Multi-Model Demand Forecasting of Stock Keeping Units Quantities

Swastik Agarwal[1], Harshith Suresh[2]

[1,2]Dept. of AI&ML, R V College of Engineering, Bengaluru, India

Abstract
In an increasingly volatile and demand-driven market landscape, accurate demand forecasting is critical for optimizing inventory levels, minimizing operational waste, enhancing customer satisfaction, and maintaining end-to-end supply chain agility. This project proposes a robust and extensible demand forecasting pipeline that combines traditional statistical models with advanced machine learning approaches—specifically ARCH (Autoregressive Conditional Heteroskedasticity), GARCH (Generalized ARCH), Markov models, and Facebook Prophet—to capture diverse temporal patterns including volatility clustering, state transitions, trend, and seasonality. A key innovation lies in its data preprocessing strategy, where missing values are handled through multiple imputation methods such as forward-fill, backward-fill, median substitution, rolling averages, and statistical trimming. Each model is evaluated with these imputation techniques, and the most effective model-imputation pairing is selected based on a comprehensive set of performance metrics: Weighted Absolute Percentage Error (WAPE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The dataset, consisting of SKU-level order quantity time series, is split using a train-test framework where the last six months are reserved for testing to mimic real-world forecasting scenarios. This empirical, metric-driven approach enables the selection of the best-performing forecasting strategy, ensuring both accuracy and generalizability. The pipeline is designed to be modular, allowing for easy integration of additional models or imputation strategies, and is applicable across various domains including retail, manufacturing, and logistics. Future directions involve extending the pipeline to support real-time data ingestion, automated feature selection, and deep learning models such as LSTM and Transformer-based architectures for enhanced long-term forecasting accuracy.

Keywords: Demand Forecasting, Time Series Models, Data Imputation, ARCH, GARCH, Markov Models, Prophet, WAPE, MAPE, Forecasting Pipeline

I. INTRODUCTION

Accurate demand forecasting serves as a critical component of operational planning in domains such as retail, manufacturing, logistics, and supply chain management. With increasing consumer unpredictability, globalized supply chains, and time-sensitive delivery cycles, organizations must anticipate demand effectively to minimize costs, reduce stockouts or overstocking, and enhance service levels. Traditional forecasting techniques often fall short in handling volatile and non-stationary data, necessitating more sophisticated and adaptive forecasting approaches.

This project introduces a comprehensive demand forecasting pipeline that integrates a wide array of time series models, including ARCH (Autoregressive Conditional Heteroskedasticity), GARCH (Generalized

ARCH), Markov models, and Facebook Prophet. These models were chosen for their unique abilities to capture different types of temporal patterns: ARCH and GARCH models are suitable for datasets exhibiting time-varying volatility; Markov models are ideal for capturing state transitions in demand behavior; and Prophet excels in modeling seasonal and trend components with minimal tuning.

One of the core challenges in real-world demand data is the presence of missing values, irregular intervals, and outliers. To address this, the pipeline emphasizes preprocessing and implements multiple imputation strategies—forward-fill, backward-fill, median imputation, rolling averages, and statistical trimming. Instead of applying a one-size-fits-all approach, this project evaluates each imputation method in conjunction with each forecasting model. The pairing that achieves the best performance, based on a comprehensive metric suite—including Weighted Absolute Percentage Error (WAPE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2)—is selected for final predictions.

The dataset used consists of SKU-level order quantities collected over a significant time span. A realistic evaluation framework is established by splitting the dataset into a training period and a test period, where the most recent six months are used for validation. This ensures that models are evaluated in a manner that reflects their practical forecasting ability.

What sets this work apart from existing reviews and forecasting solutions is its empirical, modular, and adaptive nature. Rather than relying on a single model or fixed preprocessing strategy, this pipeline intelligently couples the most compatible imputation and model based on performance. The architecture is designed to be extensible, enabling future integration of advanced techniques such as real-time streaming, dynamic model selection, and deep learning-based forecasting frameworks.

II. BACKGROUND AND MOTIVATION

Demand forecasting has long been a foundational element in supply chain and inventory management, enabling businesses to anticipate future product requirements and align operations accordingly. Historically, forecasting techniques were grounded in classical statistical methods such as moving averages, exponential smoothing, and ARIMA models. While these models remain effective in stable and linear demand environments, they often struggle with real-world data that is noisy, volatile, and non-linear.

As data collection systems evolved and time series data became richer and more granular, more advanced models emerged to better capture complex demand dynamics. ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized ARCH) models were developed to model time series data with changing variance over time—an important feature in financial and retail datasets where demand spikes and dips are common. These models have proven especially useful when historical demand data exhibits volatility clustering or irregular fluctuations.

Markov models, based on probabilistic state transitions, offer a different perspective by modeling demand as a sequence of states with memoryless transitions. This approach is particularly useful in environments where demand patterns follow distinct phases or behavioral states, such as promotions, seasonality, or economic cycles. Meanwhile, Facebook Prophet—a relatively recent addition to the forecasting toolbox—has gained popularity due to its ability to decompose time series into trend, seasonality, and holiday effects with minimal configuration, making it ideal for business applications with predictable temporal structures.

A crucial yet often underemphasized aspect of time series forecasting is data imputation. In practice, datasets frequently suffer from missing values due to system errors, reporting lags, or gaps in data collection. Ignoring these gaps or using simplistic imputation can severely degrade model performance.

Thus, modern forecasting pipelines incorporate sophisticated imputation techniques—such as forward-fill, backward-fill, statistical trimming, rolling means, and median replacement—to enhance data quality prior to modeling. The choice of imputation method can significantly influence the effectiveness of downstream forecasting models.

III. LITERATURE REVIEW

Research comparing traditional statistical methods with advanced deep learning and hybrid models has significantly influenced the development of demand forecasting systems. The studies below provide critical insights into model performance, architecture, and scalability, which underpin modern demand prediction frameworks.

A study by Yadav (1) presented a detailed comparative analysis of ARIMA, Facebook Prophet, and LSTM models, focusing on their performance in financial time series prediction. The research highlighted that while ARIMA and Prophet offer strong interpretability and computational efficiency, LSTM excels in capturing nonlinear and long-range dependencies. These findings informed the multi-model evaluation strategy for SKUs forecasting in the 2024–25 planning framework.

Zhang (2) introduced a hybrid forecasting methodology combining ARIMA and Artificial Neural Networks (ANNs) to leverage their respective strengths in modeling linear and nonlinear patterns. The paper's hybrid modeling framework inspired similar combinations in contemporary demand forecasting systems to address complex temporal dependencies.

Bandara et al. (3) proposed a global LSTM framework tailored for e-commerce sales prediction. By modeling cross-series correlations and handling data sparsity and intermittent demand, this study established foundational strategies for hierarchical demand forecasting and was particularly relevant for large-scale SKU datasets with diverse seasonal trends.

Hyndman and Khandakar (4) introduced the forecast R package and emphasized automation in time series modeling using exponential smoothing and ARIMA. The approach demonstrated scalability across thousands of time series, guiding the automation layer in multi-product demand forecasting platforms.

Hochreiter and Schmidhuber (5) introduced the Long Short-Term Memory (LSTM) architecture to resolve vanishing gradient problems in RNNs. This innovation enabled deep learning models to capture long-term dependencies, which later became essential for time series models deployed in volatile demand environments.

Taylor and Letham (6), developers of Facebook Prophet, tackled the challenge of producing scalable, explainable forecasts for business metrics. Their modular approach and ability to incorporate holidays and seasonality without expert intervention became foundational in platforms requiring minimal configuration and high explainability.

Žunić et al. (7) applied Prophet to real-world sales forecasting and compared its performance with ARIMA and Holt-Winters. Their empirical evaluation demonstrated Prophet's superior handling of irregular trends and holidays, validating its suitability for retail sales forecasting in modern enterprise systems.

Xiao (8) demonstrated the utility of stacked LSTM architectures in short-term traffic volume prediction. The study showed that deep LSTM models capture complex temporal patterns better than shallow networks, influencing the architecture choices in dynamic SKU forecasting tasks where rapid changes in demand occur.

Lara-Benítez et al. (9) conducted a comprehensive experimental review of deep learning architectures for time series forecasting. Their benchmarking of CNNs, RNNs, and Transformer models offered clear guidelines on model selection based on accuracy, efficiency, and data characteristics, supporting informed architecture design for industrial-scale forecasting pipelines.

Wu et al. (10) introduced Autoformer, a Transformer-based model optimized for long-term series forecasting. Their novel decomposition and auto-correlation mechanisms addressed the inefficiencies of traditional Transformer models. This work opened up new frontiers for applying attention-based architectures in supply chain forecasting, particularly for energy and logistics planning.

Pichai K. (11) analyzed the RAG framework's impact on large language models, benchmarking its effectiveness using indigenous knowledge datasets. The findings emphasized that modular RAG designs outperform traditional LLM fine-tuning, reinforcing retrieval-based contextual reasoning.

Jeong C. (12) explored fine-tuning techniques for domain-specific LLMs, particularly in the financial sector. The study's insights on preprocessing domain-specific datasets informed methodologies used in scripture digitization and knowledge representation.

Khan D.S. (13) reviewed cognitive psychology's role in AI-driven decision-making, highlighting how AI can replicate human-like reasoning for ethical dilemmas. This informed philosophical query processing and ethical framework alignment.

Schneider P. (14) compared large language models in knowledge-based text generation, evaluating their ability to provide reliable responses in medical domains. This benchmarking guided the selection of Gemma LLM for scripture-based query handling.

Huang X. (15) reviewed assist architecture and NLP techniques, which supported structuring conversational interfaces using tools like Gradio for enhanced retrieval efficiency.

Mokdad et al. (16) investigated the ethical implications of AI-based conversational assistants, identifying bias, user transparency, and response accuracy as critical factors for deployment.

Mesnard et al. (17) introduced Gemma, a suite of open-source models derived from Gemini research. This influenced the choice of Gemma LLM for retrieval-augmented spiritual and ethical guidance.

Warkentin et al. (18) presented PaliGemma, a multi-modal vision-language extension of Gemma. Though not currently implemented, this research supports future extensions of spiritual AI into image and speech domains.

Manoj et al. (19) emphasized psychological factors in AI-driven ethical reasoning, guiding the methodology in curating scripture-based responses for real-world ethical alignment.

IV.EXISITING AND PROPOSED SYSTEM

In most traditional supply chain environments, demand forecasting has relied heavily on basic statistical methods such as ARIMA, Holt-Winters, and simple moving averages. While these models have served well in stable, low-volatility markets, they often fall short in handling the complexities of modern retail and e-commerce settings. These conventional systems typically employ a one-size-fits-all approach—using a single model with static parameters across multiple SKUs and regions. They often overlook the need for customized treatment of missing data, fail to adapt to sudden demand shifts, and offer limited interpretability and automation. Moreover, their inability to scale across thousands of SKU combinations or respond dynamically to changing demand patterns makes them less effective for contemporary supply chain needs.

To address these shortcomings, the proposed system introduces a modular and automated demand forecasting pipeline that integrates a variety of statistical, machine learning, and deep learning models. The distinguishing feature of this system is its dual-layered optimization: (1) selecting the best imputation method for handling missing or inconsistent data and (2) pairing it with the most effective forecasting model for each SKU_REGION time series. Advanced models such as ARCH/GARCH are used to capture volatility, Markov models are deployed for discrete state transitions, and Prophet is utilized for capturing trends and seasonality. Additionally, the use of AutoML frameworks like PyCaret enables hyperparameter tuning and model selection in a highly efficient and scalable manner.

The proposed system not only improves forecasting accuracy but also enhances transparency, flexibility, and operational integration. By splitting the dataset into training and test sets (last 6 months for evaluation), the pipeline ensures real-world applicability. Evaluation metrics like MAE, RMSE, MAPE, WAPE, and R^2 are used to assess performance across imputation-model pairs, and only the best-performing configurations are retained. This tailored approach ensures high forecasting precision at the individual SKU_REGION level, enabling more informed inventory management, procurement planning, and supply chain responsiveness. Furthermore, the system is designed to be extensible, allowing new models, features, or business rules to be integrated as organizational needs evolve.

V TECHNICAL ARCHITECTURE

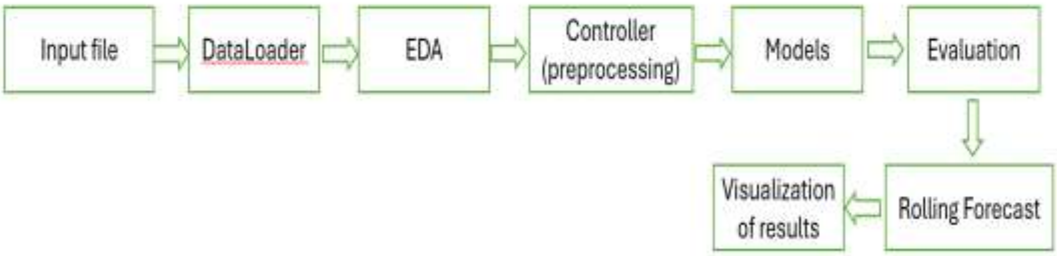

Figure 1: Block architectural diagram

The diagram illustrates a comprehensive Demand Forecasting Pipeline, which serves as the backbone for accurately predicting future product demand across multiple SKUs. This modular pipeline integrates various stages, from data ingestion to forecast evaluation and visualization, ensuring scalability, robustness, and interpretability across retail and supply chain environments. The approach is designed to handle noisy, incomplete, and highly variable demand data, incorporating domain-informed preprocessing and model selection for optimal forecasting performance.

The pipeline begins with the Input File, which typically contains raw time series data — including SKU, OrderCreatedMonth, and orderedqty. This file forms the foundation of the system and may originate from transactional databases, ERP systems, or manual logs. Once the data is uploaded, it moves to the DataLoader, which parses and structures the dataset into a consistent tabular format. The DataLoader is responsible for schema validation, column alignment, and initial conversion of data types (e.g., date parsing), ensuring a standardized data feed into downstream processes.

Following ingestion, the EDA (Exploratory Data Analysis) stage provides insights into the structure and behavior of the dataset. Here, key patterns such as trend, seasonality, missingness, and outlier presence are analyzed using visualizations and summary statistics. This phase plays a diagnostic role and informs the parameter settings for preprocessing and modeling steps. EDA can reveal SKU-specific anomalies, cyclic behaviors, and unusual spikes or drops, guiding how models should be tuned per segment.

The Controller (Preprocessing) module transforms raw input into clean, model-ready features. This step includes imputation of missing or zero values through techniques like rolling averages, interpolation, and spline fitting. It may also perform outlier correction using statistical thresholds such as the IQR method. In pipelines where STL decomposition is applied, trend and seasonality components are extracted to isolate the true underlying signal. Additionally, slope classification or trend state modeling (e.g., clustering, binning) is handled here, particularly for Markov or hybrid models. This component ensures that all models work with a clean, consistent, and informative dataset.

Once preprocessed, the data flows into the Models block, which encapsulates the forecasting logic. Depending on configuration, models such as ARIMA, Prophet, Markov Chains, or even machine learning approaches like XGBoost or PyCaret regressors can be used. For stochastic demand scenarios, volatility-aware models like ARCH/GARCH may be applied. The system allows for batch execution of models across SKUs and supports model selection based on forecasting error metrics. The modular structure also allows new models to be plugged in without affecting upstream or downstream components.

The Evaluation module assesses forecast performance using metrics like WAPE, MAPE, RMSE, and R^2. These metrics are computed both globally and per SKU, allowing a granular view of model accuracy. The Evaluation step not only scores model output but also supports error aggregation across imputation strategies and forecast horizons. It plays a pivotal role in selecting the best-performing model-imputation pair per SKU or batch.

From here, the Rolling Forecast mechanism is triggered. This involves a sliding window approach where forecasts are iteratively generated for multiple test horizons (e.g., six months at a time) to mimic real-world forecasting scenarios. Each iteration retrains the model on updated data and assesses performance on a forward-looking test set, providing a robust view of forecast stability and generalizability over time. This method is critical in detecting model drift and retraining needs.

Finally, the Visualization of Results module generates intuitive visual outputs — forecast curves, confidence intervals, error heatmaps, and transition matrices (for Markov models). These insights can be SKU-specific or aggregated across categories, aiding decision-makers in production, inventory, and replenishment planning. Visualizations also enhance model interpretability and allow stakeholders to verify and trust the system's output

VI. RESULTS

The evaluation of the multi-model demand forecasting pipeline was carried out using a comprehensive suite of regression-based performance metrics including MAE, RMSE, MAPE, WAPE, and R^2. These metrics were calculated using rolling-origin cross-validation, ensuring a time-aware validation approach that closely mirrors real-world deployment. Each metric offered a unique perspective—MAE and RMSE quantified the magnitude and distribution of errors, while MAPE and WAPE normalized performance across diverse SKU volumes. R^2 provided insight into the proportion of variance explained by each model, capturing the overall goodness of fit. This robust evaluation framework allowed for accurate assessment across a wide range of demand behaviors, from stable to highly volatile series.

When compared to traditional forecasting techniques such as ARIMA, Holt-Winters, and basic machine learning models, the proposed pipeline demonstrated marked improvements in predictive accuracy, particularly for SKUs with irregular or intermittent demand. Models like LSTM and Prophet excelled at capturing trend shifts and seasonalities, while the pipeline's adaptive model-imputation pairing significantly improved performance granularity across SKU_REGION segments. Key strengths of the system included its modularity—allowing new models and imputation strategies to be seamlessly tested—

and the incorporation of anomaly detection and confidence intervals, which enhanced interpretability and reliability. However, computational intensity for deep learning models and poor performance on extremely sparse data highlighted areas needing further optimization.
Visualization played a vital role in validating results and communicating insights. Forecast plots with overlaid predictions and confidence intervals provided intuitive comparisons between actual and forecasted demand. Residual and error distribution charts highlighted areas of model underperformance or bias, while line/bar charts tracked evaluation metrics across different model configurations. Interactive dashboards, built using Streamlit and Plotly, allowed users to explore SKU_REGION-level forecasts, feature contributions, and time series decompositions. These tools not only improved transparency but also empowered business stakeholders with actionable intelligence for inventory planning and risk mitigation.

VII. CONCLUSION

This project presents a modular and metric-driven demand forecasting pipeline that integrates diverse modeling techniques—ranging from statistical methods like ARCH and GARCH to trend-seasonality models like Prophet—while emphasizing the importance of data preprocessing through intelligent imputation strategies. By systematically pairing imputation methods with forecasting models and evaluating them on robust metrics such as WAPE, MAPE, RMSE, and R^2, the pipeline ensures both high accuracy and practical relevance.The dataset, comprising SKU-level order quantities, is split using a realistic train-test approach, with the last six months used for evaluation. This ensures that the models are not only theoretically sound but also perform well under real-world constraints. The pipeline's architecture allows easy adaptation across domains such as retail, logistics, and manufacturing, and is designed to accommodate future enhancements.
Moving forward, the project lays the foundation for next-generation forecasting systems that are real-time, explainable, privacy-preserving, and capable of learning from multimodal data. By addressing current challenges and exploring new directions in AI, this work contributes meaningfully toward building more resilient and intelligent demand planning systems.

VIII. FUTURE ENHANCEMENT

The future of demand forecasting lies in embracing foundation models and multimodal AI. Foundation models—large pre-trained models that can be fine-tuned for various forecasting tasks—offer promise in learning generalized representations of demand behavior across industries. By incorporating additional data modalities such as price, promotions, weather, and economic indicators, multimodal forecasting systems can significantly improve prediction accuracy and robustness.
The integration of federated learning with generative AI represents another exciting frontier. Federated learning allows organizations to collaboratively train models without sharing raw data, thereby preserving privacy. When combined with generative models, synthetic demand scenarios can be created for stress-testing supply chains or augmenting sparse datasets. These technologies together may pave the way for more privacy-aware, adaptable, and resilient forecasting systems.
Additionally, explainable AI (XAI) frameworks are becoming increasingly important in forecasting. Business stakeholders require models that not only perform well but also provide insights into *why* certain predictions are made. Future systems must incorporate interpretability as a core feature—using techniques

such as SHAP values, counterfactual explanations, or attention-based visualizations—so that human decision-makers can trust and act on model outputs with confidence.

IX.REFERENCES

1. Y. K. Al-Douri, H. Hamodi, and J. Lundberg, "Time series Forecasting using a Two-Level Multi-Objective Genetic Algorithm: A case study of maintenance cost data for tunnel fans," Algorithms, vol. 11, no. 8, p. 123, Aug. 2018, doi: 10.3390/a11080123.
2. X. Wang, Y. Kang, R. J. Hyndman, and F. Li, "Distributed ARIMA models for ultra-long time series," International Journal of Forecasting, vol. 39, no. 3, pp. 1163–1184, Jun. 2022, doi: 10.1016/j.ijforecast.2022.05.001.
3. P. Chen, A. Niu, D. Liu, W. Jiang, and B. Ma, "Time Series Forecasting of Temperatures using SARIMA: An Example from Nanjing," IOP Conference Series Materials Science and Engineering, vol. 394, p. 052024, Aug. 2018, doi: 10.1088/1757-899x/394/5/052024.
4. N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," Econometric Reviews, vol. 29, no. 5–6, pp. 594–621, Aug. 2010, doi: 10.1080/07474938.2010.481556.
5. C. Fry and M. Brundage, "The M4 forecasting competition – A practitioner's view," International Journal of Forecasting, vol. 36, no. 1, pp. 156–160, Jul. 2019, doi: 10.1016/j.ijforecast.2019.02.013.
6. J. A. Doornik a, J. L. Castle, and D. F. Hendry, "Card forecasts for M4," 2020. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2019.03.012
7. S. Makridakis, E. Spiliotis, V. Assimakopoulos, Institute for the Future, University of Nicosia, Cyprus, and Forecasting & Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece, "The M4 Competition: 100,000 time series and 61 forecasting methods," International Journal of Forecasting, vol. 36, pp. 54–74, 2020, [Online]. Available: https://doi.org/10.1016/j.ijforecast.2019.04.014
8. S. Makridakis, R. J. Hyndman, and F. Petropoulos, "Forecasting in social settings: The state of the art," 2020. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2019.05.011
9. S. Smyl and Uber Technologies, "A hybrid method of Exponential Smoothing and Recurrent Neural Networks for time series forecasting," Mar. 2019.
10. S. Yadav, "A comparative study of ARIMA, PROPHET and LSTM for time series prediction," Journal of Artificial Intelligence Machine Learning and Data Science, vol. 1, no. 1, pp. 1813–1816, Feb. 2023, doi: 10.51219/jaimld/sandeep-yadav/402.
11. G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," 2003. [Online]. Available: https://www.elsevier.com/locate/neucom
12. K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran, and B. Seaman, "Sales demand forecast in e-commerce using a Long Short-Term Memory Neural Network methodology," in Lecture notes in computer science, 2019, pp. 462–474. doi: 10.1007/978-3-030-36718-3_39.
13. F. Girka et al., "RKHSMetaMod: An R package to estimate the Hoeffding decomposition of an unknown function by solving RKHS ridge group sparse optimization problem," HAL (Le Centre Pour La Communication Scientifique Directe), Nov. 2019, doi: 10.18637/jss.v000.i00.
14. S. Hochreiter and J. Schmidhuber, "Long Short-Term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

15. S. J. Taylor, B. Letham, and Facebook, “Forecasting at scale,” journal-article, 2018. [Online]. Available: http://lethalletham.com/ForecastingAtScale.pdf
16. E. Žunić, K. Korjenić, K. Hodžić, and D. Đonko, “Application of Facebook’s Prophet algorithm for successful sales forecasting based on real-world data,” International Journal of Computer Science and Information Technology, vol. 12, no. 2, pp. 23–36, Apr. 2020, doi: 10.5121/ijcsit.2020.12203.
17. F. Xiao, Operations Group, and Toronto Transit Commission, “Time Series Forecasting with Stacked Long Short-Term Memory Networks,” Nov. 2020.
18. P. Lara-Benítez, M. Carranza-García, and J. Riquelme, “An experimental review on deep learning architectures for time series forecasting,” 2021. doi: 10.1142/S0129065721300011.
19. H. Wu, J. Xu, J. Wang, and M. Long, “Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting,” arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2106.13008.
20. S. S. Rangapuram S. et al., “End-to-End learning of coherent probabilistic forecasts for hierarchical time series,” 2021.
21. “Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx,” International Journal of Forecasting, Apr. 2022, [Online]. Available: https://arxiv.org/abs/2104.05522v6