iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50712

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Website Traffic Forecasting Using Deep Learning

D. Leela Dharani[1], B.Hema Bindu[2], D.Chaitanya[3], G.Yamuna[4], M. Reshwanth[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*B.Tech Students, Department of Information Technology Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India*

Abstract: ***Nowadays, there is an increasing emphasis on how to predict traffic on web pages, and there is a need to explore different methods for effectively predicting future values of multiple time series. Evaluating website traffic on a web server is crucial for web service providers because, without proper demand forecasting, customers might face long waiting times and abandon the website. However, this is a difficult task because it requires reliable predictions based on the arbitrariness of human behavior. The most effective way of transmitting information would be to predict network traffic and display it visually. Nowadays we depend too much on Google's servers, but if we wanted to host a server for many people, we might have predicted in advance the number of users to prevent server failure. Time series prediction is important in many different areas. Although there are already many systems and models for predicting Internet traffic flow, most of them use shallow traffic models and are still somewhat unsatisfactory. Therefore, we will use deep learning techniques based on current and past data to predict future traffic.***

Keywords: website traffic, servers, time series forecasting.

I. INTRODUCTION

Website traffic analysis and prediction have many applications in various fields and have been the subject of numerous studies. It is a proactive approach that helps ensure secure, reliable, and high-quality web communications. Web traffic is the amount of data sent and received by visitors to a website, determined by the number of visitors and the number of pages they visit. Website owners often use web traffic tools to monitor incoming and outgoing traffic, for example, to determine the popularity of web pages based on page views; information collected about website traffic can help structure website content and address security issues such as denial-of-service (Dos) attacks. Web traffic is measured in visits, sometimes called "sessions," and is a common way to measure the effectiveness of an online business in attracting an audience. Web traffic measurement is used to check the popularity of websites or specific web pages on that website. Each time a user visits the website, a page visit is generated. Any user who is connected to the Internet will always visit at least one page and could visit many more if they stay connected. Monitoring web traffic requires information such as the total number of visitors, average page views per visitor, most popular pages, average visits by visitors, and duration of page views, etc. , which are commonly used to predict web traffic. We will take an in-depth look at the web traffic dataset and how we can use LSTM to solve this time series prediction problem.

II. LITERATURE SURVEY

In Paper [1], Web Traffic Time Series Forecasting using ARIMA and LSTM RNN is described and published in International Conference on Advances in Computing and Communications (ICACC) 2020. They use Auto-Regressive Integrated Moving Average (ARIMA) and Long Short-Term Memory Recurrent Neural Network (LSTM RNN). LSTM RNN brings more efficiency to the system. It effectively captures seasonal patterns and long-term trends. Information about holidays, days of the week, language and region can also help our model to capture the highs and lows. The forecast results of LSTM are more accurate compared to the forecast of the ARIMA model. The study in [2] encountered so many time series forecasting models that our work was both tedious and entertaining. In the paper, we developed a time series prediction technique to predict Internet traffic using historical data. Many prediction techniques, such as ARIMA, are widely used in the literature but are best suited for time series that has a linear shape. Neural networks such as RNN, on the other hand, are particularly suitable for predicting nonlinear time series. In the proposed method, the discrete wavelet transform and a high-pass and low-pass philter are used to generate linear and nonlinear components for the time series. ARIMA and RNN are significantly outperformed by the proposed approach [3]. The technology is very easy to use in data centers due to its simplicity. In the research work [4], a novel engineering approach has been proposed to predict the exit-link traffic pattern in on-campus networks. It is also predicted that if enough historical data is available, EPTS will have the following effect in predicting network traffic. Web traffic time series prediction 1) Using past network traffic data, exit link traffic trends can be estimated so that network resource planning can be done in advance.

2) It is easy to implement and has manageable computational complexity. The effects of the LSTM network, BPNN model, and ARIMA model on time series recorded at a single point are compared in the paper [5]. Under typical conditions, the proposed LSTM network can accurately predict traffic flow based on a relatively constant time series. In contrast, the traffic system on roadways is stochastic and complex and is often affected by unusual circumstances such as severe weather, traffic accidents, and major events.

III. PROPOSED SYSTEM

We describe the proposed architecture for website traffic forecasting. This architecture is modular, distributed, and scalable so that it can be easily adapted with minimal changes and used for website traffic prediction, whether it is a closed computer network or a website. Most people have experienced a website crashing or loading very slowly when it is used by many people, e.g., when various shopping websites crash just before the holidays because more people try to log in to the website than was originally possible. So, users give the website a lower rating and use another website instead, which negatively affects their business. Therefore, a traffic management strategy or plan should be created to reduce the risk of such mishaps that could jeopardize the company's existence. Until recently, such tools were not necessary as most servers could handle the increase in traffic. However, in the age of smartphones, the demand for some websites has increased so much that companies have struggled to keep up with the variable level of customer support.

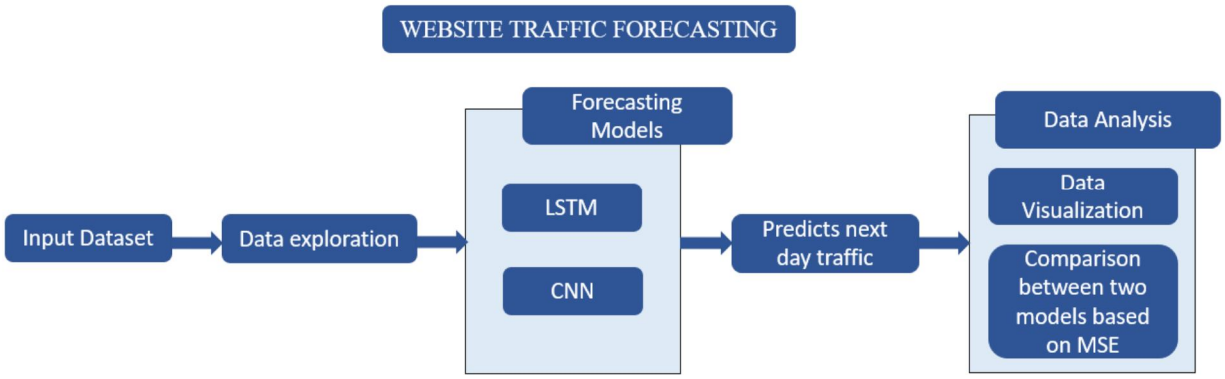


Fig:1 Architecture Diagram for Website Traffic Forecasting

The flow of the project is explained in detail in the architecture diagram. We start with a dataset containing the features hour index and sessions. To identify trends, we examine the data. Create training and validation sets from the dataset. Two models, LSTM and CNN, were created, using the training data to feed the models and the validation data to test the models. To compare the performances of the different models, determine the mean square error. Use line graphs with the hour index on the X-axis and the sessions on the Y-axis to compare the performance of the two models.

IV. IMPLEMENTATION

In the first step is to Load Dataset for Web Traffic Forecasting and that dataset is a Wikipedia data Here we are reading the dataset by using pandas. It has over 4800 observations. Check the shape of the dataset to understand the features and number of hours. The first column is the hours as in this is the first hour, this is the second hour, and so on. And the second column session is the volume of traffic at an hourly level. For example, this is the number of sessions in the second hour and so on. 4.1 Data Exploration for Web Traffic Forecasting

Examine the data and plot the entire time series. At each point on this curve, which represents an early session count, there are some recurring patterns in the time series.

After almost equal time intervals, the traffic volume decreases. That being said, there are some traffic peaks in this graph as well. Let us examine this data in more detail. Instead of using the whole time series, we can use a part of it. We can see that the repeated pattern is more apparent now that we have only shown the first week's data.

These dips in the online traffic graph can occur once every 24 hours. So it is obvious that there are two times in the day when traffic is high, such as occasionally, and when it is low.

A. Explore The Data And Plot The Entire Time Series

we can observe that at each point of this curve, which represents an early session count, there are some recurring patterns in the time series. After almost equal time intervals, the traffic volume decreases. Apart from that, there are some traffic peaks in this graph as well.

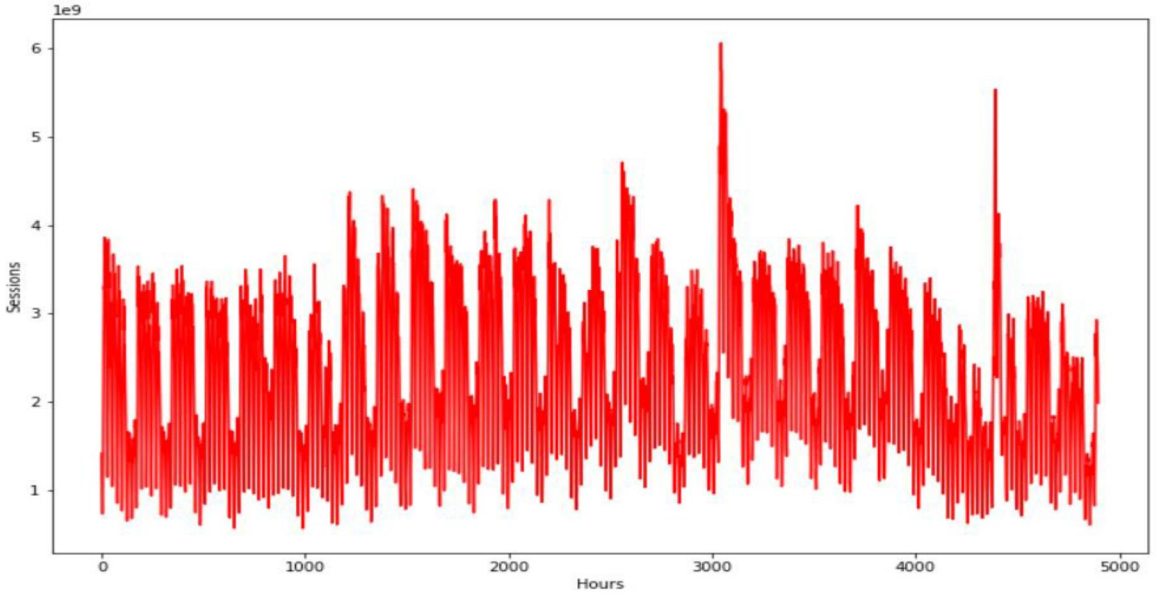


Fig 4.1 Plot of sessions and hours on the whole dataset

1) Explore And Plot The Week's Data

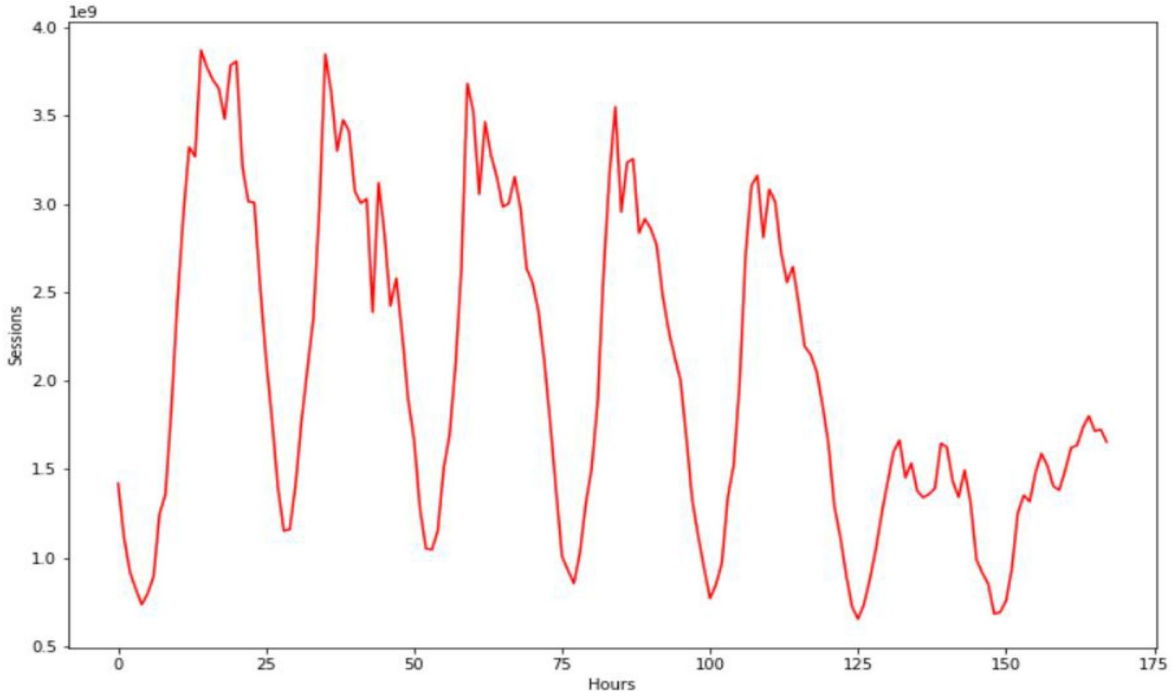


Fig 4.1.2 Plot of sessions and hours of week's data

Fig 4.1.2 says the repeating pattern is more apparent now that we have shown only the first week's data. These dips in the online traffic graph can occur once every 24 hours. So it is obvious that there are two times in the day when traffic is high, such as occasionally, and when it is low.

B. Data Preparation for Web Traffic Forecasting

We will create input sequences from the block traffic data to prepare the data for model training. This is a time series. Each cell would contain a number or value. We will create sequences of five observations each, with the first five observations being the initial sequence and the sixth observation being the target. The second sequence starts with the second element and runs through the sixth and seventh elements. The 168-hour sequence we describe corresponds to the time span of one week. Consequently, we construct one-week sequences as our entry sequences. There are now considerably more than 4700 sequences.

C. Split the Dataset

The data must then be split into a training set and a validation set, with a ratio of 90 to 10. Since there is a major timing problem, the data are split sequentially rather than randomly.

D. LSTM Model Building for Web Traffic Forecasting

We will now train a deep-learning LSTM model to predict future website traffic based on these sequences. Before making predictions, we load the weights of the best model. To shop the weights of the best model, we will again use the model checkpoint and mean square error method. Finally, the model training phase begins. In the evaluation phase, we see that the root mean square error for the validation data is only 0.014. Analyze the performance of the model using the validation data.

```
Mean Square Error: 0.014353805221617222
```

Fig 4.4.1 Mean Squared Error of Long Short-Term Memory Model

This red curve is the actual value and this yellow curve is the predicted value both are pretty much close to each other.
The below fig 4.4.2 describes the prediction of 24 hours data using LSTM in which the X-axis is the Hour index and Y-axis is the sessions

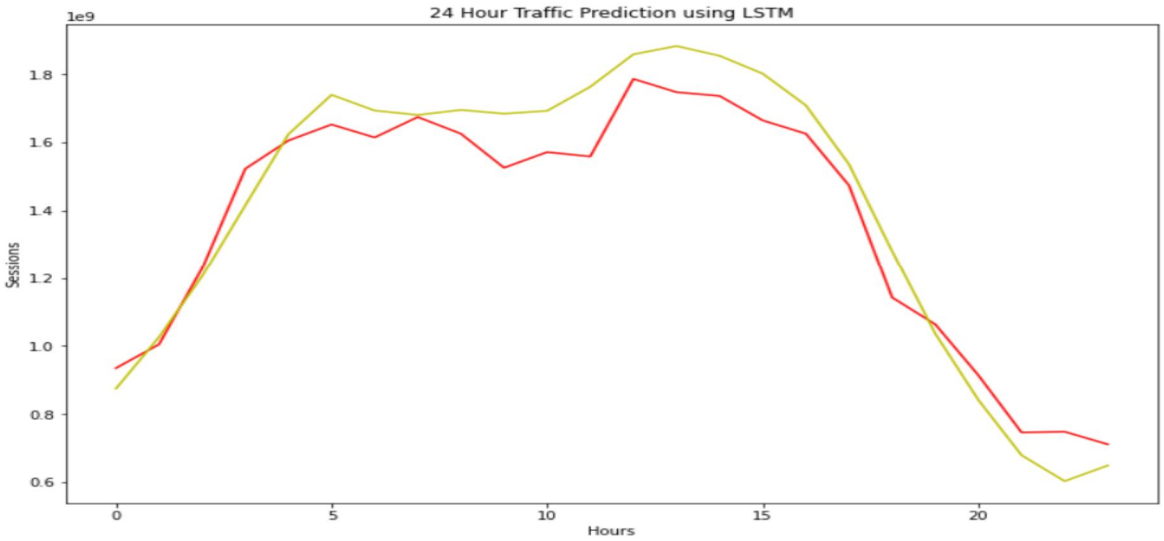


Fig 4.4.2 Traffic Prediction Using LSTM for 24 hours

E. CNN Model with Forecasting

Conv1D layers are used in the model architecture of this instance. A layer that flattens the surface follows these layers. The input is converted by this layer into a one-dimensional array, which is then passed to this collection of dense layers. We see that the root mean square error for the validation data is only 0.016.

```
Mean Square Error: 0.016352154314517975
```

Fig 4.5.1 Mean Squared Error of Convolutional Neural Network Model

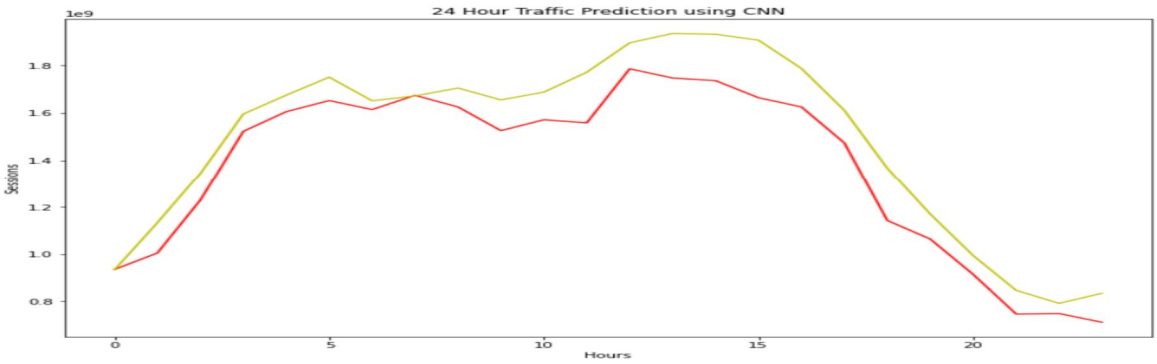


Fig 4.5.2 Traffic Prediction Using CNN for 24 hours

The above fig 4.6 describes the prediction of 24 hours data using CNN in which the X-axis is the Hour index and Y-axis is the sessions.

F. Comparison with the LSTM and CNN Model

Our observation shows that the predicted line at LSTM and CNN almost collides with the actual line. The predicted values are almost close to the actual values. The performance of LSTM seems to be very similar to that of CNN in the underlying model. The mean square error of LSTM is 0.14 and of CNN is 0.16. Comparing both errors, we can say that the LSTM performs better. In conclusion, we can say that the results of the LSTM-based model are more meaningful compared to CNN.

V. SCOPE OF FUTURE USE

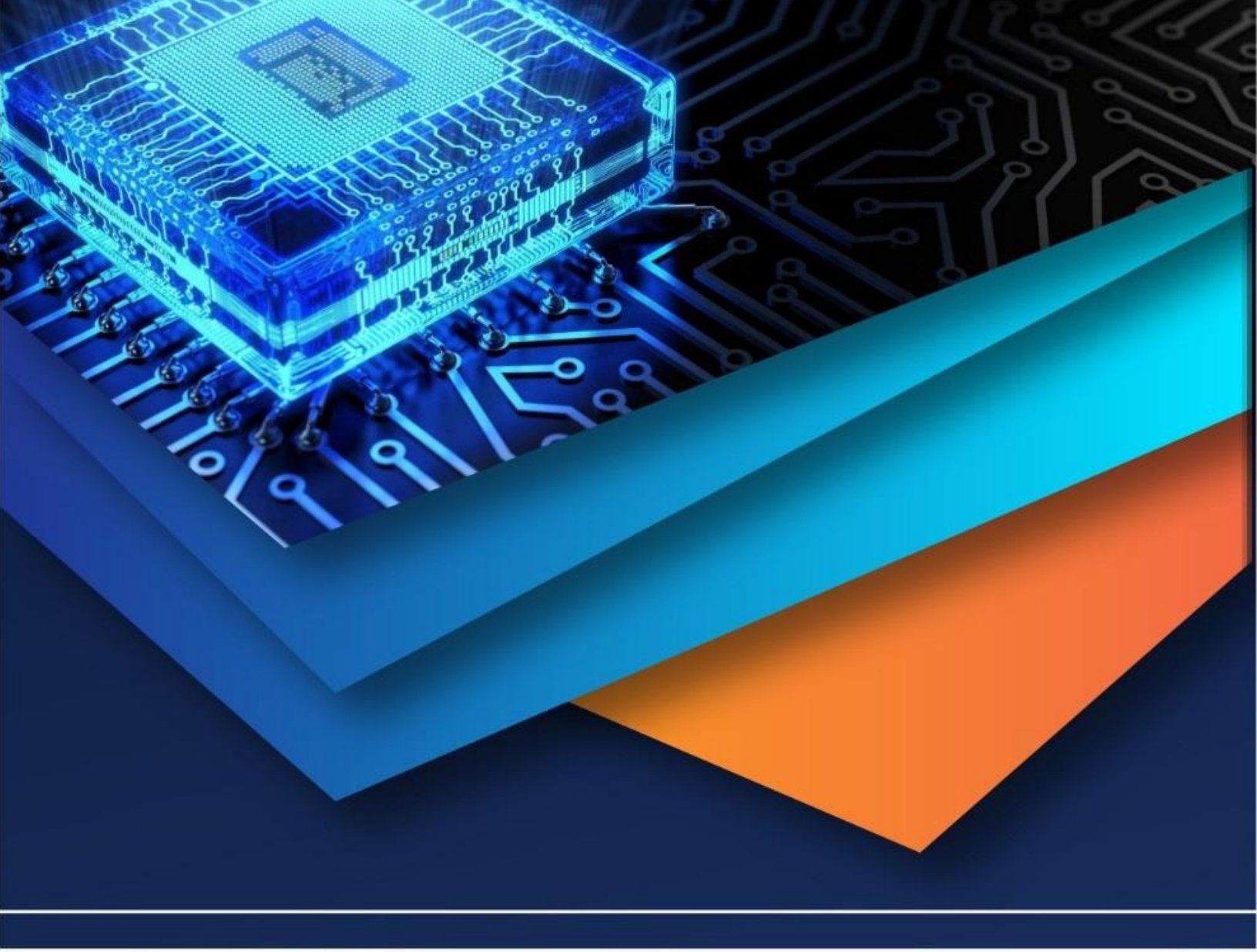
Time series forecasting is one of the least researched areas, and several models are evaluated to improve forecast accuracy. The proposal focuses on forecasting future Internet traffic to make decisions for better congestion control. Values from the past will be used to predict future values. We will also try to explore multivariate time series and propose to simplify the decision making process in real-time.In the future, we want to improve our ability to detect hidden trends so that we can study how human behavior affects online traffic. We'll look at the unsupervised model that has been used in other work to improve our model.

VI. CONCLUSION

The main goal of our research is to develop a consistent forecasting model for predicting future web page traffic. Time series prediction of web traffic can be performed more efficiently and accurately using recurrent neural networks with long-term memory and CNN. We trained the model with this data using features such as the hours and number of visits, i.e., sessions for pages for one year to predict future web traffic. It is possible to predict the number of users that will access the website in the future. The proposed system will continue to improve as more user data is fed in. Our system can be used on all websites to improve Internet load management and business analysis. LSTM RNN brings more efficiency to our system. Moreover, despite the limited amount of data we had available, we achieved quite good results in training the LSTM. In future work, we plan to deepen the extraction of hidden patterns to improve the efficiency of the LSTM and to study how human behavior affects web traffic. To improve the performance of our model, we will investigate the unsupervised model proposed in previous work

REFERENCES

[1] Navyasree Petluri and Eyhab Al-Masri, "Wikipedia Page Traffic Prediction," 2018 IEEE International Conference on Big Data (Big Data).
[2] Mohammad Asifur Rahman Shuvo, Muhtadi Zubair, Afsara Tahsin Purnota, Sarowar Hossain, and Muhammad Iqbal Hossain, "Traffic Forecasting Using Time-Series Analysis," 6th International Conference on Inventive Computation Technologies, 2021. (ICICT).
[3] Partha Sarathi Mangipudi and Rishabh Madan, "Predicting Computer Network Traffic: A Time Series Forecasting Approach Using DWT, ARIMA, and RNN," 2018 Eleventh International Conference on Contemporary Computing (IC3).
[4] Jianhu Zheng and Mingfang Huang, "Traffic Flow Forecasting Using Deep Learning and Time Series Analysis," IEEE Access, 2020. P Montero-Manso.
[5] Montero-Manso, P.; Athanasopoulos, G.; Hyndman, R.J.; Talagala, T.S. Fforma: Featurebased forecast model averaging. Int. J. Forecast. 2020,36, 86–92.
[6] Boone, T.; Ganeshan, R.; Jain, A.; Sanders, N.R. Forecasting sales in the supply chain: Consumer analytics in the big data era. Int. J. Forecast. 2019,35,170–1801



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY