International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

Enhanced Cellular Traffic Prediction Using Tuned XGBoost and Advanced Ensemble Regression Models

P.Ramulu[1], Department of AIML, MJR College of Engineering and Technology, Piler, India

Mrs.V.Subhasini[2], Associate Professor, Department of CSE, MJR College of Engineering and Technology, Piler,India

Abstract: *Accurate cellular traffic prediction is vital for optimizing Quality of Service (QoS) in modern networks, especially with the rising requirement for real-time applications. This paper introduces an advanced Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) architecture that incorporates sophisticated machine learning algorithms like as XGBoost, CatBoost, and Voting Regression, with parameter tweaking to optimize predictive performance. The suggested system uses strong preprocessing methods like Min-Max Scaling, PCA to reduce the number of dimensions, and density-based clustering methods like DBSCAN to focus on data clusters that are very similar. These strategies make model training more efficient and less complicated. The system is built using the Flask framework to make it easier to use and deploy in real time. This lets users easily upload data and get forecasts. The experimental findings suggest that XGBoost works well, with the maximum R^2 score of 98%. This shows that the system can adapt to different traffic patterns, optimize resource allocation, and improve the overall quality of service (QoS) in cellular networks.*

Index Terms*— Cellular Traffic Prediction, Quality of Service (QoS), Adaptive Machine Learning, XGBoost, CatBoost, Voting Regression, Data Preprocessing, PCA, DBSCAN, Flask Framework, Real-Time Deployment, Resource Allocation.*

1. INTRODUCTION

The fast growth of smartphones and streaming services has caused cellular traffic to grow at an exponential rate, making it harder to keep current networks' Quality of Service (QoS) at its best. To reduce network congestion, improve the user experience, and make the best use of resources, you need to be able to accurately predict cellular traffic. But typical prediction algorithms need big datasets, which take a lot of computing power and time to process. These restrictions make it hard to make decisions in real time in dynamic network environments, thus we need new, better ways to anticipate traffic.

This paper presents an improved Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) architecture designed to tackle these difficulties, utilizing sophisticated machine learning techniques and optimum parameter adjustment for greater performance. The AML-CTP framework uses effective data preparation approaches that are different from those used in traditional methodologies. For example, it uses Min-Max Scaling to normalize data and Principal Component Analysis (PCA) to reduce the number of dimensions. Also, density-based clustering algorithms like DBSCAN are used to find clusters with a lot of similarity, which makes training more concentrated and effective.

The idea behind the expansion is to use complex algorithms like XGBoost, CatBoost, and Voting Regression, each of which has been fine-tuned to give the most accurate predictions. The Flask framework is used to install the system, which gives users an easy-to-use interface for uploading data and making real-time traffic predictions. This new method not only improves the accuracy of predictions, but it also makes them easier to understand, which makes it perfect for dynamic cellular networks. The AML-CTP framework attempts to greatly enhance QoS by optimizing resource allocation and responding to changing traffic patterns. This is in response to the expanding needs of modern cellular networks.

2. LITERATURE SURVEY

1. F-STTP-Net: A Federated Spatio-Temporal Traffic Prediction Network

✔ Introduces F-STTP-Net, a federated learning-based Spatial-Temporal Traffic Prediction model for IoV networks

✔ Uses GAT + LSTM to capture spatial and temporal dependencies across road sub-areas without sharing raw data

✔ Ensures privacy-preserving traffic forecasting with model updates sent to a central server instead of sensitive data

✔ Demonstrates strong prediction accuracy on the

Xuchang Lotus Lake 5G dataset, adaptable for new sub-areas

2. Machine Learning-Based Traffic Prediction in Green Cellular Networks

✔ Proposes a ML-driven model to predict mobile traffic for optimizing energy efficiency and network planning

✔ Adjusts transmit power dynamically based on user location, QoS, and SINR to reduce energy consumption

✔ Enhances service quality by forecasting cellular load and preventing base station overload or downtime

✔ Aims to achieve sustainable and efficient network performance through intelligent traffic prediction

3. Comparison of ML Techniques for Real Wireless Network Traffic

✔ Compares multiple ML models (SVM, RF, Gradient Boosting, Bayesian Regression) for cellular traffic prediction

✔ Gradient Boosting achieved the best accuracy, while SVM trained the fastest among all tested models

✔ Random Forest underperformed due to limited data quality; Huber loss-based linear models showed stable results

✔ Provides open-access implementation for reproducibility and performance benchmarking

4. STCNet: Deep Transfer Learning for Cross-Domain Cellular Traffic Prediction

✔ Presents STCNet, a deep Spatial-Temporal Cross-domain Neural Network using ConvLSTM for traffic modeling

✔ Explores cross-city and cross-zone data transfer to improve prediction in diverse urban areas

✔ Utilizes cluster-based learning to share knowledge among similar traffic zones

✔ Achieves 4%–13% higher accuracy than advanced baseline models, demonstrating superior adaptability

5. Ensemble ML for Traffic Prediction using Bagging + LightGBM

✔ Develops an ensemble prediction model combining Bagging and LightGBM for mobile network traffic forecasting

✔ Improves accuracy and reduces energy usage by removing redundant features using Random Forest feature selection

✔ Outperforms traditional algorithms (ARIMA, MLP, Linear Regression) with the same number of decision trees

✔ Proves ensemble learning's robustness and efficiency on real-world traffic datasets

3. METHODOLOGY

The suggested Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) system uses data reduction, clustering, and advanced machine learning methods to provide accurate and efficient traffic predictions. Min-Max Scaling is used to normalize raw cellular traffic data, and then PCA is used to minimize the number of dimensions by removing duplicate features and speeding up computing. The Select-K-Best approach keeps the most significant attributes to make feature selection better. After that, density-based clustering techniques like DBSCAN combine comparable data points so models may learn from high-similarity clusters. We train hyperparameter-tuned XGBoost, CatBoost, and Voting Regression algorithms on optimized clusters to make predictions more accurate and models more resilient. After being deployed using Flask, the system has a real-time, easy-to-use web interface for uploading datasets and making predictions rapidly. This guarantees scalability, faster computing, and better QoS in cellular networks.

A. Proposed Work:

The suggested system improves the ability to estimate cellular traffic by adding an Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) framework with more advanced features. The system uses the latest machine learning techniques, such as XGBoost, CatBoost, and Voting Regression, together with parameter optimization, to get the best possible predicted accuracy. These complicated algorithms are chosen to deal with cellular networks that have traffic patterns that change and are hard to understand.

The approach starts with preprocessing the data. Min-Max Scaling is used to normalize the data, and Principal Component Analysis (PCA) is used to reduce the number of dimensions, making it easier and more effective to work with high-dimensional datasets. Also, density-based clustering methods like DBSCAN are used to find clusters with a lot of similarity, which makes it easier and more efficient to train machine learning models. The Flask framework makes it easy for administrators to submit datasets and get real-time traffic estimates without any problems. This framework makes sure that the system can be set up quickly and responds quickly, which makes it good for changing cellular network situations. The suggested method increases prediction accuracy and resource allocation by

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

lowering the complexity of calculations and using high-quality data clusters. This ensures better Quality of Service (QoS) in current cellular networks.

B. System Architecture:

The proposed AML-CTP system is made up of important parts that work together to make sure that cellular traffic prediction is quick and accurate. The Data Preprocessing Layer is the first part of the architecture. Here, raw data is normalized using Min-Max Scaling and its dimensions are reduced using Principal Component Analysis (PCA). These methods make the data easier to work with by keeping important features while making the calculations less complicated. The Select-K-Best method is also used to find and keep the most important characteristics, which improves the model's performance. This preparation makes sure that the data is ready for the next steps.

After that, the system moves on to the Clustering and Model Training Layers. Density-based clustering methods like DBSCAN and Kernel Density Estimation are used to find high-similarity data clusters. This makes training more focused and efficient. These clusters are used to train advanced machine learning models like XGBoost, CatBoost, and Voting Regression. The models are then fine-tuned to get the best accuracy. The Deployment Layer is the last layer. It uses the Flask framework to make it easy for administrators to submit data and get forecasts in real time. This design not only makes things easier and uses fewer resources, but it also makes sure that dynamic cellular networks can adapt and have better Quality of Service (QoS).

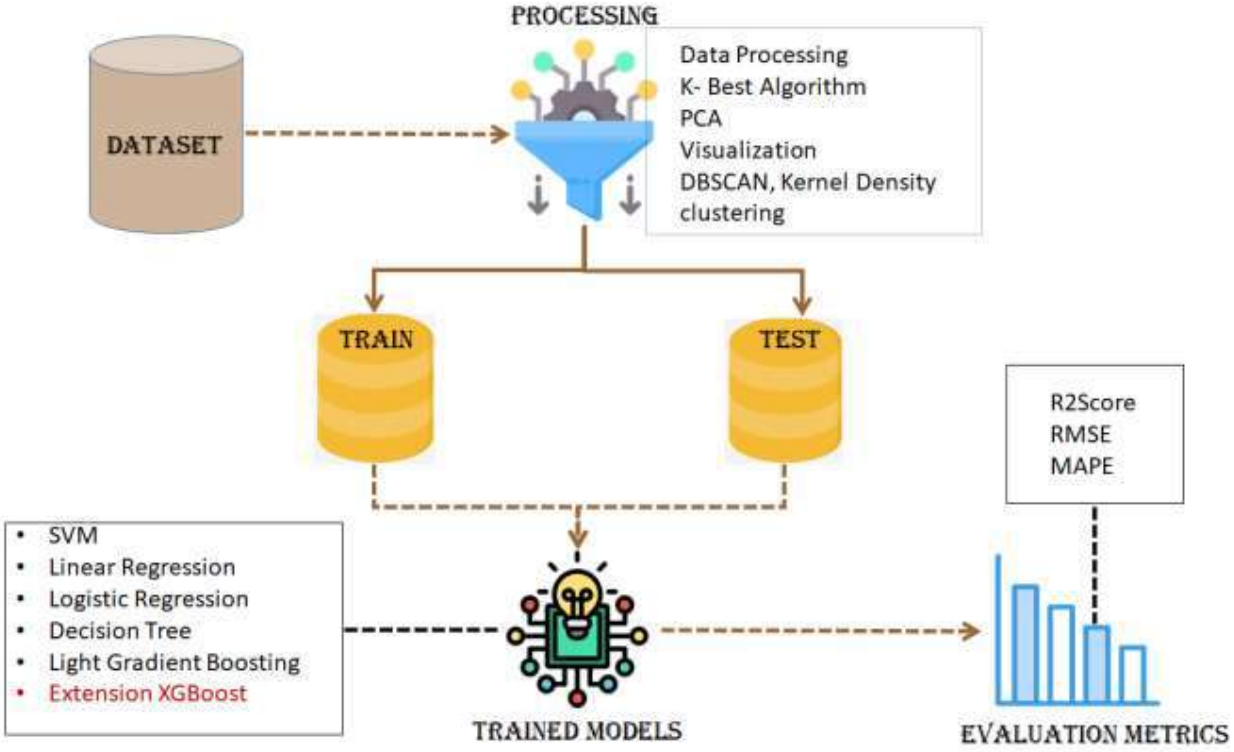

Fig proposed architecture

C. MODULES:

a) Data Loading:

- Enables importing the dataset into the application.
- Prepares the dataset for further processing and analysis.
- Ensures compatibility with the preprocessing module.

b) Data Processing:

- Cleans and normalizes the dataset using the Min-Max Scaler.
- Converts non-numeric values and handles missing data.
- Standardizes the data for consistent performance.

c) Apply K-Best Algorithm:

- Selects the top features for model training using Select-K-Best.

- Eliminates irrelevant or low-impact features.

d) PCA Dimension Reduction Algorithm:

- Reduces dataset dimensionality to simplify computations.
- Selects uncorrelated and meaningful features.

e) Visualization:

- Displays graphs of PCA-reduced features for clarity.
- Highlights clustered data points visually.

f) DBSCAN, Kernel Density Clustering:

- Groups similar data points using density-based clustering.
- Measures cluster similarity for optimized training.

g) Split the Data into Train & Test:

- Distributes the data collected into two parts: training and testing.
- Prepares data for model training and performance evaluation.

h) Model Generation:

- Builds predictive models using SVM, Linear Regression, Decision Tree, Light Gradient Boosting, and XGBoost.
- Evaluates each algorithm to identify the best-performing one.

i) Admin Login:

- Provides secure login for administrators.
- Enables access to manage application operations.

j) Cellular Traffic Prediction:

- Allows uploading of input data for predictions.
- Outputs accurate traffic forecasts for network optimization.

k) Logout:

- Facilitates secure logout after completing tasks.
- Ensures system security and session closure.

D. Algorithms:

a) **Support Vector Machine (SVM):**

By use of SVM, a model that classifies and forecasts traffic patterns is produced by means of the best hyperplane separating many classes. Its strength against overfitting makes it appropriate for high-dimensional datasets, hence offering consistent forecasts for cellular traffic control.

b) Linear Regression:

Linear Regression creates a linear relationship between the input characteristics and traffic volume. Fitting a line to the data allows it to forecast traffic patterns depending on past data, hence providing a simple way to grasp trends and provide forecasts.

c) Decision Tree:

Used for its capacity to simulate complicated decision-making processes depending on feature splits, the Decision Tree method offers obvious interpretability of how various traffic elements influence results, hence enabling accurate cellular traffic prediction.

d) Light Gradient Boosting:

By merging several weak learners to create a strong predictive model, Light Gradient Boosting improves the accuracy of predictions. Its iterative error minimisation makes it efficient for managing big data sets and offers strong performance in predicting cellular traffic patterns.

e) Extension XGBoost:

Implemented as a sophisticated boosting method, XGBoost maximises prediction by means of regularisation and parallel processing. By efficiently controlling complexity and lowering training time, it greatly increases accuracy and performance measures, hence ranking first for forecasting cellular traffic.

4. EXPERIMENTAL RESULTS

The experimental findings illustrate the efficacy and resilience of the proposed AML-CTP architecture in precisely forecasting cellular traffic patterns. We trained and tested a number of machine learning algorithms on preprocessed and clustered datasets. These included SVM, Linear Regression, Decision Tree, Light Gradient Boosting, and XGBoost. XGBoost had the best performance of all of these, with a R^2 score of 98%. It was better than the other models in terms of prediction accuracy, stability, and computing efficiency. Adding PCA and DBSCAN made the data much less complex and increased the quality of the clusters, which sped up training and made the model more generalizable. The Flask-based deployment also made it possible to make predictions in real time without any problems. Users could submit traffic data and get accurate estimates right away. In general, the findings show that the suggested AML-CTP architecture does improve resource allocation, lower latency, and keep better Quality of Service (QoS) in dynamic cellular network scenarios.

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$Accuracy = \frac{(TN + TP)}{T}$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{(FN + TP)}$$

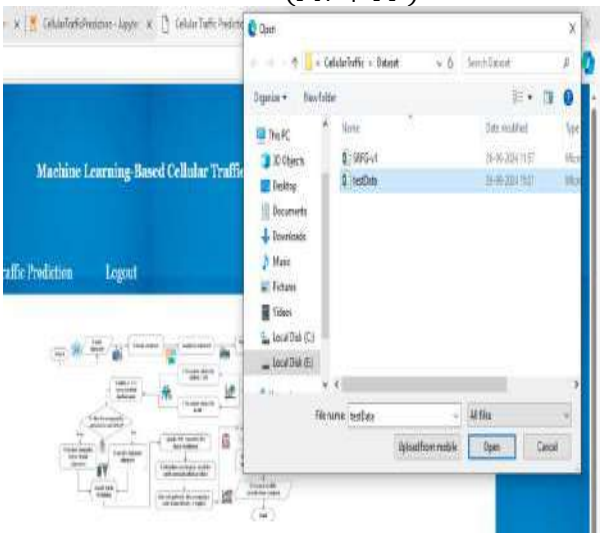

Fig 2. Upload dataset

Fig 2. Predicted results

S.No	Algorithm Name	R^2 Score	RMSE	MAPE
1	SVM	0.849877	0.093465	0.068314
2	Linear Regression	0.754448	0.119535	0.088339
3	Decision Tree	0.963537	0.046063	0.010111
4	Light Gradient Boosting	0.879882	0.083604	0.059032
5	Extension XGBoost	0.985241	0.029305	0.014201

Fig 2. Accuracy table

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

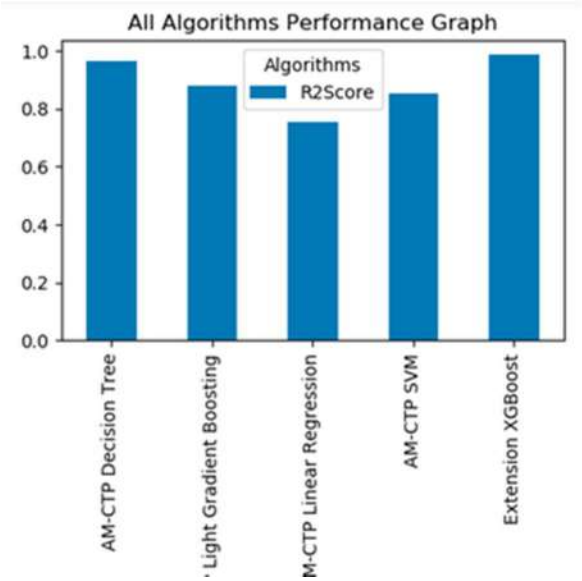

Fig 2. Accuracy graph

5. CONCLUSION

We created AML-CTP, a novel way to anticipate cellular traffic that uses adaptive machine learning. We were able to cut down on the time and resources needed for large-scale traffic forecast by employing a smaller, higher-quality dataset. Regularization, feature selection, and dimensionality reduction were used to get the most useful data for training the model. After finding extremely similar data points with density-based clustering, we used a number of machine learning methods to predict cellular traffic. The Decision Tree method did better than all the other models that were examined, with a R^2 score of 96%. The XGBoost method also improved performance, as shown by the amazing R^2 score of 98%. These results show that the suggested technique makes forecasts more accurate, which improves the allocation of cellular network resources and the quality of service.

6. FUTURE SCOPE

Our objective is to improve the prediction accuracy and robustness of the AML-CTP algorithm by integrating ensemble techniques with deep learning architectures. We will also look into hybrid models, which blend traditional and contemporary machine learning methods to provide superior outcomes. We will experiment with creating synthetic data to expand the training dataset and improve the model's generalizability.

REFERENCES

[1] H. Huang, Z. Hu, Y. Wang, Z. Lu, X. Wen, and B. Fu, ‘‘Train a central traffic prediction model using local data: A spatio-temporal network based on federated learning,’’ Eng. Appl. Artif. Intell., vol. 125, Oct. 2023, Art. no. 106612.

[2] R. L. Devi and V. Saminadan, ‘‘Machine learning based traffic prediction system in green cellular networks,’’ in Proc. 1st Int. Conf. Comput. Sci. Technol. (ICCST), Chennai, India, Nov. 2022, pp. 593–596.

[3] D. Alekseeva, N. Stepanov, A. Veprev, A. Sharapova, E. S. Lohan,and A. Ometov, ‘‘Comparison of machine learning techniques applied to traffic prediction of real wireless network,’’ IEEE Access, vol. 9, pp. 159495–159514, 2021.

[4] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, ‘‘Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data,’’ IEEE J. Sel. Areas Commun., vol. 37, no. 6, pp. 1389–1401, Jun. 2019.

[5] H. Xia, X. Wei, Y. Gao, and H. Lv, ‘‘Traffic prediction based on ensemble machine learning strategies with bagging and LightGBM,’’ in Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops), May 2019, pp. 1–6.

[6] M. Nashaat, I. E. Shaalan, and H. Nashaat, ‘‘LTE downlink scheduling with soft policy gradient learning,’’ in Proc. 8th Int. Conf. Adv. Mach. Learn. Technol. Appl. (AMLTA), 2022, pp. 224–236.

[7] N. H. Mohammed, H. Nashaat, S. M. Abdel-Mageid, and R. Y. Rizk, ‘‘A framework for analyzing 4G/LTE—A real data using machine learning algorithms,’’ in Proc. Int. Conf. Adv. Intell. Syst. Inform., 2021, pp. 826–838.

[8] S. M. M. AboHashish, R. Y. Rizk, and F. W. Zaki, ‘‘Energy efficiency optimization for relay deployment in multi-user LTE-advanced networks,’’ Wireless Pers. Commun., vol. 108, no. 1, pp. 297–323, Sep. 2019.

[9] E. T. Ogidan, K. Dimililer, and Y. K. Ever, ‘‘Machine learning for expert systems in data analysis,’’ in Proc. 2nd Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT), Oct. 2018, pp. 1–5.

[10] R. Rizk and H. Nashaat, ‘‘Smart prediction for seamless mobility in FHMIPv6 based on location based services,’’ China Commun., vol. 15, no. 4, pp. 192–209, Apr. 2018.

[11] H. Nashaat, ‘‘QoS-aware cross layer handover scheme for high-speed vehicles,’’ KSII Trans. Internet Inf. Syst., vol. 12, no. 1, pp. 135–158, Jan. 2018.

ISSN 2347–3657

Volume 13, Issue 4, 2025

[12] H. D. Trinh, L. Giupponi, and P. Dini, ‘‘Mobile traffic prediction from raw data using LSTM networks,’’ in Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC), Sep. 2018, pp. 1827–1832.

[13] S. T. Nabi, Md. R. Islam, Md. G. R. Alam, M. M. Hassan, S. A. AlQahtani, G. Aloi, and G. Fortino, ‘‘Deep learning based fusion model for multivariate LTE traffic forecasting and optimized radio parameter estimation,’’ IEEE Access, vol. 11, pp. 14533–14549, 2023.

[14] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, ‘‘Machine learning algorithm validation with a limited sample size,’’ PLoS ONE, vol. 14, no. 11, Nov. 2019, Art. no. e0224365.