Al-Farabi Journal of Engineering Sciences Volume (4), Issue (2) December (2025)

ISSN: 2617-5517 (issn.org)
Al-Farabi Journal of Engineering Sciences
https://iasj.rdd.edu.iq/journals/journal/view/97
مجلة الفارابي للعلوم الهندسية تصدرها جامعة الفارابي

Developing An AI Model To Predict Network Data Consumption

Aber aziz kareem alsaadi, Department of Computer Science, College of Basic Education, Mustansiriyah University, Baghdad, Iraq,aber8177@gmail.com

Abstract

With the rapid increase in network traffic and the connection of more devices, predicting network data consumption has become essential for effective network management and avoiding bottlenecks. This research proposes an advanced AI architecture based on cutting-edge machine learning algorithms such as Gradient Boosting Regressor and XGBoost Regressor, and employing a Stacked XGBoost model to enhance performance accuracy.

The Stacked XGBoost model demonstrated outstanding performance with an MAE of 158.399, RMSE of 394.335, and R^2 of 0.987, indicating high accuracy and reliability in predicting network data consumption. The results suggest the potential for using this model to facilitate network planning, bandwidth optimization, and real-time traffic management within local area networks (LANs) and Internet of Things (IoT) environments.

Keywords: Network Data Consumption, Machine Learning, Traffic Prediction, Stacked Ensemble, Regression Model.

1. Introduction

In the new digital age, data usage in the world networks has been on an accelerating scale at an unprecedented pace. The growth of online streamlining, cloud computing, handheld computing and Internet of Things (IoT) have resulted in a dramatic increase in the volume of data that travels over the networks. As a result, it is an ongoing problem in which network operators and service providers are hard pressured to predict, control and optimize bandwidth use in order to achieve effective network performance and avoid congestion. Traditionally, estimation of network traffic has been done by use of statistical model and analysis techniques like autoregressive integrated moving average (ARIMA) models, exponential smoothing or simple threshold based methods [9]. Artificial Intelligence (AI) and Machine Learning (ML) are the new technologies that have become revolutionary in recent years that can resolve these issues. As the pressure grows to thrive on lower energy consumption and cost of operation, intelligent forecasting models will help in dynamically changing power consumption, switching capacities, or load balancing in network segments [3]. This study helps solve these problems by implementing an AI-driven regression model that would be able to make predictions about the network data consumption using flow-level indicators.

2. Literature Review

The initial researches was mainly on statistical and rule-based approaches in order to approximate network traffic behavior. Nonetheless, as high-speed networks and data-driven services have become feasible, more advanced methods, namely, machine learning (ML) and deep learning (DL), have become prominent [10]. The paper address how artificial intelligence (AI) can be implemented to predict network traffic patterns, prevent

congestion, load balance and effective allocation of resources. The article explains the concept of machine learning (ML) algorithms and deep learning (DL) models and hybrid AI methods that have been created to predict traffic within high-demand networks [1]. In this paper, a new system is proposed to forecast the wireless network traffic of Open Radio Access Networks (O-RAN) in small temporal scales based on a transformer architecture, which is a machine learning model that is the basis of generative AI tools. the proposed prediction-based approach boosts the average energy efficiency by 39.7% relative to the one with the name of the "Always on Traffic Steering xApp" and boosts the throughput by 10.1% relative to the implementation of the name of the "Always on Cell Sleeping rApp)[2]. paper [4] developed an innovative meta-heuristic optimization framework of boosting the prediction of road traffic based on the fitness and grey wolf optimization algorithms of the dipper throat. The given algorithm is used to maximize the hyper parameters of long short-term memory (LSTM) network as one of the most efficient time series modelling strategies that is popular in sequence prediction problems.

2.2 Gaps in the Literature

Despite the development of many models of traffic prediction, there are still several constraints:

- Data Availability: A large number of studies use proprietary or simulated databases, which are not necessarily what a real network does.
- Feature Engineering: Most literature does not investigate the use of extensive feature extraction.

2.3 Contribution of This Study

The study remedies the above weaknesses, using a preprocessed flow-based data, with critical traffic features. Rather than placing emphasis on the time-based relationship, this research puts an emphasis on the association between flow characteristics and data consumption and offers a generalized and understandable model. The suggested AI-based framework of regression will integrate both analytical and interpretive capabilities and will be used to reach the balance between the performance and the simplicity of calculations.

3. Proposed System

The system will be architecture in a 3 layer fashion as shown in figure 1:

1. **Data Preprocessing Layer**

   Raw network traffic data are initially cleaned and standardized and transformed. Median values are used to inpute missing values, the duration of 0 is corrected to eliminate division errors and the logarithmic transformation is used to skew the value to zero and stabilize the variance. Also, derived features like traffic rate (Traffic Rate Bps) are calculated to promote the interpretability and predictability of the model.

2. **Base Model Layer**

   The preadjusted traits are inputted into two complementary basic models:

- Gradient Boosting Regressor (GBR)
- XGBoost Regressor

3. **Ensemble (Stacked) Layer**

   In order to improve the accuracy and stability of prediction, the meta-learner is a Stacked XGBoost model. This layer receives the predictions of the Gradient Boosting and

XGBoost base models as its input features and it learns how to combine these features in the best manner.

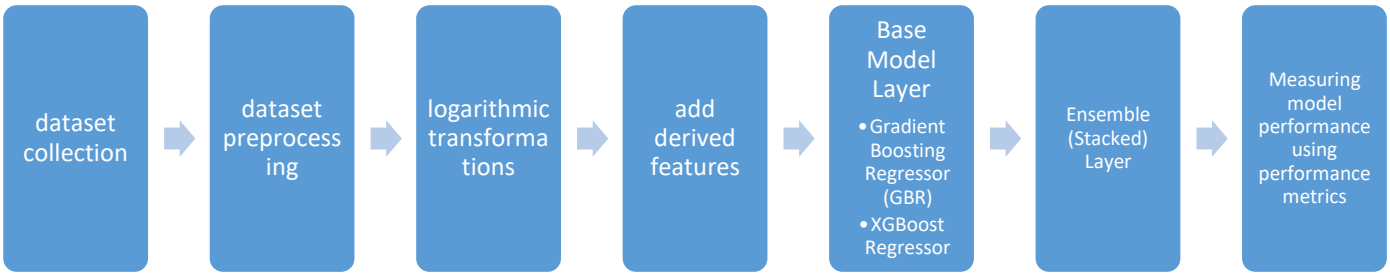

Figure 1 proposed system

System Benefits:

- High predictive accuracy, where the ensemble has R = 0.987, that is virtually all of the variance of network data consumption is predicted.
- Scalability and flexibilities, which is the ability of the system to be implemented in many network environments, such as local and IoT networks.

4. Dataset Description

The dataset employed in the study is a processed form of it and is geared towards predicting network traffic and analyzing data consumption [8].

4.1 Overview of the Dataset

Table 1 shows dataset features.

Table 1 dataset features

Feature Name	Description	Type	Example Value
Traffic Volume (Bytes)	Total number of bytes transmitted during the flow. This feature represents the total data consumption and serves as the target variable for prediction.	Continuous	3,000,000.0
Packets per Second (PPS)	The rate at which packets are transmitted per second during the flow. Indicates flow intensity.	Continuous	674.0
Packet Size (Bytes)	The average size of packets transmitted during the flow.	Continuous	252.75
Flow Duration (Seconds)	The total duration of the network flow.	Continuous	1.0
Bandwidth Utilization (%)	The percentage of the utilized bandwidth relative to the maximum link capacity (1 Gbps in this case).	Continuous	5.392

5. Data Preprocessing

In the first step, the data set was read as indicated figure 2, Reading and validating the dataset at this stage is done to verify the integrity of the data as well as to ensure that the variables are identified in a right way before any preprocessing or feature engineering takes place.

	Traffic Volume (Bytes)	Packets per Second (PPS)	Packet Size	Flow Duration	Bandwidth Utilization
0	110546.0	1.691453e+03	1417.333333	45523	0.019427
1	12.0	2.000000e+06	6.000000	1	0.096000
2	674.0	3.000000e+06	252.750000	1	5.392000
3	0.0	1.843318e+04	0.000000	217	0.000000
4	1076.0	6.404673e+01	267.500000	78068	0.000110

Figure 2 Snapshot of the dataset

Thereupon, the statistical analysis of the numerical characteristics of the dataset was conducted in detail as shown in the following figure 3 This descriptive study also involved computing of

the principal statistical characteristics of measurable features including mean, median, standard deviation, minimum, and maximum values. This move was to acquire a baseline of the distribution, scale, and variability of the network parameters including the rate at which the packet was passed, the period of a flow, and the amount of bandwidth used before any transformation or normalization was done.

	count	mean	std	min	25%	50%	75%	max
Traffic_Volume	3577296.0	1.312907e+05	2.796210e+06	0.000000	1.200000e+01	968.000000	6.441000e+03	1.346164e+09
Packets_per_Second	3577296.0	8.896338e+04	4.027620e+05	0.016667	1.128096e+00	33.937525	4.214963e+03	6.000000e+06
Packet_Size	3577296.0	1.988191e+02	3.327427e+02	0.000000	6.000000e+00	62.833333	2.500000e+02	1.070867e+04
Flow_Duration	3577296.0	2.544247e+07	4.014430e+07	1.000000	6.280000e+02	584729.500000	4.500153e+07	1.200000e+08
Bandwidth_Utilization	3577296.0	3.238967e-02	6.040832e-01	0.000000	1.505943e-07	0.000009	1.875000e-04	1.151680e+02

Figure 3 dataset description

The data set was then checked to be free of empty values as shown in table 2.

Table 2 data set with on missing value

Feature	Missing Value
Traffic_Volume	**0**
Packets_per_Second	**0**
Packet_Size	**0**
Flow_Duration	**0**
Bandwidth_Utilization	**0**

figure 4 presents the distribution analysis of Traffic Volume feature in the data set. This visualization will give one information about the way in which data is consumed within various network flows. This is because we can observe the pattern of distribution and therefore identify skewness or heavy tailedness which is normally common in network traffic data because of skewness or bursty transmission patterns.

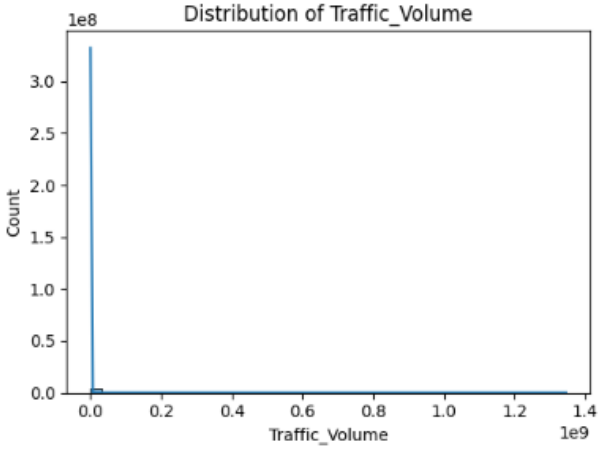

Figure 4 traffic volume distribution

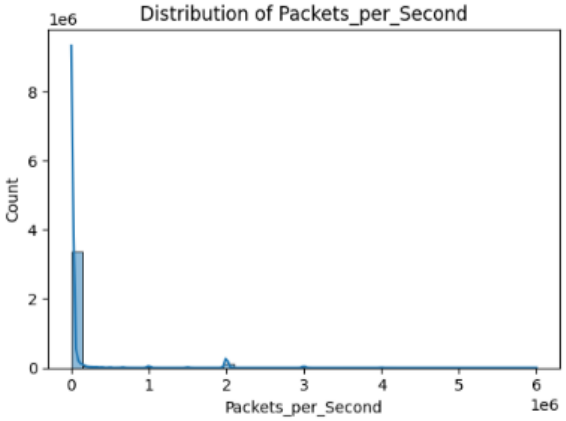

figure ٥ shows a Packet per second distribution.

Figure 5 Packet per second distribution

Figure 6 shows Packet size distribution.

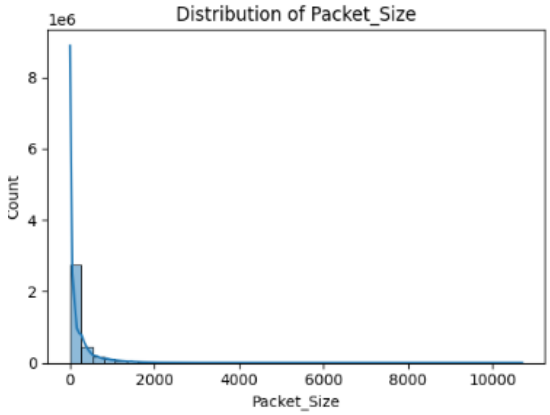

Figure 6 Packet size distribution

- **Anomaly detection**

To study the anomaly in the data set, box plots were used as shown in figure 9, Outliers appear as values outside the range.

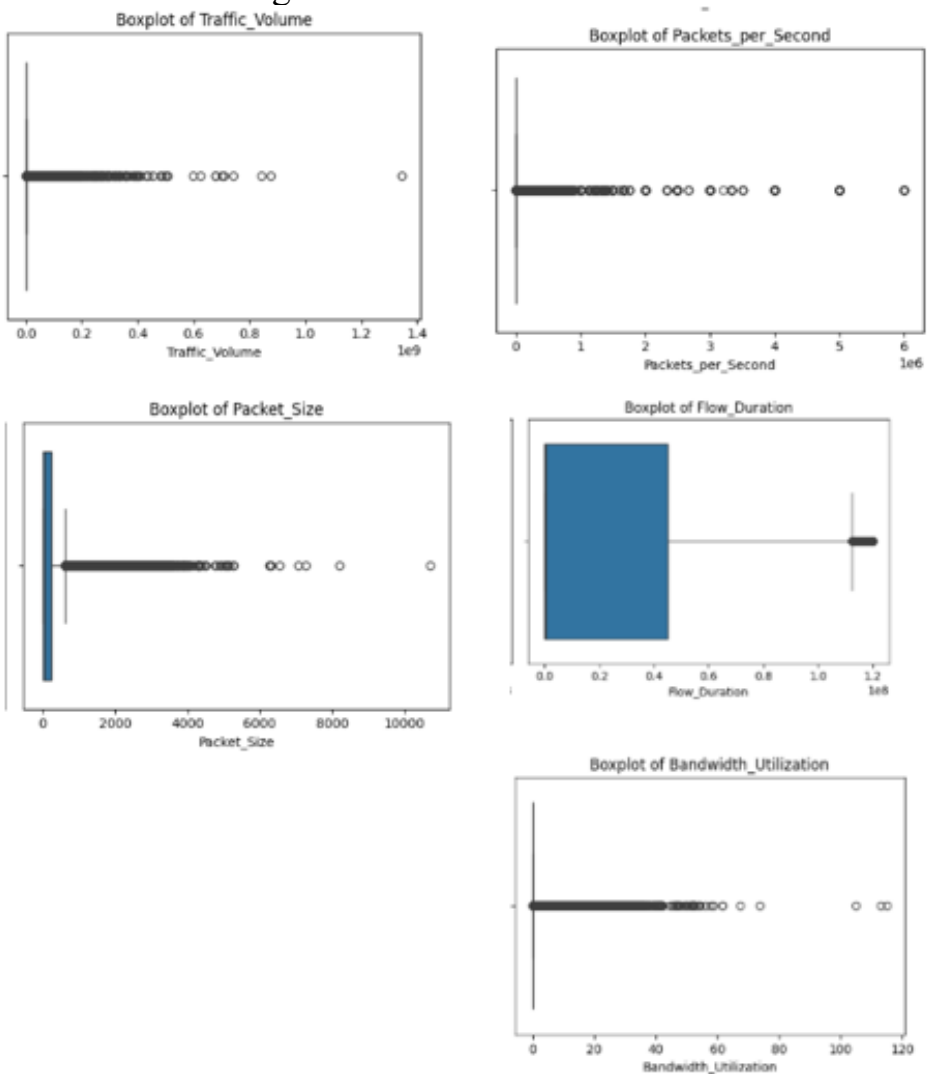

Figure 7 boxplot for anomaly detection

The analysis of the correlation between the numerical features was then provided to determine the possible dependencies or linear relationships between the network parameters. The relationship between the features as shown in figure 10 was mostly poor which implied that the variables showed little interdependence between the packet size, flow duration, and bandwidth utilization. This low correlation implies that both features have special information to add to the predictive model.

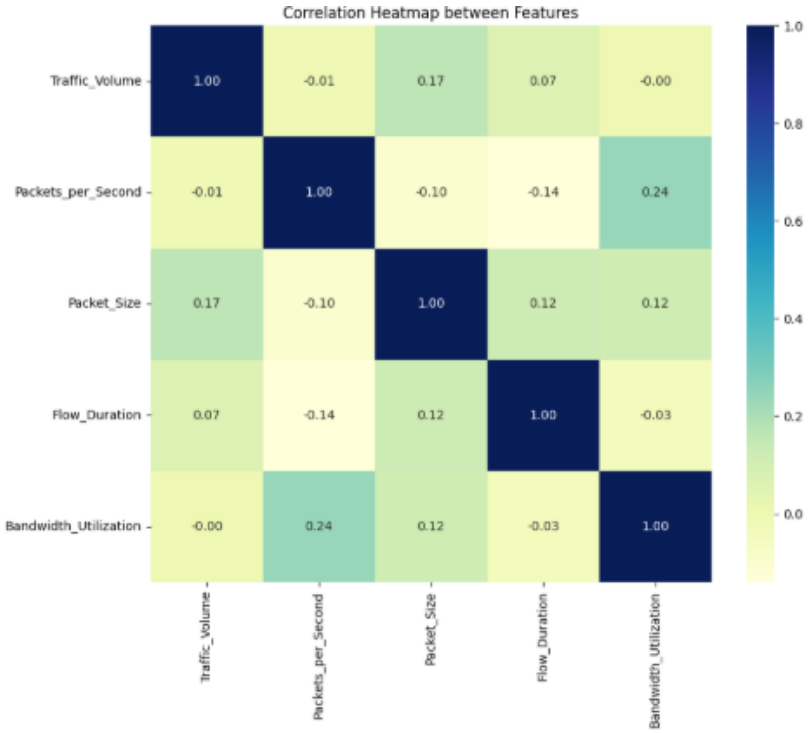

Figure 8 Correlation Heatmap between Features

A complete data preprocessing pipeline was applied before training the predictive models to consume network data to improve the quality of the data and the performance of the models. The key operations that are preprocessed include the following:

- **Data Cleaning**
  The cases of missing values were treated by the use of the median of the values in each column to reduce the impact of outliers. Also, a new option (Flow Duration safe) was developed to substitute the zero figures in (Flow Duration) with the median of the column to avoid possible division errors in further calculations..
- **Feature Engineering**
  A new feature, **Traffic Rate (Bytes per Second)**, was derived to represent the efficiency of bandwidth utilization:
  Traffic_Rate_Bps = Traffic_Volume/ Flow_Duration_safe
  This variable captures how effectively data volume is transmitted over a given duration, providing the model with a more descriptive view of traffic behavior.
- **Logarithmic Transformation**
  Since the dataset exhibited high variability and a non-uniform distribution of feature values, a **natural logarithmic transformation** was applied using: log_x = log(1 + x)The transformation was applied to the following features: Traffic_Volume, Packets_per_Second, Packet_Size, Flow_Duration_safe, Bandwidth_Utilization, and

Traffic_Rate_Bps.

This process reduced skewness and brought the data closer to a normal distribution, which is crucial for the stability and accuracy of regression models such as Gradient Boosting and XGBoost.

- **Feature and Target Selection**
  The following logarithmically transformed variables were used as input features: log_Packets_per_Second, log_Packet_Size, log_Flow_Duration_safe, log_Bandwidth_Utilization, log_Traffic_Rate_Bps, The target variable used for prediction was log_Traffic_Volume.
- **Data Splitting**
  The dataset was divided into 80% for training and 20% for testing using train_test_split with random_state=42 to ensure reproducibility of the results.
- **Feature Standardization**
  The StandardScaler was used in the training pipeline (Pipeline) to normalize all features in such a way that the mean of all features would be 0 and the standard deviation of all features would be 1.

6. Results and discussion

Three popular regression measures were used to assess the predictive ability of the models, which are the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), and the Coefficient of Determination (R squared).

1. **Mean Absolute Error (MAE) [5]:** The measurement of the average size of errors in forecasting traffic values of predicted values and real values, ignoring their direction.
2. **Root Mean Squared Error (RMSE)[6]:** Is a calculation that results in the square root of the mean squared errors of the predicted and actual outcomes.
3. **Coefficient of Determination (R squared)[7]**: This is simply the percentage of the variance in the dependent variable that is accounted by the model

Table 4 shows the results obtained in this research.

Table 3 results obtained in this research

Model	MAE	RMSE	R^2	Analysis & Interpretation
Gradient Boosting Regressor	14,663.072	703,137.604	0.934	Very good performance, capturing the main patterns in the data. However, larger errors still occur in high-volume flows, reflected in the relatively high RMSE.
XGBoost Regressor	171.855	436.066	0.984	Excellent performance: highly accurate predictions across nearly all cases, significantly reducing large errors and achieving a very high R^2. Suitable for practical network applications.
Ensemble (Stacked XGBoost)	158.399	394.335	0.987	The best-performing model. The smart combination of Gradient Boosting and base XGBoost predictions results in more stable and accurate forecasts, leveraging the strengths of both base models while minimizing extreme errors.

Figure 12 shows MAE Result of all Models, it is clear that Ensemble (Stacked XGBoost) is the best.

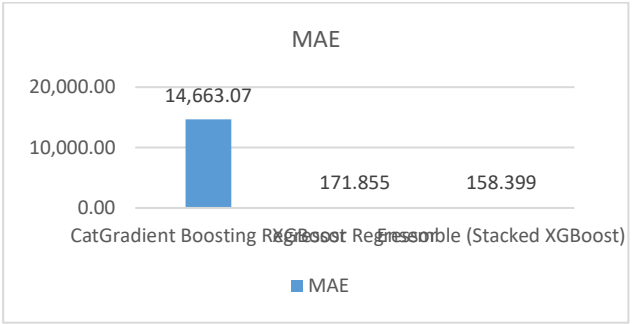

Figure 9 MAE Result of all Models

Figure 13 shows RMSE Result of all Models, it is clear that Ensemble (Stacked XGBoost) is the best.

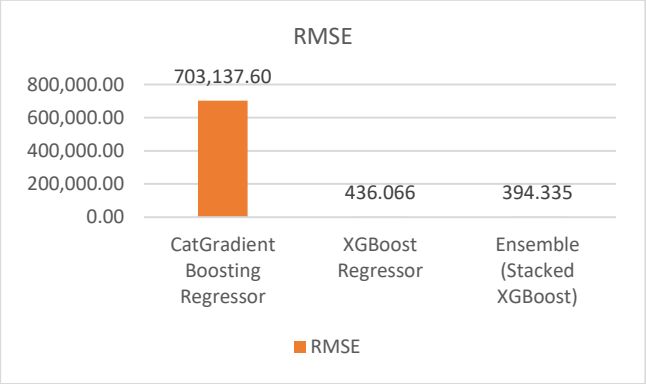

Figure 10 RMSE Result of all Models

Figure 13 shows R squared Result of all Models, it is clear that Ensemble (Stacked XGBoost) is the best.

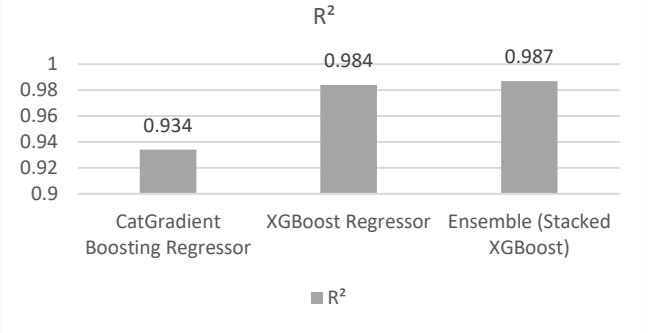

Figure 11 R squared Result of all Models

Figure 15 shows True vs. Predicted graph illustrates the performance of a Gradient Boosting. Each point represents a sample of data, and if the points are close to the diagonal line (y = x), this indicates high prediction accuracy.

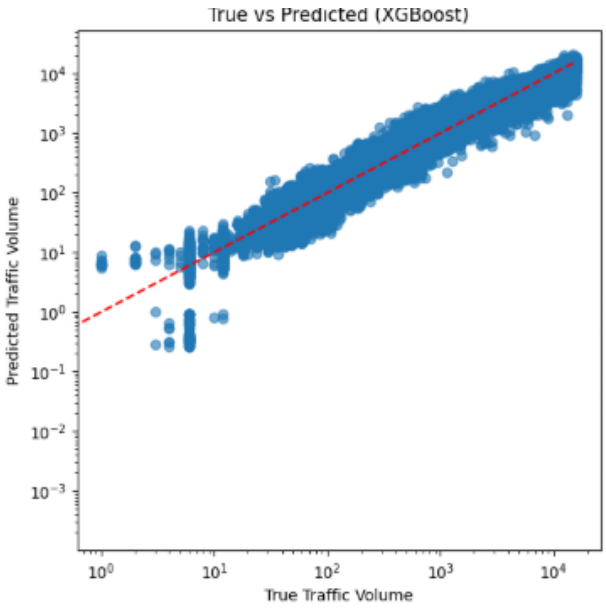

Figure 12 True vs. Predicted graph - Gradient Boosting

Figure 16 shows True vs. Predicted graph.

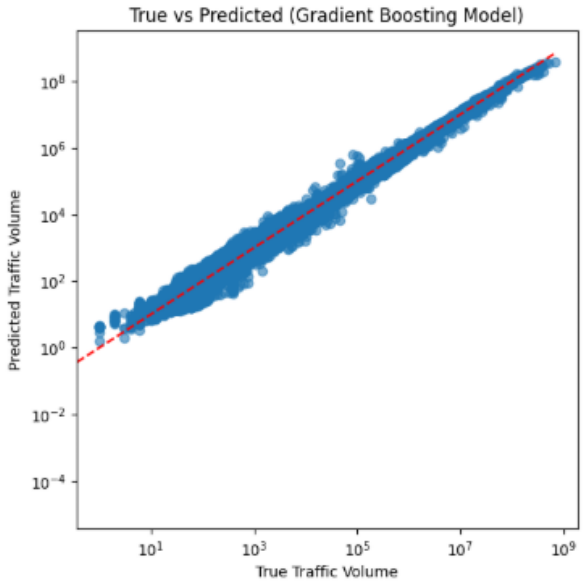

Figure 13 True vs. Predicted graph - XGBoost

Analysis & Observations

MAE (Mean Absolute Error): XGBoost and Stacked XGBoost advanced models have very much smaller MAE meaning the predictions are highly accurate in most cases.

1. **RMSE (Root Mean Squared Error):** XGBoost and Stacked XGBoost have a lower RMSE which indicates that these are not only models that do not make drastic errors but also give consistent forecasts even during peak network traffic value.
2. **(Coefficient of Determination):** Stacked XGBoost XGBoost explains 98.7% of the variance, the greatest among the models, which is followed by Gradient Boosting (0.934) and then, the XGBoost (0.984).
3. **Overall Conclusion:** The best model that fits the prediction of network data consumption is the stacked XGBoost,

7. Conclusion

This paper presents a comprehensive AI-based prediction architecture using network data and advanced set learning techniques. The data was pre-processed with rigorous data preprocessing, feature engineering techniques, and logarithmic transformation to ensure optimal predictive performance. Two core models, the step-up boost slope and the XGBoost slope, were trained to identify underlying patterns in network traffic, and their predictions were augmented using a stacked XGBoost, a multi-powered slope. As the evaluation results demonstrate, the proposed system is highly accurate and robust.

٨. Future Work

Based on the encouraging findings of the present investigation, a number of directions of future work are implied:

1. **Incorporation of Temporal Features:**
   - Introduce time-series information such as hourly, daily, or weekly traffic patterns to improve predictions for dynamic network conditions.
2. **Expansion to Real-Time Streaming Data:**
   - Modify the system to accept streaming network traffic real-time and manage it with proactive control and detect anomalies.

References

[1] Muhammad, Syed & Bukhari, Syed Muhammad Shakir & Khan, Taimoor & Siddiqui, Muhammad Ahmad & Mustafa, Umer & Batool, Waseema & Lughmani, Ibrahim & Development, Khairpur & Mirs, Pakistan. (2024). NETWORK TRAFFIC PREDICTION: USING AI TO PREDICT AND MANAGE TRAFFIC IN HIGH-DEMAND IT NETWORKS. 2. 2024.

[2] Transformer-Based Wireless Traffic Prediction and Network Optimization in O-RAN, Md Arafat Habib and Pedro Enrique Iturria-Rivera and Yigit Ozcan and Medhat Elsayed and Majid Bavand and Raimundus Gaigalas and Melike Erol-Kantarci,arXiv, 2024

[3] Al-Salim, Ali M., et al. "Energy efficient big data networks: Impact of volume and variety." IEEE Transactions on Network and Service Management 15.1 (2017): 458-474.

[4] Alkanhel, Reem, et al. "Metaheuristic optimization of time series models for predicting networks traffic." CMC-Computers Materials & Continua 75.1 (2023): 427-442.

[5] Error, Mean Absolute. "Mean absolute error." Retrieved September 19.2016 (2016): 14.

[6] Hodson, Timothy O. "Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not." Geoscientific Model Development Discussions 2022 (2022): 1-10.

[7] Di Bucchianico, Alessandro. "Coefficient of determination (R 2)." Encyclopedia of statistics in quality and reliability (2008).

[8] https://www.kaggle.com/datasets/noobbcoder2/preprocessed-dataset-for-network-traffic-analysis

[9] Zhani, Mohamed Faten, Halima Elbiaze, and Farouk Kamoun. "Analysis and Prediction of Real Network Traffic." J. Networks 4.9 (2009): 855-865.

[10] Mohammed, AysŞe Rumeysa, Shady A. Mohammed, and Shervin Shirmohammadi. "Machine learning and deep learning based traffic classification and prediction in software defined networking." 2019 IEEE International Symposium on Measurements & Networking (M&N). IEEE, 2019.