UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Machine Learning-Based Cellular Traffic Prediction Using Data Reduction Techniques

Batchu Bharat
PG Scholar,
Department of CSE,
Srinivasa Institute of Engineering and Technology,
Cheyyeru, Andhra Pradesh,
bharatlalitha@gmail.com

Saipriya Vissapragada,
Associate Professor & HoD - CSE,
Department of CSE,
Srinivasa Institute of Engineering and Technology,
Cheyyeru, Andhra Pradesh,
saipriya.vissapragada@gmail.com

Abstract: For modern networks to maximise Quality of Service (QoS), precise cellular traffic prediction is crucial, particularly given the increasing need for real-time applications. In order to increase predictive performance, this study introduces an upgraded Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) framework that incorporates cutting-edge machine learning algorithms such as XGBoost, CatBoost, and Voting Regression with parameter adjustment. The suggested system makes use of density-based clustering techniques like DBSCAN to concentrate on high-similarity data clusters, PCA for dimensionality reduction, and robust preprocessing approaches like Min-Max Scaling. These techniques minimise computing complexity while guaranteeing effective model training. The Flask framework is used in the system's implementation to improve accessibility and real-time deployment, enabling smooth user interaction for data uploading and prediction generation. With the greatest R2 score of 98%, experimental findings illustrate how successful XGBoost is at adapting to changing traffic patterns, allocating resources optimally, and improving 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 exponential growth in cellular traffic brought about by the quick spread of smartphones and streaming services makes it difficult to maintain the best possible Quality of Service (QoS) in contemporary networks. Accurate cellular traffic forecasting is necessary to optimise resource allocation, reduce network congestion, and enhance user experience. But conventional prediction techniques depend on big datasets, which

need a lot of processing power and take a long time. These drawbacks make it difficult to make decisions in real time in dynamic network environments, which is why creative and effective traffic prediction methods are essential.

This paper presents an improved Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) framework that uses cutting-edge machine learning techniques and optimal parameter adjustment for better performance in order to overcome these issues. The AML-CTP framework uses effective data pretreatment techniques, such as Principal Component Analysis (PCA) for dimensionality reduction and Min-Max Scaling for normalisation, in contrast to traditional approaches. Additionally, high-similarity clusters are found using density-based clustering techniques like DBSCAN, allowing for more targeted and effective training.

The extension idea incorporates cutting-edge algorithms that are optimised to maximise prediction accuracy, including XGBoost, CatBoost, and Voting Regression. The Flask framework, which is used to deploy the system, offers an intuitive user interface for smooth data transfer and real-time traffic forecasts. This novel method is perfect for dynamic cellular networks since it not only improves prediction performance but also lowers time complexity. The AML-CTP framework seeks to greatly enhance QoS by maximising resource allocation and adjusting to changing traffic patterns, meeting the increasing needs of contemporary cellular networks.

2. LITERATURE SURVEY

i) Train a central traffic prediction model using local data: A spatio-temporal network based on federated learning:

https://www.sciencedirect.com/science/article/abs/pii/S0952197623007960

With the increasing sophistication of onboard sensors and the widespread deployment of road sensors, deep learning offers fine-grained traffic prediction by using vast quantities of unprocessed traffic data from the Internet of Vehicles. By constructing a prediction model collaboratively utilising all local data, the majority of current research draw attention to privacy, data security, and communication concerns. In order to enable model parameter updates to the central server without disclosing sensitive information, this research introduces the F-STTP-Net, a Spatial-Temporal Traffic Prediction Network based on federated learning. In order to classify the road network based on macroscopic fundamental schematic properties, we first develop a sub-area division approach. To counteract the road network's reliance on time and space, we suggest employing GAT and LSTM to train a model locally for every sub-area. Traffic volumes at sub-area junctions are predicted using the branch structure model. A strong central model that

satisfies international data sharing and privacy standards is produced by combining local models with federated learning. Even without sub-region raw data, F-STTP-Net was able to forecast when tested on the real dataset from the Xuchang Lotus Lake 5G autonomous vehicles demonstration area. It could also be simple to expand the notion to include a new sub-area.

ii) Machine Learning Based Traffic Prediction System in Green Cellular Networks

https://ieeexplore.ieee.org/abstract/document/10040347

If networks are able to estimate consumer mobile phone traffic precisely in advance, they may improve planning and service. Regardless of the number of mobile users, the main goal is to improve the network's quality and performance. thereby lowering energy consumption and enhancing network performance. This is done when its usage is significant due to the possible impact on cellular networks. One of the world's most robust mobile networks consumes gigawatts of energy annually. With more accurate and reliable time-series models of mobile cellular traffic volume, network performance might be improved. This research project will develop a methodology to estimate cellular network load traffic intelligently. If cellular networks are made simpler and offer higher quality of service, they can handle the high traffic load more effectively. This might be used to accurately satisfy user demands. The signal-to-interference-and-noise ratio, user locations, and quality-of-service needs all influence the minimum transmit power. Don't let any base stations fall down. The transmitter power is tuned to the lowest possible level after considering the signal-to-noise ratio and user quality of service.

iii) Comparison of Machine Learning Techniques Applied to Traffic Prediction of Real Wireless Network

https://ieeexplore.ieee.org/abstract/document/9623523

These days, a growing number of gadgets are increasing network traffic. Researchers find complex connections, anomalies, and new traffic patterns to increase the system's efficacy. An expanding area of expertise in this subject is the application of both conventional and state-of-the-art Deep Learning approaches to enhance network performance in complex and diverse contexts. We create a list of the most recent machine learning applications in communications before determining the biggest problems and potential fixes. We use publicly accessible cellphone traffic to construct an ML environment. The results indicated that the SVM methodology trained faster than alternative methods. Gradient Boosting generated the most accurate estimations because of its effective data determination. Because of its restricted attributes, random forest performs badly. Probabilistic Bayesian

regression was only somewhat less successful than Gradient Boosting, while being quicker to train. The Huber loss function may be used to optimise model parameters, and linear models that made use of it did well in performance tests. We contribute by releasing the analysed algorithm's source code under Open Access.

iv) Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data:

https://ieeexplore.ieee.org/abstract/document/8667446

Precise traffic modelling and prediction enabled by machine learning is crucial for autonomous network management and service provisioning in future smart cellular networks driven by big data. The Spatial-Temporal Cross-domain Neural Network, or STCNet, is a cutting-edge deep learning architecture that can recognise intricate cellular data patterns. STCNet models spatial-temporal relationships with its convolutional long short-term memory network. STCNet actively collects and models three datasets from various locations to determine what variables outside of traffic itself cause delays. As cellular traffic from different city functional zones is both similar and different, we propose to partition the city into clusters and develop a learning system across clusters to improve information reuse. The suggested STCNet model looks into how knowledge is transferred via various types of cellular traffic. Three assessment criteria were applied in order to demonstrate that STCNet functions with real cellular traffic data. The results show that STCNet outperforms the newest and most sophisticated algorithms. For instance, performance may be improved by 4%–13% using STCNet-based transfer learning.

v) Traffic Prediction Based on Ensemble Machine Learning Strategies with Bagging and LightGBM

https://ieeexplore.ieee.org/abstract/document/8757058

One of the biggest challenges in developing mobile networks is predicting with sufficient precision how to use resources most efficiently while lowering energy consumption and enhancing quality of service. In the age of big data, machine learning techniques have been used to extract deep data in order to define network traffic instability, despite the fact that a single ML model's performance is frequently insufficient. Ensemble learning improves machine learning accuracy. In this study, we use LightGBM and RF to eliminate superfluous features from a model that forecasts traffic on mobile networks. Furthermore, a novel traffic prediction model based on bagging and the LightGBM ensemble architecture is proposed. The model is tested using real traffic data. Even with the same number of decision trees, the suggested model outperforms a single LightGBM when compared to other well-known algorithms like ARIMA, MLP, and Linear Regression..

3. METHODOLOGY

A. Proposed Work:

By providing a sophisticated Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) architecture, the suggested system improves cellular traffic prediction. To get the best predicted accuracy, the system combines state-of-the-art machine learning techniques, such as XGBoost, CatBoost, and Voting Regression, with parameter optimisation. These cutting-edge algorithms were chosen especially to manage intricate and dynamic traffic patterns in cellular networks.

To ensure effective and efficient handling of high-dimensional datasets, the technique starts with data pretreatment utilising Principal Component Analysis (PCA) for dimensionality reduction and Min-Max Scaling for normalisation. Furthermore, high-similarity clusters are found using density-based clustering algorithms like DBSCAN, which allow for targeted and resource-efficient machine learning model training. The Flask framework is used to create the system, providing administrators with an easy-to-use interface to submit datasets and easily receive real-time traffic estimates. The solution is appropriate for dynamic cellular network scenarios because of this framework's rapid deployment and reactivity. The suggested solution ensures higher Quality of Service (QoS) in contemporary cellular networks by lowering computing complexity and using high-quality data clusters to improve resource allocation and prediction accuracy.

B. System Architecture:

The suggested AML-CTP system is organised into essential parts to guarantee precise and effective cellular traffic forecasting. The Data Preprocessing Layer is the first layer of the design, where raw data is normalised using Min-Max Scaling and its dimensionality reduced using Principal Component Analysis (PCA). These methods simplify the data while maintaining important aspects and lowering computing complexity. To improve model performance, the Select-K-Best method is also used to find and keep the most pertinent characteristics. The data is optimised for further phases thanks to this treatment.

After that, the system moves on to the Layers for Clustering and Model Training. Density-based clustering methods such as DBSCAN and Kernel Density Estimation are used to identify high-similarity data clusters, which enables targeted and effective training. These clusters are used to train sophisticated machine learning models, such as XGBoost, CatBoost, and Voting Regression, with parameter adjustment to optimise accuracy. The Deployment Layer, the last layer, uses the Flask framework to give administrators an easy-to-use interface for uploading data and getting real-time forecasts. This design guarantees flexibility and enhanced Quality of Service (QoS) in dynamic cellular networks

while simultaneously lowering time complexity and resource consumption.

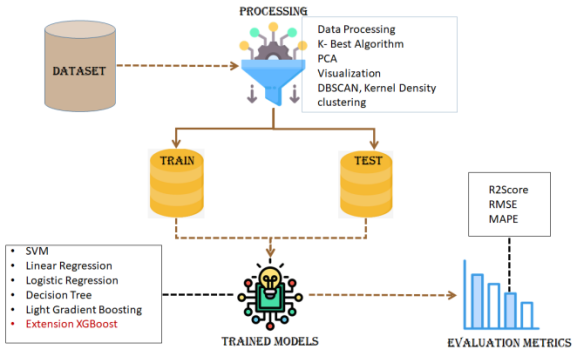

Fig 1Proposed Architecture

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

i. Support Vector Machine (SVM): The optimal hyperplane separating several classes is used in SVM to create a model that both classifies and predicts traffic patterns. It provides reliable predictions for cellular traffic control because of its resistance to overfitting, which makes it suitable for high-dimensional datasets.

ii. Linear Regression: A linear connection between the input qualities and traffic volume is produced using linear regression. By fitting a line to the data, it can predict traffic patterns based on historical data, offering a straightforward method of understanding trends and making predictions.

iii. Decision Tree: Accurate cellular traffic prediction is made possible by the Decision Tree approach, which is used to mimic complex decision-making processes based on feature splits. It provides clear interpretability of how different traffic aspects affect outcomes.

iv. Light Gradient Boosting: Light Gradient Boosting increases prediction accuracy by combining many weak learners into a powerful predictive model. It performs well in anticipating cellular traffic patterns and is effective at managing large data sets due to its iterative error minimisation.

v. Extension XGBoost: XGBoost is a complex boosting technique that uses parallel processing and regularisation to optimise prediction. It ranks #1 for cellular traffic forecasting because it effectively manages complexity and reduces training time, which significantly improves accuracy and performance metrics.

4. EXPERIMENTAL RESULTS

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

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

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

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

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

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

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

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

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

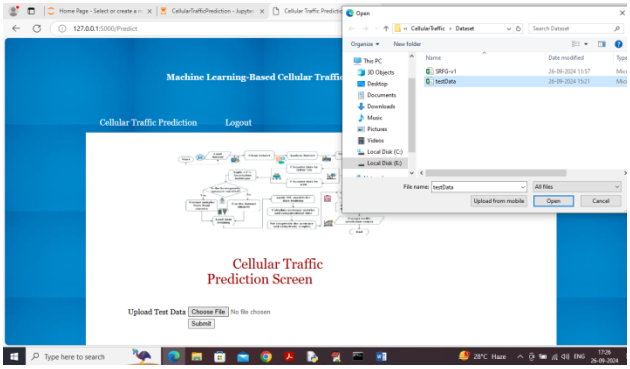

Fig 2. Upload dataset

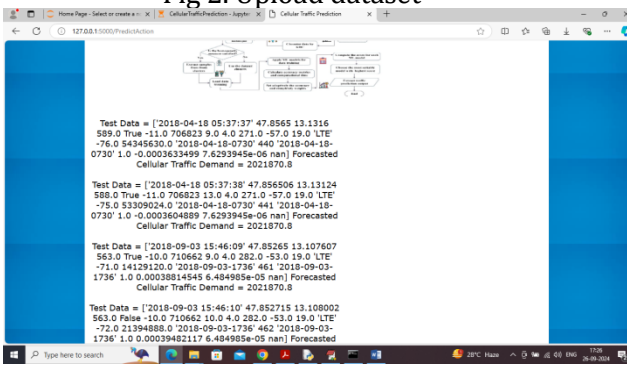

Fig 2. Predicted results

	Algorithm Name	R2 Score	RMSE	MAPE
0	SVM	0.849877	0.093465	0.068314
1	Linear Regression	0.754448	0.119535	0.088339
2	Decision Tree	0.963537	0.046063	0.010111
3	Light Gradient Boosting	0.879882	0.083604	0.059032
4	Extension XGBoost	0.985241	0.029305	0.014201

Fig 2. Accuracy table

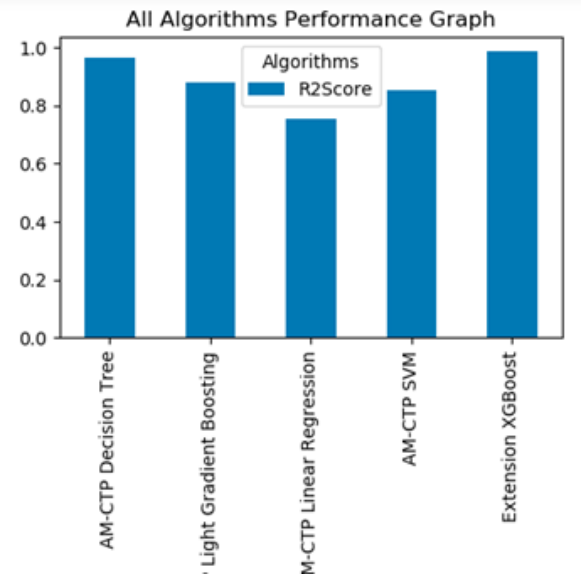

Fig 2. Accuracy graph

5. CONCLUSION

Based on adaptive machine learning, we have created a novel technique for cellular traffic prediction termed AML-CTP. We were able to cut down on the time and resources needed for large-scale traffic forecast by employing a smaller, higher-quality dataset. By using regularisation, feature selection, and dimensionality reduction, the most pertinent data for model training was used. After employing density-based clustering to identify highly similar data points, we tested a number of machine learning techniques to predict cellular traffic. The Decision Tree approach beat all other examined models with an amazing R2 score of 96%. The impressive R2 score of 98% indicates that the XGBoost method was also used to improve performance. These findings show that the suggested approach improves forecast

accuracy, which improves quality of service and cellular network resource allocation.

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 generalisability.

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.

[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.