Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X

Volume 3, Issue 5, 2025

OPTIMIZATION OF TRANSFORMER LOAD FORECASTS IN SMART GRIDS THROUGH AI-DRIVEN REGRESSION AND WEATHER DATA FUSION

Muhammad Yaseen[1], Muhammad Afnan[2], Kiran Raheel[3], Ali Mujtaba Durrani[*4], Khalid Rehman[5], Zaheer Farooq[6], Muhammad Imran[7], Abdul Aziz[8]

[1,2,3,*4,5,6,7,8]*Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar, KPK, Pakistan*

[1]ykyaseen11@gmail.com, [2]mafnankhan788@gmail.com, [3]kiran@cecos.edu.pk, [*4]ali@cecos.edu.pk, [5]khalid@cecos.edu.pk, [6]zaheer@cecos.edu.pk, [7]imran446bg@gmail.com, [8]abdulamiruet@gmail.com

DOI: https://doi.org/10.5281/zenodo.15573709

Keywords
Transformer Load Forecasting, Machine Learning, Regression Models, Energy Demand Prediction, Weather Variables, Ensemble Learning, Time Series Forecasting, Random Forest, XGBoost

Article History
Received on 22 April 2025
Accepted on 22 May 2025
Published on 31 May 2025

Copyright @Author
Corresponding Author: *
Ali Mujtaba Durrani

Abstract

This paper explores machine learning regression models for predicting maximum transformer load using historical and weather data. The growing energy demand and stress on infrastructure during peak periods motivate the need for accurate forecasting to enhance reliability and planning. Six models were evaluated: linear regression, Decision Tree, Random Forest, Support Vector Regression (SVR), K-nearest neighbors (KNN), and XGBoost. Three scenarios were tested. One year of historical data, one year of data plus weather variables, and ten years of synthetic data with weather fluctuations. Key features included connected load, date-based elements (day, month, year), and weather metrics like temperature, humidity, wind speed, and global horizontal irradiance (GHI). Data preprocessing involved merging transformers and weather datasets, feature engineering, and using Grid Search CV with Time Series Split for time-aware model tuning. Performance was evaluated using the root mean squared error (RMSE), the mean absolute error (MAE), and the R^2 coefficient. Scaled normalization facilitated visual comparison of models by plotting predicted versus actual line plots. In the one-year scenario without weather data, Linear Regression performed best ($R^2 = 0.99$), with Random Forest and KNN also performing well. When weather variables were added, Random Forest ($R^2 = 0.90$) and Linear Regression ($R^2 = 0.99$) remained strong, but SVR and KNN underperformed. With ten-year synthetic data, Random Forest (RMSE = 0.01, $R^2 = 0.97$) and XGBoost (RMSE = 0.02, $R^2 = 0.98$) outperformed others, capturing long-term and seasonal trends. Linear Regression and SVR struggled with extended forecasts. Correlation analysis revealed that transformer load had a moderate correlation with temperature ($r = 0.34$) and wind speed ($r = 0.55$), and a strong correlation with global horizontal irradiance (GHI) ($r = 0.74$). These findings validate the value of ensemble models and environmental variables in enhancing load forecasting accuracy. The study supports the integration of weather-aware machine learning for more intelligent energy grid management.

INTRODUCTION

In recent years, the increasing global demand for electricity and the rapid growth of urbanization have placed unprecedented stress on power systems, particularly on distribution transformers. Forecasting demand for transformers is a crucial aspect of both planning and operating power systems today. If the loads on a transformer exceed its designed values, the chance of overheating or insulation breakdown also rises. As a result of these risks, both service dependability and profits fall for utility companies. For this reason, creating accurate and usable methods to forecast transformer loads is now at the heart of energy research today [1], [2]. ARIMA, MLR, and seasonal trend decomposition approaches were initially employed in transformer load forecasting to statistically and linearly handle the data. Although these methods are effective under limited circumstances, they often fail to work well when input and output in real-world loads are highly complex [3], [4]. Additionally, many of these systems do not readily adapt to changes in climate, increased electricity consumption, and the integration of distributed energy resources (DERs). With the introduction of machine learning (ML), energy analytics have improved, enabling the creation of precise models that utilize data to understand how loads are used. Load data includes challenges that ML models can solve by focusing on their nonlinear aspects. For example, Random Forest, Gradient Boosting, and LSTM models are very successful at finding patterns in load [5]–[7]. They work well because they recognize different patterns and seasons within larger datasets and continually learn from this information, without relying on rigorous rule structures. This thesis examines the predictive performance of six machine learning regression models: Linear Regression, Decision Tree, Random Forest, SVR, KNN, and XGBoost. They are tested against various settings consisting of: (i) just one year of operational transformer data, (ii) the same year with connected weather data, and (iii) a ten-year dataset with both natural seasonality and climate elements. No bright grid plan is complete if transformer load forecasting is inaccurate, as accurately determining future demand is crucial for proper resource management [8]. Thanks to AMI, smart sensors, and real-time monitoring technologies, it is now possible to collect detailed information about operations, which enables the deployment of real-time forecasting models.

It is well known in energy forecasting literature that weather impacts the electrical load. The amount of load a region uses depends partly on temperature, relative humidity, wind speed, and solar irradiance (GHI), mainly because HVAC systems significantly contribute to high demand [9], [10]. Typically, higher temperatures mean more people run their air conditioning, which can increase the demand on local power transformers. Similarly, high humidity can impair the performance of cooling systems, thereby increasing the need for energy. Using these variables in modeling enhances the accuracy and adaptability of these models in various situations. The growing importance placed on renewable energy and decentralized energy systems motivates this research. The development of rooftop solar panels, wind turbines, and electric vehicles has added uncertainties to load forecasting. Transformers in the distribution sector must now handle bidirectional electricity flow and rapid changes in load [11]. Because they can learn and estimate data safely in real-time, machine learning models are well-suited for these challenges. To examine the hypothesis, this thesis focuses on cleaning the data, creating useful features, selecting and training models, optimizing hyperparameters, and evaluating performance using regression-based metrics. Using TimeSeriesSplit, the thesis ensures that the training and test parts of the data are consistently ordered. The performance of a model is estimated quantitatively using RMSE, MAE, and R^2 score. It is revealed through correlation analysis with transformer load that environmental conditions have an impact on the input features. We found that transformer load is moderately to highly correlated with temperature ($r = 0.34$), wind speed ($r = 0.55$), and global horizontal irradiance (GHI) ($r = 0.74$). These findings demonstrate the importance of incorporating weather data into operational forecasting models. Explaining the results of a model is given high priority in this thesis, as it is required for applying these results in real-world utility practice. While obtaining accurate results is important, understanding how the model generates its predictions is essential for everyone to make informed decisions. To enhance transparency, upcoming work should incorporate methods such as SHAP and LIME [12]. In short, this research provides a comparison of several machine learning techniques for predicting transformer loads. It shares ways to choose a suitable model, checks how it works in changing situations, and utilizes weather information to enhance the accuracy of the forecast. This work provides a foundation for building flexible and responsive load forecasting systems in modern power grids.

1. Literature review

To predict future energy needs in the electrical sector, many experts rely on linear regression. One analyzes the connection between an energy use variable and one or more controlling variables to foresee energy use. By applying linear regression to TANGEDCO-CBE's real-time data, researchers can predict energy consumption with good levels of precision. On July 3, 2024, researchers employed linear regression to analyze and forecast energy use based on data from PALECO for the period from 2014 to 2018. The study demonstrated that using linear regression is beneficial for predicting energy consumption and supports energy planning. Additionally, models have been developed to predict electricity usage based on GDP and population. The total power load in Hebei from 2000 to 2014 was successfully predicted using a multiple linear regression model. Linear regression was used to study electric springs (ES) under various load impedance conditions. It was found that linear regression handled cases more efficiently and provided more accurate results than simulation methods when modeling ES under-voltage and over-voltage conditions. They demonstrate that linear regression is practical for various applications in power sector energy demand prediction and management. Many experts use linear regression to predict the power demand, a process necessary for managing and planning power systems. Using this approach, a linear relationship is created between the independent variables and the dependent variable. Generally, electricity consumption or load is considered the dependent variable in electrical demand prediction, while GDP, population, temperature, and past consumption data are considered the independent variables.

If several factors determine electrical demand, this method is most appropriate. It is a method for incorporating multiple independent variables into the simple linear regression approach. In Hebei, China, Cui and Wu predicted total electricity consumption figures from 2000 to 2014 using multiple linear regression (MLR), with GDP and population chosen as the independent variables. It was discovered that the model enabled the accurate prediction of power load, which forms the basis for controlling and forecasting power demand [7]. Time series data can be analyzed through linear regression to make projections about future demand using its historical data. Peña-Guzmán and Rey built multiple linear regression models in their research to project household electricity use in Bogotá, Colombia. Three types of models were examined: a simple multiple linear regression, an econometric model, and a double-log economic regression. When compared to short-run regression, the econometric model had better precision based on its higher R^2. Utilizing climate and weather information can significantly enhance the accuracy of demand forecasts. Tassou et al. conducted a study using multiple regression analysis to determine the amount of energy used by a supermarket in the UK. The model used readings of temperature, relative humidity, and measured actual temperature to figure out the humidity ratio. Temperatures are estimated to cause a 1.7% increase in electricity demand, which is expected to lead to a 13% decrease in gas consumption in the central scenario. Demand prediction now uses linear regression more effectively, thanks to the power of real-time data and big data analytics. The scientists predicted the amount of energy that would be used using linear regression and data from TANGEDCO-CBE. Daily data on energy demands and use were used to build the model, allowing it to forecast energy use within tolerable margins of error. The study employed multiple regression analysis to investigate how energy is utilized in a supermarket in northern England. Included in the model were data for the humidity ratio, which came from temperature and relative humidity, as well as the actual temperature. According to the results, electricity use increased slightly, while gas use decreased sharply, with an expected 13% drop in gas and a rise in electricity of only 2.1%. The Palawan Electric Cooperative (PALECO) conducted a study using multiple linear regression to predict electricity consumption across the Puerto Princesa grid. The model included details on the number of consumers, peak demand levels, and energy use from 2014 to 2018. The results suggest that linear regression can provide valuable insights for supporting more effective energy planning and management. A study by Kouassi et al. examined the use of linear regression and ARIMA models to forecast electricity demand in West Africa. Most countries had their electricity demand better predicted using an ARIMA model, except for the Gambia, Ghana, Guinea, Liberia, and Nigeria, for which the multivariate linear regression model performed better. Because linear regression models are easy to grasp and interpret, they can be utilized by a wide range of users. Linear regression is easy to use and suitable for dashboards, making it an option for large and real-time projects. Many people incorporate this method into their statistical and machine learning methods to solve regression problems. Suppose the link between the independent and dependent variables isn't linear. Outliers may create a distorted slope and intercept. It is not always effective in tracking relationships that follow complex patterns.

Table 1. Applications of Decision Tree Methods in Energy Systems Analysis and Management

Study	Method	Independent Variables	Dependent Variable	Main Findings
Alizamir et al. [1]	Decision Tree	Temperature, time of day, day of the week, and occupancy	Electricity Consumption	Improved forecasting accuracy and captured non-linear relationships
Namazkhan et al. [2]	Decision Tree	Building-related factors, socio-demographic factors, psychological parameters	Gas Consumption	Identified key factors affecting gas consumption, useful for customer categorization and fraud detection
Bugaje et al. [3]	Decision Tree	Grid parameters, load conditions, and weather data	Security Status	Predicted potential security issues, aiding proactive grid management
Zhang et al. [4]	Decision Tree	Key drivers, motivators, and barriers	Policy Decisions	Facilitated analytical decision-making in energy policy, helping to design effective demand response programs
Zhang et al. [5]	Decision Tree	Building characteristics, usage patterns	Building Energy Use Intensity (EUI)	Accurately classified and predicted building energy use intensity, improving building energy management

A Random Forest combines the decisions of several decision trees and chooses either the most common prediction for classification problems or averages the outcomes for regression models. It gathers information from various decision trees to enhance and improve the model's reliability, ensuring it does not overfit to its training set. Multiple copies of the original dataset, with some records repeated, are generated through bootstrapping using Random Forests. A separate decision tree is generated using every bootstrap sample. At every step, a random selection of features is used to see if the data can be separated, which adds different paths to the tree. Finally, the prediction is determined by averaging all the regression trees or by counting the tree that received the most votes in the classification. Random

Forests have a lower risk of overfitting than a single decision tree. Features in high numbers do not cause difficulties for them, and they still work well even when most features are unrelated. While every decision tree can be understood on its own, Random Forests make it easier to see what drives the prediction process. Some models can be used for either classification or regression of data. Training multiple trees can require a significant amount of computing power. Although feature importance is included, the collections of trees involved make it more challenging to study the detailed decisions made during inference. Forecasting short-term electricity use with Random Forests supports effective operations and response programs within the grid. For example, Liu and his colleagues used Random Forests to estimate future electricity demand in the residential area [1]. Using all models together yielded better and more accurate predictions than relying solely on a single decision tree model.

Table 2. Electricity Demand Forecasting with Random Forest Methods

Study	Method	Independent Variables	Dependent Variable	Main Findings
Liu et al. [1]	Random Forest	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Outperformed single decision tree models, providing more accurate and reliable predictions
Al-Naji et al. [2]	Random Forest	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Improved short-term load forecasting accuracy
Al-Musaylh et al. [3]	Random Forest	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Effective in smart grids, improved forecasting accuracy
Hung et al. [4]	Hybrid (Random Forest + LSTM)	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Achieved high accuracy and robustness
Nguyen et al. [5]	Random Forest	Temperature, time of day, day of the week, historical consumption, and renewable energy data	Electricity Demand	Effective in handling high renewable penetration, improved forecasting accuracy

Handling the changing and uncertain demand and supply in high-renewable energy grids is achieved using Random Forests. Nguyen et al. employed a Random Forest model to forecast short-term load in smart grids, demonstrating its effectiveness for high levels of renewable power. Often, forecasting accuracy is increased by merging Random Forests with other machine learning methods. Hung and his colleagues [4] developed a model combining Random Forests and Long Short-Term Memory networks to achieve reliable and effective short-term load forecasting. Distribution companies apply Random Forests to estimate demand and optimize resource allocation. Souza et al. demonstrated that Random Forests are more suitable for short-term forecasting than other statistical methods in a case study conducted at a Brazilian distribution utility [6].

Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) used for regression tasks. SVR works by finding a hyperplane that best fits the data points while minimizing the error within a specified margin. Unlike traditional regression methods, SVR focuses on the points that are closest to the hyperplane (support vectors) and tries to maximize the margin around this hyperplane. SVR uses kernel functions to transform the input data into a higher-dimensional space where it is easier to find a hyperplane. Standard kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid. SVR uses an epsilon-insensitive loss function, which ignores errors that are within a certain threshold (epsilon) and only penalizes errors that exceed this threshold. It balances how level the hyperplane stays against how wide a tolerance is given to points farther than epsilon from it. This value indicates the size of the epsilon-insensitive zone allowed. Robustness to Outliers: The main reason SVR is resistant to outliers is that it stresses the points that define the margin of error. Making sense of complex and massive datasets is easy for SVR.

SVR is able to model complex patterns in the data thanks to the use of different kernel functions. SVR is computationally complex, as it becomes prolonged for large datasets. The performance of SVR is strongly affected by the chosen kernel function, C, and ε. Because it's excellent for short-term predictions, SVR is essential for helping manage grids and demand response.

Table 3. SVR-Based Models for Electricity Demand Forecasting

Study	Method	Independent Variables	Dependent Variable	Main Findings
Hong et al. [1]	SVR (RBF Kernel)	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Accurate predictions, effective in handling non-linear relationships
Zhang et al. [2]	Hybrid (SVR + ANN)	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Outperformed single SVR and ANN models, achieving higher accuracy
Wang et al. [3]	SVR (RBF Kernel)	Temperature, humidity, historical consumption, seasonal data	Electricity Demand	Captured seasonal patterns and weather effects, improving overall accuracy
Li et al. [4]	SVR (RBF Kernel)	Real-time data (temperature, time of day, day of the week)	Electricity Demand	Provided accurate and timely predictions, supporting dynamic

A paper by Hong et al. [1] employed a radial basis function (RBF) kernel in support vector regression (SVR) to predict electricity demand over a short period. Thanks to the model's capabilities, it was able to anticipate trends from unpredictable types of data. Many times, SVR is used in conjunction with other methods to enhance the accuracy of the forecasts. For example, Zhang et al. [2] introduced a new model that combines support vector regression (SVR) and artificial neural networks (ANN) for short-term forecasting of energy loads. The results show that the hybrid model outperformed both the single SVR and ANN models. Better predictions of electrical demand can be made using SVR because it is good at working with seasonal and weather data. The authors of [3] attempted to forecast electricity demand using SVR, taking into account temperature, humidity, and historical data. The model was able to capture variations in weather and seasons, which enhanced its prediction accuracy. SVR is used to estimate present and future electricity needs. Real-time smart grid load forecasting was addressed in a study by Li et al. [4] using support vector regression (SVR). The model enabled grid operators to predict loads, making load management significantly more straightforward and accurate.

K-Nearest Neighbors (k-NN) can easily handle both classification and regression processes. With k-NN in regression, the model identifies the k nearest data points (neighbors) to the input and predicts the outcome by averaging their targets. It is especially critical for forecasting load over a period. For instance, Singh et al. [1] demonstrated the application of k-NN in predicting hourly electricity demand. Because the model handles non-linear relationships well, the predictions were accurate for electrical demand. A research group led by Li et al. [2] applied the k-NN algorithm to forecast electricity demand, based on measurements of temperature, humidity, and previous usage patterns. The model incorporated both seasonal trends and weather effects, thereby enhancing the accuracy of its predictions. As a result, k-NN is now used for immediate demand forecasting to improve grid operation. Wang et al. conducted a study to apply k-NN for immediate forecasting of smart grid energy levels. The predictions from the k-NN model were quick and accurate, making it easier for grid operators to monitor the system's load. For instance, Zhang et al. [4] proposed a model that combines k-NN and ANN to forecast short-term energy demand. Assembling k-NN and ANN led to more accurate results than when the models were used separately.

Table 4. KNN-Based Forecasting of Electricity Demand

Study	Method	Independent Variables	Dependent Variable	Main Findings
Singh et al. [1]	KNN	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Accurate predictions, effective in handling non-linear relationships
Li et al. [2]	KNN	Temperature, humidity, historical consumption, seasonal data	Electricity Demand	Captured seasonal patterns and weather effects, improving overall accuracy
Wang et al. [3]	KNN	Real-time data (temperature, time of day, day of the week)	Electricity Demand	Provided accurate and timely predictions, supporting dynamic grid management
Zhang et al. [4]	Hybrid (KNN + ANN)	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Outperformed single k-NN and ANN models, achieving higher accuracy

XGBoost (EXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms. The system is designed to be highly efficient, flexible, and easily relocatable. XGBoost builds the model by repeatedly adding small predictors (usually decision trees) to the overall model, allowing each one to address the errors identified by earlier learners. It is called gradient boosting within machine learning. Each time, XGBoost adds a new learner, stresses on the mistakes left by previous learners, and updates the model. XGBoost relies on L1 and L2 regularization to prevent overfitting. The process employs a structured model approach to avoid overfitting, which contrasts with the greedy algorithm used in standard decision trees. Because XGBoost can handle parallelization, it is highly effective for working with large datasets. Short-term prediction is the primary use of XGBoost, ensuring better Grid management and demand response programs. For instance, Zhang et al. [1] investigated the use of XGBoost to forecast short-term electricity demand. Because it could handle non-linear relationships, the model made accurate predictions. Weather and seasonal data handling is one of the strengths of XGBoost, which is essential for predicting electrical demand. A team of researchers [2] utilized XGBoost to make future predictions regarding electricity demand by incorporating temperature, humidity, and previous statistics. The inclusion of seasonal trends and weather effects in the model raised the accuracy level. XGBoost enables real-time demand forecasting to support grid management and operations. In 2023, Wang et al. [3] investigated XGBoost as a tool for real-time power usage forecasting in innovative grid systems. The way the model worked provided grid operators with accurate knowledge of the load, enabling them to manage it effectively. A combination of XGBoost and additional techniques consistently yields more precise forecasting. For instance, according to research by Liu et al. [4], a model based on XGBoost and LSTM networks was developed to predict short-term future power loads. A hybrid model produced better results and more accurate outcomes than the other two models.

Table 5. XGBoost-Based Electricity Demand Forecasting

Study	Method	Independent Variables	Dependent Variable	Main Findings
Zhang et al. [1]	XGBoost	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Accurate predictions, effective in handling non-linear relationships
Li et al.	XGBoost	Temperature, humidity,	Electricity	Captured seasonal patterns and

[2]		historical consumption, seasonal data	Demand	weather effects, improving overall accuracy
Wang et al. [3]	XGBoost	Real-time data (temperature, time of day, day of the week)	Electricity Demand	Provided accurate and timely predictions, supporting dynamic grid management
Liu et al. [4]	Hybrid (XGBoost + LSTM)	Temperature, time of day, day of the week, and historical consumption	Electricity Demand	Outperformed single XGBoost and LSTM models, achieving higher accuracy

2. Data Collection and Exploration

Exploratory data analysis (EDA) is used to take a multistage approach to predicting flooding. First, single-source transformer data is modeled; then, multivariate forecasting with weather is performed; and finally, a 10-year forecast is made. This setup aligns with the wide range of information required and the varying time spans that forecasting and optimizing transformer load in the grid necessitate. The historical transformer information includes the highest load, the lowest load, the total load, and the times at which these values were recorded. Using additional information, such as temperature, humidity, wind speed, and global horizontal irradiance (GHI), made our weather data more useful for forecasting. A multivariate dataset was created by combining the two datasets, using the same time and date fields. The data for transformer loads focuses on key aspects, including the maximum load, minimum load, connected load, and associated data and time. Weather data was linked to measurements of transformers to show how real-time conditions impact the transformer. The connections between weather changes and variations in electrical load were identified using the merged data. The data was prepared to guarantee both its consistency and accuracy. All date fields were formatted consistently using standard datetime formats, while numerical fields such as load values were converted into numeric data types. Month names were also mapped to numeric values to support time-series modelling. The datasets were cleaned of any missing or invalid entries in key columns to ensure high-quality input for modelling. Time-series line plots display the changes in connected load over time, revealing potential seasonal trends and peak usage periods. Histograms of load distributions provide insights into the standard operational ranges of transformer loads. Annual average connected load plots were generated to analyze long-term consumption patterns. Distributions of transformer loads were illustrated using histograms.

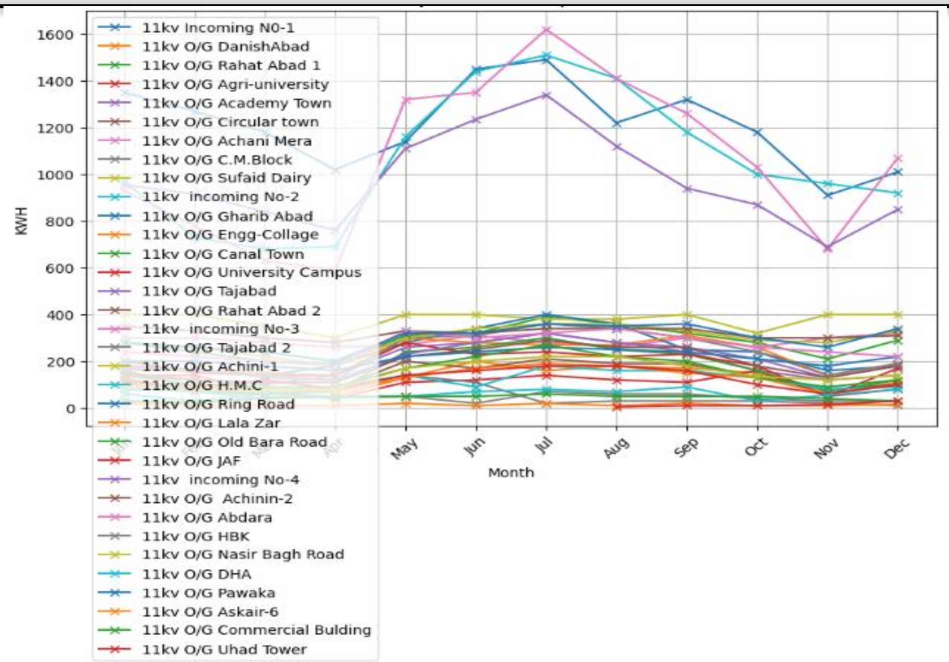


Figure 1: Monthly Load Trends for Various Feeders

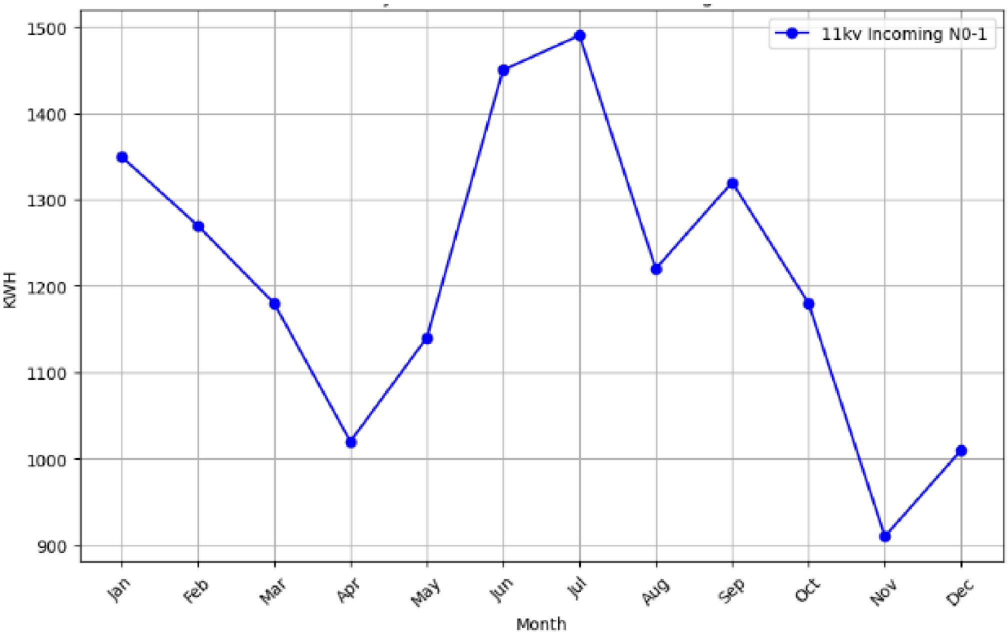


Figure 2: Monthly Load Trends for Single Feeder Incoming 1

Figure 1 illustrates the load trends of all feeders over one year, while Figure 2 focuses on a single feeder, specifically incoming 1. The graph illustrates the variations in load demand throughout the year.

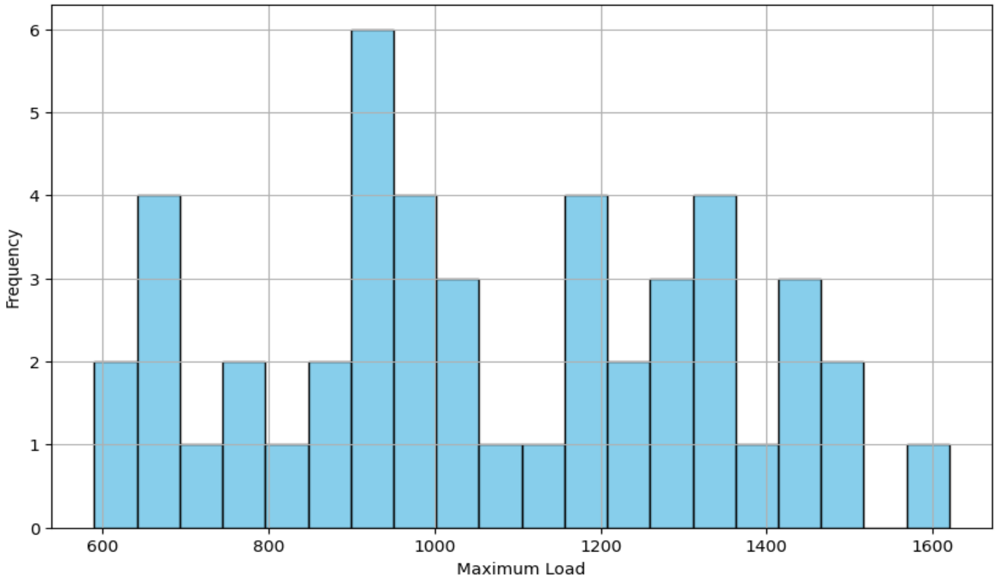


Figure 3. Distribution of Maximum Load

Figure 3 illustrates the distribution of maximum loads, which peaks around 1000–1200 units. Figure 4 highlights the minimum load distribution, which is right-skewed, indicating that many transformers operate under low-load conditions during off-peak hours.

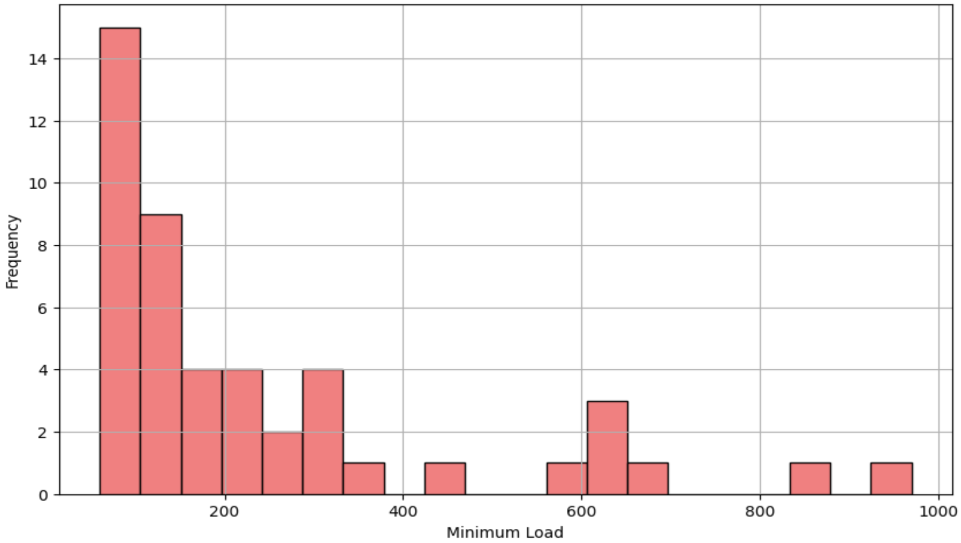


Figure 4. Distribution of the Minimum Load

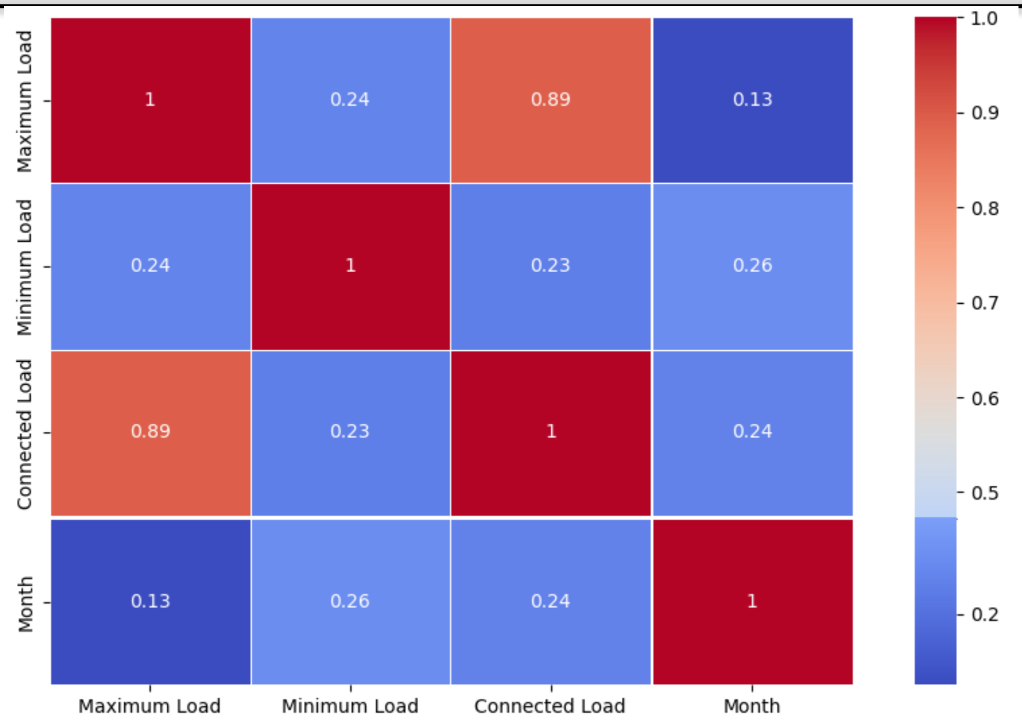


Figure 5. Correlation Matrix of the Transformer Data

Figure 5 summarizes the correlation matrix, which shows the relationships among the load variables. Maximum and connected loads exhibit a strong correlation (r = 0.89), while the relationships involving minimum load and time-related features remain weak. The correlation matrix supports this finding, while also indicating weak relationships between the minimum load and other variables.

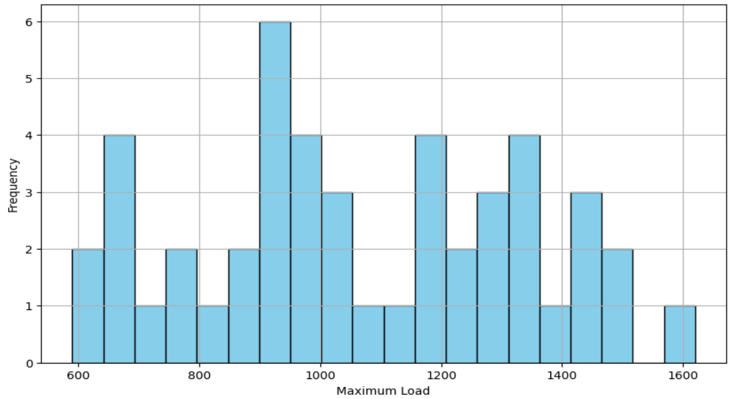


Figure 6. Distribution of Transformer Maximum Load

Distribution plots further clarify that most transformers operate around 1000–1200 units for maximum load and below 200 units for minimum load, reflecting a broad operational range and underutilization during off-peak times.

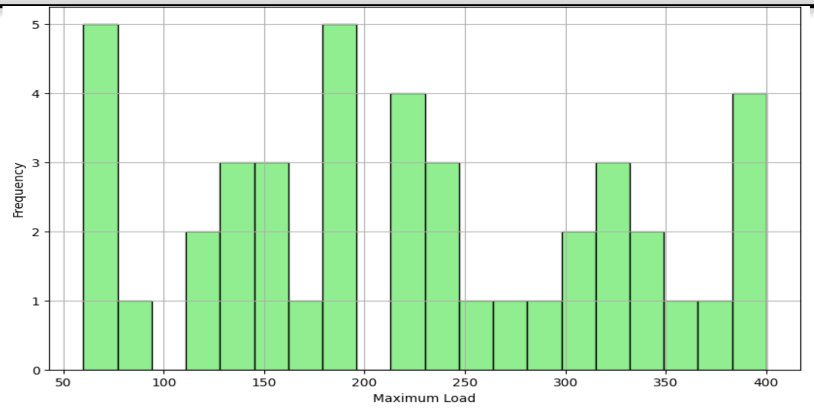


Figure 7. Distribution of Transformer Minimum Load

The histogram of transformer maximum load distribution in Figure 4 reveals a bell-shaped curve, while transmission lines in Figure 6 show an even distribution between 150–250 kW. The histogram of transformer maximum load is bell-shaped, with most values ranging from 900 to 1300 kW. Figure 8 shows the Minimum Transformer Load, which has a narrower, more uniform distribution, typically ranging from 150 to 250 kW.

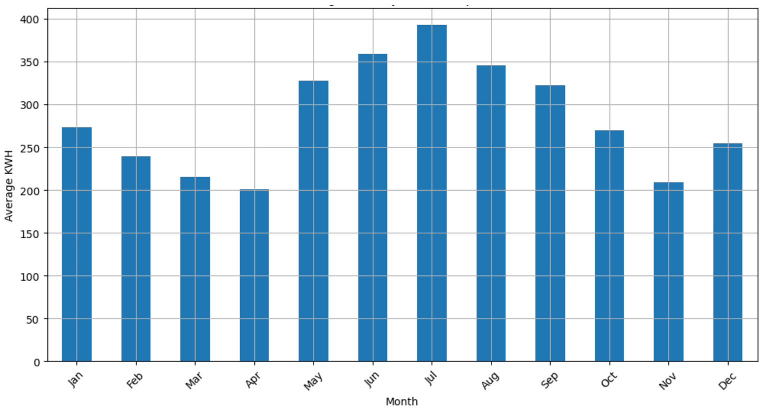


Figure 8. Average Monthly Data Comparison

Figure 8 shows the average monthly data comparison in kWh over the 1 year. The overall analysis of the data revealed the maximum and minimum loads at various times of the year, along with the dependencies of different load variables on each other under different load conditions, as well as anomalies within the data.

3. Data Training and Load Prediction

The data was split into training and testing sets to allow unbiased model evaluation. Linear regression and similar basic models were applied to estimate the load values. Prediction accuracy and strength were enhanced by combining transformer data with weather data to create a multivariate model. Optimizing the hyperparameters of algorithms was done with GridSearchCV and other techniques. This enabled the prediction models to take account of both technical and environmental factors when estimating transformer performance. During the final stage, the model examined predictions that extended ten years into the future. For this task, both advanced time-series models, including Random Forest and Gradient Boosting, were analyzed. They were designed to identify and utilize repeated load patterns, trends, and seasonal impacts over extended periods of time. Model results were measured using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 as the primary metrics.

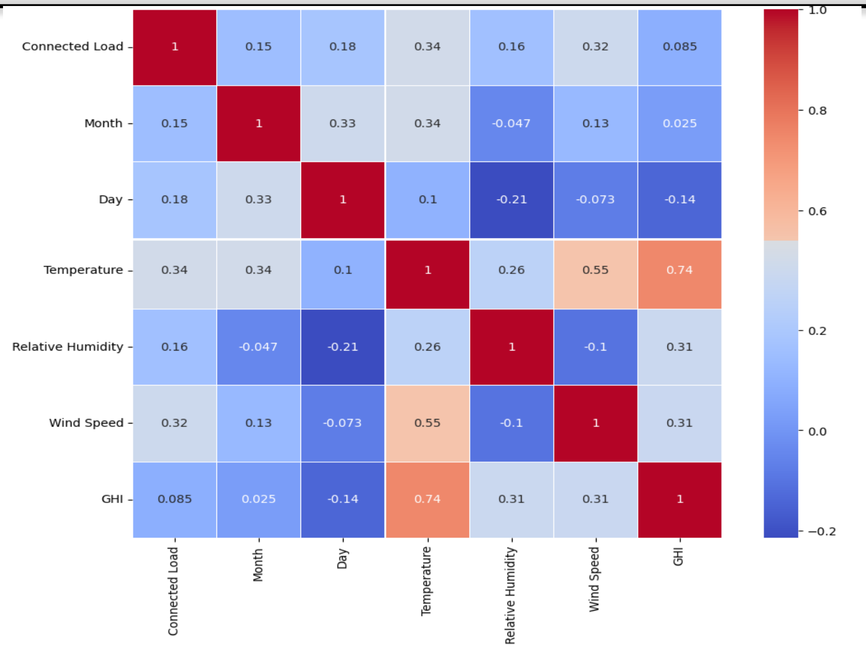
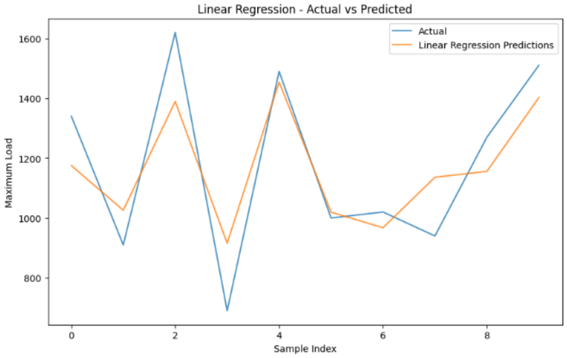
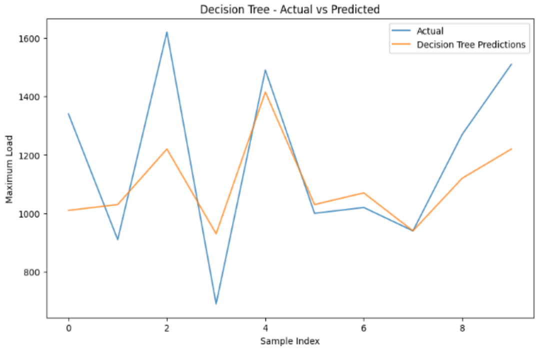


Figure 9. Correlation of the Matrix of Load Data and Weather Data

These metrics were used to determine which approach was most accurate and could be generalized. The process begins by collecting data from various sources and combining them. The transformer data contains the main information, including the highest and lowest loads each transformer handled, along with the time of each event. To analyze how environmental conditions affect transformer load, the dataset is linked to temperature, humidity, wind speed, and horizontal irradiance (GHI). The original datasets are connected using shared date fields to create a single, improved dataset. Various AI/ML models are applied to predict transformer load based solely on the connected load. With weather features included, models are retrained. The correlation heat maps in Figure 9 show that temperature and GHI have strong correlations with load.





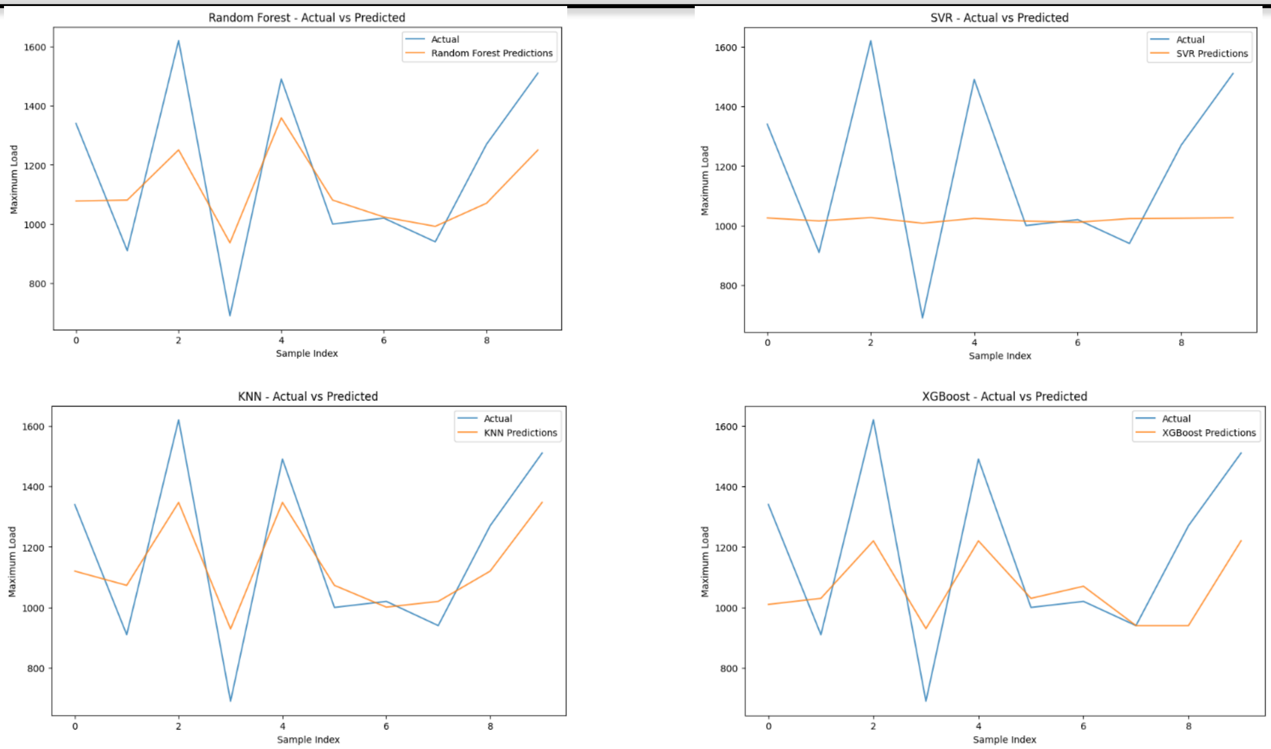


Figure 10. Actual vs Predicted Values for 1 year

This modeling process utilizes linear regression, Decision Trees, Random Forests, SVR, KNN, and XGBoost. Linear Regression predictions align closely with actual values. Figure 10 shows the actual vs. predicted values graph for 1 year. Here, the linear regression model performs well and accurately predicts the load, closely matching the actual values over a one-year period. However, the model does not accurately predict the load spike. The Decision Tree slightly over fits to the actual values the Random Forest offers the most balanced performance with the prediction and it can be seen that the SVR and KNN perform poorly under high variability with XGBoost showing moderate success.

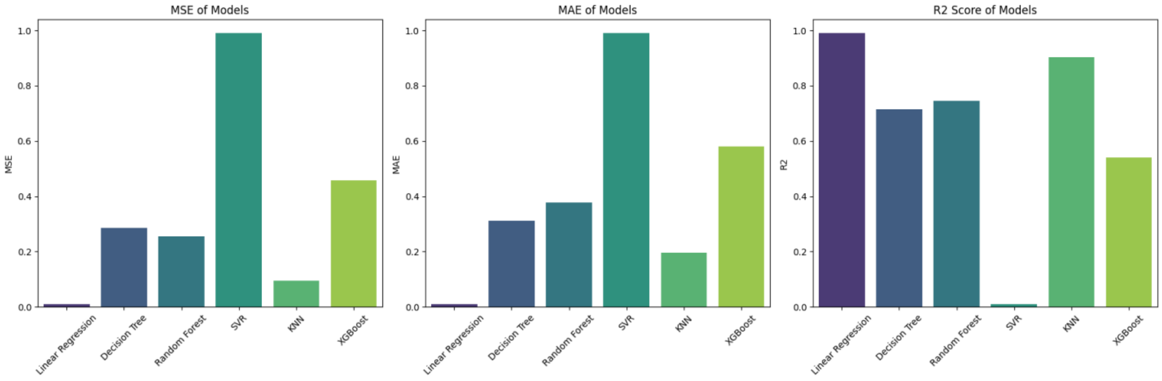


Figure 11. Performance Comparison of all the models for 1 year

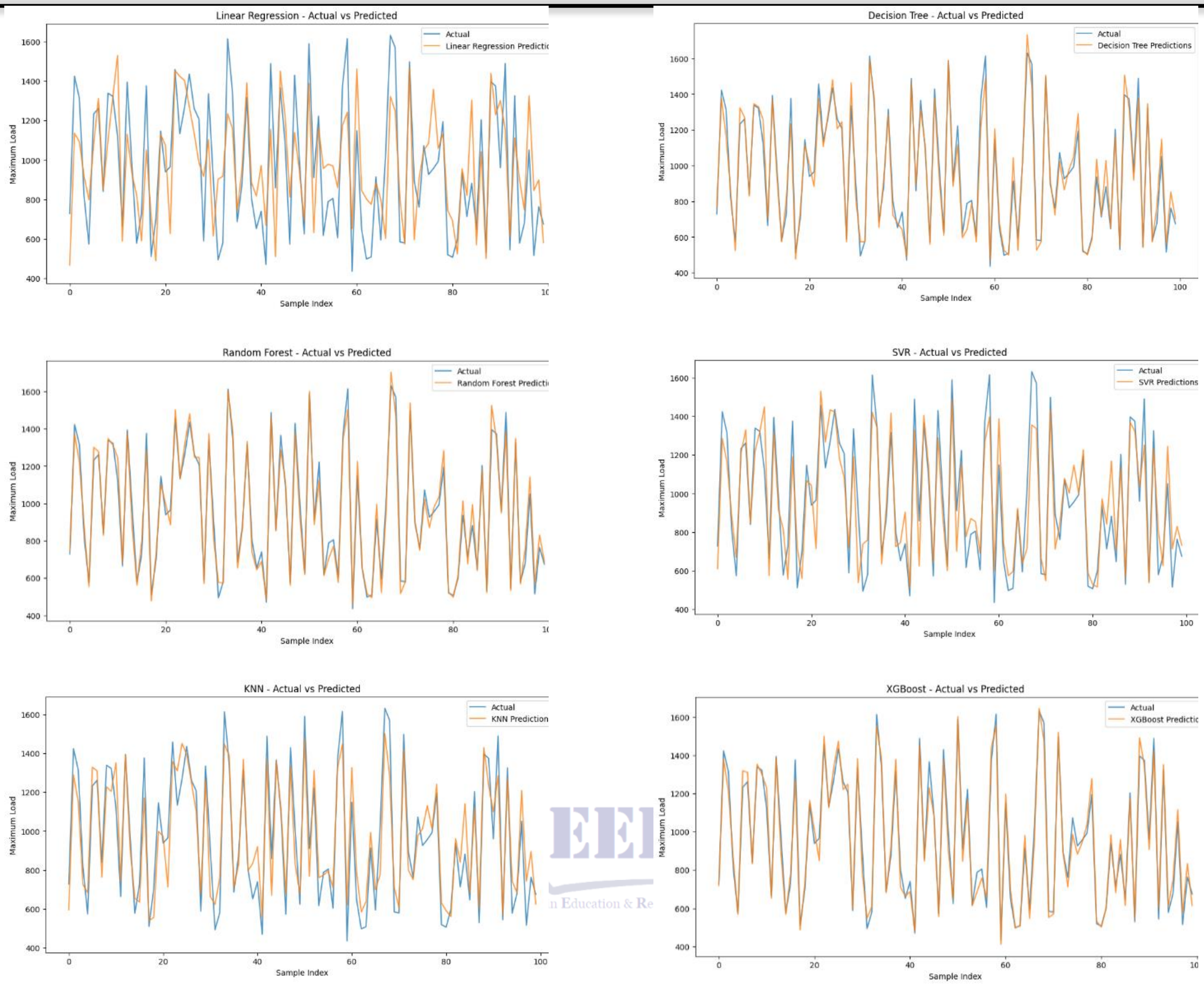


Figure 12. Actual vs Predicted Values for 10 years

Performance metrics such as MSE, MAE, and R² are summarized in figure 11 here the bar plots for the MSE show that the Linear Regression has the least error the Decision Tree, Random Forest and KNN also performs well but SVT and XGBoost show very high values. Similar is the case in the MAE bar plots with Linear Regression with the minimum value and SVR and XG Boost with high values while the other models performs reasonably well than these two but not better that the Linear Regression and Random Forest. Next up we predict the data for 10 years here the actual and predicted values of all the models are seen in figure 12. The graphs show Linear Regression with significant errors in prediction and high error with the Decision Tree and Random Forest along with XGBoost.

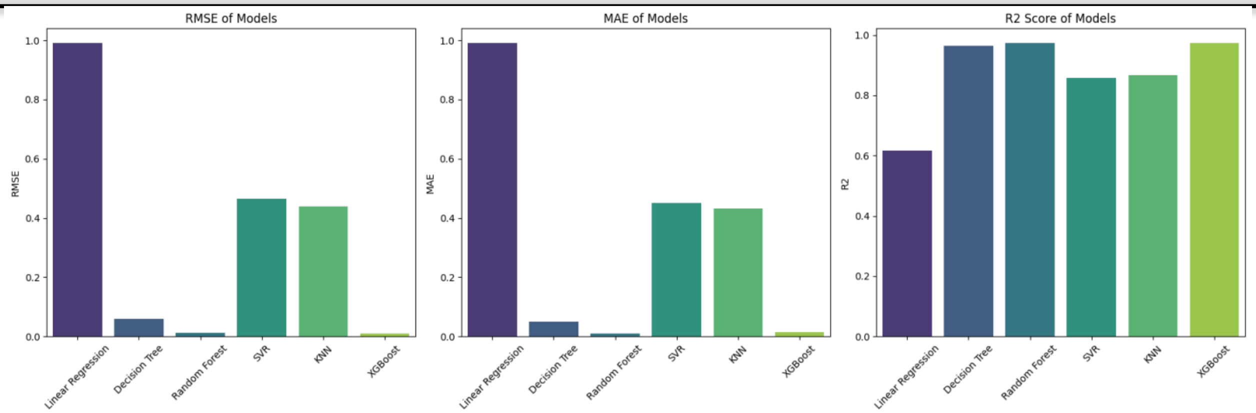


Figure 13. Performance Comparison of all the models for 10 year

Figure 13 shows the performance comparison of all the models for the span of 10 years here the linear regression has highest RMSE and MAE values showing a high error ratio in the predictions the SVR and KNN also perform poorly for the 10 year span of the data. The other models like Decision Tree, Random Forest and XG Boost show very low error in RMSE and MAE plots showing a significant performance in the overall accuracy. These values are also depicted in the table 6 with linear regression showing the highest percentage of errors and lowest accuracy of 62% the SVR and KNN perform considerably well with accuracy of 86% and 87% also the Decision Tree and Random Forest have efficiencies of 96% and 97% respectively with XGboost outshining all the models with an accuracy of 98% overall.

Table 6 Model Evaluation: RMSE, MAE, and R² Scores

Linear Regression	**0.99**	**0.99**	**0.62**	**62%**	**Performs poorly due to a lack of flexibility in modeling non-linearity**
Decision Tree	0.06	0.06	0.96	96%	Interpretable and straightforward, but slightly less effective than ensemble methods
Random Forest	0.01	0.01	0.97	97%	Best performer overall, with the lowest error and a high R²
SVR	0.46	0.45	0.86	86%	Performs reasonably well, but is computationally expensive
KNN	0.44	0.43	0.87	87%	Moderate performance and sensitive to local patterns
XGBoost	0.02	0.02	0.98	98%	Competitive with Random Forest with slightly higher error

The results highlight the superior generalization ability of ensemble models, such as Random Forest and XGBoost, particularly in handling long-term, multivariate, and weather-influenced transformer load forecasting scenarios. While Linear Regression is fast and interpretable, its simplicity makes it unsuitable for capturing complex relationships. SVR and KNN, though theoretically capable, fail to generalize well due to sensitivity to parameter tuning and local variations. Consequently, this study recommends ensemble techniques, especially Random Forest and XGBoost, for operational transformer load forecasting.

4. Conclusion

This research highlights the significant potential of machine learning, particularly ensemble-based models, in accurately forecasting transformer load across both short and long-term horizons. Employing a rigorous and phased methodological approach, the study effectively bridged the challenges of classical electrical load forecasting with modern data-driven strategies, resulting in improved predictive accuracy and enhanced system resilience. Initially, one-year historical transformer data were used to benchmark the base model's performance. Linear Regression performed surprisingly well ($R^2 = 0.99$) within this window, given the linear characteristics of the dataset, while Decision Tree and Random Forest models showed competitive results. However, the ensemble advantage was less pronounced due to the simplicity of the data. SVR and KNN demonstrated moderate performance, with some limitations in capturing patterns accurately and avoiding overfitting. The second phase integrated meteorological data temperature, humidity, wind speed, and solar irradiance exposing the limitations of simpler models, such as linear regression, and emphasizing the superiority of ensemble methods. Random Forest and XGBoost exhibited strong performance (RMSE ≈ 0.01, $R^2 \approx 0.90$–0.98), effectively modeling nonlinear and multivariate dependencies. Correlation analysis confirmed that weather variables, such as GHI (0.74), wind speed (0.55), and temperature (0.34), had a significant impact on transformer load, validating the inclusion of environmental factors in the predictive model.

5. Future work

The results from this study lead researchers to consider various advanced machine learning frameworks for enhancing transformer load forecasting. Such prospective paths include new ways to collect data, new research tools, and systems combining them, as well as strategies for applying these in everyday use. By collaborating, their goal is to enhance the way predictive systems address smart grid challenges. It looks promising to add deep learning methods such as LSTM, GRU, or TCN, as they can handle long-term links in data that appears in sequence. They can work together with standard machine learning models to enhance learning from data that follows a time series pattern. At the same time, using fast data from IoT smart meters and sensors allows models to update themselves as more data becomes available. Geospatially, it's clear that the way transformers react depends on city density and the condition of the supporting infrastructure. Using geospatial metadata and satellite images in these models can enhance the accuracy of their predictions. It would also be helpful to include broader environmental factors (such as rainfall and air pressure) and socioeconomic markers (including tariff changes and holidays) in the model to understand each situation better and increase forecasting accuracy. Accordingly, transformer load forecasting will depend on new technologies, practical operations, and preparedness for climate change. Working on this subject provides a secure foundation for developing intelligent and robust forecasting systems. Following these recommendations will encourage the design of flexible and reliable energy solutions for today's power grids.

References

C. Zhichu *Et Al.*, "Offshore Wind Farms Interfacing Using Hvac-Hvdc Schemes: A Review," *Computers Electrical Engineering,* Vol. 120, P. 109797, 2024.

M. Z. Yousaf, S. Khalid, M. F. Tahir, A. Tzes, A. J. I. J. O. E. P. Raza, And E. Systems, "A Novel Dc Fault Protection Scheme Based On Intelligent Network For Meshed Dc Grids," *International Journal Of Electrical Power Energy Systems,* Vol. 154, P. 109423, 2023.

M. Z. Yousaf, H. Liu, A. Raza, A. J. C. J. O. P. Mustafa, And E. Systems, "Deep Learning-Based Robust Dc Fault Protection Scheme For Meshed Hvdc Grids," *Csee Journal Of Power Energy Systems,* Vol. 9, No. 6, Pp. 2423-2434, 2022.

M. Z. Yousaf, M. F. Tahir, A. Raza, M. A. Khan, And F. Badshah, "Intelligent Sensors For Dc Fault Location Scheme Based On Optimized Intelligent Architecture For Hvdc Systems," *Sensors,* Vol. 22, No. 24, P. 9936, 2022.

M. Z. Yousaf *Et Al.*, "Multisegmented Intelligent Solution For Mt-Hvdc Grid Protection," *Electronics,* Vol. 12, No. 8, P. 1766, 2023.

M. F. Tahir, M. Z. Yousaf, A. Tzes, M. S. El Moursi, T. H. J. R. El-Fouly, And S. E. Reviews, "Enhanced Solar Photovoltaic Power Prediction Using Diverse Machine Learning Algorithms With Hyperparameter Optimization," *Renewable Sustainable Energy Reviews,* Vol. 200, P. 114581, 2024.

M. F. Tahir, A. Tzes, And M. Z. Yousaf, "Enhancing Pv Power Forecasting With Deep Learning And Optimizing Solar Pv Project Performance With Economic Viability: A Multi-Case Analysis Of 10 Mw Masdar Project In Uae," *Energy Conversion Management*, Vol. 311, P. 118549, 2024.

Aziz, Abdul, Et Al. "Advanced Ai-Driven Techniques For Fault And Transient Analysis In High-Voltage Power Systems." *Scientific Reports* 15.1 (2025): 5592.

Khan, Romaisa Shamshad, Et Al. "Advanced Battery Management For Electric Vehicles: Charge Monitoring And Fire Security." *Pakistan Journal Of Engineering And Technology* 8.1 (2025): 1-13.

Aziz, Abdul, Et Al. "Simulation-Based Conductor Optimization For Power Distribution Feeders, A Comparative Study Using Etap." *Spectrum Of Engineering Sciences* 3.4 (2025): 430-444.

Renhai, Feng, Et Al. "Adaptive Non-Parametric Kernel Density Estimation For Under-Frequency Load Shedding With Electric Vehicles And Renewable Power Uncertainty." *Scientific Reports* 15.1 (2025): 11499.

Aziz, Abdul, Umar Siddiqueb, And Mehran Ahmad. "Comprehensive Analysis And Optimization Of Radial Distribution Feeder Using Etap."

Sulieman, Muhammad, Et Al. "Revolutionizing Fault Detection In High-Voltage Transmission Lines Through Ann Models." *Spectrum Of Engineering Sciences* 3.4 (2025): 445-468.

X. Liang And M. Abbasipour, "Hvdc Transmission And Its Potential Application In Remote Communities: A Review," 2021 Ieee Industry Applications Society Annual Meeting (Ias), Vancouver, Bc, Canada, 2021, Pp. 1-9, Doi: 10.1109/Ias48185.2021.9677171.

E. Tsotsopoulou, X. Karagiannis, T. Papadopoulos, A. Chrysochos, A. Dyśko, And D. Tzelepis, "Protection Scheme For Multi-Terminal Hvdc System With Superconducting Cables Based On Artificial Intelligence Algorithms," International Journal Of Electrical Power & Energy Systems, Vol. 149, P. 109037, Jul. 2023, Doi: 10.1016/J.Ijepes.2023.109037.

M. Callavik, A. Blomberg, J. Hafner, And B. Jacobson, "The Hybrid Hvdc Breaker – An Innovation Breakthrough Enabling Reliable Hvdc Grids," Technical Paper, Abb Grid Systems, Nov 2012.

X. Zhao, J. Xu, G. Li, J. Yuan And J. Liang, "Coordinated Control Of Dc Circuit Breakers In Multilink Hvdc Grid," In Csee Journal Of Power And Energy Systems, Doi: 10.17775/Cseejpes.2020.05170.

J. Xu, X. Zhao, N. Han, J. Liang And C. Zhao, "A Thyristor-Based Dc Fault Current Limiter With Inductor Inserting–Bypassing Capability," In Ieee Journal Of Emerging And Selected Topics In Power Electronics, Vol. 7, No. 3, Pp. 1748-1757, Sept. 2019, Doi: 10.1109/Jestpe.2019.2914404.

A. Safaei, M. Zolfaghari, M. Gilvanejad, And G. B. Gharehpetian, "A Survey On Fault Current Limiters: Development And Technical Aspects," International Journal Of Electrical Power & Energy Systems, Vol. 118, P. 105729, Jun. 2020, Doi: 10.1016/J.Ijepes.2019.105729.

A. Alassi, S. Bañales, O. Ellabban, G. Adam, And C. Maciver, "Hvdc Transmission: Technology Review, Market Trends And Future Outlook," Renewable And Sustainable Energy Reviews, Vol. 112, Pp. 530–554, Sep. 2019, Doi: 10.1016/J.Rser.2019.04.062.

S. Khalid Et Al., "Technical Assessment Of Hybrid Hvdc Circuit Breaker Components Under M-Hvdc Faults," Energies, Vol. 14, No. 23, P. 8148, Dec. 2021, Doi: 10.3390/En14238148.

C. M. Franck, "Hvdc Circuit Breakers: A Review Identifying Future Research Needs," In Ieee Transactions On Power Delivery, Vol. 26, No. 2, Pp. 998-1007, April 2011, Doi: 10.1109/Tpwrd.2010.2095889.

.