International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

# Terminal Area Traffic Situation Prediction Using a Transformer-Based Model with Multi-Head Self-Attention

**Mr. Aniket Pal**
Department of Computer Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, India
aniketpal.edu@gmail.com

**Dr. Nisha Wandile**
Department of Computer Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, India
nishakimmatkar@gmail.com

***Abstract*— The constant movement of vehicles and unpredictable traffic patterns present significant challenges at terminal areas like airports, seaports, and logistics hubs. Managing these challenges is becoming increasingly complex, making it essential to have precise and timely traffic forecasts to reduce congestion, enhance safety, and optimize resource use. This research focuses on developing a Multi-Head Attention Transformer model for terminal traffic forecasting. By leveraging self-attention, the model can capture intricate relationships between different traffic elements over time, accurately predicting both short-term fluctuations and long-term trends. It employs a multi-head approach to process traffic data from multiple angles, uncovering key connections across various timeframes. Compared to traditional methods like ARIMA, LSTM, and GRU, the proposed model significantly improves prediction accuracy in trials using traffic datasets. Due to its scalability and adaptability, the model holds great potential for real-time traffic management across different terminal locations, although it does require substantial data and computational resources to operate effectively.**

***Keywords*— *Traffic situation prediction, transformer, temporal convolutional network, Intelligent transportation system, air transportation.***

## I. INTRODUCTION

Traditional air traffic management (ATM) methods have struggled to keep up with the rapid growth of the airport transportation industry. In recent years, the field of Intelligent Transportation Systems (ITS) has gained increasing attention, aiming to improve navigation efficiency and enhance safety [1]. The integration of advanced artificial intelligence technologies is pushing the evolution of ATM toward greater intelligence and efficiency.

Smart ATM focuses on understanding the current terminal traffic situation. By optimizing aircraft operations and ensuring safety within terminal areas, situational awareness plays a crucial role. This awareness consists of two main aspects: accurately identifying the current situation and predicting future developments [2]. Researchers have explored air traffic conditions within terminal areas and routes using approaches like complex network theory [3, 4], [5], [6]. These studies primarily aim to analyze and understand traffic dynamics by examining air traffic patterns in conjunction with related theories.

A novel approach to airspace complexity prediction was developed by Du et al. [7], who introduced a spatio-temporal hybrid deep learning model. This model effectively captured the geographic correlations and temporal dependencies of airspace complexity data. In another study, Sui et al. proposed a spatiotemporal graph convolutional network [8], which was designed to predict traffic conditions and examine the correlations between changes in airspace operational states. However, despite the focus of these studies on forecasting air traffic conditions in the broader airspace, there is a lack of in-depth research into predicting operational scenarios specifically within the terminal area, the busiest part of the air traffic system.

Recently, transformer-based deep learning models have become a key area of research for time-series forecasting tasks [9], [10]. Given that traffic in the terminal area follows a time-series pattern, this study proposes a model called ConvTrans-TCN for predicting terminal traffic conditions. The model is divided into three main parts: part one encodes the data, part two synthesizes it, and part three computes situational values. The model generates a final prediction of the traffic situation by processing multidimensional scenario data across multiple computation layers. The accuracy of these predictions could aid ATM decision-making, improving traffic management and policy implementation.

## II. LITERATURE SURVEY

Time- series information $Xt \in R^{(N \times D)}$, where N is the number of nodes and D is the number of scenario characteristics at time step t, is used for traffic situation prediction [8]. The following equation [26] expresses the

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

mapping function $f$, which allows one to forecast future circumstances by examining past scenario data.

$$[X_{t-T'+1}, \ldots, X_t] \xrightarrow{f} [X_{t+1}, \ldots, X_{t+T'}] \quad (1)$$

The attention mechanism is widely used in various domains such as natural language processing (NLP), image recognition, protein identification, and recommendation systems. These methods are effective for understanding the relationships between variables and target objectives [11], [12]. One specific attention technique, called self-attention, focuses on modeling many-to-many relationships. By uncovering hidden correlations among features, it can detect intricate patterns [13]. Studies suggest that combining self-attention with different network structures can enhance predictions of traffic flow. For instance, Fang et al. [14] applied attention mechanisms to overcome the limitations of LSTM models, which struggle to capture long-term dependencies in traffic flow data, resulting in better short-term predictions. Despite these advances, many existing approaches still fail to address both short- and long-term forecasting tasks adequately. To tackle this, the Long Short-Term Orient Graph Convolutional Network [15] was developed, merging attention mechanisms with graph neural networks to better capture complex spatial relationships. Additionally, Kong et al. [16] introduced a graph-based talking-heads attention layer, which efficiently models spatial dependencies. By leveraging the shared patterns between terminal area traffic data and traffic flow data, multi-head self-attention is used to extract meaningful features and reveal complex patterns.

Transformers have become a cornerstone in NLP and visual tasks, with attention mechanisms being at their core [17]. Unlike RNNs, transformers excel in processing long sequences more efficiently by utilizing parallelization. Their powerful sequence modeling capabilities have made transformers popular in traffic prediction tasks. Researchers are continually enhancing transformer architectures by incorporating diverse network components to boost performance. Examples of such architectures include GMAN [18] and NAST [19], which combine the transformer's encoder-decoder framework with neural networks to capture dependencies more effectively. However, current methods still face limitations, such as restricted input sequence lengths and inadequate modeling of local data features. This study proposes a model that integrates a temporal convolutional network [20] with transformers and introduces a self-attention module based on causal convolution. This combination results in improved forecasting performance for terminal area traffic conditions.

## III. METHODOLOGY

### A. *Transformer*

In 2017, Google introduced the Transformer model to address the limitations of recurrent neural networks (RNNs) in capturing long-range dependencies within extended sequences. The Transformer gained significant attention for its exceptional performance in machine translation tasks.

The core components of the Transformer architecture are the Encoder and Decoder, each consisting of six blocks. The Encoder is made up of two primary layers: the self-attention layer and the feedforward neural network layer. These layers process and encode the input sequences. On the other hand, while the Decoder's Attention layer is structurally similar to the Encoder, it also incorporates a Masking mechanism. This Mask ensures that during the prediction of each token, future tokens remain hidden, allowing the model to generate outputs in a sequential manner without peeking at the subsequent tokens.

As part of this study, the suggested prediction model makes use of the Transformer's Encoder component. Consequently, the two layers that are crucial to the Encoder's operation—the Self-Attention and the Feedforward Brain Network—are the ones that will be discussed here.

#### 1. Multi head and self attention mechanism

Involving the following actions, self-attention is among the most common attention mechanisms: At the outset, the query grid $Q \in RT \times Dq$, the key of the matrix $K \in RT \times Dk$, and the resulting matrix $V \in RT \times Dv$ are all created by linearly transforming the input matrix $X \in RT \times Dx$. Next, an attention-scoring function is used to calculate the attention matrix $MA \in RT \times T$ [26]. The attention balance matrix $WA \in RT \times T$ is generated by normalizing the result with the softmax function. Finally, the value matrix and the weight for attention matrix are multiplied by each other to yield the output $H \in RT \times Dv$. The following equations show how the attention-scoring function uses the scaled dot-product method. where [26]:

$$Q = XW_q \quad (2)$$

$$K = XW_k \quad (3)$$

$$V = XW_v \quad (4)$$

$$M_A = \frac{QK^T}{\sqrt{D_K}} \quad (5)$$

$$H = \text{softmax}(M_A)V = W_A V \quad (6)$$

The matrices Wq, Wk and Wv represent parameter matrices used for linear transformations. The function softmax(·) normalizes each row vector to ensure the sum equals one. A query vector's dimensionality (Dk) in the

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

queries matrix K is used to express the scaled dot-product function through Equation (2). To deal with the problem of high dimensionality and big numerator variation, the denominator is the cubic root of Dk.. This adjustment prevents the gradient from becoming too small, which would otherwise make training the model more challenging.

The Multi-Head Self-Attention Mechanism enhances the model's capability to extract diverse feature information from various positions by combining the outputs of multiple attention calculations through a linear transformation.

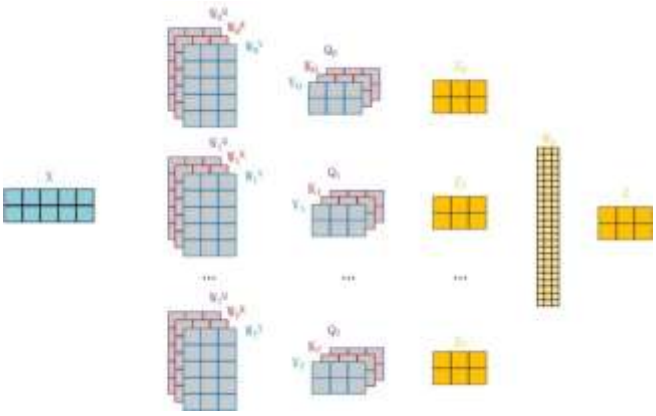


Fig. 1. The calculation process of the multi-head self-attention mechanism.

The computation involves projecting the query, key, and value vectors (Q, K and V) into multiple matrices using distinct linear projection matrices. Each set of Q, K and V matrices undergoes the attention calculation independently. The results from all attention calculations are concatenated and linearly transformed with the parameter matrix W0 to produce the final output, as illustrated in Figure. 1.

If there are " h " projection spaces, the multi-head self-attention mechanism is calculated as follows:

$$\text{Multihead}(H) = W^{O}[h_1; h_2; \ldots; h_h] \qquad (7)$$

the calculation of each head component is done in accordance with point (6).

### 2. Position by position and feed forward networks

The feed forward neural network component consists of two fully connected layers with a ReLU activation function applied between them. This component is designed to integrate and synthesize all the encoded information, as expressed in:

$$F_{\text{FFN}}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \qquad (8)$$

where the weight matrices are W1 and W2. Bias exists in b1 and b2. The input is x.

### 3. Enhanced transformer model

Terminal traffic situation data, being time-series in nature, requires accurate modeling of both long-term and short-term patterns while handling outliers caused by emergency data. Zhou et al. [22] addressed this by proposing a causal convolutional based self attention module, which improves local feature extraction in sequence data. Unlike traditional multi-head self-attention modules that rely solely on linear projections of individual time points, this module utilizes a Conv1D layer to generate the query and key matrices, effectively capturing local dependencies and mitigating the impact of outliers.

$$Q = \text{Conv1D}(X) \qquad (9)$$

$$K = \text{Conv1D}(X) \qquad (10)$$

$$H = \text{softmax}(QK^T / \sqrt{D_K} + \text{Mask})V \qquad (11)$$

An essential The terminal transport structure forecasting system includes the Conv1D layer, which employs causal padding to guarantee consistent sequence length during encoding and avoid future information leaks. To further improve prediction accuracy, it employs a mask to convert the attention-scoring matrix to a smaller triangular matrix. The structure and operations of this module, including causal convolution and masking, are detailed in Figures 2 and 3.

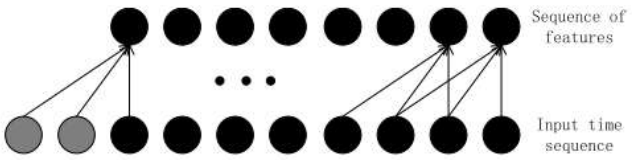



Fig. 2. Causal convolution

*B. Information fusion module based on tcn*

An input data sequence with proportions matching the input is encoded into features by the causal convolutional self-attention module in order to forecast terminal-area traffic circumstances. The TCN is used for data fusion and ultimate prediction, whereas this Transformer-adapted module acts as the feature extractor [23].

To maintain temporal directionality and avoid data leakage, the TCN design stacks 1D convolutional layers and employs causal convolutions.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

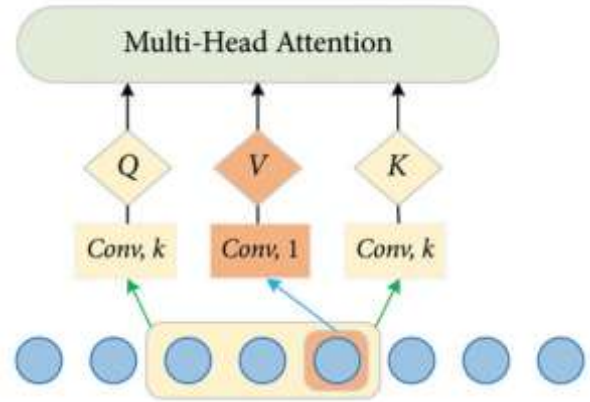



Fig. 3. Multi head causal convolutional self attention

It uses Residual Connections and Dilated Causal Convolution kernels [24] to improve performance while expanding the receptive field without using very deep networks. TCN provides enough capacity for traffic prediction while also delivering simplicity by successfully capturing both short-term and long-term interdependence in time sequences. Below is a summary of the calculation details for Residual Connections and Dilated Causal Convolution[26].

**1. Dilated causal convolution**

The model's receptive field can be exponentially expanded with dilated causal convolution [26].

$$F(X) = \sum_{i=0}^{k-1} f(i) \cdot X_{s-d \cdot i} \qquad (12)$$

The convolution is calculated with the following parameters: d is the dilation rate, k is the kernel size, and s − d · i represents the matching position in the input sequence, given an input X ∈ Rn and a one-dimensional causality dilated convolution kernel f ∈ Rk. As shown by[26], the dilation rate d usually grows as the network layer i does:

$$d = O(2^i) \qquad (13)$$

This exponential growth of the dilation rate accelerates the expansion of the receptive field compared to adjusting the kernel size, ensuring that higher-level convolution kernels can capture all relevant inputs in the sequence. As a result, it enhances information fusion and effectively models long-term dependencies, as illustrated in Fig. 4.

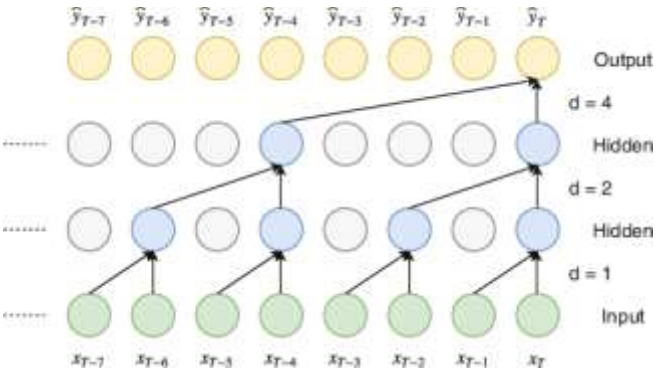



Fig. 4. Dilated causal convolution.

**2. Residual connection**

The first branch, the original network branch, applies a function F to the input; the second branch, the residual connect branch, takes the input and adds it to the network's output immediately [26]. Here is the final output:

$$o = \text{Activation}(x + F(x)) \qquad (14)$$

Deeper networks can adjust to data distribution, maintain consistent performance, and avoid problems like gradient vanishing with the aid of this structure, which allows multi-layer networks to learn identity mappings easily rather than complicated transformations. The input is adjusted using a 1D convolution with a 1 kernel size to ensure that the input and outputs tensor shapes are identical [26].

*C. ConvTrans-TCN-based situation prediction model*

With the introduction of the ConvTrans-TCN model, neural network traffic scenario prediction becomes more accurate. It has three parts: extracting features and encoding data, fusing data, and calculating scenario prediction values [26]. A causal convolutional self-attention module encodes past traffic data in the feature extraction component.

X=[x1;x2;… ;xT]X = [x_1; x_2; \dots; x_T] as follows[26]:

$$MH = ConvSA(X) \qquad (15)$$

Next, the encoded features are processed using two layers of Fully Connected Neural Networks (FNN) with ReLU activation:

$$\text{FM} = \text{ReLU}(\text{MH} \cdot W_1 + b_1) \cdot W_2 + b_2 \qquad (16)$$

A residual connection is applied with layer normalization:

$$\text{FX} = \text{LayerNorm}(X + \text{FM}) \qquad (17)$$

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

The information fusion part employs the Temporal Convolutional Network (TCN) architecture, which uses dilated convolution for extracting features and fusion in the time dimension. The processing involves:

$$\mathrm{TX} = \mathrm{ReLU}(\mathrm{FX} \cdot W + b) \qquad (18)$$

Then, a dilated convolution layer with weight normalization and dropout is applied:

$$\mathrm{TY}^1 = \mathrm{Dropout}(\mathrm{ReLU}(\mathrm{WN}(\mathrm{DConv}(\mathrm{TX})))) \qquad (19)$$

The final situation prediction value is obtained from the last vector in the sequence:

$$y = \mathrm{ReLU}(\mathrm{ty}_T \cdot W_1 + b_1) \cdot W_2 + b_2 \qquad (20)$$

The model employs the mean squared error (MSE) loss function:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{observed}_i - \mathrm{predicted}_i)^2 \qquad (21)$$

The ConvTrans-TCN model optimizes prediction accuracy with techniques like residual connections, normalization, and regularization, ensuring fast convergence and superior generalization.

## IV. EXPERIMENTS

This section first evaluates the prediction model's feasibility, then tests the impact of information encoding and situation data causality. It also compares the proposed model with common prediction methods to demonstrate its superiority.

### *A. Preparation transformer*

This study uses terminal area traffic data from Tianjin Binhai International Airport (ZBTJ), collected between June 3–16, 2019 [1]. ZBTJ, a busy dual-runway airport, operates at full capacity starting in June. Each data sample spans 10 minutes and includes 13 features: 12 for traffic conditions and one expert-labeled category (smooth, normal, congested, or standstill). Labels were provided by experienced air traffic controllers and aviation researchers.

Data was normalized and split into training (70%), validation (10%), and test (20%) sets. Training data was further augmented via interpolation. A 10-time-step sliding window was applied to create multistep sequences for input, with the next time step's situation level as the target output.

The model includes a causal convolutional encoder with kernel size 3 and 28 filters, and a 4-head multi-head attention module [2]. Its temporal convolutional network (TCN) uses six 1D dilated causal convolutional layers (kernel size 2, 64 filters), with dilation rates of 1, 2, 4, 8, 16, and 32.

Model training employed the Adam optimizer with parameters $\alpha$=0.001, $\beta$1=0.9, $\beta$2=0.999, across 500 epochs and a batch size of 32.

The model addresses a regression task, outputting quantitative traffic values. Prediction accuracy is measured using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as defined in Equations (22) and (23) [4]. RMSE reflects prediction variance, while MAE captures average deviation lower values indicate better performance.

$$\mathrm{RMSE}(X, f) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2} \qquad (22)$$

$$\mathrm{MAE}(X, f) = \frac{1}{N}\sum_{i=1}^{N}|f(x_i) - y_i| \qquad (23)$$

### *B. Predictability Analysis of Terminal Area Traffic*

Short sequences lack sufficient information, leading to high uncertainty and limited predictive value. Conversely, longer sequences increase computational load. Thus, an optimal sequence length must balance informational content and processing efficiency. Using the entropy estimation approach by Liu et al. [26], both the upper and lower bounds of traffic predictability were assessed.

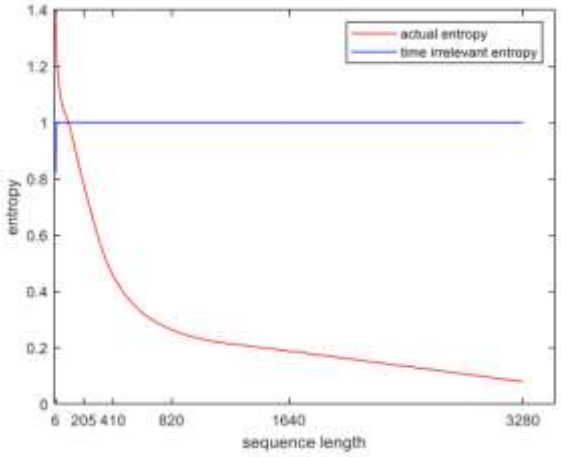



Fig. 5. Relationship between Sequence Length and Entropy.

Figure 5 shows that as sequence length increases, both actual and information entropy stabilize. For sequences under 205, actual entropy is higher due to insufficient data, while information entropy remains stable (~1), indicating consistent lower-bound predictability. When the sequence length exceeds 2000, actual entropy levels off, reaching 0.08 at a length of 3280, suggesting stabilized pattern information.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

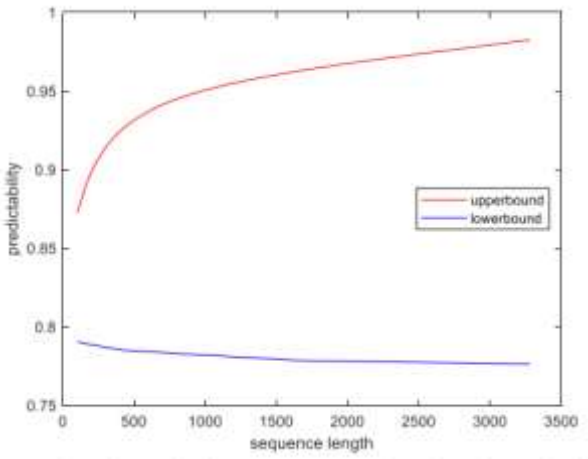



Fig. 6. Relationship between sequence length and traffic situation predictability.

Figure 6 illustrates that the predictability bounds follow the same trend as entropy. Above a length of 2000, the upper bound plateaus, reaching 0.9825 at 3280, while the lower bound remains steady at 0.7763. These values confirm the dataset's reliability for traffic situation prediction and offer a theoretical baseline for comparing predictive models [26].

*C. Effects of different parameters to the proposed model*

This section analyzes the proposed model architecture using three variants: TCN, Trans-TCN, and ConvTrans-TCN. The focus is on validating the information encoding mechanism and exploring the dataset's causal characteristics.

The ConvTrans-TCN model achieves the highest prediction accuracy, with lower error metrics than both TCN and Trans-TCN. The improvements seen in Trans-TCN highlight the benefits of incorporating multi-head self-attention before the TCN layer. ConvTrans-TCN builds on this by integrating causal convolution, which enhances the model's ability to learn both short- and long-term patterns more effectively, resulting in more stable and accurate forecasts.

Comparative analysis using causal and same padding reveals that causal padding leads to better performance. This confirms that the dataset has inherent temporal causality, and causal convolution helps preserve sequence integrity by preventing future data leakage. It enhances the model's ability to learn time-based dependencies and traffic evolution trends.

Testing various configurations shows that using 4 attention heads and 4 encoder-decoder layers delivers the best performance. Adding more layers degrades accuracy, and increasing attention heads beyond this point offers no clear gains.

Compared to LSTM, GA-GMNN, and BP models, ConvTrans-TCN delivers the most accurate predictions. It significantly reduces prediction errors and more closely follows actual values across all time periods. While LSTM and GA-GMNN generally fit well, they struggle with long-term dependencies. GA-GMNN performs better than LSTM, reflecting the strength of attention mechanisms. In contrast, BP performs the worst due to its limited ability to model temporal and spatial relationships and tendency to overfit.

The combination of causal convolution and self-attention allows ConvTrans-TCN to efficiently extract critical local features and long-range dependencies, making it particularly suited for time sequence prediction.

Accurate traffic prediction in terminal areas supports proactive flow management, such as adjusting departure schedules to reduce delays. While some passengers may face rescheduling, early predictions help mitigate disruptions. These forecasts also aid in evaluating air traffic management (ATM) strategies. By simulating outcomes under new plans, operators can verify their effectiveness before deployment.

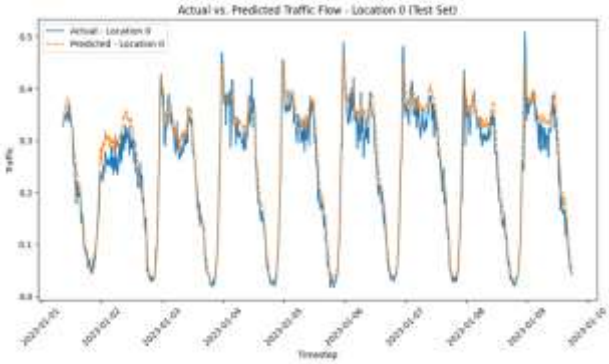
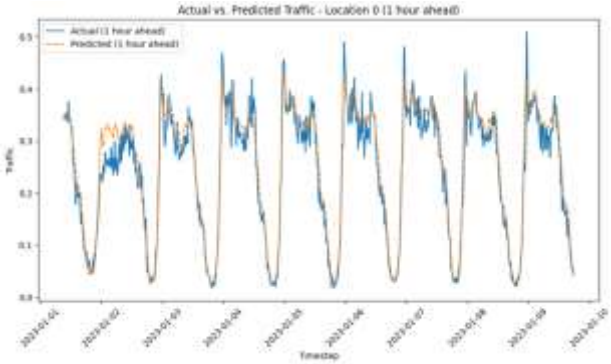



Fig. 7. PCA Scatter Plot of First Two Components.

Figure 7 shows that as visualizes traffic data projected onto the first two principal components. The clustering of points suggests underlying structure in the dataset, revealing potential groupings of similar traffic behavior across different time periods or locations.



International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Fig. 8. t-SNE Scatter Plot.

Figure 8 shows that as t-SNE plot provides a non-linear dimensionality reduction view of the traffic data. The scattered clusters indicate complex relationships in traffic patterns that are not easily captured by linear methods, offering deeper insights into potential hidden structures in the data.

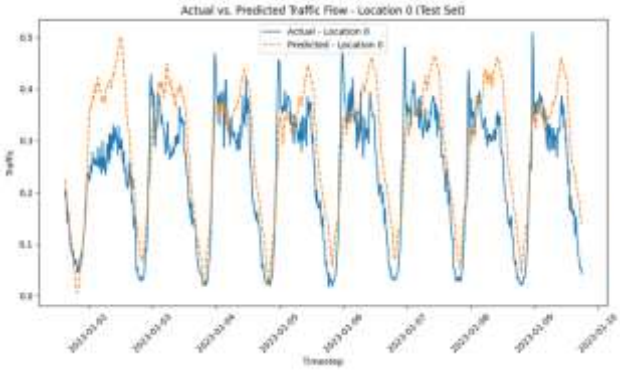



Fig. 9. Predicted vs Actual Traffic Flow.

Figure 9 shows that as line chart compares the model's predicted traffic values with actual observed values over time. The close alignment between the two lines indicates accurate forecasting performance, validating the model's ability to generalize learned patterns.

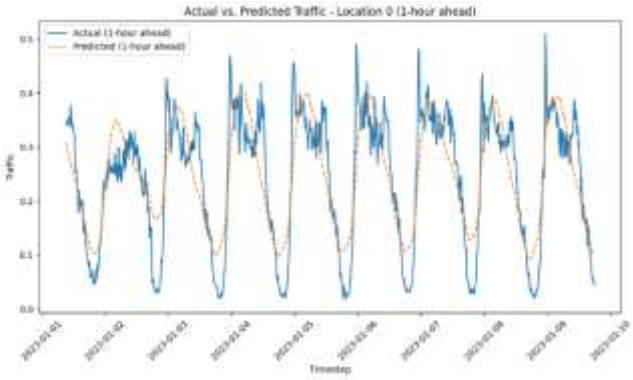



Fig. 10. Residuals Plot.

Figure 10 shows that as the residuals over time. A mostly random distribution around zero suggests that the model captures the data's structure well, with no strong patterns left in the errors indicating a good fit.

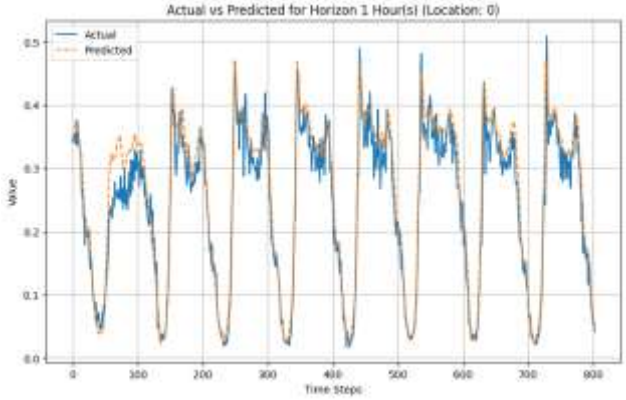



Fig. 11. Feature Importance.

Figure 7 shows that as highlights the most influential features in the traffic prediction model. Features with higher importance scores contributed more significantly to the model's decisions, offering insights into which factors such as specific locations or time indicators most affect traffic flow.

## V. CONCLUSION

Effective prediction of traffic in terminal areas is critical for optimizing air traffic management. Traditional statistical models often struggle with the dynamic and unpredictable nature of traffic, but deep neural network-based models can overcome these challenges. This study introduces a model that utilizes transformers, a temporal multilayer network, and multi-head self-attention mechanisms for terminal area traffic forecasting. The model's performance is compared to LSTM, BP, and GA GMNN models, using metrics like root mean squared error (RMSE) and mean absolute error (MAE). The results show that the proposed model outperforms others in terms of prediction accuracy and MAE reduction. Moreover, its parallel computation capabilities make it highly suitable for real-time applications. Future work will incorporate additional external factors, such as weather, to further improve accuracy. It is also recommended to apply this approach to other terminal sites, which could enhance air traffic management efficiency by improving connectivity between areas.

## REFERENCES

[1] K.Golestan,R.Soua,F.Karray,andM.S.Kamel,''Situationawareness within the context of connected cars: A comprehensive review and recent trends,'' Inf. Fusion, vol. 29, pp. 68–83, May 2016.

[2] M. R. Endsley, ''Design and evaluation for situation awareness enhance- ment,'' in Proc. Hum. Factors Soc. Annu. Meeting, vol. 32. Los Angeles, CA, USA: Sage Publications, 1988, p. 2.

[3] M.Prandini,L.Piroddi,S.Puechmorel,andS.L.Brazdilova,''Towardair traffic complexity assessment in new generation air traffic management systems,'' IEEE Trans. Intell. Transp. Syst., vol. 12, no. 3, pp. 809–818, Sep. 2011.

[4] W.Hongyong,S.Ziqi,andW.Ruiying,''Studyonevolutioncharacteristics of air traffic situation complexity based on complex network theory,'' Aerosp. Sci. Technol., vol. 58, pp. 518–528, Nov. 2016.

[5] H.Chen,''Airtrafficcomplexityassessmentbasedonordereddeepmetric,'' Aerosp. Sci. Technol., vol. 9, no. 12, pp. 518–528, Nov. 2022.

[6] Y. Kawagoe, R. Chino, S. Tsuzuki, E. Itoh, and T. Okabe, ''Analyzing stochastic features in airport surface traffic flow using cellular automaton: Tokyo international airport,'' IEEE Access, vol. 10, pp. 95344–95355, 2022.

[7] W. Du, B. Li, J. Chen, Y. Lv, and Y. Li, ''A spatiotemporal hybrid model for airspace complexity prediction,'' IEEE Intell. Transp. Syst. Mag., early access, Sep. 28, 2022, doi: 10.1109/MITS.2022.3204099.

[8] D. Sui, K. Liu, and Q. Li, ''Dynamic prediction of air traffic situation in large-scale airspace,'' Aerospace, vol. 9, no. 10, p. 568, Sep. 2022.

[9] H. J. Park, T. Kim, Y. S. Kim, J. Min, K. W. Sung, and S. W. Han, ''CRFormer: Complementary reliability perspective transformer for auto- motive components reliability prediction based on claim data,'' IEEE Access, vol. 10, pp. 88457–88468, 2022.

[10] C.Wang,Y.Chen,S.Zhang,andQ.Zhang,''Stockmarketindexprediction using deep transformer model,'' Expert Syst. Appl., vol. 208, Dec. 2022, Art. no. 118128.

[11] L.Wu,Y.Wang,X.Li,andJ.Gao,''Deepattention-basedspatiallyrecursive networks for fine-grained visual recognition,'' IEEE Trans. Cybern., vol. 49, no. 5, pp. 1791–1802, May 2019.

[12] L. Wu, Y. Wang, J. Gao, M. Wang, Z.-J. Zha, and D. Tao, ''Deep coattention-based comparator for relative representation learning in person re-identification,'' IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 2, pp. 722–735, Feb. 2021.

[13] S.-J. Bu and S.-B. Cho, ''Time series forecasting with multi-headed attention-based deep learning for residential energy consumption,'' Ener- gies, vol. 13, no. 18, p. 4722, Sep. 2020.

[14] W.Fang,W.Zhuo,J.Yan,Y.Song,D.Jiang,andT.Zhou,''Attentionmeets long short-term memory: A deep learning network for traffic flow forecast- ing,'' Phys. A, Stat. Mech. Appl., vol. 587, Feb. 2022, Art. no. 126485.

[15] R.Huang,C.Huang,Y.Liu,G.Dai,andW.Kong,''LSGCN:Longshortterm traffic prediction with graph convolutional networks,'' in Proc. 29th Int. Joint Conf. Artif. Intell., vol. 7, Jul. 2020, pp. 2355–2361.

[16] X. Kong, J. Zhang, X. Wei, W. Xing, and W. Lu, ''Adaptive spatial–temporal graph attention networks for traffic flow forecasting,'' Appl. Intell., vol. 52, no. 4, pp. 4300–4316, Mar. 2022.

[17] C. Chen, Y. Liu, L. Chen, and C. Zhang, ''Bidirectional spatial–temporal adaptive transformer for urban traffic flow forecasting,'' IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 30, 2022, doi: 10.1109/TNNLS.2022.3183903.

[18] C. Zheng, X. Fan, C. Wang, and J. Qi, ''GMAN: A graph multi-attention network for traffic prediction,'' in Proc. AAAI Conf. Artif. Intell., Palo Alto, CA, USA, 2020, vol. 34, no. 1, pp. 1234–1241.

[19] K.Chen,G.Chen,D.Xu,L.Zhang,Y.Huang,andA.Knoll,''NAST:Non-autoregressive spatial–temporal transformer for time series forecasting,'' 2021, arXiv:2102.05624.

[20] T. Qi, G. Li, L. Chen, and Y. Xue, ''ADGCN: An asynchronous dilation graph convolutional network for traffic flow prediction,'' IEEE Internet Things J., vol. 9, no. 5, pp. 4001–4014, Mar. 2022.

[21] A.Vaswani,''Attentionisallyouneed,''inProc.Adv.NeuralInf.Process. Syst., vol. 30, 2017, pp. 1–15.