Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Data Reduction Techniques-Driven ML-Based Cellular Traffic Prediction

Dr. M. PRAVEEN KUMAR [1], THOTA NAGA SUJITH [2]
#1 Working As Professor, Department Of CSE, PBR Visvodaya Institute Of Technology & Science, Kavali, Andhra Pradesh 524201
#2 PG Scholar, Department Of CSE, PBR Visvodaya Institute Of Technology & Science, Kavali, Andhra Pradesh 524201

Abstract: Especially with the growing need for real-time applications, accurate cellular traffic prediction is essential for optimising Quality of Service (QoS) in contemporary networks. This study presents an enhanced Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) framework that integrates advanced machine learning algorithms, including XGBoost, CatBoost, and Voting Regression, with parameter tuning for improved predictive performance. The proposed system employs robust preprocessing techniques such as Min-Max Scaling, PCA for dimensionality reduction, and density-based clustering methods like DBSCAN to focus on high-similarity data clusters. These methods ensure efficient model training while reducing computational complexity. To enhance accessibility and real-time deployment, the system is implemented using the Flask framework, allowing seamless user interaction for uploading data and obtaining predictions. Experimental results demonstrate the effectiveness of XGBoost, achieving the highest R^2 score of 98%, thereby showcasing the system's capability to adapt to varying traffic patterns, optimize resource allocation, and enhance overall 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 rapid proliferation of smartphones and streaming services has led to an exponential increase in cellular traffic, posing challenges for maintaining optimal Quality of Service (QoS) in modern networks. Reducing network congestion, improving user experience, and optimising resource allocation all need precise cellular traffic forecasting. However, traditional prediction methods rely on large datasets, which demand high computational resources and result in time-intensive processes. These limitations hinder real-time decision-making in dynamic network environments, making the need for innovative, efficient traffic prediction techniques imperative.

To address these challenges, this study introduces an enhanced Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) framework, which leverages advanced machine learning algorithms and optimized parameter tuning for superior performance. Unlike conventional methods, the AML-CTP framework employs efficient data preprocessing techniques, including Min-Max Scaling for normalization and Principal Component Analysis (PCA) for dimensionality reduction. Furthermore, density-based clustering methods like DBSCAN are utilized to identify high-similarity clusters, enabling more focused and efficient training.

The extension concept integrates advanced algorithms such as XGBoost, CatBoost, and Voting Regression, each tuned to maximize prediction accuracy. The system is deployed via the Flask framework, providing a user-friendly interface for seamless data upload and real-time traffic predictions. This innovative approach not only enhances predictive performance but also reduces time complexity, making it ideal for dynamic cellular networks. By optimizing resource allocation and adapting to fluctuating traffic patterns, the AML-CTP framework aims to significantly improve QoS, addressing the growing demands of modern 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

Deep learning provides fine-grained traffic prediction as onboard sensors are more sophisticated and road sensors are widely deployed by using large amounts of raw traffic data in the Internet of Vehicles. Most of the present studies highlight privacy, data security, and communication issues by building a prediction model jointly using all local data. This paper presents the F-STTP-Net, a Spatial-Temporal Traffic Prediction Network built on federated learning, therefore allowing model parameter updates to the central server without revealing sensitive data. We initially create a sub-area division technique to categorise the road network according to macroscopic essential schematic characteristics. We propose training a model locally for each sub-area using GAT and LSTM to offset the road network's dependence on time and space. The branch structure model predicts traffic volumes at sub-area intersections. Combining local models with federated learning creates a robust central model meeting global data sharing and privacy requirements. Testing F-STTP-Net on the actual dataset from the Xuchang Lotus Lake 5G autonomous cars demonstration area revealed it could forecast even without sub-area raw data. A new sub-area might also be easily added to the concept.

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

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

Networks may enhance planning and service. if they can accurately predict consumer mobile phone traffic in advance. Improving the network's performance and quality is the primary objective, regardless of how many mobile users there may be. While improving network performance and decreasing energy usage. Because of the potential impact on cellular networks, this is done when its usage is

considerable. Gigawatts of energy are used up every year by one of the strongest mobile networks in the world. Better network performance may be achieved with more precise and trustworthy time-series models of mobile cellular traffic volume. In order to intelligently predict cellular network load traffic, this research project will build a model. Cellular networks can better manage the heavy traffic load if they are made simpler and have better quality of service. Accurately meeting user needs may be achieved with this. The minimum transmit power is determined by quality-of-service requirements, user locations, and the signal-to-interference-and-noise ratio. Keep all base stations from going down. After taking into account the signal-to-noise ratio and the quality of service for users, the transmitter power is set to the lowest achievable level.

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

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

Increasing numbers of devices are driving up network traffic nowadays. In order to improve the system's effectiveness, researchers uncover complicated linkages, irregularities, and novel traffic patterns. The use of both traditional and cutting-edge Deep Learning techniques to improve network performance in complex and heterogeneous settings is an emerging area of expertise in this field. Before identifying the most significant challenges and possible solutions, we compile a list of current Machine Learning applications in communications. We build an ML environment with a publicly available cellular traffic. When compared to other approaches, the findings demonstrated that the SVM methodology trained more quickly. Due to its efficient data determination, Gradient Boosting produced the most accurate estimates. The limited qualities lead random forest to perform poorly. Despite being faster to train, probabilistic Bayesian regression was only slightly less effective than Gradient Boosting. Model parameters may be optimised using the Huber loss function, and linear models using this function performed well in performance evaluations. We make one contribution by making the source code of the algorithm that was examined available 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

For autonomous network management and service provisioning in smart cellular networks powered by big data of the future, precise traffic modelling and prediction made possible by machine learning is essential. One novel deep learning design that can capture complex cellular data patterns is the Spatial-Temporal Cross-domain Neural Network, or STCNet. Using its convolutional long short-term memory network, STCNet simulates spatial-temporal connections. To find out what factors outside of traffic itself cause delays, STCNet actively gathers and models three datasets from different places. We suggest dividing the city into groups and establishing a system for learning amongst clusters to enhance the reuse of information, as cellular traffic from various city functional zones is both comparable and distinct. The proposed STCNet model investigates the transmission of knowledge across different kinds of cellular traffic. In order to prove

that STCNet works with actual cellular traffic data, three evaluation criteria were used. Results demonstrate that STCNet achieves better results than the most recent and advanced algorithms. For example, a 4%-13% improvement in performance is achieved with STCNet-based transfer learning.

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

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

Predicting with enough accuracy how to use resources most efficiently while reducing energy consumption and improving quality of service is a major challenge in mobile network development. Even though the performance of a single ML model is often inadequate, ML approaches have been utilised to extract deep data in the era of big data in order to characterise network traffic instability. Accuracy in ML is enhanced by ensemble learning. In this research, we train a model to forecast traffic on mobile networks using LightGBM and RF to remove unnecessary characteristics. Additionally, a new model for traffic prediction is suggested, which is based on bagging and the LightGBM ensemble architecture. Using actual traffic data, the model is evaluated. The proposed model achieves better results than a single LightGBM when compared to other popular algorithms such as ARIMA, MLP, and Linear Regression, even while using the same amount of decision trees.

3. METHODOLOGY

A. Proposed Work:

The proposed system enhances cellular traffic prediction by introducing an Adaptive Machine Learning-based Cellular Traffic Prediction (AML-CTP) framework with advanced capabilities. The system integrates cutting-edge machine learning algorithms, including XGBoost, CatBoost, and Voting Regression, with parameter tuning to achieve optimal predictive accuracy. These advanced algorithms are specifically selected to handle dynamic and complex traffic patterns in cellular networks.

The methodology begins with data preprocessing using Min-Max Scaling for normalization and Principal Component Analysis (PCA) for dimensionality reduction, ensuring efficient and effective handling of high-dimensional datasets. Additionally, density-based clustering techniques such as DBSCAN are employed to identify high-similarity clusters, enabling focused and resource-efficient training of machine learning models. The system is implemented through the Flask framework, which facilitates a user-friendly interface for administrators to upload datasets and obtain real-time traffic predictions seamlessly. This framework ensures quick deployment and responsiveness, making the system suitable for dynamic cellular network environments. By reducing computational complexity and leveraging high-quality data clusters, the proposed system not only improves prediction accuracy but also enhances resource allocation, ensuring superior Quality of Service (QoS) in modern cellular networks.

B. System Architecture:

The proposed AML-CTP system is structured into key components to ensure efficient and accurate cellular traffic prediction. The architecture begins with the Data Preprocessing Layer, where raw data undergoes normalization using Min-Max Scaling and

dimensionality reduction through Principal Component Analysis (PCA). These techniques streamline the data, reducing computational complexity while preserving critical features. Additionally, the Select-K-Best algorithm is employed to identify and retain the most relevant features, enhancing model performance. This preprocessing ensures that the data is optimized for subsequent stages.

The system then transitions to the Clustering and Model Training Layers. High-similarity data clusters are identified using density-based clustering techniques like DBSCAN and Kernel Density Estimation, allowing for focused and efficient training. Advanced machine learning models, including XGBoost, CatBoost, and Voting Regression, are trained on these clusters with parameter tuning to maximize accuracy. The final layer, the Deployment Layer, leverages the Flask framework to provide a user-friendly interface for administrators to upload data and receive real-time predictions. This architecture not only reduces time complexity and resource usage but also ensures adaptability and improved Quality of Service (QoS) in dynamic cellular networks.

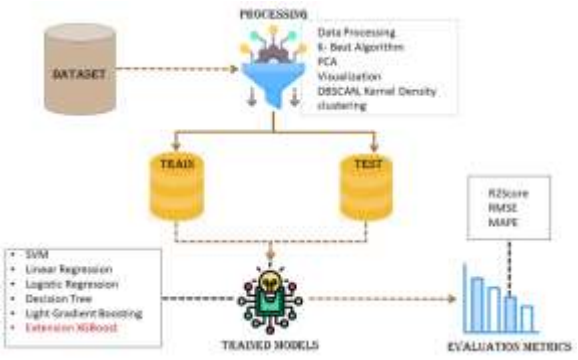

Fig 1Proposed Architecture

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): 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.

ii. 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.

iii. 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.

iv. 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.

v. 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

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

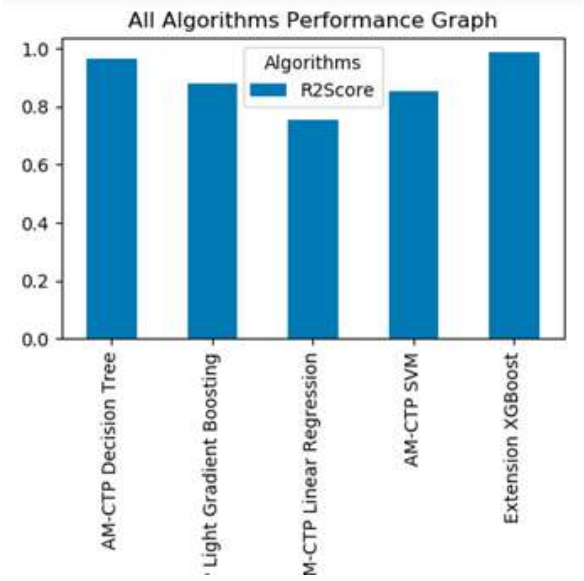

Fig 2. Accuracy graph

5. CONCLUSION

We have developed a new method for cellular traffic prediction called AML-CTP, which is based on adaptive machine learning. By using a smaller, higher-quality dataset, we were able to reduce the time and resources required for large-scale traffic prediction. The most relevant data for model training was employed through regularisation, feature selection, and dimensionality reduction. Following the identification of very comparable data points using density-based clustering, we conducted tests using several machine learning algorithms to forecast cellular traffic. With an impressive R^2 score of 96%, the Decision Tree technique outperformed all other models that were tested. The use of the XGBoost algorithm also enhanced performance, as seen by the remarkable R^2 score of 98%. These results demonstrate that the proposed strategy enhances the accuracy of predictions, leading to better cellular network resource allocation and quality of service.

6. FUTURE SCOPE

By incorporating ensemble methods and deep learning architectures into the AML-CTP algorithm, we aim to enhance its prediction accuracy and durability. In addition, we will investigate hybrid models, which combine old and modern machine learning techniques to get better results. In order to enhance the model's generalisability and increase the size of the training dataset, we will experiment with synthetic data generation.

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.

Author’s profile

Dr M. PRAVEEN KUMAR is an Associate Professor in the Department of Computer Science and Engineering at PBR Visvodaya Institute of Technology and Science in Kavali, Andhra Pradesh, India. With 22 years of academic experience.

His qualification B.E., M.E.,M.Tech, Ph.D in Computer science from SVU, Uttar Pradesh

I am **Naga Sujith Thota**, currently pursuing a M.Tech in Computer Science and Engineering at PBR VITS, Kavali, SPSR Nellore, Andhra Pradesh, India. My areas of interest include Artificial Intelligence/Machine learning and Data Analytics. I have earned certifications from Institute. I have certificate on International Conference on Detection of Ransomware Attacks Using Processor and Disk Usage Data