A Project Report

On

**Forecasting Metro Rail Ridership: A data-driven approach utilizing historical data**

BY

**Nikhitha Punati**

**SE21UARI100**

**Vismay C V**

**SE21UCSE250**

Under the supervision of

Dr. Raja Rao Tripuraneni

**SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR**

**DEGREE OF BACHELOR OF TECHNOLOGY**

**PR 301: PROJECT TYPE COURSE**



**ECOLE CENTRALE SCHOOL OF ENGINEERING**

**MAHINDRA UNIVERSITY**

**HYDERABAD**

**(June 2024**)

CONTENTS

Acknowledgements……………………………………………………..1

Certificate…………………………………………………………….....2

Title page………………………………………………………….…….3

Abstract………………………………………………………………....4

1.Chapter I……………………………………………….…………......5

2.Chapter II……………………………………………………………..9

3.Chapter III.…………………………………………………………….13

4.Chapter IV……………………………………………………………17

Conclusion……………………………………………………………..25

References……………………………………………………………...26

ACKNOWLEDGMENTS

**This project work would not have been complete without the mention of the following. We express our hearty gratitude to Mahindra University for providing us with the opportunity and platform to work on the project and our project mentor Dr. Raja Rao Tripuraneni has supported and guided us throughout our project.**

**1**



**Ecole Centrale School of Engineering**

**Mahindra University**

**Hyderabad**

Certificate

This is to certify that the project report entitled “**Forecasting Metro Rail Ridership: A data-driven approach utilizing historical data”** submitted by Ms. Nikhitha Punati (SE21UARI100) and Mr. Vismay C V (SE21UCSE250) in partial fulfillment of the requirements of the course PR 301, Project Course, embodies the work done by him/her under my supervision and guidance.

**(Dr. Raja Rao Tripuraneni& Signature)**

Ecole Centrale School of Engineering, Hyderabad.

Date: 07 June 2024

**2**

A train on the tracks

Description automatically generated**Forecasting Metro Rail Ridership: A data-driven approaching utilizing historical data historical data**

**3**

ABSTRACT

This research endeavor seeks to develop a robust model capable of accurately predict metro rail passenger flows by applying advanced analytics and machine learning techniques, enabling metro authorities to optimize operations, enhance passenger experience, and make informed decisions regarding infrastructure and resource allocation. The central premise of this study is to leverage historical data to assess and anticipate metro rail activity, providing valuable insights for efficient management and planning.

The proposed methodology involves the integration of diverse data sources, including real-time sensor readings, historical ridership patterns, demographic information, and external factors such as weather conditions, events, and socioeconomic indicators. Through the application of advanced analytics and machine learning algorithms, the research team aims to uncover hidden patterns, identify key drivers of passenger demand, and construct a predictive model that can reliably forecast future passenger flows. The development of such a model will empower metro operators to make proactive adjustments to service schedules, resource allocation, and infrastructure maintenance, thereby improving overall system efficiency, reducing operational costs, and enhancing the commuter experience.

Furthermore, the insights gained from this research can inform strategic planning decisions, such as the expansion of metro networks, the optimization of station layouts, and the allocation of resources to address evolving transportation needs.

**4**

**CHAPTER-I**

**5**

**Introduction**

Imagine a Metro System That Runs Like Clockwork: Optimizing Operations and Passenger Experience with AI-powered Traffic Prediction. We propose a revolutionary approach that leverages the power of Artificial Intelligence (AI) and historical data to predict metro traffic with exceptional accuracy. Imagine a system that can:

Forecast passenger flow: By analyzing historical ridership data, station usage patterns, and even external factors like weather and events, our model can predict how many passengers will be at each station at any given time.

Optimize train scheduling: Armed with these insights, metro authorities can dynamically adjust train schedules to match predicted demand. This reduces overcrowding, minimizes waiting times, and ensures a smoother flow of passengers throughout the network.

Enhance passenger experience: Real-time information on wait times and platform congestion empowers passengers to plan their journeys effectively. This reduces stress and frustration, making metro travel a more pleasant choice.

Informed decision-making: Metro authorities gain valuable insights into ridership patterns and infrastructure needs. This allows for data-driven resource allocation, targeted maintenance efforts, and informed planning for future expansion.

**Real-World Application:**

This project holds immense potential for congested metro systems. Here's how it translates to tangible benefits:

Reduced Operational Costs: Efficient scheduling and resource allocation minimize unnecessary train runs, leading to cost savings for metro authorities.

Increased Passenger Satisfaction: Shorter wait times, less crowded platforms, and improved predictability translate to a happier and more loyal ridership.

Smarter Infrastructure Development: Data-driven insights guide future expansion plans, ensuring infrastructure investments cater to actual ridership patterns.

Sustainable Urban Mobility: Improved efficiency encourages commuters to choose metro over personal vehicles, reducing traffic congestion and pollution in cities.

**6**

**The Future of Metro Travel:**

This project represents a significant step towards a future where metro systems are intelligent, adaptable, and passenger-centric. By harnessing the power of AI and historical data, we can transform the daily commute into a seamless and stress-free experience for millions of urban citizens.

**Current Challenges:**

Real-time metro traffic monitoring is expensive, labor-intensive, and relies on non-scalable sensor technologies.

Existing methods lack accuracy due to factors like passenger behavior, train operations, and passenger access routes.

Traditional methods don't capture the dynamics of waiting times, headways (train arrival intervals), and delays, which significantly impact passenger experience.

**Objective:**

Develop a system for accurate metro rail traffic estimation and prediction using readily available historical data from existing network infrastructure (fiber optic network, VDU systems, and centralized traffic management systems).

**Benefits:**

Improved operational efficiency through better traffic prediction for scheduling and resource allocation.

Reduced passenger waiting times and congestion by proactively managing traffic flow.

Enhanced passenger experience through real-time information on headways and wait times.

Informed decision-making for metro rail system maintenance, development, and optimization.

Improved planning for abnormal conditions like system failures, evacuations, and high-demand events.

**7**

Focus Areas:

Leverage historical data to capture patterns and trends in metro rail traffic.

Develop a model that accurately predicts traffic volume, headways, waiting times, and delays.

Integrate the prediction model with existing systems for real-time information dissemination to passengers.

Account for factors like passenger behavior, train operations, and access routes in the prediction model.

Expected Outcome:

A cost-effective and scalable system for metro rail traffic estimation and prediction, leading to improved operational efficiency, passenger satisfaction, and informed decision-making for metro rail authorities.

**8**

**CHAPTER-II**

**9**

**PROBLEM DEFINITION**

This project aims to develop an AI-powered model that predicts metro traffic patterns using historical data. This will answer crucial questions for metro authorities, leading to better decision-making and improved passenger experience. Here are the key questions our project will address:

1. **How many passengers will be at each station at a given time?** This involves predicting passenger flow throughout the metro network, considering factors like time of day, day of the week, special events, and weather conditions.
2. **What is the optimal train schedule to minimize wait times and congestion?** By knowing the predicted passenger flow, we can recommend train schedules that match demand, reducing overcrowding and wait times at stations.
3. **How can we leverage passenger flow predictions to optimize current metro management?** The deployment of station staff, such as ticket agents, security personnel, and cleaning crews can be made according to the data obtained through our predictions
4. **How can we improve passenger experience through real-time information?** The model's predictions can be used to provide passengers with real-time information on wait times, platform congestion, and potential delays. This empowers them to plan their journeys effectively and avoid crowded trains or stations.
5. **What insights can be gained about ridership patterns to inform future planning?** The model can analyse historical and predicted data to identify trends in ridership patterns. This knowledge helps metro authorities make informed decisions regarding infrastructure development, resource allocation, and expansion plans.

 **Passenger Flow Prediction:**

* We aim to predict the number of passengers entering and exiting each station at any given time (t). This can be represented as P(s, t), where:
  + P: Predicted number of passengers
  + s: Specific station within the metro network
  + t: Time (e.g., hour, day)
* Factors influencing P(s, t) can include:
  + Historical ridership data (e.g., daily, weekly, monthly patterns) for station s
  + Time-based factors (e.g., weekdays vs. weekends, rush hour vs. off-peak hours)
  + External factors (e.g., weather conditions, holidays, special events)

 **Optimal Train Schedule:**

* Given predicted passenger flow (P(s, t)), we want to determine the most efficient train schedule to minimize wait times and congestion. This involves:
  + Train arrival frequency at each station (considering passenger volume)
  + Train capacity (number of passengers a train can hold)
  + Platform capacity (maximum number of passengers a platform can safely accommodate)

**10**

* The mathematical formulation for optimal scheduling can involve optimization algorithms that consider factors like:
  + Minimizing the difference between predicted passenger flow (P(s, t)) and train capacity
  + Minimizing platform wait times based on predicted passenger arrivals and departures

 **Real-Time Information for Passengers:**

* Based on the prediction model's output (P(s, t) and potentially real-time sensor data), we can provide passengers with real-time information such as:
  + Estimated wait times at platforms
  + Predicted platform congestion levels
  + Potential delays and alternative routes
* This information can be displayed on station screens, mobile apps, or announcements, empowering passengers to make informed decisions about their journeys.

 **Insights for Future Planning:**

* By analysing historical and predicted data (P(s, t)), we can identify trends in ridership patterns. This includes:
  + Identifying peak hours and stations with consistently high passenger volume
  + Analysing the impact of external factors (e.g., Sporting event near a station)
* These insights can be translated into mathematical models for:
  + Infrastructure development (e.g., platform expansion, additional train carriages)
  + Resource allocation for maintenance and staff deployment
  + Long-term planning for network expansion based on predicted ridership growth

**Mathematical Representation:**

Here's a general mathematical framework to illustrate the concept:

**Prediction Model:** f(x) = y

* **f(x):** Represents the prediction model that takes various factors (x) as input. These factors can include historical ridership data (e.g., time series data for passenger entries/exits at each station), external factors (e.g., weather, events), and potentially even real-time data (e.g., sensor readings).
* **y:** Represents the predicted outcome, which can be passenger flow at a specific station and time, optimal train schedule, or real-time wait time information.

**11**

**1. Time Series Analysis:**

* **Applications:** Identifying trends, seasonality, and patterns in historical ridership data (e.g., daily, weekly, monthly variations) for each station. This helps predict future passenger flow (P(s, t)) based on past behaviour.
* **Example:** By analysing historical data for a specific station (s), we might identify a strong seasonal pattern (e.g., higher ridership on weekdays during rush hour). This can be used to predict higher P(s, t) values during those times.

**2. Machine Learning Techniques:**

* **Algorithms:**

**Neural Networks:** Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM)

* **Applications:** These algorithms can learn complex relationships between various factors like historical ridership data (P(s, t) for different times), time-based factors (weekdays, holidays), and external factors (weather, events). The model learns to predict P(s, t) for future scenarios based on these learned relationships.
* **Example:** An LSTM network could be trained on historical data for all stations (s) and various time periods (t). It would learn complex relationships between factors and predict P(s, t) for upcoming hours, considering daily and seasonal variations as well as potential external influences.

**12**

**CHAPTER-III**

**13**

**BACKGROUND AND RELATED WORK**

We referred to several research papers related to do this and analyzed them to improve this project. Thus we came up with the following reports for their research papers before we came up with what models we are going to use for this project.

Research paper 1: <https://www.researchgate.net/publication/309755490_A_Study_on_Traffic_Forecast_for_Metro_Railway_Expansion_in_Chennai>

Report:

The work wiped out the paper "A Consider on Activity Estimate for Metro Railroad Development in Chennai" includes a point by point examination of the achievability and request for metro rail extension in Chennai, especially centering on the Porur to Kamarajar Salai corridor. The creators conducted a web study utilizing Google shapes to gather information on travel characteristics such as trip reason, length, taken a toll, and mode. They moreover analyzed populace and vehicle development patterns to venture future transportation needs.

With respect to the utilize of neural systems, the paper notices the application of neural organize methods in determining short-term traveler stream in high-speed railroad frameworks and the utilize of a three-phase back-propagation neural network approach to estimate short-term railroad traveler request. Be that as it may, it does not indicate the precise show of neural systems utilized in their ponder.

The shortcomings in the ideas proposed in the paper include:

1. Limited Scope: The study is focused on a specific corridor in Chennai, which may limit the generalizability of the findings to other areas or cities.

2. Data Collection Method: The online survey method may not reach all potential metro users, potentially leading to a biased sample.

3. Model Specificity: The paper does not provide detailed information on the neural network model used, which makes it difficult to assess the robustness and accuracy of the forecasting methods.

**14**

4. Infrastructure and Operational Considerations: While the paper discusses the demand and feasibility of metro rail, it does not delve into the infrastructure requirements, operational challenges, or the cost-benefit analysis of implementing such a system.

5. Policy and Planning Implications: The paper does not discuss the policy implications or the planning process required to integrate metro rail into the existing transportation network effectively.

6. Dynamic Changes: The study may not account for dynamic changes in travel behavior, technological advancements, or shifts in economic conditions that could affect the demand for metro rail in the future.

Overall, while the paper provides valuable insights into the potential for metro rail expansion in Chennai, it could benefit from a more comprehensive approach that includes detailed modeling, broader data collection, and consideration of various factors that could influence the success of such a project.

Research Paper 2:

<https://www.researchgate.net/publication/368159069_Forecasting_metro_rail_transit_passenger_flow_with_multiple-attention_deep_neural_networks_and_surrounding_vehicle_detection_devices>

Report:

The paper presents a consider on determining traveler stream in metro rail travel (MRT) frameworks employing a profound learning show called the Multiple-Attention Profound Neural Organize (MADNN). The MADNN demonstrate coordinating authentic MRT traveler stream information with information from encompassing vehicle discovery (VD) gadgets to upgrade expectation exactness.

**15**

The show highlights three consideration layers:  
the MRT Consideration Layer (MRT-AL), the Encompassing VD Consideration Layer (SVD-AL), and the MRT-SVD Consideration Layer (MRT-SVD-AL), which together produce covered up highlights and consideration weights for MRT stations and VD devices.The neural organize demonstrate used in this inquire about could be a profound learning engineering that combines Repetitive Neural Systems (RNNs), particularly Long Short-Term Memory (LSTM) systems, with consideration components. LSTM systems are capable at taking care of time-series information and can capture long-term conditions, which are significant for foreseeing passenger flow. The consideration components permit the model to center on the foremost important highlights and time focuses within the information.  
  
The investigate was approved utilizing information from the Taipei MRT framework in Taiwan, and the comes about appeared that the MADNN demonstrate outperformed other well-known models like ARIMA, CNN\_2D, CNN\_3D, GCN, and T-GCN in terms of forecast precision.

However, the study does have some limitations:

1. The model could be improved by incorporating MRT route information, as passenger flow at a station is influenced by the flow from previous stations on the same route.

2. Time-based elements such as monthly or weekly cycles and workday schedules were not included in the analysis, which could further enhance the model's performance.

**16**

**CHAPTER-IV**

**17**

**IMPLEMENTATION**

For this project the following models are used and the following explanation is for how we incorporated those models:

## **1)Why LSTMs are Ideal for Metro Rail Traffic Prediction**

Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN), are particularly well-suited for metro rail traffic prediction due to their ability to handle:

* **Sequential Data:** Metro ridership data is inherently sequential, with patterns unfolding over time (e.g., daily rush hour, weekend fluctuations). LSTMs excel at capturing these temporal dependencies within the data.
* **Long-Term Dependencies:** Unlike traditional RNNs, LSTMs can learn long-term dependencies between data points, crucial for capturing the complex relationships between historical ridership (past weeks/months) and future passenger flow (P(s, t)).
* **Variable Factors:** LSTMs can effectively handle the influence of various factors on ridership. These can include time-based features (weekdays vs. weekends), external factors (weather, events), and potentially even real-time sensor data.

## Incorporating LSTM Model in Metro Rail Traffic Prediction

Here's how an LSTM model can be incorporated into this project:

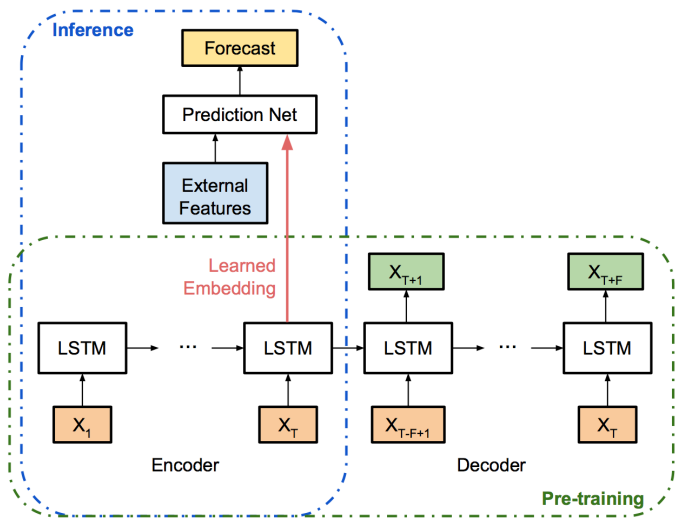
1. **Data Preprocessing:**
   * We can Collect historical ridership data (entry/exit counts) for each station (s) at different time intervals (t). This forms the core input (P(s, t)) for the model.
   * Gather data on relevant factors like time of day, day of the week, holidays, and weather conditions.
   * Clean and pre-process the data: handle missing values, normalize numerical features, and potentially create new features based on domain knowledge (e.g., hourly ridership for weekdays).
2. **Model Architecture:**
   * We can Design an LSTM model with an input layer that accepts the preprocessed data (P(s, t) and other factors).
   * Utilize one or multiple LSTM layers to capture temporal dependencies and relationships within the data. The number of layers and hidden units within each layer can be optimized through experimentation.
   * Implement a dense output layer with a single neuron to predict the passenger flow (P(s, t)) for a future time period (e.g., next hour, next day).

**18**

1. **Model Training:**
   * Split the preprocessed data into training, validation, and testing sets.
   * Train the LSTM model on the training data, allowing it to learn the patterns and relationships between input factors and passenger flow (P(s, t)).
   * Use the validation set to fine-tune hyperparameters (learning rate, number of epochs) to optimize model performance.
2. **Evaluation and Deployment:**
   * Evaluate the model's performance on the testing set using metrics like Mean Squared Error (MSE) or R-squared to assess prediction accuracy.
   * If satisfied with the results, deploy the trained model to predict future passenger flow (P(s, t)) for a specified station (s) and time (t).

## Benefits of LSTM Model

* **Accurate Predictions:** LSTMs can potentially achieve high accuracy in predicting passenger flow (P(s, t)), leading to better scheduling and resource allocation.
* **Adaptability:** The model can be continuously retrained with new data, adapting to changing ridership patterns and external factors over time.
* **Scalability:** The architecture can be scaled by adding more layers or increasing the number of hidden units within each layer to handle more complex data or larger network sizes.

Architecture of LSTM for time series forecasting.

**19**

## **2)** **Why TCN is a Strong Candidate for Metro Rail Traffic Prediction**

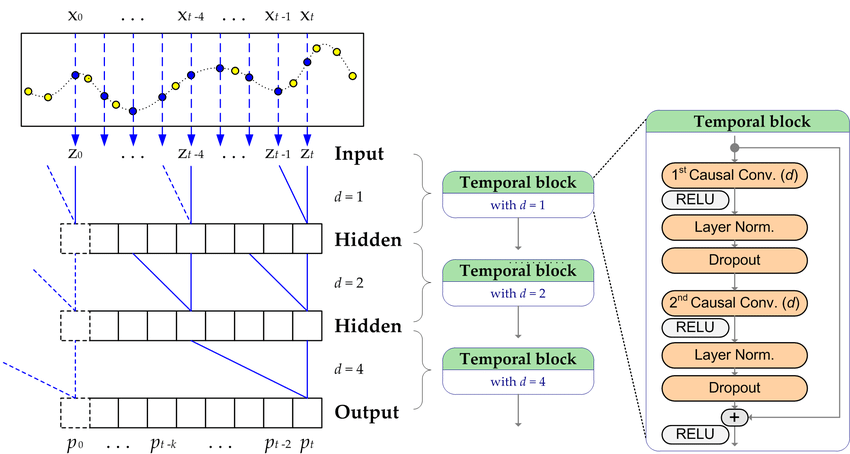
In our metro rail traffic prediction project, a Temporal Convolutional Network (TCN) emerges as a compelling choice due to its strengths in handling sequential data like historical ridership information. Here's a breakdown of why TCN excels in this context:

* **Modeling Temporal Dependencies:** TCNs are specifically designed to capture long-term dependencies within sequences. This is crucial for this project, as ridership patterns often exhibit recurring trends over time (e.g., daily rush hour peaks, weekly variations). TCN's architecture effectively learns these dependencies from historical data, leading to more accurate predictions of future passenger flow (P(s, t)) at each station.
* **Long-Range Dependencies:** Unlike standard Convolutional Neural Networks (CNNs) that struggle with capturing long-term dependencies in sequences, TCNs address this limitation. They incorporate dilated causal convolutions, which allow the network to "see" further back in the sequence without information leakage (future data influencing past predictions). This is essential for capturing the nuances of ridership patterns that might unfold over longer timeframes.
* **Efficiency:** TCNs offer a computationally efficient way to model temporal dependencies compared to recurrent architectures like LSTMs. This can be advantageous for handling large datasets of historical ridership data, especially when dealing with extensive metro networks with many stations.

## Incorporating TCN into the Project

Here's a breakdown of how we could incorporate a TCN model into this project:

1. **Data Preprocessing:**
   * Prepare our historical ridership data (P(s, t)) for each station. This might involve cleaning, normalization, and potentially feature engineering to create additional features relevant to prediction (e.g., day of the week, time of day, weather data).
2. **TCN Model Architecture:**
   * Design a TCN model with the following components:
     + **Input Layer:** This layer takes the preprocessed sequence data (P(s, t)) for a specific station (s) and time window as input.
     + **Dilated Causal Convolutional Layers:** Stack multiple dilated causal convolutional layers to extract temporal features from the sequence. Each layer uses filters with dilated kernels to capture dependencies at different time scales.
     + **Activation Function:** Apply a non-linear activation function (e.g., ReLU) after each convolutional layer to introduce non-linearity and improve model capacity.
     + **Output Layer:** The final layer predicts the future passenger flow (P(s, t)) for a specific station and time frame. This could be a single value or a sequence of predicted values depending on the desired prediction horizon **20**
3. **Model Training:**
   * Train the TCN model on a portion of our historical ridership data. This involves feeding the model with sequences of past passenger flow data (P(s, t)) and the corresponding actual passenger flow values for the target time period. The model learns to map the input sequences to the desired outputs through backpropagation and optimization algorithms.
4. **Model Evaluation and Prediction:**
   * Evaluate the trained TCN model's performance on a separate validation dataset. Use metrics like Mean Squared Error (MSE) or R-squared to assess prediction accuracy.
   * Once satisfied with the model's performance, use it to predict future passenger flow (P(s, t)) for various stations and timeframes based on unseen historical data.



Architecture of TCN model for time series prediction.

**21**

## **3)** Why Transformers are Well-Suited for Metro Rail Traffic Prediction

While various algorithms can be used for metro rail traffic prediction, Transformer-based models offer several advantages that make them a compelling choice for this project:

* **Long-Term Dependencies:** Unlike traditional RNNs, Transformers can effectively capture long-term dependencies within the data. This is crucial for metro traffic prediction, as factors like seasonality and special events can have a significant impact on ridership patterns even weeks or months later.
* **Parallel Processing:** Transformers leverage parallel processing architectures, making them computationally efficient for handling large datasets of historical ridership data and external factors.
* **Multi-Head Attention:** The core mechanism of Transformers, multi-head attention, allows the model to focus on specific aspects of the input data relevant to the prediction task. This is particularly beneficial for considering various factors influencing passenger flow, like time of day, station location, and weather conditions.
* **Ability to Handle Multiple Input Types:** Transformers can effectively integrate data from diverse sources. This is advantageous for incorporating not only historical ridership data but also external factors like weather patterns, event schedules, and even social media sentiment (related to public transportation disruptions).

## Incorporating a Transformer Model in the Project

Here's how a Transformer model can be integrated into our metro rail traffic prediction project:

**1. Data Preprocessing:**

* Prepare historical ridership data (passenger entries/exits) for each station at different time intervals (e.g., hourly, daily).
* Include data on external factors like weather, holidays, and events.
* Preprocess the data by handling missing values, normalization, and potentially encoding categorical features.

**2. Model Architecture:**

* **Encoder-Decoder Structure:** Utilize an encoder-decoder architecture. The encoder processes the historical ridership data and external factors, capturing the underlying patterns and relationships. The decoder then uses this encoded information to predict the future passenger flow (P(s, t)) for a specific station and time.
* **Transformer Blocks:** Within the encoder and decoder, employ multiple Transformer blocks with multi-head attention layers. These layers allow the model to focus on relevant aspects of the historical data and external factors for accurate prediction.

**22**

* **Positional Encoding:** Since Transformers lack inherent understanding of sequential order, incorporate positional encoding to represent the temporal relationship between data points in the historical ridership data.

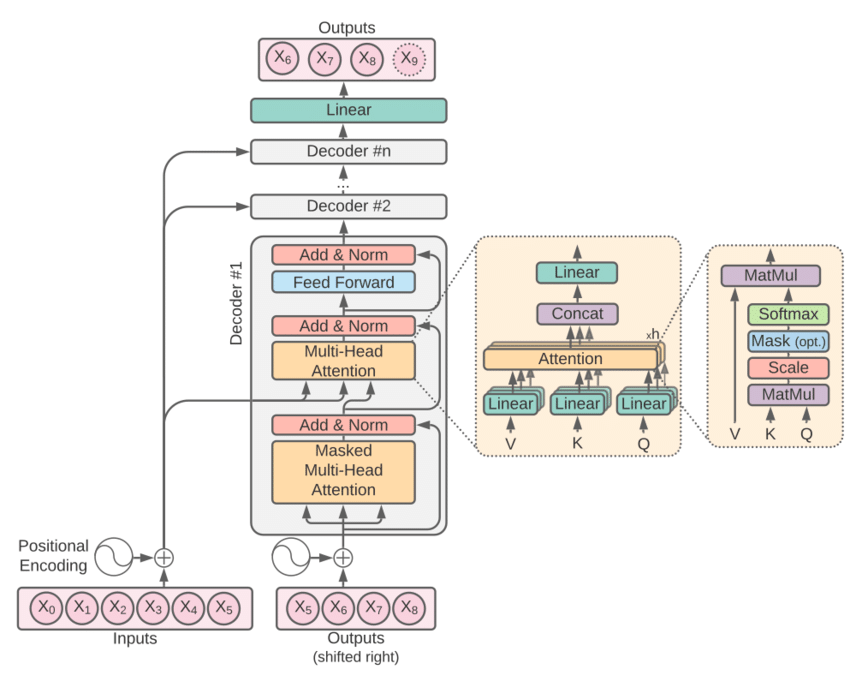
**3. Training and Evaluation:**

* Train the Transformer model on a large portion of the historical data, splitting it into training, validation, and testing sets.
* Use appropriate loss functions (e.g., Mean Squared Error) to measure the difference between predicted and actual passenger flow (P(s, t)).
* Monitor the model's performance on the validation set during training, adjusting hyperparameters as needed to optimize prediction accuracy.
* Evaluate the final model's performance on the testing set using metrics like R-squared or Mean Absolute Error (MAE).

**4. Integration and Deployment:**

* Integrate the trained Transformer model into the metro rail system's infrastructure. This might involve connecting it to data feeds for real-time updates on external factors.
* Utilize the model's predictions (P(s, t)) to:
  + Generate real-time information for passengers regarding wait times and platform congestion.
  + Assist in creating dynamic train schedules that optimize resource allocation and minimize wait times.

**23**



Architecture of Transformer based model for time series prediction.

**24**

**CONCLUSION**

These are the different models which can be incorporated into the metro rail ridership forecasting system which is driven by historical data of the passenger flow patterns. The suggested models aim to optimize operations, enhance passenger experience and help the metro transport management make informed decisions.

**25**

**REFERENCES**

Reference links:

<https://www.researchgate.net/publication/255710610_Critical_Issues_Related_to_Metro_Rail_Projects_in_India>

<https://www.researchgate.net/publication/309755490_A_Study_on_Traffic_Forecast_for_Metro_Railway_Expansion_in_Chennai>

<https://www.researchgate.net/publication/368159069_Forecasting_metro_rail_transit_passenger_flow_with_multiple-attention_deep_neural_networks_and_surrounding_vehicle_detection_devices>

<https://arxiv.org/pdf/1904.01280>

<https://journals.sagepub.com/doi/10.1177/03611981211037553>

**26**