SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I11P117

Volume 11 Issue 11, 196-212, November 2024

Original Article

Efficient Traffic Routing with Temporal Fusion Transformers: Addressing Urban Congestion Challenges

Sreelekha M[1], Midhunchakkaravarthy Janarthanan[2]

[1]*Faculty of Engineering, Lincoln University College, Malaysia*
[2]*Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia.*

[1]*Corresponding Author : sreelekha.edu@gmail.com*

Received: 19 September 2024 Revised: 18 October 2024 Accepted: 16 November 2024 Published: 03 December 2024

Abstract *- The exponential increase in urban population necessitates the emergence of transportation systems that are both effective and sustainable, using the potential modern technology. The issue of dynamic traffic flow significantly impedes the movement of vehicles. Traffic congestion is a critical issue affecting urban mobility and efficiency in cities worldwide, with Bangalore no exception. This study addresses the challenge of leveraging advanced predictive analytics and intelligent transport systems to manage traffic congestion. The proposed research aims to address the limitations of traditional traffic management strategies by integrating the Temporal Fusion Transformer (TFT) model into an Intelligent Transport System (ITS) framework. The research employs rigorous data preprocessing techniques to leverage extensive data from multiple online map service providers and traffic monitoring platforms, spanning from January 1, 2019, to December 31, 2023. The TFT model forecasts traffic congestion with notable precision, achieving a Mean Absolute Error (MAE) of 0.39, Mean Squared Error (MSE) of 0.30, Root Mean Squared Error (RMSE) of 0.55, Mean Absolute Percentage Error (MAPE) of 7.2%, and an R-squared (R^2) value of 0.87. The outcomes obtained clearly illustrate the model's superior accuracy and efficacy. Integrating TFT predictions into the ITS framework enhances real-time traffic control by improving the timings of traffic signals, recommending alternative routes, and improving incident management. This proactive approach significantly reduces traffic congestion and enhances travel efficiency, substantially advancing urban traffic management solutions.*

Keywords *- Temporal Fusion Transformer, Intelligent Transport System, Traffic congestion, Traffic volume, Smart city.*

1. Introduction

In order to create "smarter" cities and improve people's quality of life, the emergence of smart cities is combined with a significant transformation in urban design and the adoption of innovative technologies. The European Commission introduced pioneering and noticeable innovation in the field of smart cities, focusing on four crucial aspects: buildings, electricity, cooling and heating facilities, and transportation [1]. An intelligent transportation technology can improve traffic flow in smart cities by analyzing traffic patterns and regulating traffic signal timing.

The aim is to identify and promote sustainable transportation methods to improve Intelligent Transportation Systems that utilize up-to-date information. This system includes Traffic Management Systems (TMSs) to prevent traffic congestion and ensure safety, as well as green applications that strive to reduce gas, fuel, and electricity usage [2]. ITS utilizes innovative and developing methods to enhance the comfort and affordability of transportation in an intelligent urban environment, as depicted in Figure 1.

Recently, a key issue in transportation systems has been the problem of traffic congestion. This matter must be resolved to reduce fuel consumption, prevent accidents, alleviate traffic jams, and decrease driver dissatisfaction [3]. The substantial volume of automobiles is the primary source of traffic congestion in urban areas. The implementation of traffic regulations has become essential in urban areas owing to the limited availability of land assets and the overcrowded infrastructure for transportation. Due to the excessive number of people, urban areas are experiencing various traffic-related issues that hinder the movement of individuals between different locations [4].

Sustaining economic expansion and enhancing the convenience of road users are two essential requirements for the progress of the nation, which cannot be achieved without the smooth movement of traffic. With the advancement of the transportation industry, authorities are increasingly prioritizing monitoring traffic volume on traffic information systems. Traffic forecasts provide authorities with the opportunity to strategically allocate resources in order to maximize the efficiency of travel. Congestion imposes

limitations on the utility of street transport systems. These reductions entail indirect as well as direct expenses for the community. The impact of congestion on the economic system and social structure has been extensively studied.

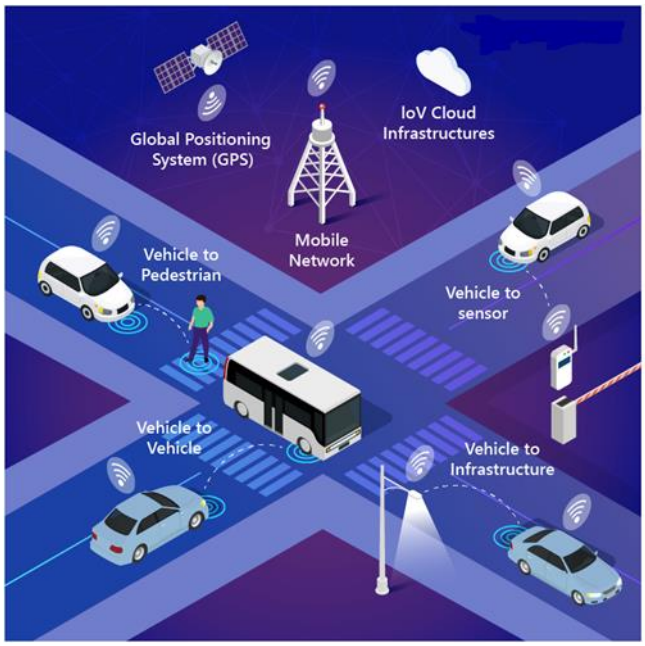


Fig. 1 Intelligent transportation system

Late working hours are an obvious result of traffic congestion. Subsequent calculations revealed that the United States experienced an annual loss of 8.8 billion labor hours attributable to traffic and congestion. The task of traffic prediction involves estimating parameters associated with traffic levels, ranging from 15 minutes to several hours, via the utilization of several artificial intelligence techniques on the gathered traffic data. Five factors are commonly assessed in congestion monitoring and prediction: occupancy, traffic volume, trip time and congestion rate. The overload parameter evaluation depends on the gathered data type and the particular AI methods employed.

Innovative developments in Artificial Intelligence (AI) and Deep Learning (DL) have enabled smart environmental monitoring devices in smart cities. These systems allow for exact monitoring of various aspects that influence the environment, such as pollution levels and traffic congestion. This advanced monitoring enables optimal control and mitigation of environmental adverse impacts. Excessive traffic congestion adversely affects the quality of life for individuals by reducing transportation efficiency and worsening substantial environmental pollution. Therefore, traffic congestion significantly impacts the nation's productivity, economic progress, and human endeavors. The primary concern in urban planning is to find a suitable approach to successfully tackle traffic congestion [5]. Managing traffic congestion is a prominent research subject, with numerous ideas arising from academic efforts undertaken in this field in recent decades [6]. Over time, the collection of traffic data and the development of ITS have advanced in order to address these concerns [7].

The primary aspects of the proposed research are outlined below:

- Employs the TFT model to forecast traffic congestion, leveraging its ability to handle time-series data and generate accurate predictions.
- Integrates TFT predictions into the ITS framework, enabling real-time updates, alternate route recommendations, and enhanced incident management.
- Enhances urban traffic management by embedding predictive insights into the ITS, allowing for proactive traffic flow management, congestion minimization, and improved travel efficiency.
- Assesses the effectiveness of the proposed model using relevant evaluation metrics to ensure robust and reliable results.

The ensuing portions of this paper are organized as follows: Section 2 provides an extensive literature assessment of existing methodologies, highlighting their limitations and identifying gaps that the current research aims to address. Section 3 elaborates on the proposed methodology in detail, outlining the data preparation, model implementation, and integration into the ITS. Section 4 discusses the outcomes obtained from the research, signifying the efficacy of the proposed approach in forecasting traffic congestion and optimizing urban traffic management. Section 5 wraps up the work by summarizing the main findings and contributions while proposing prospective avenues for future research.

2. Related Works

J. Prakash et al. (2024) [8] examined the integration of smart cities with the Internet of Things (IoT) to enhance traffic management. The study specifically addressed the challenges posed by inadequate infrastructure and connection in developing nations. The researchers developed an ITS for vehicle networks based on the Internet of Vehicles (IoV), utilizing various Machine Learning (ML) algorithms. Their strategy utilized ensemble learning and feature selection techniques to enhance detection accuracy, resulting in greater classification accuracy utilizing the stacking method.

Nevertheless, despite its impressive precision, the system encountered challenges such as an uneven distribution of classes, incomplete data, and computational intricacy. The findings indicated that tree-based algorithms with feature selection achieved superior performance compared to conventional ML methods. However, the model's effectiveness depended on the CIC-IDS2017 dataset and its emphasis on certain types of attacks, which could have restricted its applicability to more diverse real-world situations.

An Unmanned Aerial Vehicle (UAV)-guided Emergency Management System (EMS) has been proposed by Abdullahi Chowdhury et al. (2023) [9] to enhance the effectiveness of Emergency Vehicle (EV) routing in densely populated metropolitan regions. The concept included adaptive travel route selection based on real-time traffic data, where drones directed EVs to minimize speed loss at crossings and optimize the timing of traffic signals. The technology was designed to reduce both the instantaneous response time of electric vehicles and the interference caused by non-emergency vehicles. Analysis of simulation data revealed a decrease of 8% in EV response time and an enhancement of 12% in clearance time at crossings. Nevertheless, the research acknowledged several constraints, such as the increased density in neighbouring cells when several signals were activated, potentially resulting in extended clearing durations in situations of severe congestion.

Sura Mahmood Abdullah et al. (2023) [10] employed Gated Recurrent Units (GRUs) to develop a Bidirectional Recurrent Neural Network (BRNN) for predicting traffic congestion in smart cities. The research employed real-time data obtained from sensors and connected devices to categorize traffic into congested and non-congested conditions. The proposed methodology enhanced congestion prediction by integrating supplementary inputs such as road and meteorological conditions. The results demonstrated that the BRNN model surpassed current benchmark approaches regarding precision, MAE, MAPE, and RMSE. However, constraints encompassed a decline in the effectiveness of predictions over extended forecasting durations and difficulties in considering temporal and spatial correlations.

Muhammad Saleem et al. (2022) [11] developed a Fusion-based Intelligent Traffic Congestion Management System for Vehicular Networks (FITCCS-VN) by employing machine learning methods. This technology was designed to tackle the issue of transportation congestion in smart cities. The methodology comprised data collection using IoV-enabled devices, followed by preprocessing and subsequent application of Artificial Neural Networks (ANN) and SVM to predict and manage traffic congestion. The system demonstrated a 95% accuracy and a 5% error rate, surpassing previous methods. Nevertheless, the research encountered constraints, such as the possibility of imprecise predictions of traffic congestion caused by noisy data and the difficulty of extrapolating findings to various metropolitan settings.

Majumdar et al. (2021) [12] investigated the influence of traffic congestion on the sustainability of metropolitan areas, underscoring its role in intensifying air pollution. In order to forecast the propagation of congestion on road networks, the researchers employed Long Short-Term Memory (LSTM) networks and utilized vehicle speed data acquired from IoT traffic sensors. This study carried out a comparison of univariate and multivariate models. The univariate model exclusively considered vehicle speed as the input, whereas the multivariate model additionally incorporated vehicle flow and headway. Each model attained an accuracy ranging from 84% to 95%, contingent upon the route configuration. Significantly, the univariate model performed similarly to the multivariate model, indicating that vehicle speed alone was adequate for rendering precise predictions.

G. Kothai et al. (2021) [13] investigated a hybrid DL model that integrates Convolutional Neural Networks (CNN) and Bayesian Linear Stochastic Mean Ensembles (BLSTME) to tackle the task of predicting traffic congestion in Vehicular Ad-hoc Networks (VANETs). The study's objective was to enhance road safety through improving traffic management. The CNN model utilized spatial characteristics derived from traffic images, while the BLSTME model was employed to train and strengthen weak classifiers to predict traffic behaviour. Model construction and testing were conducted using real-time data obtained from Seoul's artery network. The simulations were performed using Simulation of Urban Mobility (SUMO) and OMNeT++. The suggested BLSTME-CNN method exhibited remarkable results in comparison to existing models, achieving a 10% enhancement above other DL models.

A. Ata et al. (2020) [14] investigated the application of Radio Frequency Identification (RFID) technology for traffic congestion management. They proposed a method that dynamically modifies traffic signal timings in response to the current vehicle density, as measured in real-time. The study utilized RFID readers and tags as sensors to track the car count between two locations on the road. The provided data was inputted into a fuzzy logic system in order to forecast traffic congestion and subsequently modify signal timings accordingly. The simulation results, utilizing MATLAB R2012b, demonstrated that the suggested system efficiently mitigated congestion by adjusting signal durations in accordance with traffic flow. The model showcased its capacity during periods of high demand, providing a substantial enhancement compared to stationary signal systems. Nevertheless, the study's focus was restricted due to its reliance on RFID tags for each vehicle, which could not have been practical in all areas. The system's efficacy also relied on precise sensor positioning and meticulous data processing, which could present issues in practical scenarios.

A. Ata et al. (2019) [15] designed a traffic congestion management system utilizing ANN, specifically the MSR2C-Artificial Back Propagation Neural Networks (ABPNN) model. The neural network was trained using a backpropagation technique to predict areas of traffic congestion. This information was then utilized to dynamically control the flow of traffic. The study utilized a dataset from M1 junction 37 in England, combining meteorological data and traffic speed to forecast congestion. The results showed that the MSR2C-ABPNN system performed better than existing models in terms of MSE and accuracy in training and validation. It achieved regression values higher than 0.90. Nevertheless, the system's efficiency was subject to data latency and interference, potentially compromising its efficacy in real-time situations.

S. Muthuramalingam et al. (2019) [16] utilized IoT and big data analytics to create an ITS specifically designed for smart cities in India. The methodology entailed constructing an ecosystem consisting of sensor systems, monitoring systems, and display facilities. The system employed IoT technology to track vehicles, enable smart parking, and monitor traffic. It developed several analytical methods, including multiple regression analysis, cluster analysis, and logistic regression. Empirical evidence showed that the suggested s-ITS system surpassed current systems regarding data transfer rate, packet delivery efficiency, and network latency. The system effectively optimized vehicle routing to minimize traffic congestion and facilitated intelligent parking by providing real-time updates on the availability status.

Sen Zhang et al. (2019) [17] devised a systematic approach to gather and analyze extensive traffic congestion data by means of image analysis. The researchers generated the Seattle Area Traffic Congestion Status (SATCS) dataset by compiling and modifying snapshots of traffic congestion maps obtained from the Washington State Department of Transportation. The authors introduced the Deep Congestion Prediction Network (DCPN), a network based on deep autoencoders designed to forecast traffic congestion. The results demonstrated that DCPN effectively obtained temporal correlations within the transportation network compared to current benchmark models, resulting in superior prediction accuracy and computing efficiency. Nevertheless, the research was constrained by the information loss introduced during data preprocessing, particularly in the grid-based depiction of congestion levels, and the computing efficiency resulting from incorporating non-road regions in the analysis.

The growing interest in road traffic prediction emphasizes the critical demand for efficient congestion management techniques driven by the escalating issues caused by infrastructural developments worldwide. Traditional methods for forecasting traffic congestion often rely on static data from limited sources, such as single sensor networks or historical traffic records. These approaches face significant limitations in capturing the dynamic nature of traffic flow, especially when incorporating diverse influencing factors like weather conditions, social media updates, and special events.

The complexity of probabilistic models further escalates when these variables are considered, resulting in models that struggle to provide accurate predictions in real time. Existing research predominantly utilizes static datasets that do not adequately represent the dynamic changes in traffic patterns. The challenge is exacerbated by the limited time frame for data collection, often spanning only a few days, which is insufficient to accurately capture and model the evolving traffic conditions at congestion points. Moreover, integrating multiple data sources introduces additional complexity in evaluating and analyzing traffic flow patterns, complicating the effectiveness of existing models. Addressing these gaps requires the development of novel models that leverage comprehensive, real-time data sources and consider long-term traffic patterns. Incorporating a broader range of variables and extending the data collection period will enable more accurate forecasting and better road traffic congestion management. This approach will enhance predictive accuracy and improve the effectiveness of intelligent transportation systems in mitigating congestion.

3. Materials and Methods

The proposed approach leverages data from multiple online map service providers and traffic monitoring platforms, encompassing the timeframe spanning from January 1, 2019, to December 31, 2023, in the city of Bangalore. To ensure accurate and reliable traffic congestion predictions, the data undergoes a comprehensive preparation process, including time synchronization, spatial alignment, missing data management, and data cleaning, followed by feature extraction and normalization. The dataset is subsequently divided into training and testing sets in an 80:20 ratio.

The ITS architecture incorporates the predictions of the TFT technique, which is used to forecast traffic congestion. This integration involves using TFT forecasts to update real-time traffic data, improve traffic signal timings, recommend alternate pathways to drivers, and enhance incident management. By embedding these predictive insights into the ITS, the framework can proactively manage the flow of vehicles, minimize congestion, and improve travel effectiveness, ultimately leading to a more adaptive and efficient urban traffic management solution. The efficacy of the suggested model is evaluated via pertinent evaluation metrics. The workflow of the proposed research is depicted in Figure 2.

3.1. Dataset

Data from several online map service providers and traffic monitoring platforms, such as Google Maps, MapmyIndia, HERE Technologies, Waze Live Map, Live Traffic Cams, and the Ministry of Road Transport and Highways (MORTH), is employed in the proposed research. These platforms give up-to-date information about traffic congestion in Bangalore, including various route types, such as highways, arterial roads, junctions, and other essential elements of the transportation network. An illustration of the traffic congestion scenario in Bangalore City is depicted in Figure 3.

To achieve the goal, data on traffic congestion is gathered in the timeframe spanning from 1 January 2019 to 31 December 2023. The collection comprises recordings of traffic conditions taken at regular intervals, usually every 10 minutes, throughout the peak rush hour. The raw data comprises congestion levels obtained from multiple sources, such as real-time mobility data collected from mobile map applications used by travelers, together with traffic flow data extracted from cameras and sensors placed across the city. The image samples from the dataset are shown in Figure 4.

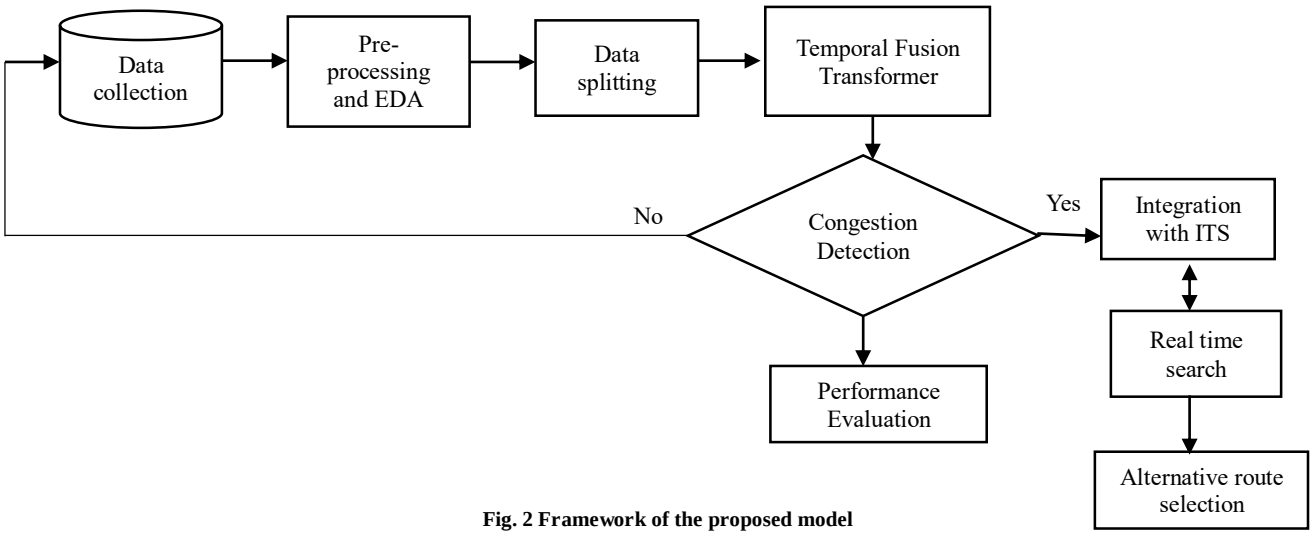


Fig. 2 Framework of the proposed model

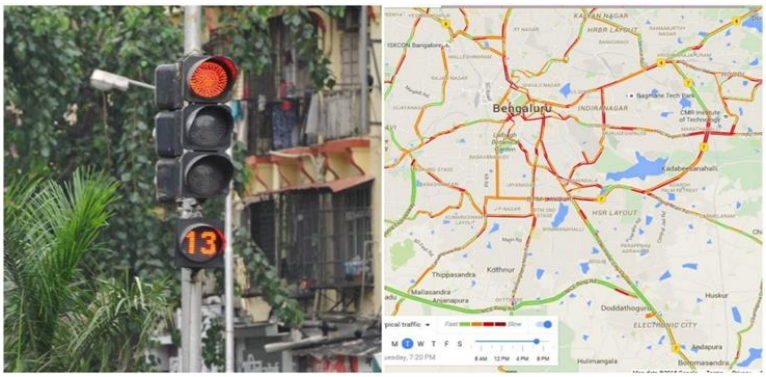


Fig. 3 Google map live traffic update illustrating traffic congestion scenario in Bangalore city

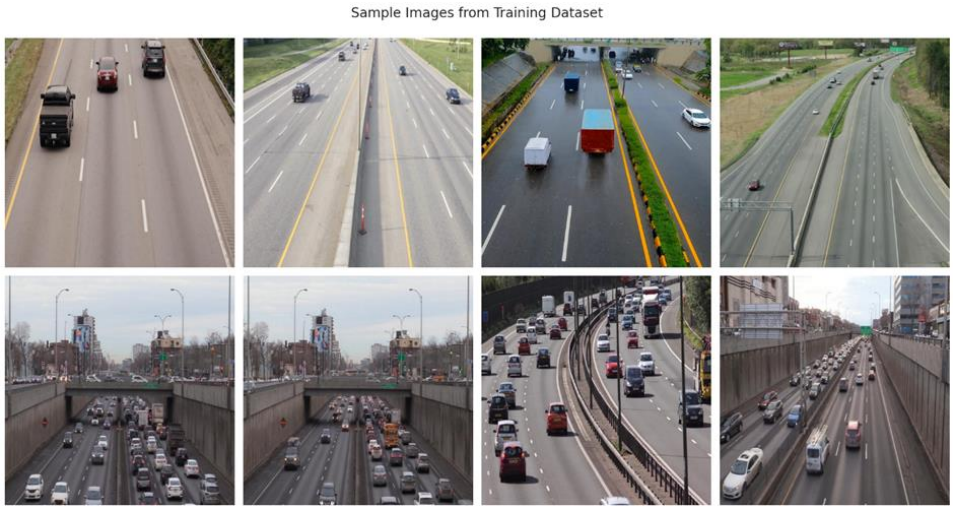


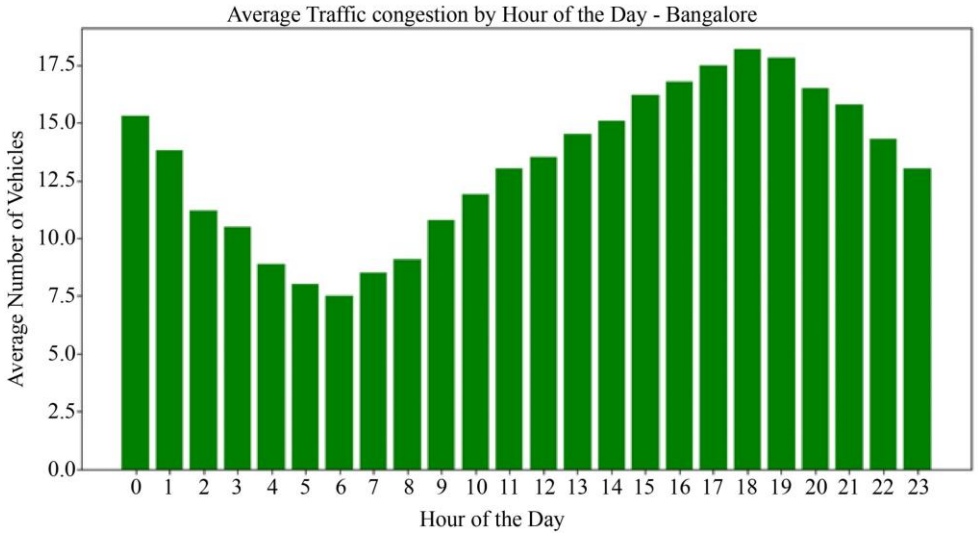
Fig. 4 Sample images from the dataset

Considering several elements contributing to congestion in Bangalore's dynamic urban environment, this extensive and varied dataset allows the model to detect patterns and anomalies in traffic flow. By utilizing this comprehensive dataset, the research endeavors to construct a resilient and dependable traffic congestion forecasting model capable of predicting current traffic conditions. This model will assist in efficiently routing emergency services and generally mitigating traffic congestion in Bangalore city.

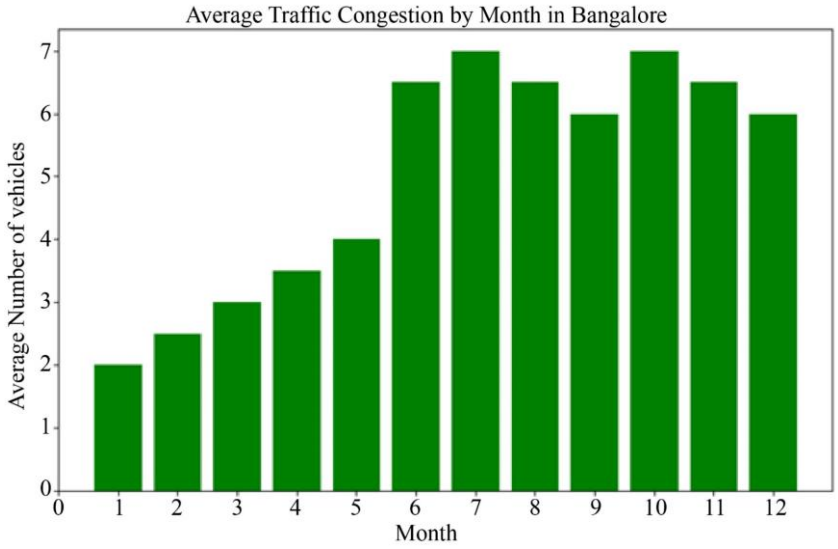
3.2. Data Preprocessing and Exploratory Data Analysis (EDA)

Effective data preprocessing is essential to ensure that the data utilized for traffic congestion forecasts is free from errors, consistent and prepared for model training. The preprocessing procedures consist of multiple crucial phases, each focusing on distinct aspects of the gathered information. Due to the heterogeneous representation of data providers, it is imperative to synchronize the data across different sources. Implementing data synchronization ensures that every time interval from several sources aligns with the same period, therefore facilitating precise comparisons and analysis. After achieving time synchronization, the subsequent task is to ensure the spatial alignment of the traffic data. This task entails comparing the geographic coordinates specified by several map service providers in Bangalore. Considering the variable degrees of specificity and precision in data obtained from various sources, spatial alignment ensures that the recorded congestion levels remain consistent among providers for a given place. Achieving this alignment is crucial to establish a cohesive perspective of traffic congestion throughout the city. Traffic data collecting is inherently imperfect, and it is typical to encounter missing data points or intervals. The initial stage in managing missing data is the identification of these discrepancies. There exist multiple approaches to address this issue efficiently. An effective method is to employ imputation methods, which involve estimating missing values using available data. One possible approach is to leave the missing data intact, therefore enabling the model to acquire knowledge from the partial dataset and maybe reveal patterns associated with the missing intervals. The process of data cleaning entails the elimination of noise and the preservation of the dataset's integrity. Traffic data noise can arise from sensor faults, causing abrupt increases or decreases in congestion levels that do not accurately represent real traffic conditions. To avoid affecting the learning process of the model, such outliers are detected and eliminated. Furthermore, the process of identifying and removing duplicate data entries ensures that every time interval contains a distinct collection of data points.

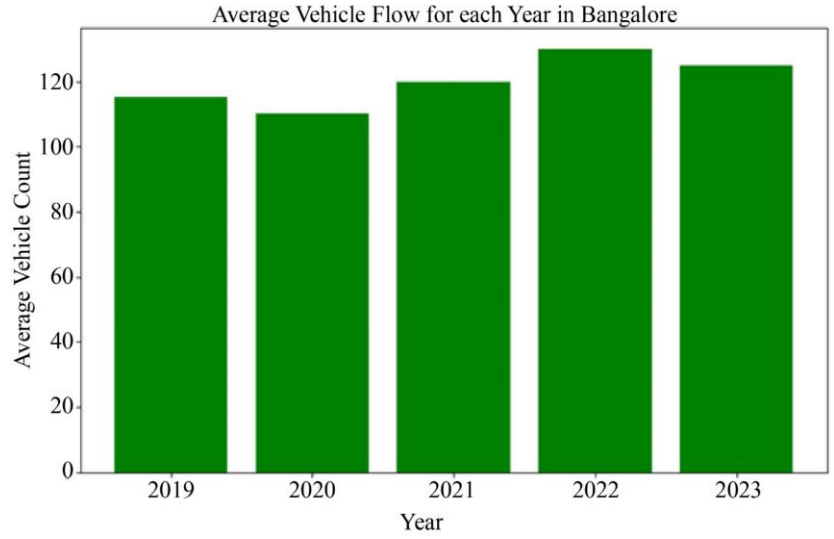
Feature extraction is the mathematical procedure of extracting significant variables from the unprocessed data utilized by the model to generate predictions. Timing-related patterns in traffic congestion are captured by extracting temporal variables like the hour of the day, day of the week, and month. The extraction of spatial data includes identifying road categories (such as highways and arterial roads) and assessing the distance between the roads and significant intersections, which might impact the congestion levels. Furthermore, previous congestion patterns are taken into account, together with external variables like weather conditions, which can have a substantial influence on traffic. To ensure an equal contribution of all features to the model, normalization and scaling techniques are used. Continuous variables such as traffic congestion levels are standardized to a set scale, usually ranging from 0 to 1. This procedure mitigates the dominance of characteristics with greater numerical ranges in the learning process of the model. One-hot encoding is a statistical method employed to transform categorical information, such as road kinds and weather conditions, into a numerical format suitable for processing by the model. The plots obtained after performing EDA are depicted in Figure 5 (a-h). The analysis emphasizes key congestion considerations, such as the hourly traffic distribution, where the highest congestion level is shown during peak hours. Monthly variations also exhibit seasonal patterns, characterized by periods of increased congestion, possibly associated with weather conditions or local activities. The data visualization of vehicle traffic over several years shows a consistent upward trend, suggesting a growing congestion problem over time.



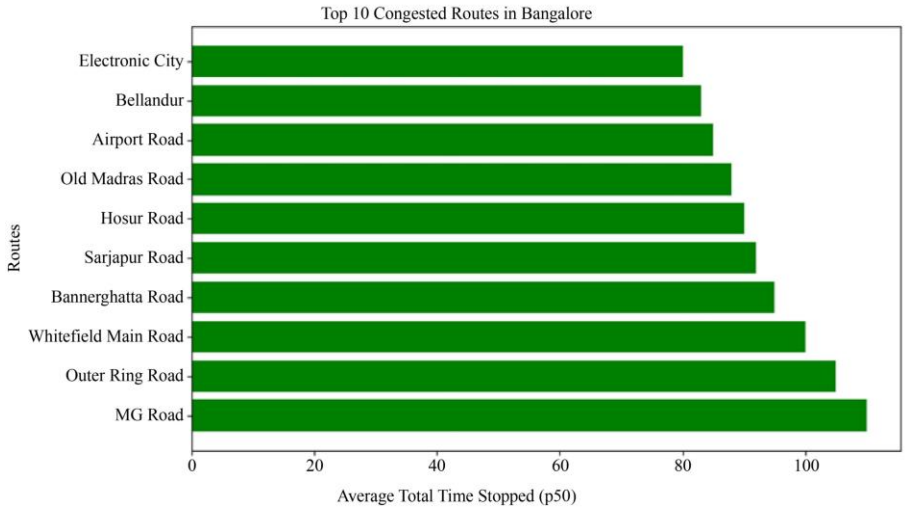
(a) Average traffic congestion by hour of the day



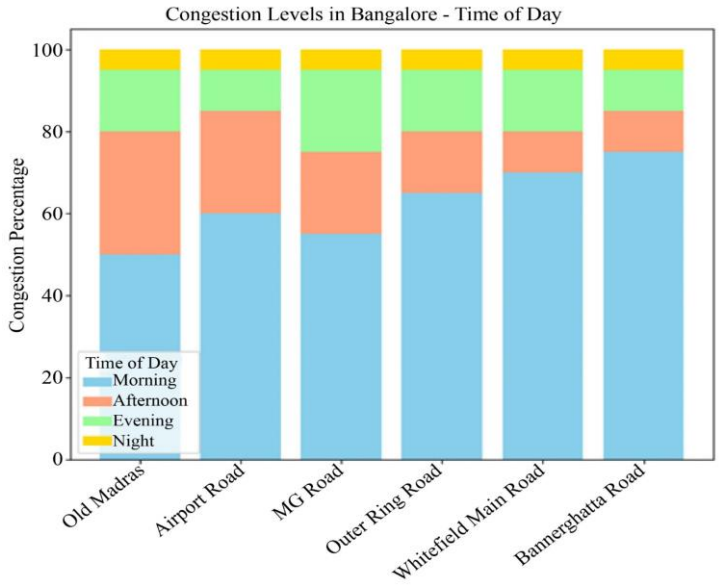
(b) Average traffic congestion by month



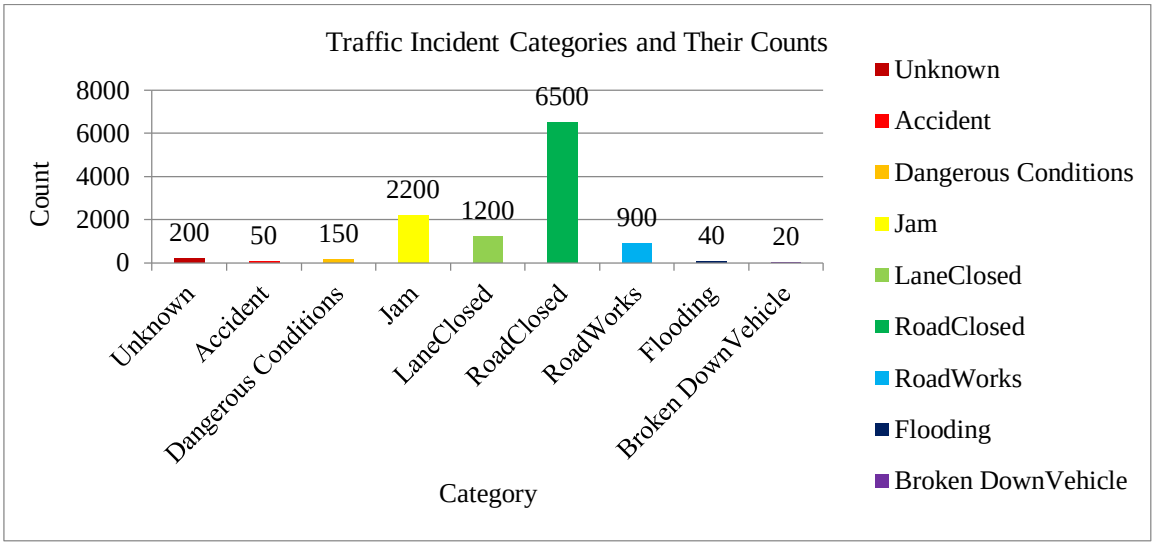
(c) Average vehicle flow year-wise



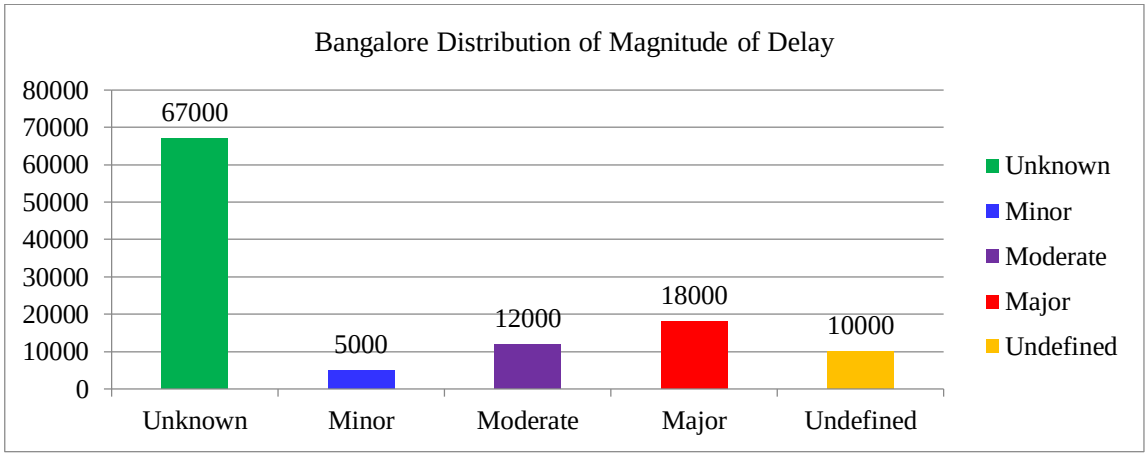
(d) Top 10 congested routes



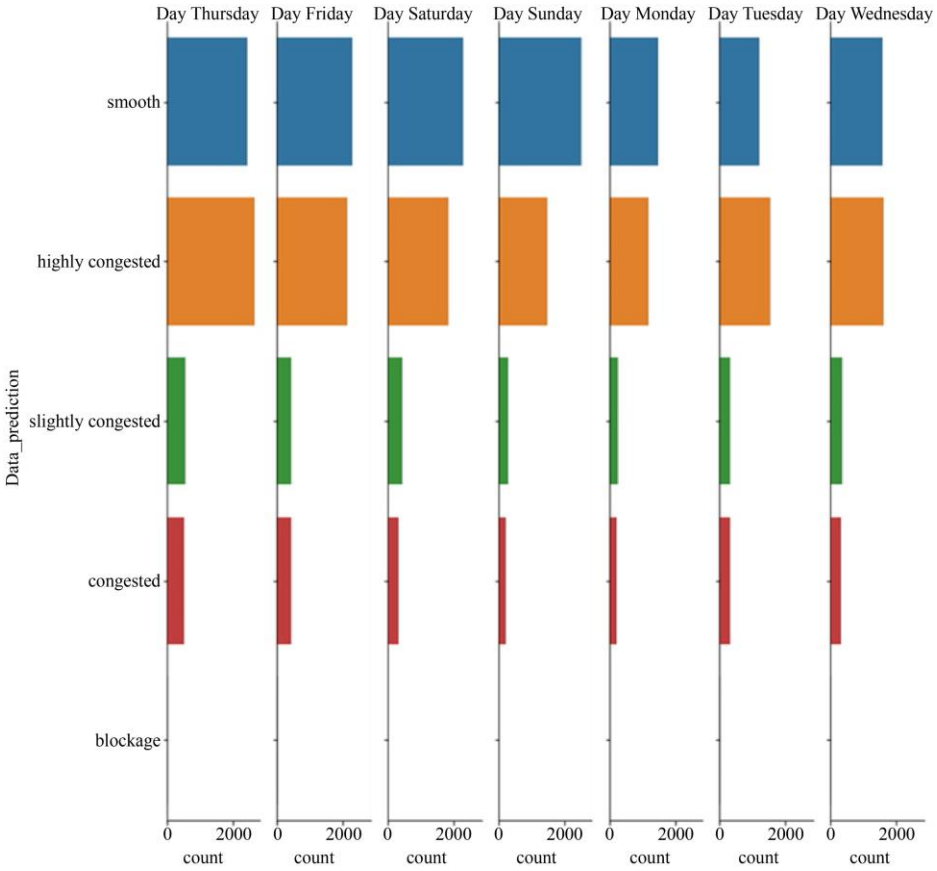
(e) Congestion levels



(f) Traffic incident categories and their counts



(g) Magnitude of delay



(h) Rate of congestion day-wise

Fig. 5 Data set visualization

By identifying the top 10 most crowded paths, the study precisely identifies areas of high congestion requiring immediate infrastructure and policy involvement. Analyzing congestion levels throughout the day gains a more comprehensive view of how traffic varies throughout various periods, providing valuable information for improving traffic management efficiency. In addition, the study categorizes traffic events, therefore offering a more well-defined understanding of the factors contributing to delays. Moreover, the analysis of the distribution of delay magnitude in Bangalore highlights the seriousness and frequency of these disturbances. Overall, the EDA provides a strong basis for creating data-driven solutions to detect traffic congestion in Bangalore.

3.3. Proposed Temporal Fusion Transformer Framework

Attention mechanisms are vital in enabling models to selectively concentrate on specific segments of the input sequence [18]. This focus is necessary for capturing intricate temporal patterns and interdependencies across distinct time steps when predicting tasks. Traditionally, Recurrent Neural Networks (RNNs), specifically LSTM networks [19], have been employed to analyze sequential data and detect underlying patterns in sequences. In these cases, the order of each element in the sequence is crucial for the prediction process. However, Transformers, a deep learning architecture focused on attention processes, provides benefits over RNNs and LSTMs by substantially decreasing training durations and effectively handling lengthy sequences via parallelization. Transformers demonstrate exceptional proficiency in capturing patterns throughout large sequences, therefore establishing themselves as a highly effective framework for time series modelling.

The proposed research selects the TFT due to its exceptional capabilities in managing time-series forecasting tasks, which are crucial for accurately predicting traffic congestion. The TFT model is adept at handling sequential data, allowing it to capture complex temporal patterns and historical trends essential for traffic prediction. The system's capacity to include many data sources, such as real-time traffic updates and static features like road conditions, enhances the accuracy of congestion forecasts. Moreover, the attention mechanisms within TFT enable the model to focus on significant features and time steps, identifying critical traffic patterns and sudden changes. Additionally, TFT's interpretability offers valuable insights into how various factors contribute to congestion predictions, facilitating better traffic management and decision-making.

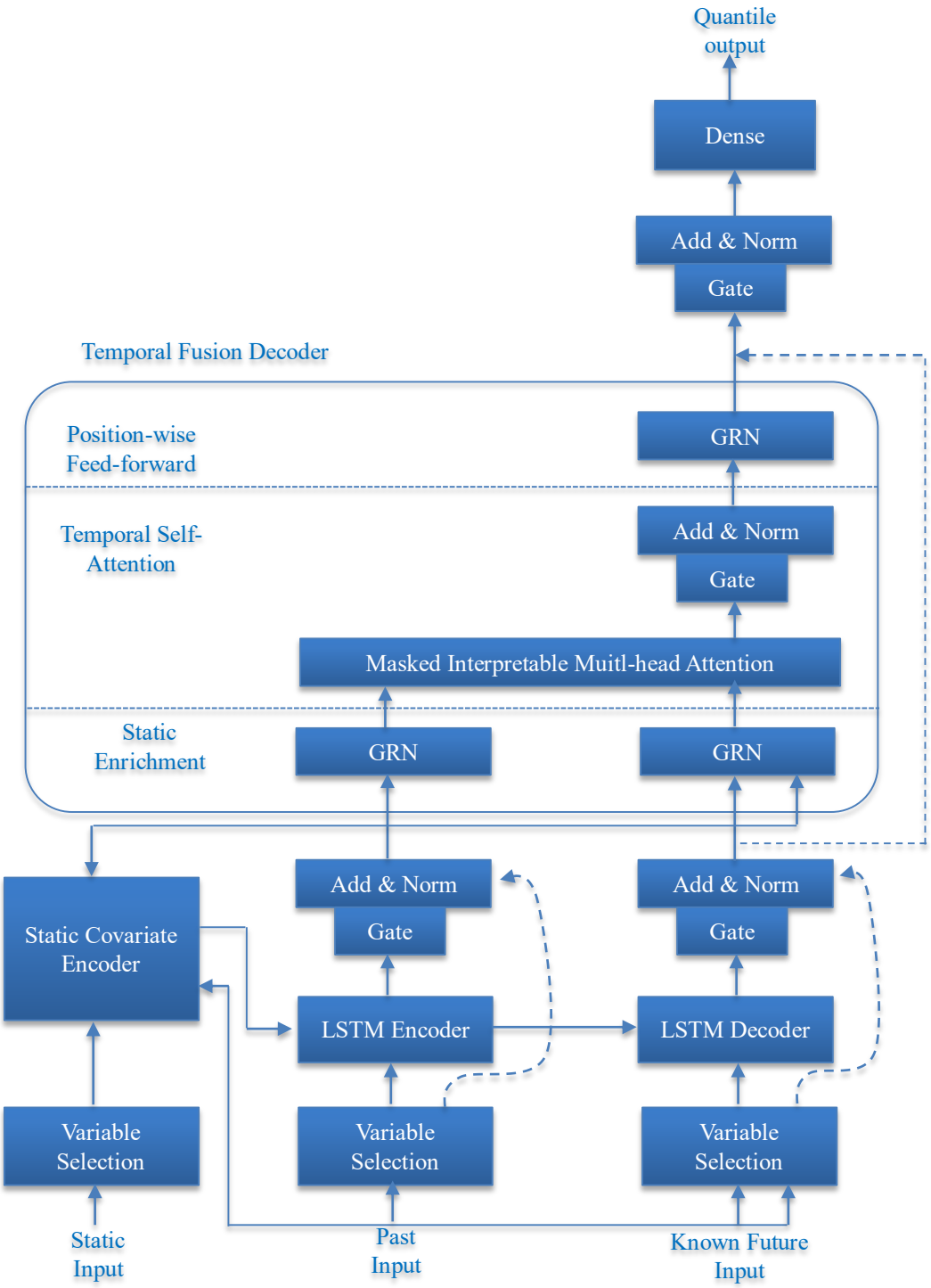


Fig. 6 Architecture of TFT

The TFT architecture, as shown in Figure 6, integrates sophisticated methods for extracting features and ensuring interpretability. Prior inputs, such as historical traffic data, undergo processing using a sequence of LSTM encoders, which excel at capturing temporal relationships in time-series data. These LSTM encoders collect fundamental characteristics from the input sequences, vital for precise congestion prediction. Concurrently, LSTM decoders are used to process known future inputs, such as scheduled events or traffic forecasts, allowing for extracting features that determine future traffic patterns.

Efficient model performance is achieved by including both static and dynamic data through the use of temporal variable selection. Static covariate encoders facilitate this procedure by extracting context vectors from static metadata, which are subsequently included in different sections of the TFT network. The integration of static data in temporal representation learning enables the conditioning of temporal patterns by enriching them with pertinent static information.

The variable selection network is structured with separate sets of blocks for each input type: static covariates, historical inputs (both unknown and known over time), and anticipated future inputs. In order to effectively manage reweighted sums of modified inputs at each time step, each block of the Sequence-to-Sequence layer acquires the ability to assess the significance of its related features. This method incorporates acquired linear transformations of continuous data and structured representations of categorical attributes. An external context vector derived from the output of the static covariate encoder block is specifically excluded from the static covariate block. This selective exclusion ensures that the model concentrates on the most pertinent characteristics, therefore enhancing the integration of time-based and fixed data for more precise prediction.

The temporal self-attention mechanism [20], an essential element of the TFT, critically contributes to enhancing the comprehensibility of the model. Through the assessment of the significance of each input vector, this process facilitates the identification of the most pertinent features and time steps for formulating predictions. Prior to the attention calculation, the Gate and Add & Norm layers are used to improve the hidden states generated by the LSTMs. The Gate layer, which implements Gated Linear Units (GLUs) [21], provides adaptability by enabling the model to eliminate superfluous elements, therefore customizing the architecture to suit the particular dataset being used. Equation (1) provides a mathematical expression for the GLUs.

$$GLU(X) = \sigma(W_1 X + b_1) \odot (W_2 X + b_2) \quad (1)$$

Where X denotes the input of the Gated Linear model, where W_1 and W_2 are parameters of the learnable weight matrix and b_1 and b_2 are the related bias parameters, σ is the sigmoid activation function and $\odot$ denotes the element-wise Hadamard product, ensuring that only the most relevant features are retained for accurate and interpretable traffic congestion predictions.

An essential element is the Add and Norm layer, which integrates residual connections with layer normalization [22]. This combination has demonstrated significant efficacy in extracting features from different transformer architectures, therefore ensuring the framework's very efficient capture of pertinent patterns in the data. Furthermore, the Gated Residual Network (GRN) is integrated into the model to offer versatility when implementing non-linear processing. Figure 7 shows the basic architecture of GRN.

Equations (2) – (4) construct the GRN, in which the Layer Norm (.) function carries out conventional layer normalization. The inputs indicated as a and c, correspond to the main input and an external context vector, respectively. The Exponential Linear Unit (ELU) [23] function incorporates non-linearity, enabling the model to acquire knowledge of intricate relationships. Intermediary layers, denoted as η_1 and η_2, together with weight matrices W_3, W_4, W_5, and related bias parameters b_3, b_4, enhance the model's information processing capabilities by selectively applying non-linear transformations as needed.

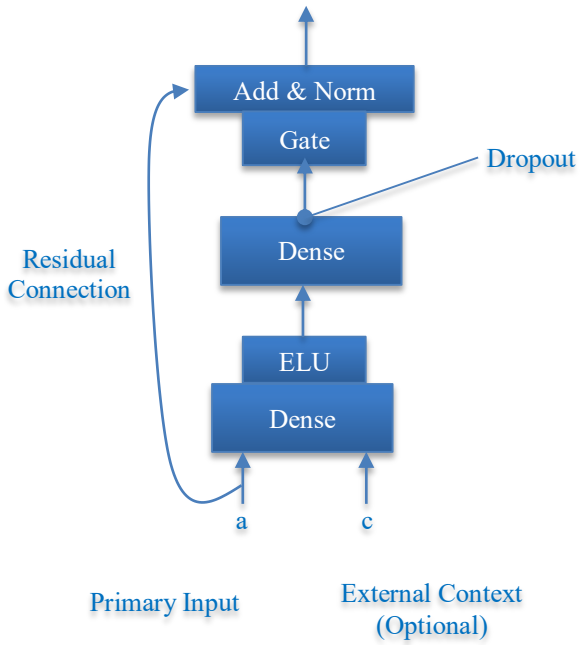


Fig. 7 GRN architecture

$$GRN(a,c) = LayerNorm\big(a + GLU(\eta_1)\big) \quad (2)$$

$$\eta_1 = W_3 \eta_2 + b_3 \quad (3)$$

$$\eta_2 = ELU(W_4 a + W_5 c + b_4) \quad (4)$$

GRN enables effective transmission of information through skip connections and gating layers, therefore ensuring strong feature extraction and data representation. Furthermore, the model has Masked Interpretable Multi-Head Attention (MIMHA) layers to enhance interpretability and focus on crucial temporal patterns [24]. This rectified multi-head attention method enables the model to detect and highlight important characteristics in the data, hence enhancing the transparency and comprehensibility of the attention process. Moreover, using the quantile loss function allows the model to produce precise predictions across several quantiles.

Implementing the (MIMHA) method greatly improves the model's capacity to selectively highlight various segments of the input sequence. This enhancement enables a more profound comprehension and analysis of the patterns exhibited by the model. By incorporating a masking technique into the attention mechanism, MIMHA ensures that the attention coefficients are both efficient and easily understandable. The typical multi-head attention formulation, which forms the basis for MIMHA, is given in Equation (5).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Wherein Q, K, and V are the matrices for query, key, and value, correspondingly, and d_k is the key dimension.

Within the attention mechanism, a masking matrix M, is incorporated to improve the interpretability of the model's focus during prediction. The masking matrix generates significant negative values at positions where attention should be inhibited, appropriately directing the model to disregard irrelevant or insignificant data points.

The model is directed toward the most pertinent, unmasked locations by forcing the softmax function to provide nearly-zero values at these masked areas. Equation (6) defines the masked attention mechanism.

$$Masked\ Attention(Q, K, V, M) = softmax\left(\frac{QK^T+M}{\sqrt{d_k}}\right)V \quad (6)$$

Within the MIMHA mechanism, the model incorporates a rectification technique to enhance interpretability by ensuring that the attention score remains non-negative. This alteration streamlines the analysis of the attention distribution, thereby facilitating the identification of the specific features and time steps that influence the model's predictions. The improvements made to the attention mechanism allow the TFT model to accurately and transparently concentrate on important parts of the data, thereby enhancing the precision of traffic congestion forecasts. Equation (7) depicts the formulation of Rectified Attention.

$$Rectified\ Attention(Q, K, V, M) = ReLU(Masked\ Attention(Q, K, V, M)) \quad (7)$$

Wherein ReLU denotes the Rectified Linear Unit Function.

The TFT model in the proposed research has exceptional proficiency in incorporating a wide range of data, encompassing both static inputs (such as road conditions) and dynamic inputs (such as real-time traffic data from several sources). This functionality enables a thorough comprehension of traffic patterns by utilizing both past data and present circumstances.

The TFT algorithm improves its capacity to discover important patterns and abnormalities in traffic data by successfully identifying and focusing on the most relevant features and time steps through the use of attention mechanisms. This is especially advantageous in the field of urban traffic management, for which sudden changes and congestion during peak hours occur regularly. Table 1 shows the TFT parameters employed in the study. The algorithm for the proposed framework is depicted below.

Table 1. Transformer network parameters

Parameter	Description	Value
d_{model}	Size of embedding output and dimensions of Q, K, and V vectors	512
encoder	Total transformer encoder stacks	6
num_heads	Number of heads in the attention mechanism	12
ffn_units	Units in the feed-forward neural network layer	2048
conv_filters	Number of convolution filters in the feed-forward part	5
dropout_rate	Dropout rate applied during training	0.1
mlp_dropout	Dropout rate for the feed-forward part	0.3

Algorithm 1: Traffic Congestion Prediction using TFT

Input: Dataset [Traffic Volume, Traffic Congestion Level, Time, Date,]
Output: Prediction Result [Traffic Congestion Level]

1. **Start:**
2. Load the dataset from sources.
3. Preprocess the dataset.
4. Split the dataset into data (train) data (test) in an 80:20 ratio.
5. Define window size (WS) and prediction horizon (H).
6. **Model ← build_model(TFT)**: Construct the TFT model architecture.
7. **Model ← train_model (data (train))**: Train the TFT model using the training dataset.

8. **Model ← optimize_hyperparameters(data(test))**: Fine-tune hyper parameters using the test dataset.
9. **Model ← evaluate (model, data (test))**: Assess model performance using metrics (MAE, MSE, RMSE, MAPE, R^2).
10. **Model ← save_best_model()**: Save the best-performing model.
11. **MAE, MSE, RMSE, MAPE, R^2 ← (Model, data (test))**: Calculate and record performance metrics.
12. **Prediction ← (Model (WS, H), data (test))**: Predict traffic congestion levels for the given window size and horizon.
13. Return Prediction.

End

The ITS seamlessly integrates the predictions from the TFT into its conceptual framework to deliver real-time traffic updates, enhance signal timings, recommend alternative paths to drivers, and enhance incident management. The ITS could actively control traffic flow, enhance route efficiency, and decrease overall congestion by incorporating these sophisticated forecasts.

3.4. Performance Metrics

A number of appropriate metrics are employed to evaluate the efficacy of the suggested forecasting model. These statistical measures aid in evaluating the forecasts' accuracy, reliability, and efficiency.

MAE is a metric that quantifies the average size of discrepancies among predicted and actual values, disregarding the direction. MSE quantifies the mean of the squared discrepancies between predicted and observed values. The RMSE is the square root of the MSE. It quantifies the magnitude of the error in the same units as the data and is particularly responsive to significant errors.

The MAPE is a statistical metric that expresses the error as a percentage of the actual values. This metric is valuable for assessing the error in relation to the magnitude of the real values. The R-squared (R^2) statistic quantifies the amount of variability in the dependent variable that can be explained by the independent variables. It quantifies the degree of fittingness of the model. Equation (8-12) depicts the performance evaluation metrics.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (8)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (10)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (12)$$

Where, $\hat{y}_i$ and y_i is the predicted and actual value for the i^{th} observation, n is the total number of observations, $\bar{y}$ is the mean of the actual values.

3.5. Hardware and Software Setup

This research employed a high-performance computing configuration consisting of an Intel Core i7 processor, 32GB of RAM, and the robust NVIDIA GeForce GTX 1080Ti GPU. The model was developed using the Keras package, which functions as a prototype built on the TensorFlow framework and implemented in Python. Renowned for its intuitive interface and powerful functionalities, Keras was instrumental in developing intricate neural network structures. The present methodology ensures optimal resource allocation across CPU, GPU, and TPU system configurations. The deployment was executed using Google Colab, a cloud-based Python notebook platform, to utilize significant computer resources and enhance model training. Hyperparameters are crucial in defining the behavior of a learning framework during the training phase. Unlike model parameters, which are derived from the data, hyperparameters are established by the user prior to the training process. Table 2 shows the hyperparameter specifications employed in the research.

Table 2. Hyperparameter specifications

Hyperparameters	Values
Learning rate	0.001
Optimizer	Adam
Loss function	Sparse categorical cross-entropy
Batch size	32
Epochs	50
Number of folds	5

4. Results and Discussion

The efficiency of the proposed model is evaluated using a range of performance metrics to ensure its accuracy and reliability in traffic congestion prediction, as depicted in Table 3 and Figure 8. The model's accuracy in predicting congestion levels is inferred from the low MAE, MSE, and RMSE values, which indicate minimum error margins. Furthermore, the MAPE of 7.2% indicates that the model retains a comparatively small percentage error, which is essential for real-time traffic control applications where accuracy is of utmost importance. The R^2 value of 0.87 provides additional evidence for the model's efficacy, indicating that the model can account for 87% of the variability in traffic congestion. Overall, these findings emphasize the superiority of the suggested methodology in providing dependable and precise traffic forecasts, hence establishing it as an essential tool for improving urban traffic management systems.

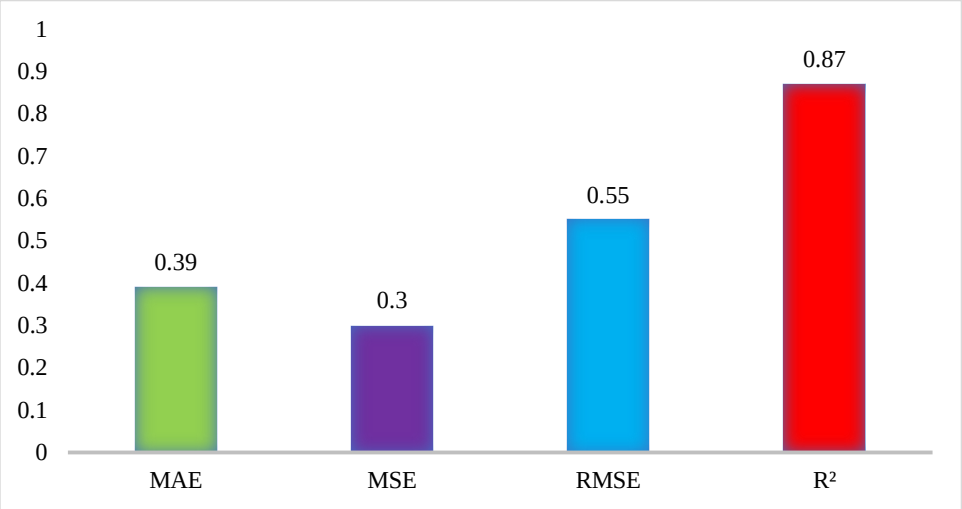


Fig. 8 Performance evaluation plot

Table 3. Performance evaluation

Evaluation Metrics	Results obtained
MAE	0.39
MSE	0.30
RMSE	0.55
MAPE	7.2%
R^2	0.87

Figure 9 illustrates the predicted traffic congestion levels, highlighting areas of low, medium, and high congestion based on the forecasted data. This stratified depiction helps understand the varying degrees of traffic congestion, facilitating better route planning and traffic management decision-making.

(a) High congestion

(b) Medium congestion

(c) Low congestion

Fig. 9 Prediction outputs

Figure 10 provides a comprehensive visualization of traffic routes in Bangalore, highlighting both the congested and alternative routes. The red line depicts the congested route, which represents the most direct path between the start and end points but is currently experiencing heavy traffic. In contrast, the blue line illustrates the alternative route, which strategically bypasses the congested area by diverting around it, thereby reducing potential delays and improving travel efficiency. A heatmap layer is included, showing congestion levels across various points in the city, further aiding in identifying traffic hotspots. This visual representation effectively demonstrates how the proposed method can recommend more efficient travel options by avoiding high-traffic zones, thereby enhancing overall route management and congestion mitigation.

The predicted traffic flows for Bangalore City between 2024 and 2032, depicted in Figure 11, indicate a notable change in urban mobility trends for the next decade. Extending beyond 2028, the graph shows a resumption of the declining pattern, as traffic flow consistently diminishes to around 106 units by 2032. The continuous decrease observed indicates the possible long-term effectiveness of strategies implemented to alleviate traffic congestion. The persistent decline indicates that the city may have implemented more resilient and enduring measures to control its traffic, such as incorporating intelligent traffic management systems, increasing investment in alternate transportation modes, or expanding the public transit network.

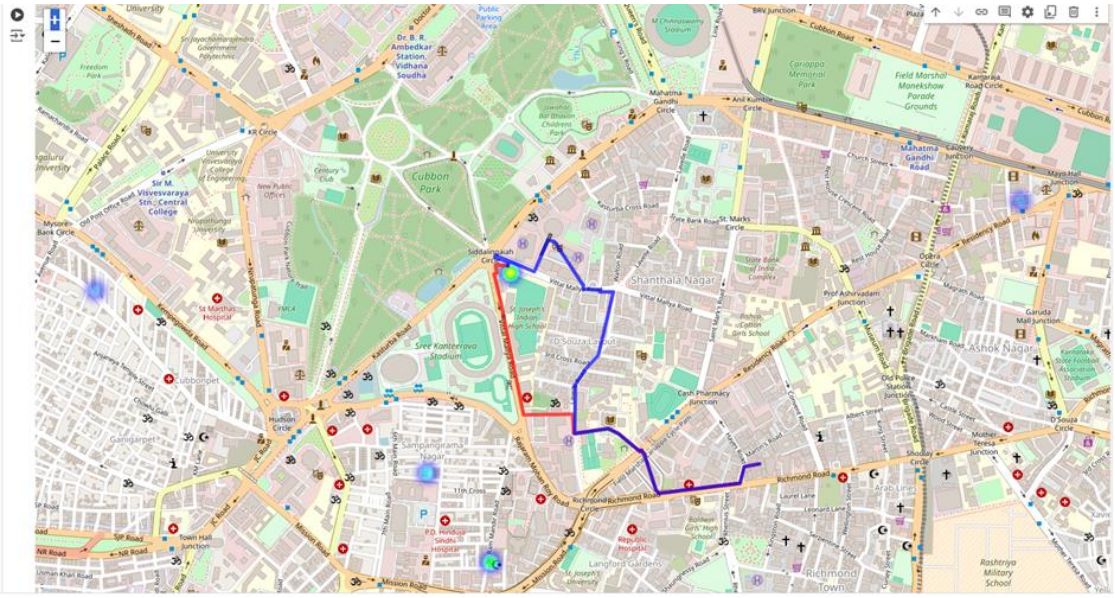


Fig. 10 Traffic congestion visualization and route optimization

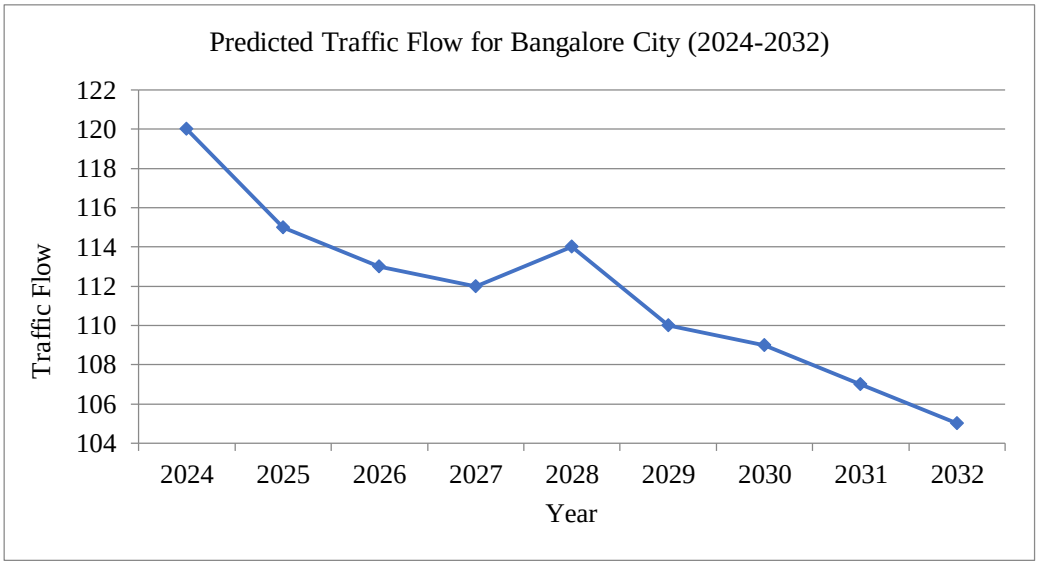


Fig. 11 Predicted traffic flow trends

Overall, the estimated traffic flow highlights a future in which implementing planned measures could result in a steady but significant decrease in traffic congestion. The overall declining trend in the graph emphasizes the need for ongoing urban planning endeavors, with a specific emphasis on sustainable transportation solutions and smart city projects. These measures are crucial to prevent traffic congestion from hindering the city's development and quality of life in the coming years.

5. Conclusion

Traffic congestion poses significant challenges to urban mobility, leading to delays, increased emissions, and reduced quality of life. The proposed research addresses the critical issue of traffic congestion in urban environments by integrating advanced forecasting techniques with ITS. The study employs a robust methodology by utilizing the Temporal Fusion Transformer (TFT) model to predict traffic congestion, leveraging a diverse dataset from various online map services and traffic monitoring platforms. The comprehensive data preparation process, including synchronization, alignment, and normalization, ensures the accuracy and reliability of the forecasts. The outcomes validate the superiority of the proposed approach, with performance metrics such as MAE of 0.39, MSE of 0.30, RMSE of 0.55, MAPE of 7.2%, and R^2 of 0.87, indicating high prediction accuracy. Integrating TFT predictions into the ITS framework enhances real-time traffic management by optimizing signal timings, suggesting alternate routes, and improving incident management. This research highlights the potential of combining predictive analytics with intelligent systems to create a more adaptive and efficient traffic management solution, ultimately reducing congestion and improving urban mobility.

As cities continue to grow and traffic patterns become increasingly complex, there is significant potential for expanding TFTs to incorporate additional data sources such as real-time traffic incident reports, weather conditions, and socio-economic factors. Future research could explore the integration of TFT predictions with advanced vehicular communication systems and autonomous vehicle technology, creating a more adaptive and proactive traffic management ecosystem. Exploring the integration of predictive analytics with public transportation systems to improve route planning and efficiency also presents a valuable opportunity.

Acknowledgements

The author expresses profound appreciation to the supervisor for providing guidance and unwavering support throughout the course of this study.

References

[1] Vito Albino, Umberto Berardi, and Rosa Maria Dangelico, "Smart Cities: Definitions, Dimensions, Performance, and Initiatives," *Journal of Urban Technology*, vol. 22, no. 1, pp. 3-21, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[2] Roopa Ravish, and Shanta Ranga Swamy, "Intelligent Traffic Management: A Review of Challenges, Solutions, and Future Perspectives," *Transport and Telecommunication Journal*, vol. 22, no. 2, pp. 163-182, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Wesley E. Marshall, and Eric Dumbaugh, "Revisiting the Relationship between Traffic Congestion and the Economy: A Longitudinal Examination of US Metropolitan Areas," *Transportation*, vol. 47, pp. 275-314, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ghosh Banishree, *Intelligent Mobility for Minimizing the Impact of Traffic Incidents on Transportation Networks*, 1st ed., Augmented Intelligence Toward Smart Vehicular Applications, CRC Press, pp. 175-194, 2020. [Google Scholar] [Publisher Link]

[5] Guofa Li et al., "Influence of Traffic Congestion on Driver Behavior in Post-Congestion Driving," *Accident Analysis & Prevention*, vol. 141, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Elias C. Eze et al., "Advances in Vehicular Ad-Hoc Networks (VANETs): Challenges and Road-Map for Future Development," *International Journal of Automation and Computing*, vol. 13, pp. 1-18, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[7] K. Ramesh, A. Lakshna, and P.N. Renjith, "Smart Traffic Congestion Model in IoT - A Review," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, pp. 651-658, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] J. Prakash et al., "A Vehicular Network Based Intelligent Transport System for Smart Cities Using Machine Learning Algorithms," *Scientific Reports*, vol. 14, no. 1, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Abdullahi Chowdhury et al., "IoT - Based Emergency Vehicle Services in Intelligent Transportation System," *Sensors*, vol. 23, no. 11, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Sura Mahmood Abdullah et al., "Optimizing Traffic Flow in Smart Cities: Soft GRU-Based Recurrent Neural Networks for Enhanced Congestion Prediction Using Deep Learning," *Sustainability*, vol. 15, no. 7, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Muhammad Saleem et al., "Smart Cities: Fusion-Based Intelligent Traffic Congestion Control System for Vehicular Networks Using Machine Learning Techniques," *Egyptian Informatics Journal*, vol. 23, no. 3, pp. 417-426, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sharmila Majumdar et al., "Congestion Prediction for Smart Sustainable Cities Using IoT and Machine Learning Approaches," *Sustainable Cities and Society*, vol. 64, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] G. Kothai et al., "A New Hybrid Deep Learning Algorithm for Prediction of Wide Traffic Congestion in Smart Cities," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ayesha Atta et al., "An Adaptive Approach: Smart Traffic Congestion Control System," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 9, pp. 1012-1019, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] A. Ata et al., "Modelling Smart Road Traffic Congestion Control System Using Machine Learning Techniques," *Neural Network World*, vol. 29, no. 2, pp. 99-110, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] S. Muthuramalingam et al., "IoT Based Intelligent Transportation System (Iot-ITS) for Global Perspective: A Case Study," *Internet of Things and Big Data Analytics for Smart Generation*, pp. 279-300, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sen Zhang et al., "Deep Autoencoder Neural Networks for Short-Term Traffic Congestion Prediction of Transportation Networks," *Sensors*, vol. 19, no. 10, pp. 1-19, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Derya Soydaner, "Attention Mechanism in Neural Networks: Where It Comes and Where It Goes," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13371-13385, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Sepp Hochreiter, and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. [CrossRef] [Google Scholar] [Publisher Link]

[20] Bryan Lim et al., "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748-1764, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yann N. Dauphin et al., "Language Modeling with Gated Convolutional Networks," *Proceedings of the 34th International Conference on Machine Learning*, Sydney NSW Australia, vol. 70, pp. 933-941, 2017. [Google Scholar] [Publisher Link]

[22] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer Normalization," *Arxiv*, pp. 1-14, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[23] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *Arxiv*, pp. 1-14, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ashish Vaswani et al., "Attention is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach California USA, pp. 6000-6010, 2017. [Google Scholar] [Publisher Link]