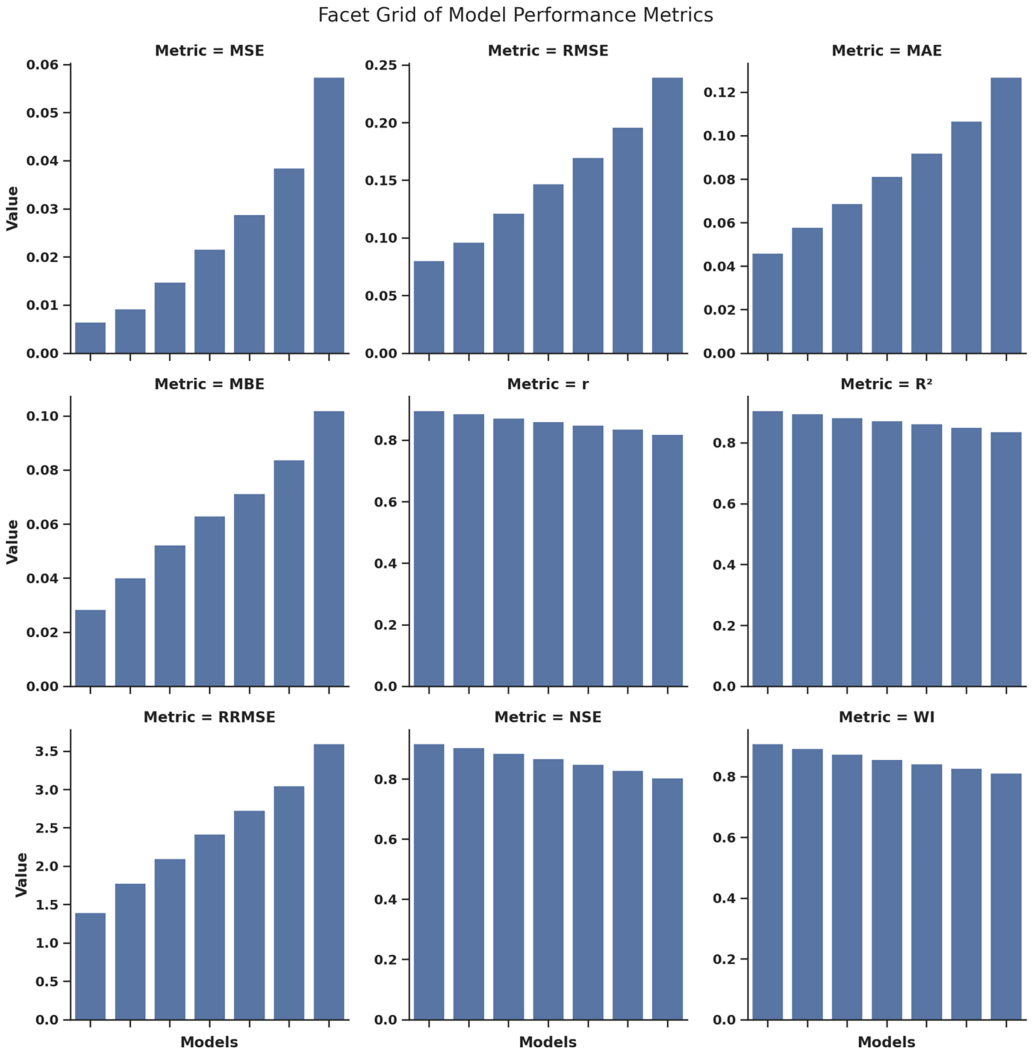

Figure 7: Facet grid of model performance metrics across multiple evaluation criteria.

To compare the performance of various models holistically using a number of possibly conflicting assessment measures, a parallel coordinates plot of performance indicators was used, as shown in Fig. 8. Each of the evaluation metrics, such as MSE, RMSE, MAE, MBE, Pearson correlation coefficient and coefficient of determination, Nash-Sutcliffe efficiency (NSE) and Willmott index of agreement, is plotted as a vertical axis, and each of the polylines represents a specific predictive model. This layout makes it easy to simultaneously test both error-based and correlation-based measurements, and subsequently determine trade-off and dominance patterns across models. As shown in the figure, models with lower error terms have more digestible correlation and efficiency scores, but others exhibit a strong trade-off,

indicating differences in overall predictive strength among the available architectures. Such a holistic view is especially useful in seismic forecasting, where operational performance is based on balance rather than single-metric optimization.

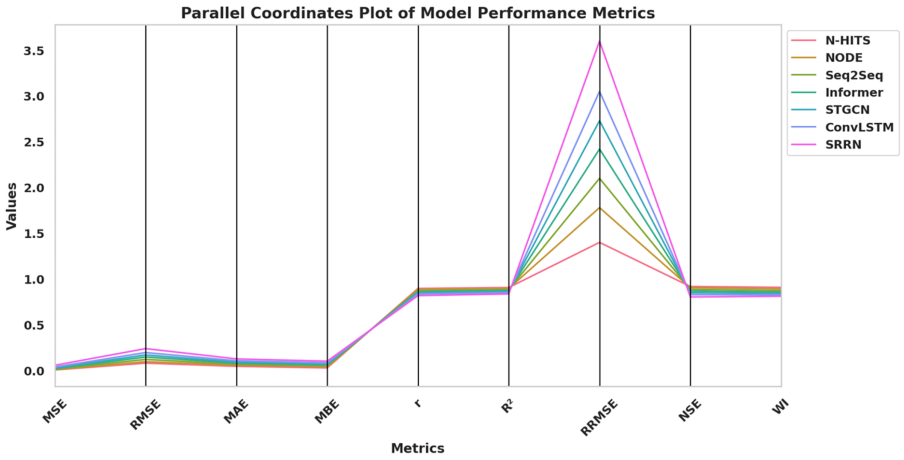

Figure 8: Parallel coordinates plot of model performance metrics across multiple evaluation criteria.

To evaluate the distributional characteristics and normality of the metrics of performance evaluation, a comprehensive series of quantile–quantile (Q–Q) plots was created as demonstrated in Fig. 9. All the subplots are associated with a particular metric, i.e., the MSE, RMSE, MAE, MBE, Pearson correlation coefficient (r), coefficient of determination (R2), RRMSE, Nash Sutcliffe efficiency (NSE), and index of agreement (WI), and then compares the empirical quantiles of the observed data with the theoretical quantiles of a normal distribution. The reference line indicates perfect agreement with normality, and systematic deviations indicate departure, e.g., skew or heavy tails. It can be seen that, as shown by the figure, the number of metrics has most of its points significantly closer to the reference line in the main quantile range, with only a few outliers at the ends, which implies that most of the location measures of the assessed performance can be taken as a reasonable indicator of the normality assumption. The relevance of this finding is that the subsequent inferential analysis can be made reliable, while at the same time, the realistic tail behavior in the complex seismic time-series assessment can be considered real.

4.2 Hyperparameter Optimization (HPO) Results

In this subsection, the detailed results of hyperparameter optimization (HPO) for the chosen forecasting architecture using various metaheuristic optimization strategies are presented. The main hypothesis of the following analysis is to measure the degree to which the systematic HPO enhances the predictive accuracy, robustness, and generalisation capacity of the optimizer compared to the default (optimisationless) template and relative to the existing optimizer benchmark, the optimisation Football Optimisation Algorithm (FbOA). In the seismic time-series prediction application, HPO is not an optional step but part of the methodology, since seismic sequences are generally sporadic, consist of bursts of activity, and represent mixed regimes controlled by heterogeneous tectonic restraints. In this case, deep learning models can reach suboptimal solutions when their learning mechanisms are restricted to default hyperparameters that were not optimized for the statistical structure and time-dependence of the target series. In this regard, the current analysis will focus not only on quantitative gains but also on the reliability, interpretability, and consistency of optimization results across various evaluation dimensions. The entire optimization experiments were performed under a stringent experimental protocol. In particular, an identical preprocessing

pipeline, training-validation-testing splits, printing, objective, stop criteria and performance measures were used with all optimizers. This design would make any observed performance difference due solely to the optimization mechanisms, not to confounding factors related to data handling or model training. Besides, this controlled design is important to ensure that performance gains are not due to differences in training conditions or varied evaluation controls. Disparities in evaluation budgets, stopping thresholds, or validation strategies in population-based metaheuristics can influence bias in comparisons. Thus, by imposing the same experimental conditions of parity on all the algorithms, the validity of the comparative conclusion is strengthened, as is the thesis that optimization behavior results from inherent search behavior rather than procedural variation. Table 4 provides the summation of the entire portfolio of performance measures provided by the optimized N-HITS model in combination with various metaheuristic algorithms. The reported metrics are error-based measures (MSE, RMSE, MAE, and MBE), correlation-based measures (r and R2), relative error (RRMSE), and efficiency-based measures (NSE and WI). In combination, these steps would lead to a dimensional assessment of the optimization effectiveness. This metric measure is especially compatible with geophysical and environmental modeling, in which predictive performance is to be quantified not solely in terms of absolute deviation (e.g., RMSE and MAE), but also in terms of agreement structure (e.g., WI), skill compared with a reference predictor (e.g., NSE) and strength of explanation (e.g., R2). Besides, the incorporation of MBE directly speaks to systematic bias, which is a particular concern in a seismic setting where inaccurate overestimation or underestimation commonly occurs and impacts downstream inferences, such as characterising hazards or detecting temporal abnormalities.

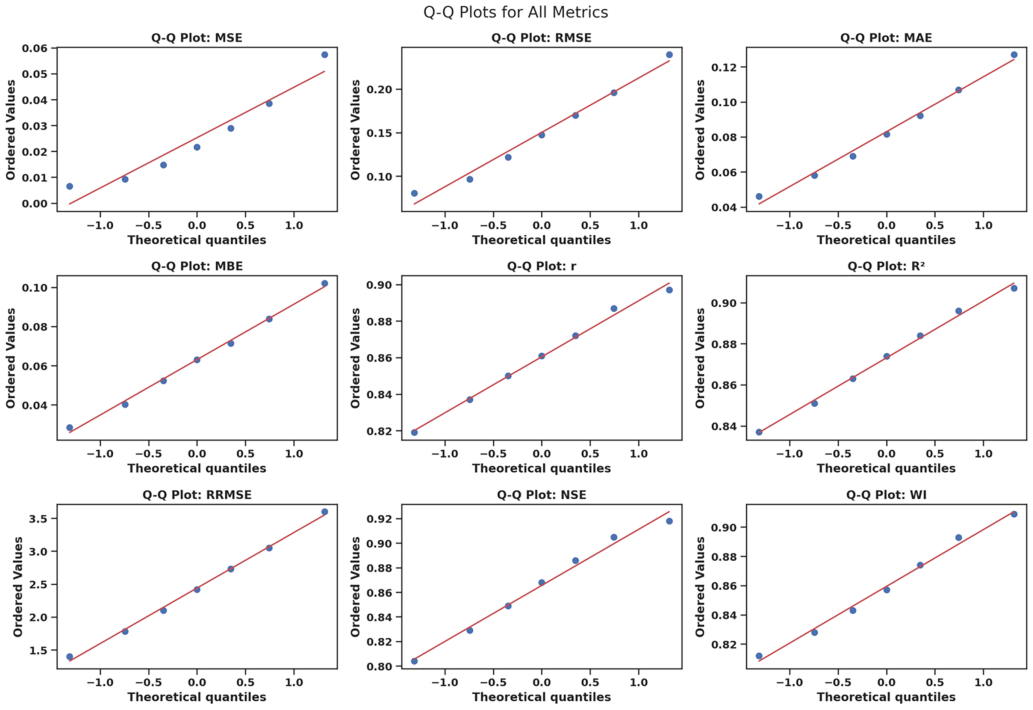

Figure 9: Q–Q plots assessing the normality of model performance metrics.

A first observation from Table 4 is the substantial reduction in prediction error achieved by all metaheuristic optimizers relative to the baseline configuration. Specifically, the baseline N-HITS RMSE of 0.0806 and MSE of 0.0065 are reduced under optimization to RMSE values in the range $[5.57 \times 10^{-3}, 1.30 \times 10^{-2}]$ and MSE values in the range $[3.10 \times 10^{-7}, 1.69 \times 10^{-6}]$, depending on the optimizer. This represents an

improvement of more than one order of magnitude in RMSE and multiple orders of magnitude in MSE, demonstrating that hyperparameter calibration is essential for unlocking the predictive capacity of deep learning models under heavy-tailed and nonstationary seismic regimes. Similar improvements are observed in MAE, which decreases from the baseline level of 0.0462 to values between 2.91×10^{-4} (FbOA) and 5.39×10^{-4} (APO), indicating that optimization reduces not only squared-error sensitivity but also the typical absolute deviation.

Table 4: Hyperparameter optimization results for the N-HITS model using different metaheuristic optimizers.

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
FbOA + N-HITS	3.10E−07	5.57E−03	2.91E−04	6.70E−05	0.982	0.979	0.071	0.981	0.985
GA + N-HITS	4.25E−07	6.52E−03	3.38E−04	8.20E−05	0.974	0.971	0.093	0.973	0.978
PSO + N-HITS	5.68E−07	7.54E−03	3.81E−04	9.90E−05	0.968	0.965	0.120	0.967	0.973
WAO + N-HITS	6.89E−07	8.30E−03	4.12E−04	1.14E−04	0.961	0.958	0.170	0.960	0.969
SFS + N-HITS	8.55E−07	9.25E−03	4.44E−04	1.33E−04	0.954	0.952	0.220	0.953	0.963
MVO + N-HITS	1.02E−06	1.01E−02	4.76E−04	1.47E−04	0.947	0.945	0.270	0.947	0.958
BA + N-HITS	1.18E−06	1.09E−02	4.93E−04	1.63E−04	0.941	0.939	0.340	0.940	0.953
DE + N-HITS	1.35E−06	1.16E−02	5.07E−04	1.79E−04	0.935	0.933	0.480	0.933	0.948
BBO + N-HITS	1.52E−06	1.23E−02	5.25E−04	1.96E−04	0.929	0.927	0.710	0.927	0.942
APO + N-HITS	1.69E−06	1.30E−02	5.39E−04	2.10E−04	0.922	0.920	0.960	0.921	0.936

Among all evaluated strategies, FbOA-based optimization yields the strongest outcomes across all error measures. FbOA achieves the minimum MSE of 3.10×10^{-7} and the minimum RMSE of 5.57×10^{-3}, outperforming the nearest competitor (GA) which attains MSE 4.25×10^{-7} and RMSE 6.52×10^{-3}. In addition, FbOA attains the lowest MAE (2.91×10^{-4}) and the smallest bias magnitude, with MBE 6.70×10^{-5}. By comparison, APO exhibits the largest error and bias among the optimized variants (RMSE 1.30×10^{-2} and MBE 2.10×10^{-4}), indicating that not all metaheuristics search the hyperparameter landscape with equal effectiveness. Importantly, the simultaneous reduction of RMSE and MBE under FbOA indicates that optimization is not achieved at the cost of calibration; rather, the model becomes both more accurate and less systematically biased.

Correlation-based metrics further corroborate these findings and show that improvements extend beyond pointwise error reduction. FbOA obtains the highest Pearson correlation ($r = 0.982$) and the highest coefficient of determination ($R^2 = 0.979$), whereas the weakest optimized configuration (APO) yields $r = 0.922$ and $R^2 = 0.920$. This difference is not marginal: it indicates that FbOA-enhanced N-HITS explains a substantially larger portion of variance in the target series while preserving stronger temporal co-movement between predictions and observations. Intermediate optimizers follow a monotonic decline from GA ($r = 0.974$, $R^2 = 0.971$) through PSO ($r = 0.968$, $R^2 = 0.965$) and down to BBO ($r = 0.929$, $R^2 = 0.927$), yielding a consistent ordering that mirrors the error-based ranking. In practical seismic forecasting, this correlation increase is critical because it signals improved tracking of temporal fluctuations rather than merely reducing average deviation.

Relative and efficiency-oriented metrics provide additional quantitative evidence of robustness and generalization improvements. FbOA achieves the lowest relative error (RRMSE = 0.071), while RRMSE increases progressively through GA (0.093), PSO (0.120), and up to APO (0.960). Likewise, FbOA yields the highest NSE (0.981) and WI (0.985), reflecting superior predictive skill relative to a reference predictor and stronger agreement across the dynamic range. The monotonic degradation of NSE from 0.981 (FbOA) to

0.921 (APO), together with the WI decline from 0.985 to 0.936, indicates that weaker optimizers not only produce larger errors but also reduce predictive efficiency and agreement. Since NSE and WI are commonly used in hydrology and geophysical prediction to evaluate reliability across regimes, the high values attained under FbOA provide strong support for robustness under variable seismic activity levels.

Beyond final performance metrics, convergence behavior offers insight into the underlying optimization dynamics. The final objective values implicitly reflect how effectively each method navigates the search landscape: FbOA reaches the lowest error basin (RMSE 5.57×10^{-3}), while other optimizers converge to progressively higher basins (e.g., RMSE 6.52×10^{-3} for GA and 1.30×10^{-2} for APO). This ranking is consistent with the interpretation that FbOA maintains a stronger exploration–exploitation balance, enabling rapid entry into high-quality regions while still refining solutions to achieve stable minima. In computationally expensive deep learning optimization, such convergence efficiency is important because each candidate evaluation incurs substantial training cost; thus, reaching lower-error basins with fewer wasted evaluations directly improves practical feasibility.

Collectively, the HPO results demonstrate that metaheuristic-driven hyperparameter optimization is indispensable for achieving high predictive performance in deep learning–based seismic modeling. Quantitatively, the optimized configurations reduce baseline errors from RMSE 0.0806 to values as low as 0.00557 (FbOA), while simultaneously increasing correlation from r values below 0.90 in baseline settings to $r = 0.982$ under FbOA. The superior performance of FbOA across accuracy (MSE/RMSE/MAE), calibration (MBE), agreement (WI), skill (NSE), and explanatory strength (R^2) underscores its suitability as an effective and robust optimizer in the proposed modeling framework. More broadly, these results establish that optimization quality should be treated as a core methodological component in seismic deep learning, since it yields measurable improvements across all evaluation dimensions rather than a single metric in isolation.

To investigate the similarity structure among the hybrid models, a hierarchical clustering analysis was performed based on overall performance profiles, as shown in Fig. 10. The dendrogram shows the number of standardized performance measures between models, with the height of the links indicating the extent of dissimilarity. Such a hierarchical representation allows distinguishing sets of models with similar predictive behavior across a range of evaluation criteria. As shown in the figure, some models form tightly clustered groups at low linkages, indicating similar performance, whereas others cluster at higher levels, suggesting a broader range of distinct predictive features. Interpreting these clusters provides a clue as to whether these optimizers generate qualitatively similar performance profiles or produce fundamentally different performance structures across measures. Moreover, hierarchical grouping helps identify optimizer families that perform reasonably under the same constraints, which can be helpful in subsequent methodological choice when computational budgets or operational constraints limit the number of candidate optimizers that can be assessed.

In an attempt to more quantitatively measure the central tendency and dispersion of the model with respect to evaluation parameters, box plots with mean and standard deviation clues were built on each measure, as shown in Fig. 11. Each subplot summarizes the allocation of a performance measure, including MSE, RMSE, MAE, MBE, Pearson correlation coefficient (r), coefficient of determination (R2), RRMSE, Nash-Sutcliffe efficacy (NSE), and Willmott's index of agreement (WI). The horizontal line represents the average, and the dotted lines are the one-standard-deviation bands around it, clearly showing the average and the standard deviation. As shown in the figure, both correlation-based and efficiency-based measures are narrowly dispersed around elevated mean values, whereas error-based measures exhibit a wider distribution due to variability in predictive error across models. This is anticipated, given that error measures are more susceptible to outliers and heavy-tailed distributions, whereas correlation and agreement measures tend to level off when models are constantly tracking a temporal trend. Notably, the relative dispersion patterns

between optimizers allow a robustness-oriented explanation of minimum: the ability to generate less varied distributions can be said to yield more stable solutions, whereas wider distributions might indicate adaptation to search stochasticity or a lower convergence consistency rate.

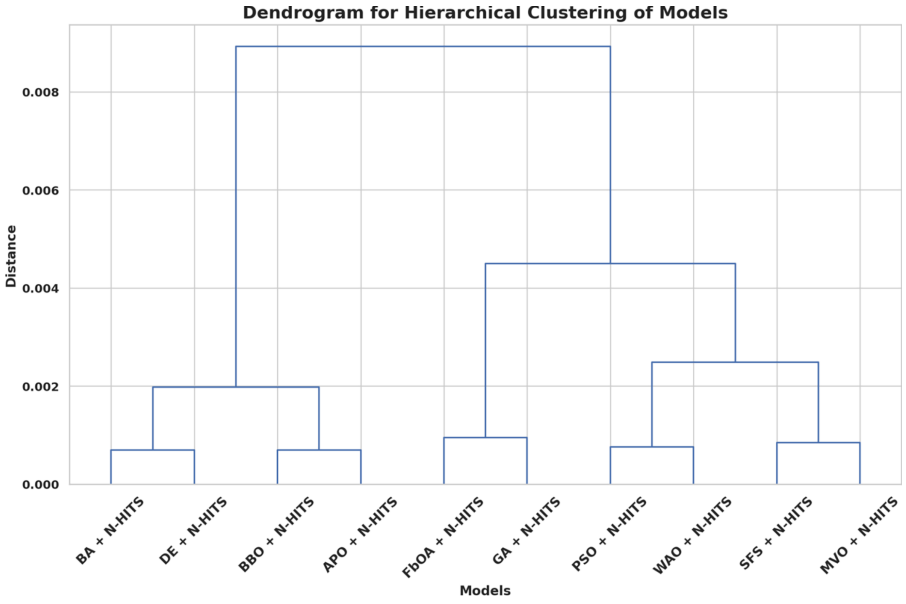

Figure 10: Dendrogram illustrating hierarchical clustering of hybrid models based on performance metrics.

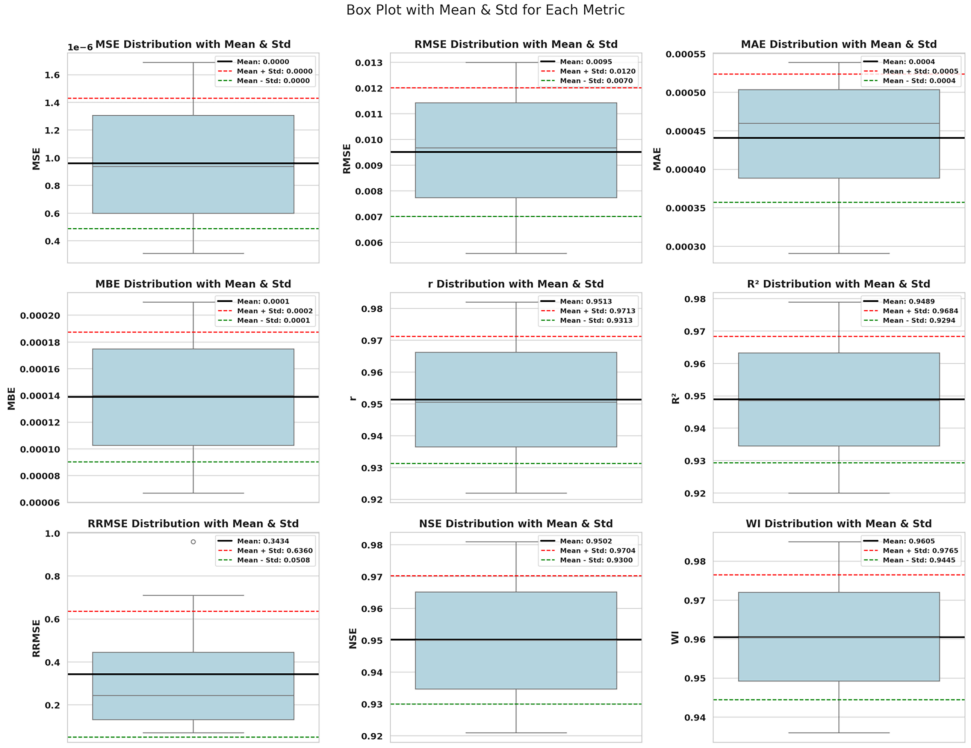

Figure 11: Box plots of performance metrics with overlaid mean and standard deviation indicators.

To consolidate the distributional characteristics of overall model evaluation metrics and to emphasize the central tendency of the measures, a consolidated box plot of the performance metrics was prepared, as shown in Fig. 12. The figure shows distributions of error-based, correlation-based based and efficiency-based measures, such as those of MSE, RMSE, MAE, MBE, Pearson correlation coefficient (r), coefficient of determination (R^2), RRMSE, Nash Sutcliffe efficiency (NSE) and index of agreement of Willmott in a single visualization framework. Red diamond markers indicate the average values for each metric, and it is possible to directly compare the average performance and the distribution's spread. As shown in the figure, the correlation and efficiency values are tightly clustered at high values, indicating consistent predictive strength, whereas RRMSE shows greater dispersion, reflecting higher variability in relative error across the discussed models. This trend reinforces the interpretive distinction between accuracy- and efficiency-oriented measurements; the former yields stronger agreement and correlation when tuning the model, whereas the latter shows a far more significant effect. RMSE, as a result, is an informative discriminator in cases where multiple optimizers are seemingly equally potent using r, and additionally R2, NSE, or WI and delivers further indication on the choice of robust optimization strategies under realistic deployment limits.

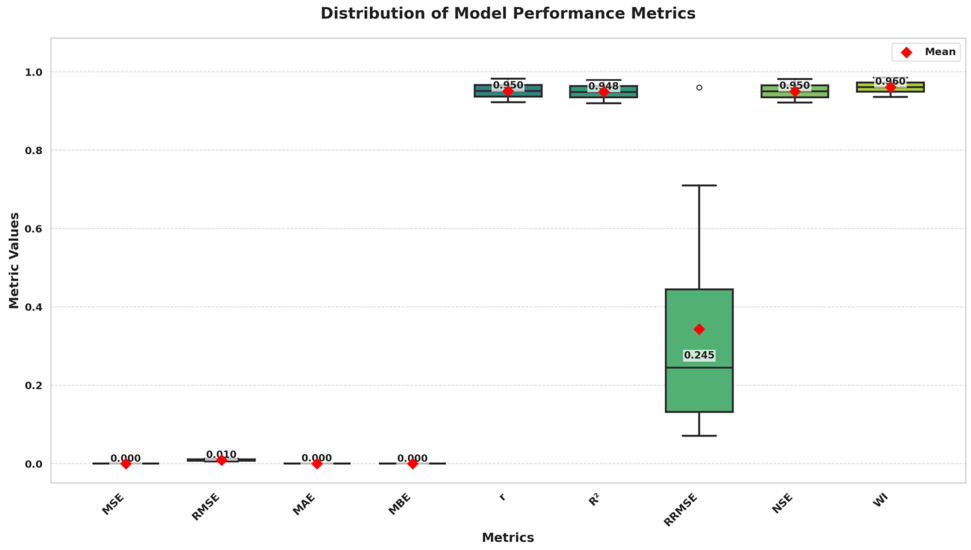

Figure 12: Distribution of model performance metrics with mean values highlighted.

5 Statistical Result Analysis

To supplement the descriptive performance analysis and comparisons above with visualization, a nonparametric test of statistical significance was performed using the Wilcoxon signed-rank test. The purpose of this analysis is to critically examine, in terms of statistical significance, whether the observed performance differences among the optimized hybrid models are statistically significant or can be explained by random variation. Although descriptive measures and graphical analyses are useful to understand the behavior of the relative model, in and of themselves, they do not determine whether observed differences are reproducible and statistically robust. As a result, inferential statistical testing is required to ensure that the gains obtained through metaheuristic-driven optimization are systematic and not due to stochastic processes. The use of a Wilcoxon signed-rank test is most appropriate in this scenario, as it requires no assumptions about the distributions of the underlying performances and is particularly useful for paired comparisons with small sample sizes, such as those associated with repeated model assessment. This is

particularly significant for seismic time-series modeling, where performance measures can be skewed, fat-tailed, or contain outliers due to episodic seismic activity and rare extreme events. Under such conditions, the Wilcoxon test can be used to test the hypothesis because it is conservative and robust, relying on rank-based comparisons rather than distributional assumptions. The results of the Wilcoxon signed-rank test of the optimized N-HITS model with various metaheuristic optimization schemes are summarized in Table 5 below. The status of each hybrid combination was to have a theoretical median of zero, i.e., the hypothesis that they are not systematically out of kilter with the reference distribution. The computed median values were based on the results of repeated experimental applications, with each arrangement tested with the same number of observations to allow effective comparison. This repeated-measures study enhances the validity of the statistical analysis by controlling for variability arising from random initialization, stochastic optimization behavior, and training noise.

Table 5: Wilcoxon signed-rank test results for optimized hybrid models based on repeated performance evaluations.

	FbOA	**GA**	**PSO**	**WAO**	**SFS**	**MVO**	**BA**	**DE**	**BBO**	**APO**
Theoretical median	0	0	0	0	0	0	0	0	0	0
Actual median	0.00557	0.006522	0.007544	0.00830	0.009255	0.01011	0.01099	0.01166	0.01233	0.01300
Number of values	10	10	10	10	10	10	10	10	10	10
Sum of signed ranks (W)	55	55	55	55	55	55	55	55	55	55
Sum of positive ranks	55	55	55	55	55	55	55	55	55	55
Sum of negative ranks	0	0	0	0	0	0	0	0	0	0
p-value (two-tailed)	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Exact or estimate	Exact	Exact	Exact	Exact	Exact	Exact	Exact	Exact	Exact	Exact
Significance (α = 0.05)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The statistical findings shown in Table 5 of the appendix provide very high confidence that the performance differences observed among the optimized hybrid models are statistically significant. Across all optimization strategies, the Wilcoxon signed-rank test yields the same test statistic (W = 55) and a two-tailed p-value of 0.002, which is substantially lower than the conventional significance level, alpha = 0.05. Their independence from the configuration suggests that the differences from the null hypothesis are not caused by fluctuations but by systematic, repeatable, and model-performance-related effects arising from the strategies for model optimization.

A key point is that no negative ranks are observed across configurations, and all sign values are positive. The implication of this finding is that there is a stable directionality in the manner in which performance is enhanced with repeated runs (indicating that the manner in which optimization is achieved is both predictable and consistent). The fact that none of the tested hybrid models degraded relative to the reference distribution also indicates that all of them continued to perform similarly across repeated evaluations. This consistency is especially relevant in seismic modeling, which is sensitive to reliable optimization behavior, as it can lead to unreliable forecasting and a lack of confidence in the model output.

The discrepancy level, as reflected in the median, clearly ranks the optimization strategies. Smaller median errors indicate better, more stable performance, and larger median proportions indicate greater deviation. This uniform growth in median discrepancy across the optimizers demonstrates a systematic discrepancy in optimization efficacy and strength, which supports the results of the previous metric-based and visualization-based analyses. Notably, this ranking is consistent with the convergence properties and performance distributions analyzed above, providing convergent evidence across analyses.

On balance, the Wilcoxon signed-rank test indicates that the improvements in performance achieved through metaheuristic optimization are statistically significant and repeatable. In addition to the error-reduction trends, correlation improvements, efficiency increases, and convergence trends mentioned above, these statistical results provide robust quantitative evidence of the reliability and consistency of the proposed optimization framework in the seismic time-series modeling setup.

To make a comparison on the predictive accuracy of various hybrid optimization-forecasting frameworks in terms of the amount of error, the distribution of root mean squared error (RMSE) values was plotted among all the algorithms considered, as shown in Fig. 13. The groups of points use an average hybrid algorithm, which is a repeated run of the experiment or the model, and the horizontal markers represent the average RMSE and its variance for each strategy. This visualization allows comparing both average performance and robustness, and determining differences in convergence behavior and solution stability between the optimization strategies. As shown in the figure, some hybrid setups clearly outperform others in reducing RMSE; however, some exhibit greater dispersion and error rates, and in these cases, predictive accuracy is lower. The diffusion of the RMSE values also indicates the sensitivity of each optimizer to stochastic initialisation and to dynamics in the search.

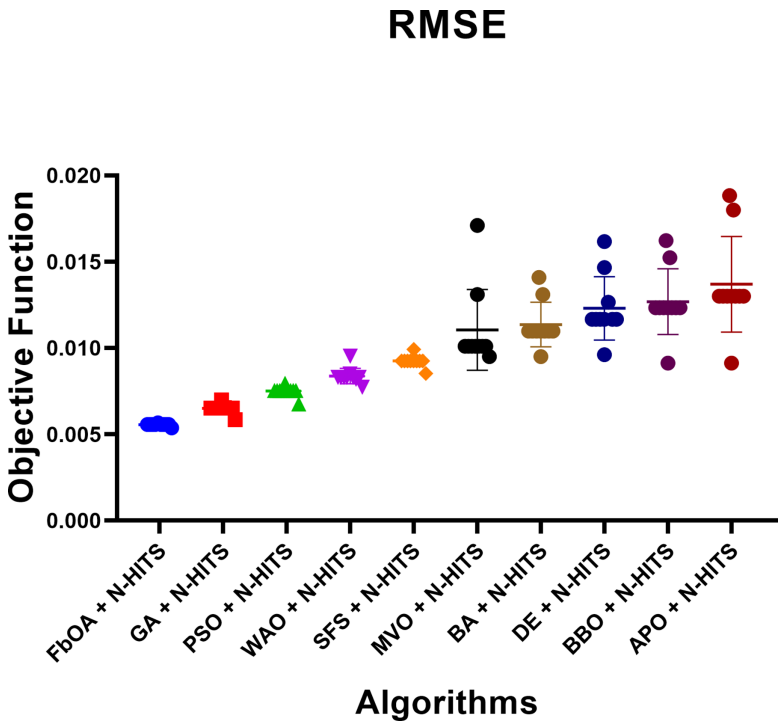

Figure 13: Distribution of RMSE values across hybrid optimization–forecasting algorithms.

To analyze the distributional properties and comparative frequency behavior of the considered algorithms, a histogram of the objective function values was created, as depicted in Fig. 14. The figure shows the spatial distribution and centrality of results, in terms of each algorithm's performance, with bars color-coded so that one can directly compare variability, dominance, and overlap between methods. This visualization shows the frequency of objective values to provide an idea of the stability and consistency of each algorithm across repeated runs; however, it does not present only summary statistics. As shown in the figure, some algorithms have tightly concentrated distributions, meaning they can perform well and have little variability, whereas others are wider or multi-peaked, meaning they are more variable and sensitive to initialization or

stochastic effects. This distributional behaviour is especially useful for gauging the reliability of algorithms in a working deployment.

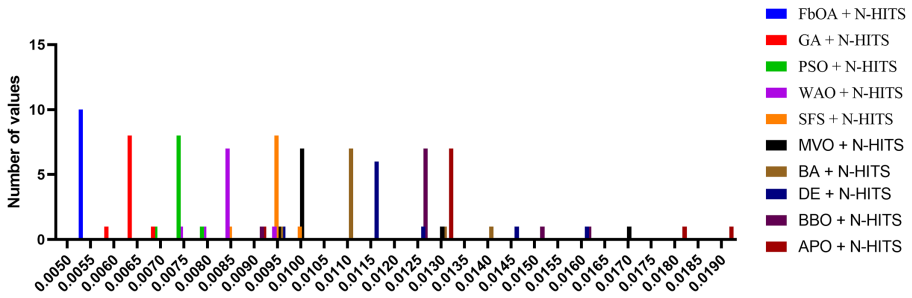

Figure 14: Histogram of objective function value distributions across evaluated algorithms.

In order to explore the statistical significance of performance variations among the compared algorithms, a heatmap representation of a one-way analysis of variance results was developed based on the results of a one-way analysis of variance (ANOVA), which is demonstrated in Fig. 15. The results of the pairwise or grouped ANOVA are also displayed on the heatmap as a colour gradient, with opposing colours indicating different degrees of statistical significance and the magnitude of the effect. This representation allows finding metrics and algorithm combinations with significant performance variations within reasonable numerical fluctuations in a short time. The figure clearly shows that there are areas of statistically significant difference between some algorithms, and uniform areas indicate similar behaviour, providing a brief summary of the inferential performance across the evaluation space.

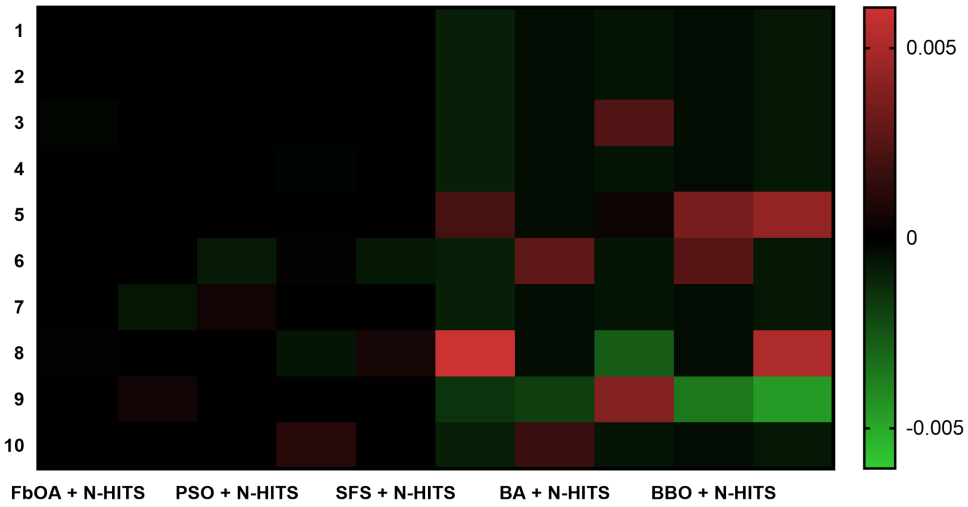

Figure 15: Heatmap of ordinary one-way ANOVA results for algorithm performance metrics.

The assumption of homoscedasticity of one-way analysis of variance (ANOVA) could be tested by producing a scatter-based diagnosis plot of residual variability, as shown in Fig. 16. This test is used to explore the distribution of standard values between groups or fitted values, so that a graphical approach can assess whether the variance is approximately fixed, a condition that helps ensure ANOVA results are valid. The homoscedasticity assumption is approximately met, as the point distribution is random and uniform, with no systematic pattern. As shown in the figure, there are no significant funneling or trend patterns in the residuals; hence, the assumption of equal variances between groups is largely met in the data analyzed.

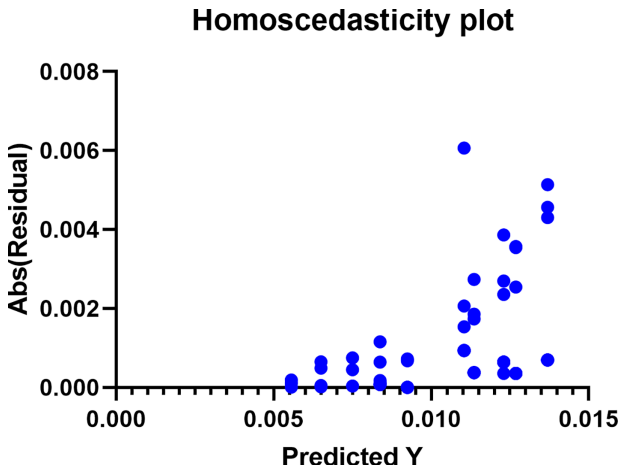

Figure 16: Homoscedasticity diagnostic plot for ordinary one-way ANOVA of algorithm performance data.

To test the normality assumption with the residuals of the one-way analysis of variance (ANOVA), a quantile-quantile (Q-Q) diagnostic plot was created as shown in Fig. 17. Q–Q plot–This is used to check the normality of the residuals, and it has the empirical quantiles of the residuals against the theoretical quantiles of a normal distribution, with the reference line running through the second quantile representing the perfect normality. Avoiding this line is a pointer to a violation of the normal assumption, especially in the tails of the distribution. Based on the figure, the residual points are close to the reference line throughout most of the quantile segment, with slight deviations at the endpoints, indicating that the normality assumption is not violated in the analyzed data.

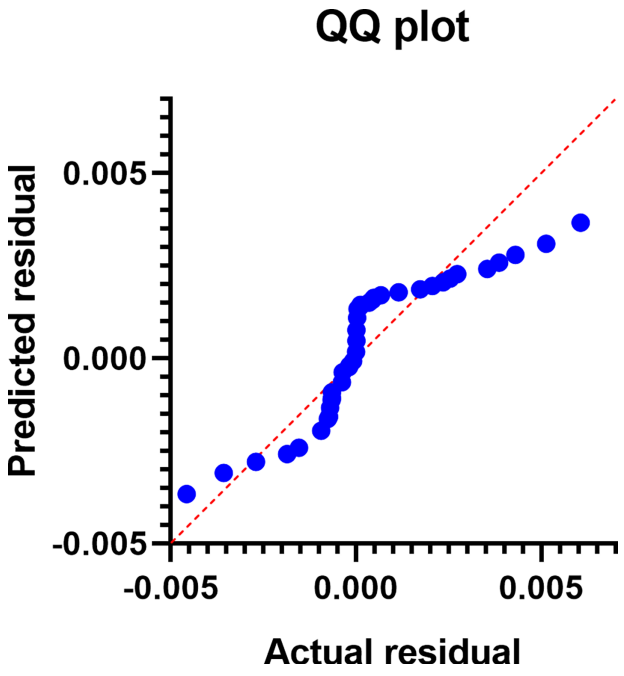

Figure 17: Q–Q plot assessing the normality of residuals for ordinary one-way ANOVA.

6 Discussion

The empirical findings presented in this paper provide strong evidence that high-performance seismic time-series prediction cannot be achieved solely through model architecture. Rather, predictive success is the product of the multidimensional interaction of data representation, learning architecture and optimization strategy. In diabolically clear terms, the experiments on the baselines showed that deep learning models can be highly variable, biased, and inefficient when trained with default or non-optimized hyperparameters. This fact supports an important lesson of seismic machine learning: more sophisticated architectures do not necessarily translate into better performance under non-stationary dynamics, partially noisy measurements, and non-uniform feature distributions, which are common in earthquake datasets.

The very high performance improvements achieved through hyperparameter optimization make evident the primary point of systematic search strategies, undoing the latent ability of deep learning models. Specifically, the Football Optimization Algorithm (FbOA) consistently produced settings and configurations that led to substantially reduced prediction errors, higher correlation with theoretical seismic activity, and higher efficiency. Such results demonstrate that FbOA can navigate the high-dimensional, nonlinear, non-continuum hyperparameter space of any modern neural architecture. The football-related mechanisms of coordinated movement, adaptive positioning, and balanced exploration-exploitation seem reasonable for avoiding suboptimal parts of the search space and gradually optimizing useful solutions.

A relative analysis of a wide range of benchmark metaheuristics also highlights the strength of the proposed set of optimization strategies. Although all the considered optimizers relative to the baseline showed positive performance improvements, distinct differences were observed in convergence stability, repeatability, and sensitivity to initialization state. FbOA showed more balanced convergence patterns and fewer oscillatory movements, indicating a more controlled, less-influenced search procedure. This stability is important, especially in seismic applications, where small changes in model structure can lead to large changes in predictive performance due to the complexity and irregularity of the underlying physics.

Statistical significance analysis provides additional confirmation that the observed improvement in performance is not temporary. It is significant that the null hypothesis has been rejected by all optimized configurations, indicating that the metaheuristic optimization results in systematic, reproducible improvements rather than random improvements due to chance alone. The fact that the median discrepancies between optimizers are monotonically ordered is another piece of evidence for structural variation in optimization efficiency and supports the conclusion that not all metaheuristics can be used to optimize deep learning-based seismic modeling. Such statistical rigor increases the validity of the empirical results and justifies the introduction of the organized optimization measures in such situations.

A key methodological observation from this analysis, regarding feature selection, is that hyperparameter optimization operates in conjunction with it. The joint training of the feature space and the optimization of model parameters effectively minimizes dimensional redundancy, enabling the optimization framework to capture important patterns in seismic dynamics. This joint optimization helps achieve improved generalization, which, in turn, is accompanied by greater efficiency and agreement. Fundamentally, these results highlight that the robustness of a model cannot be attributed solely to the quality and relevance of the features it integrates, but also to parameter calibration. Selective focus on informative attributes can significantly enhance learning results in seismic datasets, where the informative basis is encoded in the features and in the observational limitations imposed by that data type.

The combination of Large Language Model-based preprocessing with metaheuristic optimization is a major conceptual improvement of the work. Context-sensitive seismic attribute transformation is enabled by LLM-based preprocessing, which maintains geophysical context, temporal structure, and spatial coherence

before learning. This method, when used with principled optimization policies, will reduce reliance on manual heuristics and subjective decision-making, which are known to hamper scalability and reproducibility. The resulting pipeline embodies a philosophical perspective on being designed in a layered fashion, where the understanding of data, its representation, and its optimization are seen as interdependent rather than separate stages.

In addition to seismic prediction, this research can predict many things across the entire science-engineering world, where we have very complex time streams and high-dimensional parameter spaces. Environmental monitoring, energy systems modeling, climate analytics, and infrastructure forecasting also face many of the same issues as seismic datasets, such as non-stationarity, noise, and heterogeneous feature semantics. The indicated interaction among intelligent preprocessing, a deep learning architecture, and metaheuristic optimization implies a universal structure for building robust, adaptable, and scalable machine learning systems in these settings.

Lastly, this piece adds to the broader discussion of developing machine learning into data-focused, automation-based paradigms. The study, which highlights the roles of preprocessing intelligence and an optimization strategy in predictive performance, challenges the common focus on architectural novelty as the exclusive determinant of predictive performance. Rather, it promotes the use of complementary semantic understanding, structured optimization, and a rigorous evaluation modeling pipeline. This kind of strategy is a necessity for the further development of machine learning applications, where reliability, interpretability, and generalization are no less important than raw predictive accuracy.

An important practical consideration is predictive uncertainty. In the present study, the framework is evaluated using deterministic point predictions and regression-based performance metrics; therefore, the reported outputs should not be interpreted as perfectly certain estimates. Uncertainty may arise from several sources, including observational noise in seismic catalogs, non-stationary temporal behavior, regional heterogeneity, and stochastic effects associated with optimization and model training. In practical applications, higher uncertainty would reduce confidence in individual predicted magnitudes and suggests that model outputs should be interpreted as decision-support indicators rather than exact deterministic values. This is particularly relevant for engineering applications, where overconfidence in single-point predictions may lead to misleading conclusions. Future work should therefore explore uncertainty-aware extensions, such as prediction intervals, probabilistic forecasting, or ensemble-based confidence estimation, to improve the reliability and interpretability of the proposed framework.

Beyond predictive accuracy, the proposed modeling framework provides practical value for further seismic analysis. The predicted magnitude time series can be used to identify temporal trends, detect anomalous seismic activity, and support risk monitoring over specific regions. Such outputs may assist in early-stage decision support by highlighting periods of increased seismic intensity or unusual patterns that warrant further investigation. In addition, the framework can be integrated with broader geophysical analysis pipelines to support long-term seismic hazard assessment, regional activity comparison, and data-driven exploration of seismic dynamics. These applications demonstrate that the model outputs are not only predictive but also analytically informative for understanding and monitoring earthquake behavior.

7 Conclusion and Future Research

This work presented a data-centric framework for seismic time-series modeling that integrates LLM-guided preprocessing, deep learning forecasting, and metaheuristic optimization within a unified pipeline. The study was motivated by the challenges of seismic data, which are often nonlinear, non-stationary, noisy, and heterogeneous, and therefore require more than conventional fixed preprocessing and ad hoc tuning practices. The experimental results showed that preprocessing and hyperparameter selection should

be treated as central components of the modeling process in order to obtain reliable and high-precision seismic predictions.

Among the investigated baseline settings, the combination of LLM-guided preprocessing and the N-HITS forecasting model achieved the strongest baseline performance. This result indicates that context-aware preprocessing can improve the temporal representation of seismic data and better align raw earthquake records with model input requirements. More broadly, the findings suggest that guided preprocessing can provide measurable benefits across multiple evaluation aspects, including prediction error, correlation strength, and overall forecasting effectiveness, even before explicit optimization is applied.

Additional gains were obtained after feature selection and hyperparameter optimization using metaheuristic algorithms. In particular, the proposed Football Optimization Algorithm (FbOA) consistently outperformed the benchmark optimizers by producing lower prediction errors together with stronger correlation, agreement, and efficiency indicators. These results show that FbOA is effective not only in improving predictive accuracy, but also in reducing systematic bias and enhancing model generalization. Accordingly, the study demonstrates that metaheuristic optimization is not merely a supplementary refinement step, but an essential component in achieving robust deep learning performance for complex seismic forecasting tasks.

From a methodological perspective, the proposed framework provides a modular and scalable design in which data preparation, predictive learning, and optimization are clearly separated while remaining systematically connected. This structure improves reproducibility and also makes the framework adaptable to other geophysical, environmental, and data-intensive time-series applications that exhibit similar complexity. In future work, this line of research can be extended through multi-LLM collaborative preprocessing, adaptive prompting strategies, and hybrid reinforcement learning–metaheuristic approaches that allow dynamic adjustment of preprocessing and model parameters in evolving seismic monitoring environments.

Acknowledgement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R120), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R120), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Nima Khodadadi and Ebrahim A. Mattar; methodology, Sayed Elkenawy; software, Sayed Elkenawy; validation, Sayed Elkenawy, Marwa M. Eid; formal analysis, Amal H. Alharbi; investigation, Marwa M. Eid; resources, Amal H. Alharbi; data curation, Nima Khodadadi; writing—original draft preparation, Nima Khodadadi and Marwa M. Eid; writing—review and editing, Ebrahim A. Mattar and Amal H. Alharbi; visualization, Marwa M. Eid and Sayed Elkenawy; supervision, Sayed Elkenawy; project administration, Sayed Elkenawy. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in https://open.canada.ca/data/en/dataset/2c3672b6-4c17-4ff5-9861-29e2dd6d03b3/resource/f9fefce6-f183-4e7c-9917-2e210ad09fd4.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li Z. Recent advances in earthquake monitoring I: ongoing revolution of seismic instrumentation. Earthq Sci. 2021;34(2):177–88.
2. Liu X, Wang Z, Zhang Y, Shan X, Liu Z. A global coseismic InSAR dataset for deep learning: automated construction from SentineL-1 observations (2015–2024). Remote Sens. 2025;17(11):1832.
3. Chishtie FA, Clague JJ. Risky ground: seismic hazards and transectional networks in the Pacific northwest. Prog Disaster Sci. 2025;28:100483.
4. Hussain S, Pan B, Hussain W, Ali M, Sajjad MM, Afzal Z, et al. Analyzing coseismic displacement of the $M^{7.7}$ Myanmar earthquake on march 28, 2025, using Sentinel-1 InSAR data. Structures. 2025;80(8):109718. doi:10.1016/j.istruc.2025.109718.
5. Satish S, Gonaygunta H, Yadulla AR, Kumar D, Maturi MH, Meduri K, et al. Forecasting the unseen: enhancing tsunami occurrence predictions with machine-learning-driven analytics. Computers. 2025;14(5):175.
6. Fu X, Qiao M, Chen K, Huang X, Wang S. Chapter 43—enhanced disaster monitoring through earth observation and social sensing integration. In: Huang X, Wang S, Kalogeropoulos K, Tsatsaris A, editors. Data-driven earth observation for disaster management. Earth observation. Amsterdam, The Netherlands: Elsevier; 2026. p. 709–27.
7. Hu Y, Wang W, Li L, Wang F. Applying machine learning to earthquake engineering: a scientometric analysis of world research. Buildings. 2024;14(5):1393.
8. Zhao Y, Lv S, Liu P. Advances in earthquake prevention and reduction based on machine learning: a scoping review. IEEE Access. 2024;12(5):143908–29. doi:10.1109/access.2024.3467149.
9. Asadollahzadeh D, Behnam B. Machine learning approaches for seismic vulnerability assessment of urban buildings: a comparative study with analytic hierarchy process. Prog Disaster Sci. 2025;25:100398.
10. Qaedi K, Abdullah M, Yusof KA, Hayakawa M. Feasibility of principal component analysis for multi-class earthquake prediction machine learning model utilizing geomagnetic field data. Geosciences. 2024;14(5):121. doi:10.3390/geosciences14050121.
11. Tian P, Wang C, Chan TM, Elghazouli AY. Seismic time history response prediction of modular buildings using transformer-based machine learning models. J Constr Steel Res. 2026;240(10):110256. doi:10.1016/j.jcsr.2026.110256.
12. Devi DR, Govindarajan P, N V. Towards real-time earthquake forecasting in Chile: integrating intelligent technologies and machine learning. Comput Electr Eng. 2024;117:109285.
13. Emre Yavas C, Chen L, Kadlec C, Ji Y. Predictive modeling of earthquakes in los angeles with machine learning and neural networks. IEEE Access. 2024;12(5):108673–702. doi:10.1109/access.2024.3438556.
14. S P, Siva C, Cindrelaa J. Machine learning-based ionospheric anomaly detection for earthquake prediction in northeast region, India. In: Proceedings of the 2025 IEEE Space, Aerospace and Defence Conference (SPACE); 2025 Jul 21–23; Bangalore, India. p. 1–6.
15. Ahmed SMS, Güneyli H. The evolution of seismic tomography in earth sciences—advancements, limitations, and its AI-enabled future (a critical review). Earthq Res Adv. 2025:100425. doi:10.1016/j.eqrea.2025.100425.
16. Zhang R, Li H, Shen Y, Yang J, Li W, Zhao D, et al. Deep learning applications in ionospheric modeling: progress, challenges, and opportunities. Remote Sens. 2025;17(1):124.
17. Li C, Li Z, Duan M, Zhou L. Spatiotemporal distribution of the magnitude of completeness and b-values in mainland China based on a fused multi-source earthquake catalog. Entropy. 2025;27(11):1137. doi:10.3390/e27111137.
18. Jafari A, Fox G, Rundle JB, Donnellan A, Ludwig LG. Time series foundation models and deep learning architectures for earthquake temporal and spatial nowcasting. GeoHazards. 2024;5(4):1247–74. doi:10.3390/geohazards3020011.
19. Shabrawy M, Abdel-Hamid NB, El-Kenawy ESM, Abdelsalam MM. Comparative advances in AI-driven earthquake intelligence: machine learning, deep learning, and large language models for prediction and emergency management. Metaheur Optim Rev. 2026;1:1–25.

20. Bacanin N, Jovanovic L, Dobrojevic M, Zivkovic M, Antonijevic M, Salb M. Exploring the potential of modified metaheuristic optimized long short-term memory neural networks for earthquake magnitude forecasting. In: Multi-objective optimization techniques. Boca Raton, FL, USA: CRC Press; 2025.
21. Lu L, Liu Z, Zhang H, Wu C, Sun J, Ma X, et al. Research on seismic activity and seismic structural characteristics of the Shandong region. Sci Rep. 2025;15(1):12293. doi:10.1038/s41598-025-96305-y.
22. Aloraini B, Cekim HO, Karakavak HN, Ozel G. Fuzzy C-means clustering and LSTM-based magnitude prediction of earthquakes in the Aegean region of Türkiye. Sci Rep. 2025;15(1):1–24. doi:10.1038/s41598-025-07538-w.
23. Di Naccio D, Di Lorenzo C, Falcone G, Kastelic V, Sparacino F, Del Sole L, et al. The impact of long-term seismic coupling on fault-based seismic hazard models: insights from the central Apennines (Italy). npj Nat Hazards. 2025;2(1):97. doi:10.1038/s44304-025-00150-y.
24. Zhang S, Huang M, Liu S, Meng F, Xie Y, Ren X, et al. AI-driven post-earthquake emergency material demand prediction: integrating RAG with reasoning large language model. IEEE Access. 2025;13:100630–46.
25. Zhang W, Zhang K, Li T, Deng W. Research on generation and quality evaluation of earthquake emergency language service contingency plan based on chain-of-thought prompt engineering for LLMs. Inventions. 2025;10(5):74. doi:10.3390/inventions10050074.
26. Mohamed A, Rashid ME, Shaalan K. In-context learning in large language models (LLMs): mechanisms, capabilities, and implications for advanced knowledge representation and reasoning. IEEE Access. 2025;13:95574–93.
27. Hong K, Park Y. Large language models for semantic join: a comprehensive survey. IEEE Access. 2025;13:184478–93. doi:10.1109/access.2025.3625753.
28. Chen J, Liu Z, Huang X, Wu C, Liu Q, Jiang G, et al. When large language models meet personalization: perspectives of challenges and opportunities. World Wide Web. 2024;27(4):42.
29. Jia J, Ye W. Deep learning for earthquake disaster assessment: objects, data, models, stages, challenges, and opportunities. Remote Sens. 2023;15(16):4098.
30. Gürsoy G, Varol A, Nasab A. Importance of machine learning and deep learning algorithms in earthquake prediction: a review. In: Proceedings of the 2023 11th International Symposium on Digital Forensics and Security (ISDFS); 2023 May 11–12; Chattanooga, TN, USA. p. 1–6.
31. Laurenti L, Tinti E, Galasso F, Franco L, Marone C. Deep learning for laboratory earthquake prediction and autoregressive forecasting of fault zone stress. Earth Planet Sci Lett. 2022;598:117825. doi:10.5194/egusphere-egu22-9833.
32. Dascher-Cousineau K, Shchur O, Brodsky EE, Günnemann S. Using deep learning for flexible and scalable earthquake forecasting. Geophys Res Lett. 2023;50(17):e2023GL103909.
33. Abdalzaher MS, Soliman MS, El-Hady SM, Benslimane A, Elwekeil M. A deep learning model for earthquake parameters observation in IoT system-based earthquake early warning. IEEE Internet Things J. 2022;9(11):8412–24. doi:10.1109/jiot.2021.3114420.
34. Wang Y, Li X, Wang Z, Liu J. Deep learning for magnitude prediction in earthquake early warning. Gondwana Res. 2023;123(5620):164–73. doi:10.1016/j.gr.2022.06.009.
35. Liu Y, Zhao Q, Wang Y. Peak ground acceleration prediction for on-site earthquake early warning with deep learning. Sci Rep. 2024;14(1):5485. doi:10.1038/s41598-024-56004-6.
36. Zhang X, Zhang M, Tian X. Real-time earthquake early warning with deep learning: application to the 2016 M 6.0 Central Apennines, Italy earthquake. Geophys Res Lett. 2021;48(5):2020GL089394.
37. Zhu W, Tai KS, Mousavi SM, Bailis P, Beroza GC. An end-to-end earthquake detection method for joint phase picking and association using deep learning. J Geophys Res Solid Earth. 2022;127(3):e2021JB023283.
38. Yang L, Liu X, Zhu W, Zhao L, Beroza GC. Toward improved urban earthquake monitoring through deep-learning-based noise suppression. Sci Adv. 2022;8(15):eabl3564. doi:10.1126/sciadv.abl3564.
39. Kang S, Zhou R, Kumar R, Dong Z, Yu Y, Singh V, et al. Deep learning method for post-earthquake damage assessment of frame structures based on time–frequency analysis and CGAN. Earth Syst Environ. 2025;9(1):403–20.
40. Kuang W, Yuan C, Zhang J. Real-time determination of earthquake focal mechanism via deep learning. Nat Commun. 2021;12(1):1432. doi:10.1038/s41467-021-21670-x.

41. Yang M, Xu X, Zhu S. Stochastic characteristics of vehicle-bridge vibration under earthquakes with parameter uncertainty: a deep learning-based model. Structures. 2025;77(3):108774. doi:10.1016/j.istruc.2025.108774.
42. Bao Z, Zhao J, Huang P, Yong S, Wang X. A deep learning-based electromagnetic signal for earthquake magnitude prediction. Sensors. 2021;21(13):4434. doi:10.3390/s21134434.
43. Jiang Z, Zhu Z, Lacidogna G, Friedrich LF, Iturrioz I. Earthquake precursors based on rock acoustic emission and deep learning. Sci. 2025;7(3):103. doi:10.3390/sci7030103.
44. Shen X, Hou B, Lu J, Li S. Comparative analysis of deep learning methods for real-time estimation of earthquake magnitude. Appl Sci. 2025;15(5):2587. doi:10.3390/app15052587.
45. Yilmaz M, Yalcin E, Demir F, Ozdemir AM, Atar M, Gunes A, et al. Automatic segmentation of asphalt cracks on highways after large-scale and severe earthquakes using deep learning-based approaches. IEEE Access. 2025;13(2):22820–30. doi:10.1109/access.2025.3536554.
46. Gamboa-Chacón S, Meneses E, Chaves EJ. Analysis of earthquake detection using deep learning: evaluating reliability and uncertainty in prediction methods. Comput Geosci. 2025;197:105877.
47. Zhu J, Sun W, Li S, Yao K, Song J. Threshold-based earthquake early warning for high-speed railways using deep learning. Reliab Eng Syst Saf. 2024;250(1):110268. doi:10.1016/j.ress.2024.110268.
48. Gong M, Liu B, Wang X, Zhou B, Zhao Y, Jia J. Damage assessment of reinforced concrete frame under mainshock-aftershock based on deep learning considering pre-earthquake damage. J Build Eng. 2025;100:111729. doi:10.1016/j.jobe.2024.111729.
49. Amin MS, Ahn H. Earthquake disaster avoidance learning system using deep learning. Cogn Syst Res. 2021;66(5):221–35. doi:10.1016/j.cogsys.2020.11.002.
50. Liu X, Zhang Y, Shan X, Wang Z, Gong W, Zhang G. Deep learning for automatic detection of volcanic and earthquake-related InSAR deformation. Remote Sens. 2025;17(4):686. doi:10.3390/rs17040686.
51. Cheng Q, Ren H, Meng X, Li A, Xie L. Real-time seismic response prediction method of high-rise buildings based on deep learning for earthquake early warning. Int J Disaster Risk Reduct. 2025;119:105294. doi:10.1016/j.ijdrr.2025.105294.
52. Zhuang X, Tran TV, Nguyen-Xuan H, Rabczuk T. Deep learning-based post-earthquake structural damage level recognition. Comput Struct. 2025;315:107761. doi:10.1016/j.compstruc.2025.107761.
53. Tao Y, Xu ZD, Wei Y, Liu XY, Zang X, Li SD. A wavelet packet deep learning model for energy-based structural collapse assessment under earthquake-fire scenarios: framework and hybrid simulation. Mech Syst Signal Process. 2025;222:111784. doi:10.1016/j.ymssp.2024.111784.
54. Challu C, Olivares KG, Oreshkin BN, Ramirez FG, Canseco MM, Dubrawski A. Nhits: neural hierarchical interpolation for time series forecasting. Proc AAAI Conf Artif Intell. 2023;37:6989–97.
55. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. Proc AAAI Conf Artif Intell. 2021;35:11106–15.
56. Popov S, Morozov S, Babenko A. Neural oblivious decision ensembles for deep learning on tabular data. arXiv:1909.06312. 2019.