SHORT TERM PASSENGER DEMAND FORECASTING USING DEEP LEARNING TECHNIQUES

Thesis

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF TECHNOLOGY in

TRANSPORTATION ENGINEERING

By

DUDAM AMRUTH

(212TS012)

DEPARTMENT OF CIVIL ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE -575025

June, 2022

SHORT TERM PASSENGER DEMAND FORECASTING USING DEEP LEARNING TECHNIQUES

Thesis

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF TECHNOLOGY in

TRANSPORTATION ENGINEERING

By

DUDAM AMRUTH

(212TS012)

Under the guidance of

Dr. RAVIRAJ H.M.

DEPARTMENT OF CIVIL ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE -575025

June, 2022

DECLARATION

I hereby declare that the Report of the P.G. Project Work (CV759/Major Project) report entitled **"SHORT TERM PASSENGER DEMAND FORECASTING USING DEEP LEARNING TECHNIQUES"** which is being submitted to the **National Institute of Technology Karnataka, Surathkal**, in partial fulfilment of the requirements for the award of the Degree of **Master of Technology** in **Transportation Engineering**, is a bonafide report of the work carried out by me. The material contained in this Report has not been submitted to any University or Institution for the award of any degree.

DUDAM AMRUTH
(212TS012)
Department of Civil Engineering

Place: NITK, SURATHKAL
Date: June, 2023

C E R T I F I C A T E

This is to certify that the Major Project (CV759) entitled **"SHORT TERM PASSENGER DEMAND FORECASTING USING DEEP LEARNING TECHNIQUES"** submitted by **DUDAM AMRUTH** (**Register No. 212TS012**) as a record of the work carried out by him, is accepted as the P.G. Major Project Report submission in partial fulfilment of the requirements for the award of degree of Master of Technology in Transportation Engineering in the Department of Civil Engineering, National Institute of Technology Karnataka, Surathkal during the academic year 2022-2023.

Dr. RAVIRAJ H. MULANGI
Assistant Professor
Department of Civil Engineering

Chairman – DPGC
Department of Civil Engineering
NITK Surathkal

ACKNOWLEDGEMENT

I would like to thank **Dr. B. R. Jayalekshmi**, Head of the Department, Civil Engineering for giving an opportunity to work with the Project. I would also like to thank my guide **Dr. Raviraj H. Mulangi.**, Professor, Department of Civil Engineering, NITK, Surathkal for his valuable guidance and suggestions, throughout the completion of the work.

I extend my gratitude to **Mr. Nithin K. S.,** Research Scholar**,** Department of Civil Engineering, National Institute of Technology, Karnataka, for the guidance and help throughout my project work.

I acknowledge the contributions of KSRTC, Udupi and Indian Meteorological Department, Bangalore for providing data required for the study.

I express my sincere gratitude to Heera George, Kasturidas Acharya and Vipin P.R M. Tech Transportation Engineering students, Department of Civil Engineering for all their help, support and motivation for the successful completion of this work.

I also extend my sincere thanks to each and every faculty member and friends who had helped me directly or indirectly, in the preparation of this report.

DUDAM AMRUTH

(212TS012)

ABSTRACT

Public bus transportation plays a very important role in urban transportation system. Passenger flow are of great significance to bus scheduling and route optimization. Passenger flow depends on lot of external factors. The close relationship between historical weather conditions, temporal characteristics and the corresponding passenger flow has been widely analyzed by researchers. This study aims to explore the issue of how to use temporal characteristics and historical weather data to make passenger flow forecasting more accurate. Emerging deep learning models provide a good insight into improving prediction precision. This study aims to combine the modelling skills of deep learning architectures and the domain knowledge in transportation into prediction of public bus passenger flow. To this end, an hourly bus passenger flow forecasting model using a deep long short-term memory neural network (RPTW-LSTM) was developed.

The optimized traditional input variables, including the different temporal data and historical passenger flow data, were combined with weather variables for data modeling. A comprehensive analysis of the weather impacts on short-term bus passenger flow forecasting is discussed in this study. The experimental results confirm that weather variable rainfall has a significant effect on passenger flow forecasting. It is interesting to find out that the temporal characteristics and rainfall are the two most important variables to obtain more accurate forecasting results on Kelusanka route which is one of the routes in Udupi city. Compared to the widely used base line models and different combination of temporal and rainfall variables, the RPTW-LSTM is a better performing algorithm, which has the capability of making more accurate forecasts.

Keywords: RPTW-LSTM; forecasting; short-term passenger flow

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

BRTS – Bus Rapid Transit System

KSRTC - Karnataka State Road Transport Corporation

ARIMA - Autoregressive Integrated Moving Average

SARIMA - Seasonal Autoregressive Integrated Moving Average

RNN – Recurrent Neural Networks

ANN – Artificial Neural Networks

LSTM – Long Short-Term Memory

CNN - Convolution Neural Networks

GNN - Graph Neural Networks

GCN - Graph Convolution Networks

RPTW-LSTM – Recent intervals, daily Periodicity, weekly Trend, and Weather-LSTM

RPT-LSTM – Recent intervals, daily Periodicity, and weekly Trend-LSTM

RPW-LSTM – Recent intervals, daily Periodicity, and Weather LSTM

RTW-LSTM – Recent intervals, weekly Trend, and Weather LSTM

RMSE – Root Mean squared Error

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

CHAPTER 1

INTRODUCTION

1.1. General

In a modern transportation system, urban public transportation plays an important function. Comparing with other modes of travel, public transportation has the advantages of large passenger capacity, low pollution discharge, and low cost. In order to ensure the efficient and orderly operation of urban buses, not only is a good bus operation management plan required, but effective operation scheduling is also essential. Using public transportation–related information and data to make accurate public traffic passenger flow predictions can provide effective decision support for the operation and dispatch of urban buses, help transit operators control ridership inflow to avoid congestion, or adjust train. Short term passenger flow forecasting can provide real-time traffic information to help passengers make rational scheduling decisions and timetables to accommodate more passengers in peak hours.

1.2. Impact of various parameters on public bus passenger demand

Some of the factors that affect the public bus passenger flow includes travel time, ticket fares, weather conditions, route choice model and land use. Among which weather condition is considered in this study to find out the influence on public bus passenger flow.

Weather and travel habits are intimately connected. According to a Gallup poll conducted in 2002, 40% of respondents ranked weather and road conditions as the most important piece of information in their daily lives (ITSA, 2002). Given that the severity and frequency of extreme weather conditions are expected to increase as a result of climate change and global warming, it is critical to gain a better understanding of the dynamics of the relationship between weather and travel behavior in order to inform urban planners and transportation operators on how to (re)design more weather-resilient transportation systems capable of managing and adjusting transportation services in real-time in response to variations in weather.

To capture the impact of various parameters on public bus passenger flow time series statistical analysis like Auto Regressive Integrated Moving Average Method (ARIMA), Linear Regression, etc. need to be developed.

1.3. Statistical Analysis

Statistical analysis is the collection and interpretation of data in order to uncover patterns and trends. Statistical analysis gives how the relation between the various parameters is affecting the public bus passenger flow using techniques like Linear regression (LR), Auto Regressive Integrated Moving Average Method (ARIMA), Seasonal Auto regressive Integrated moving Average Method (SARIMA) etc.

1.3.1. Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

1.3.2. Auto Regressive Integrated Moving Average Method

An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values.

An ARIMA model can be understood by outlining each of its components as follows:

Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.

Integrated (I): represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).

Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

1.3.3. Seasonal ARIMA

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

1.4. Machine Learning Algorithms

Many machine-learning approaches (MLA) have been proposed to improve the prediction accuracy of short-term traffic/passenger flow, such as linear regression and feedforward neural networks. In addition, various deep learning methods have been employed in the literature. Despite the fact that MLA has become an emerging and significant transportation technology, the challenge of efficiently using it for bus passenger flow prediction persists due to the complexity of transportation systems, and this is the focus of this study.

From the beginning of forecasting models, the passenger flow was modelled using Linear Regression. The regression works well with data that has linear relationships. However, most real-world patterns have a complex structure, making linear models insufficient. The pattern of passenger flow has nonlinear temporal dependencies. Regression models such as linear regression, support vector regression with linear kernel, and ARMA are not appropriate in this case.

Historical passenger flow data and the accompanying spatiotemporal data are the most often used input variables. Despite the fact that experimental results demonstrate that weather data can accurately represent passenger flow, only a small amount of research has been focused on the problem of forecasting passenger flow using the weather factor due to the challenges associated with data collection and fusion for weather data, spatiotemporal data, and historical passenger flow data.

To capture the temporal characteristics, Long Short-Term Memory (LSTM) model is developed and to extract spatial correlation, Graph Neural Network (GNN) is developed.

1.4.1. Long Short-Term Memory

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of Recurrent Neural Networks (RNN), capable of learning long-term dependencies. They work tremendously well on a large variety of problems, and are now widely used. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn.

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

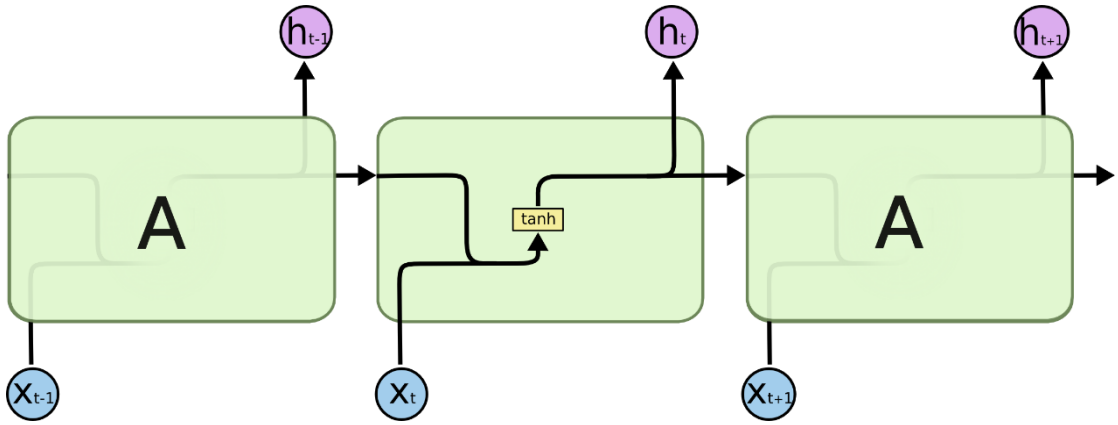


Fig 1: The repeating module in a standard RNN contains a single layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way

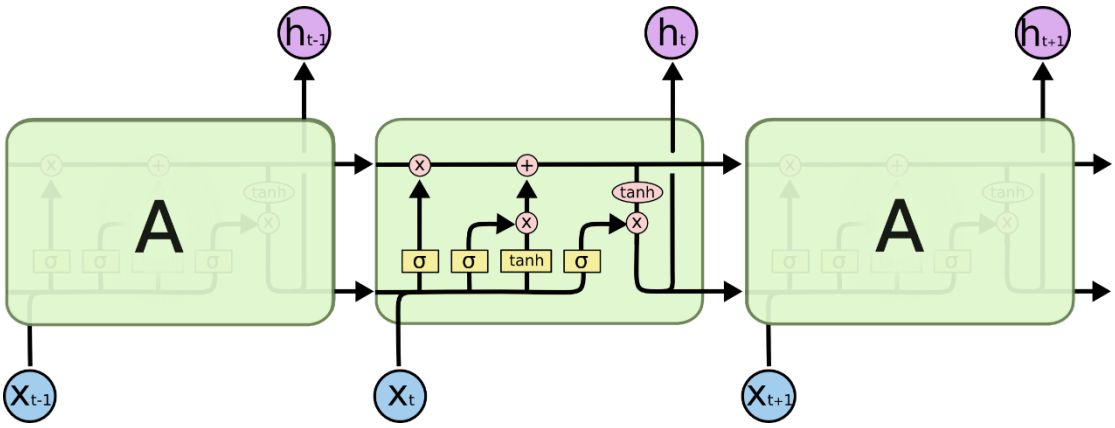


Fig 2: The repeating module in an LSTM contains four interacting layers.

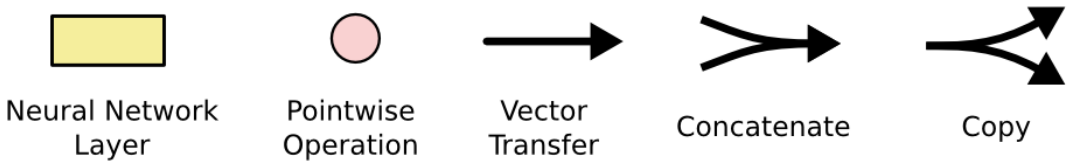


Fig 3: Operators used in LSTM networks

In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denotes its content being copied and the copies going to different locations.

1.4.2. Graph Neural Network

Graph Neural Networks (GNNs) are a class of deep learning methods designed to perform inference on data described by graphs. GNNs are neural networks that can be directly applied to graphs, and provide an easy way to do node-level, edge-level, and graph-level prediction tasks.

GNN is applied in Node Classification, the task here is to determine the labeling of samples (represented as nodes) by looking at the labels of their neighbors. Usually, problems of this type are trained in a semi-supervised way, with only a part of the graph being labeled. GNN is also applied in Link prediction, here the algorithm has to understand the relationship between entities in graphs, and it also tries to predict whether there's a connection between two entities. It's essential in social networks to

infer social interactions or to suggest possible friends to the users. It has also been used in recommender system problems and in predicting criminal associations.

1.4.3. Standard Models in Deep Learning

Apart from LSTM and GNN there are other standard models in deep learning which are used in time series problems such as ANN, CNN, GCN, RNN etc.,

Artificial Neural Networks (ANN)

ANNs are nonlinear statistical models that demonstrate a complex relationship between inputs and outputs in order to uncover a new pattern. Artificial neural networks are used for a range of applications, including image recognition, speech recognition, machine translation, and medical diagnosis.

The fact that ANN learns from sample data sets is a significant advantage. The most typical application of ANN is for random function approximation. With these types of technologies, one can arrive at solutions that specify the distribution in a cost-effective manner. ANN can also offer an output result based on a sample of data rather than the complete dataset. ANNs can be used to improve existing data analysis methods due to their high prediction capabilities.

Convolution Neural Networks (CNN)

CNNs can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. It has various applications some of which are decoding facial recognition, analyzing documents, understanding climate etc.,

Recurrent Neural Networks (RNN)

RNN works on the principle of saving the output of a particular layer and feeding this back to the input in order to predict the output of the layer. An RNN can handle sequential data, accepting the current input data, and previously received inputs. It can be used for any time series prediction, natural language processing and machine translation.

Graph Convolution Networks (GCN)

Convolutional Networks are 3-dimensional neural networks. Most practical uses of Convolutional Neural Networks include image classification and recognition, natural

language processing and speech recognition. These models are usually more complex than the usual 2-dimensional neural network models.

1.5. Need for the study

Following are the needs for forecasting the public bus passenger flow

- To understand the various factors affecting the public bus transit passenger flow.
- To predict the bus passenger flow with good accuracy by adding the factors that affects the bus passenger flow.
- To develop prediction models using state of the art techniques.

1.7. Objectives

The objectives proposed to develop a methodology for predicting the bus passenger flow are as follows:

- To examine the influence of weather on public bus passenger flow.
- To predict the future public bus passenger demand by incorporating temporal characteristics like recent time periods, daily periodicity, weekly trend and weather parameters using deep learning architectures.
- To compare the results of the proposed model with the state-of-the-art models.

1.6. Scope of the study

In earlier chapters various research works were studied and gaps in literature were identified. The BRTS in India is being developed at a greater rate and there are many upcoming projects of BRTS in India. As the BRTS projects are boosting the performance of these projects should be efficient, they should have minimum delay. Passengers' anxiety is reduced when they get accurate information on bus arrival and departure times at bus stops, and they may plan their trip accordingly. Estimating travel time and stopped delays at that portion of road, which includes the bus stop, can provide accurate information regarding bus arrival times. By understanding the causes of bus stop delays, suitable efforts to alleviate these delays can be taken. If the delay is caused

by travel time, then adequate road network design is required, if the delay is caused by bus stop design, then proper bus stop design is required.

Public transit can be made more appealing to passengers if delays at bus stops are adequately addressed and solutions are implemented. As a result, traffic congestion on Indian highways could be decreased.

Making accurate public traffic passenger flow predictions can provide effective decision support for the operation and dispatch of urban buses and help transit operators control ridership inflow to avoid congestion, or adjust train. And also, short term passenger flow forecasting can provide real-time traffic information to help passengers make rational scheduling decisions and timetables to accommodate more passengers in peak hours.

CHAPTER 2

LITERATURE REVIEW

2.1. General

Sui Tao et al (2018) has modelled the effect of local weather conditions on hourly bus ridership. They have derived a suite of time-series regression models (Auto Regressive Integrated Moving Average Method (ARIMAX), Seasonal Auto Regressive Moving Average Method (SARIMAX) which are computed to capture the concurrent and lagged effects that weather conditions exert on bus ridership. Their study area was bus network in Brisbane, Australia. Brisbane is the capital of Queensland and the third most populated city in the country, with around one million population within its local government area. The data were collected from transit smart card data and weather data were acquired from the Australian Bureau of Meteorology (BOM) for the same period of time as the smart card data. The information contained in a single smart card record includes date, route, direction (i.e., inbound and outbound trips in relation to the city center), smart card ID, boarding time and stop, and alighting time and stop and journey ID for linked trips made within a one-hour transfer limit. Measurements of four weather variables, i.e., temperature, rainfall, relatively humidity and wind speed on a 30-min interval are included for 14 weather stations located across the study context. They first visually inspected hourly system-wide ridership patterns on weekdays and weekends then marked decline in ridership in late March, early and late April is in parallel with the public holidays during these days. Except for this pattern of low ridership, a strong recurring pattern of hourly ridership persists across both weekdays and weekends. They next modelled the hourly weather-ridership relationships for their four selected destinations, Following the system wide analysis, univariate ARIMA (or SARIMA) models were separately estimated for each destination. Their findings highlight that changes in particularly temperature and rainfall were found to induce significant hour-to-hour changes in bus ridership, with such effects varying markedly across both a 24-hour period and the transit network.

Md Sami Hasnine et al (2021) investigated the effects of built environment and weather on the demands for the Transportation Network Companies (TNC) in Toronto.

Their research was based on a historical dataset of Uber trips from September 2016 to September 2018 in Toronto. A wide range of built environments, sociodemographic, and weather data were generated at the dissemination area-level and fused with the monthly aggregated Uber dataset. To provide insight into the underlying factors that affect TNC demand, a series of aggregate demand models were estimated using log transformed constant elasticity demand functions, with consideration of the seasonal lag effect. To capture the weather effect, an autoregressive moving average model was estimated for the downtown core of Toronto. Their model results show that the influence of lagged ridership and seasonal lag effect have a positive correlation with TNC demand. The trip generation and attraction models revealed that TNC trips increase when the commuting trip duration is longer than 60 min. And also founded that the number of apartments in a dissemination area is positively correlated with TNC trip generation, while the number of single-detached houses has a negative correlation. Developed a time-series model which indicated that temperature and total daily precipitations are positively correlated with TNC demand.

Ming Wei (2022) also studied the influence of local weather conditions on public transit ridership in Brisbane, Australia. Based on the statistical distribution of transit ridership, this study applied a suite of geographically weighted negative binomial regression models to capture the weather–transit ridership relationship at both daily and half-hourly levels. The results revealed that weather exerts significant effects on transit ridership and its effects vary by passenger type and are not fixed across locations and temporal periods.

Yang Liu et al (2019) combined the modeling skills of deep learning and the domain knowledge in transportation into prediction of metro inbound/outbound passenger flow. They used the standard metro service/transaction data from Nanjing Metro System, the data consists of weekday records of 103 days. The information contained from a single record are user id, inbound station, outbound station, inbound time, outbound time, and type of card. Based on Long Short-Term Method (LSTM) they proposed an end-to-end deep learning based architecture which can reasonably address the input features of the metro passenger flow prediction problem. This architecture comprised multiple extensible components, including modeling external environmental factors such as weather and holidays, temporal dependencies, spatial

characteristics, and metro operational properties respectively. The results of this study shows that the accuracy can be improved when the daily cyclicality characteristic and the weekly trend characteristic are incorporated in the model, indicating that it has a fixed behavior pattern.

Lijuan Liu et al (2020) also developed a deep long short-term memory neural network (LSTM_NN) model for predicting the metro passenger flow. The optimized traditional input variables, including the different temporal data and historical passenger flow data, were combined with weather variables for data modeling. They Constructed a Metro passenger flow forecasting (LSTM_NN) model by historical hourly passenger flow data and the corresponding temporal and whether data in a station level study with Endogenous Input variables: passenger flow direction, date, day, week, hour and Exogenous Input variables: Temp, rainfall, relative humidity, wind speed. They have compared the proposed model with weather variables and without weather variables and found that the deep LSTM_NN is a more powerful method to make the more accurate forecasts when suitable weather variables are included.

Jaison Mulerikkal et al (2022) has fed the passenger flow parameter into the layers of the deep neural network using the ST-LSTM (Spatio-Temporal Long Short-Term Memory) architecture. This architecture was evaluated with passenger movement data collected from automated fare card (AFC) information from metro rail. They have used the One-Class SVM-based (Support Vector Machine) outlier detection and elimination algorithm to reduce the impact of irregular flow. The results are compared with the existing system of regression models like SVR, ANN and LSTM. The architecture has obtained a performance improvement with an error of 0.00026.

Jinlei Zhang et al (2021) proposed a deep learning architecture combining the residual network (ResNet), graph convolutional network (GCN), and long short-term memory (LSTM) (called "ResLSTM") to forecast short-term passenger flow in urban rail transit on a network scale. Firstly, they improved the methodologies of ResNet, GCN and atttention LSTM, then the model architecture was proposed wherein ResNet is used to capture deep abstract spatial correlations between subway stations, GCN is applied to extract network topology information, and attention LSTM is used to extract temporal correlations. This model architecture includes four branches for inflow, outflow, graph-network topology, as well as weather conditions and air quality. Finally,

ResLSTM is applied to the Beijing subway using three time granularities (10, 15, and 30 min) to conduct short-term passenger flow forecasting. And the results confirmed that the weather conditions and air quality were proven to have considerable influence on prediction precision.

Shengnan Guo et al (2020) proposed novel approaches in modelling the dynamics of traffic data along both the temporal and spatial dimensions, as well as consider the periodicity and spatial heterogeneity of traffic data which is Attention based Spatial-Temporal Graph Neural Network (ASTGNN) for traffic forecasting. Attention mechanism which captures local context in time series. GCN to capture the dynamics along the spatial dimension. They have considered two kinds of datasets; the first kind of datasets are about highway traffic flow in California. The second kind of datasets is about metro crowd flow of the Hangzhou metro system and then the raw data was converted to 5-minute interval. Their model outperformed all the state-of-the-art techniques.

Feifei Zhao et al (2020) proposed a model SEHNN (station-embedding-based hybrid neural network) which utilizes the VGAE (Variational Graph auto-encoder) module to embed and extract the spatial and temporal information of order interactions and geographic neighbors among the stations of a carsharing network, while the LSTM module to capture the time sequence information of the embeddings to complete the prediction of rental and return vehicles. The results from the real data of Lanzhou, China demonstrate that, compared with ElasticNet, ARIMA, LSTM, and ConvLSTM, the mean absolute error of the proposed model targeted at hourly demand forecasting is reduced by 56.5%, 47.2%, 38.7%, and 38.5%, respectively, and also found that it outperforms some widely used models among different intervals and scales including main stations and subset carsharing system

Tao Chen et al (2021) proposed a novel algorithm, namely the Spatial–Temporal Graph Sequence with Attention Network (STGSAN), to predict transit passenger flow. The algorithm mainly focused on the following three aspects: (1) a graph attention network (GAT) was used to capture the spatial correlation of various bus stops; (2) to make full use of the historical and real-time data, a bidirectional long short-term memory and attention mechanism was conducted to extract the temporal correlation of historical ridership at bus stations; and (3) external factors that affect passenger choices

were taken into account. they conducted an experiment using field data collected in Urumqi, China. They compared their model with five other models, the proposed model was proven to have excellent performance prediction.

Can Li et al (2022) provided a confidence interval-based demand forecasting, which can help transport planning and operation authorities to better accommodate demand uncertainty/variability. The proposed Origin Destination (OD) demand prediction approach well captures and utilizes the correlations among spatial and temporal information. They proposed a Probabilistic Graph Convolution Model (PGCM) which consists of two components: (i) a prediction module based on Graph Convolution Network and combined with the gated mechanism to predict OD demand by utilizing spatio-temporal relations; (ii) a Bayesian-based approximation module to measure the confidence interval of demand prediction by evaluating the graph-based model uncertainty. They used a large-scale real-world public transit dataset from the Greater Sydney area to test and evaluate the proposed approach. The experimental results demonstrated that the proposed method is capable of capturing the spatial-temporal correlations for more robust demand prediction. The proposed approach is compared with several benchmark algorithms including ARIMA, LR, HA, GRU, and LSTnet, GCRN, and STGCN. The experiments on the real-world dataset show that the proposed approach outperforms other state-of-the-art methods.

2.2. Gaps in the literature

All the research papers have been studied thoroughly. The following gaps were identified:

- Studies related to the influence of weather conditions for the prediction of passenger demand are very less.
- Deep learning models have not been extensively used in India for predicting the passenger flow demand.
- Very few works in India have considered the impact of weather to enhance the performance of the models.

CHAPTER 3

METHODOLOGY

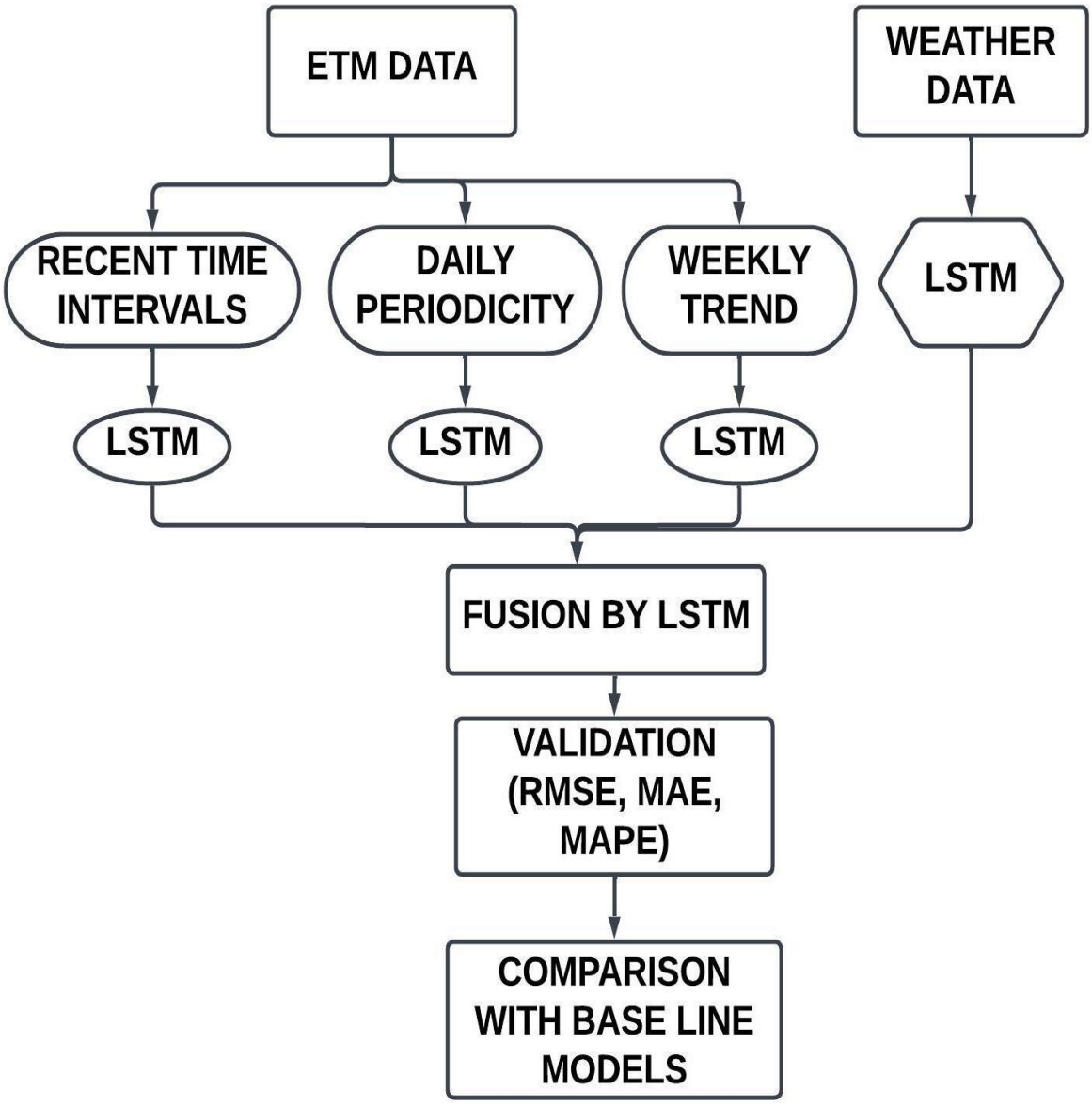
3.1. General

The prediction of public bus passenger flow is done using the historical data of bus passenger data and weather data. First step is to evaluate the various parameters that are affecting the bus passenger flow. Next step is to develop the deep learning architectures to predict the passenger flow where LSTM is used to extract the temporal correlations. Temporal correlations include hourly variation, daily periodicity and weekly trend characteristics. Moreover, finding the impact of weather like rainfall, humidity and temperature on passenger flow. Considering temporal and weather parameters, prediction of bus passenger flow is done. Validation of the results for proposed model are done with Root Mean Squared Error (rmse), Mean Absolute Error (mae) and Mean Absolute Percentage Error (mape).

3.2. Proposed Methodology

A detailed description of the proposed methodology is given in this section. After thoroughly analyzing the past research works, the most suitable method for predicting the bus passenger flow is identified. ETM (Electronic Ticket Machine) data for the year 2022 has been collected, this data is used for predicting the bus passenger flow for recent time intervals, daily periodicity, and weekly trend using Long Short-Term Memory (LSTM) algorithm. Weather data is used for predicting the bus passenger flow with respect to rainfall variation using LSTM algorithm. The output of all the four models (predictions) are fed as input to the final model to find the final predictions, which includes recent time interval, daily periodicity, weekly trend, and also weather parameter rainfall variation using LSTM algorithm.

METHODOLOGY FRAMEWORK



3.2.1. Impact of weather on bus passenger flow

Rainfall, temperature and relative humidity are the most commonly used parameters used to describe weather conditions. The impact that weather has on the quality of public transportation services and ridership has been emphasized as a crucial subject in transport scholarship. Heavy precipitation, temperatures and strong winds, for example, are known to have the ability to interrupt service schedules and damage service quality and passenger experience. These conditions also have the potential to cause both short-term and long-term drops in ridership. In order to mitigate the negative effects and probable decline in ridership, it is crucial to take into account how weather affects the regular functioning of public transportation networks. To achieve this, the effects that weather impose on public transport ridership first need to be understood to provide the necessary evidence from which planning and operation strategies can be founded.

3.2.2. Determination of predictions for bus passenger flow

Prediction for bus passenger flow is a time series problem. The analysis based on time are termed to be a time series problem. Here, the predictions are done based on the historical data. Supposing that each bus departs from the starting station as a time step, there are $\boldsymbol{l}$ time steps per day. Assume that the current time is $\boldsymbol{t}$, and the size of predicting window is $\boldsymbol{T_{(t+l)}}$. We extract historical time-series data $\mathbf{X_h}$, $\mathbf{X_d}$, $\mathbf{X_w}$ and $\mathbf{X_{wr}}$ as input data, which represent public bus passenger flows at different periods along the time axis, corresponding to the recent period, daily period, weekly period and weather, respectively.

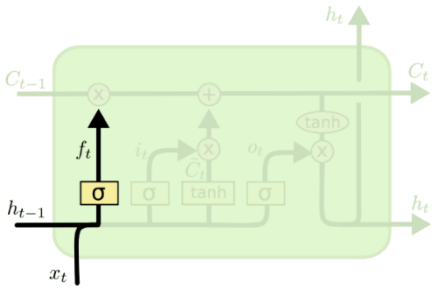
The movement of passengers on public buses is significantly influenced by temporal correlation. The majority of temporal distributions of passenger flow on public buses are unbalanced. Consequently, it is important to take into account temporal recurrent and repetitive patterns while estimating passenger flow for public buses. For handling sequential input, the recurrent neural network (RNN), which effectively employs feedback nerve cells, has been widely used. However, while training, the RNN model will run into the gradient descent or gradient explosion issue. In other words, as the time interval grows, the RNN will be less and less able to learn distant information.

Some researchers picked the LSTM model to handle the RNN problem because it can do so by using memory units.

Step by step LSTM walkthrough

1. Forget gate

The first step in LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer." It looks at h_{t-1} and x_t, and outputs a number between 0 and 1 for each number in the cell state C_{t-1}. Where 1 represents "completely keep this" while a 0 represents "completely get rid of this."

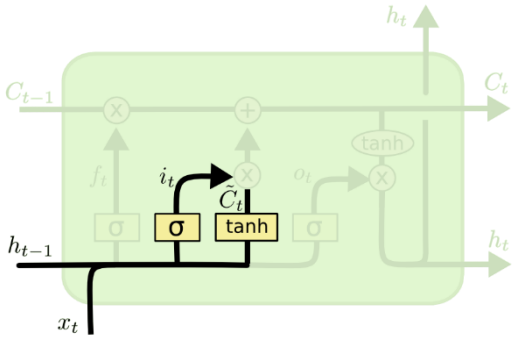


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Fig 4: Forget Gate

2.Input Gate

The next step is to decide what new information we're going to store in the cell state. This has two parts. First, a sigmoid layer called the "input gate layer" decides which values we'll update. Next, a tanh layer creates a vector of new candidate values, c_t, that could be added to the state. In the next step is combining these two to create an update to the state.



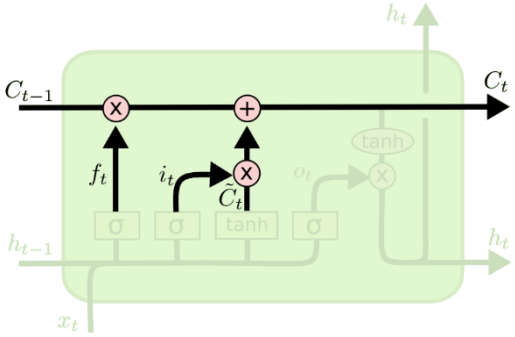
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Fig 5: Input Gate

3.Memory Cell

It's now time to update the old cell state, $\boldsymbol{C_{t-1}}$, into the new cell state C_t. Next, multiplying the old state by f_t, forgetting the things we decided to forget earlier. Then we add $i_t * \boldsymbol{\tilde{C}_t}$t. This is the new candidate values, scaled by how much we decided to update each state value.

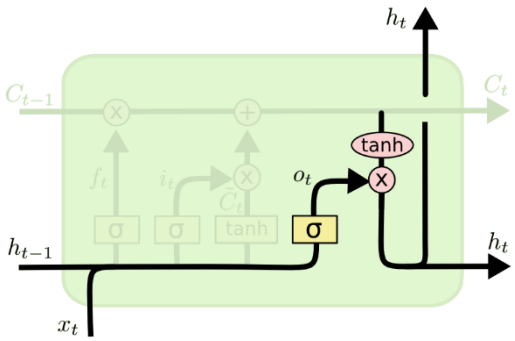


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Fig 6: Memory Cell

4.Output Gate

In the output cell first, we run a sigmoid layer which decides what parts of the cell state we're going to output. Then, we put the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Fig 7: Output Gate

3.2.3. Proposed LSTM model step by step walk through

To cover and include all the temporal characteristics, RPTW-LSTM (Recent, Periodicity, Trend and Weather – LSTM) is proposed with a combination of recent time intervals, daily periodicity, weekly trend and weather characteristics. The proposed model takes care of all the anomalies of the temporal changes.

Recent period bus passenger flow

Here, $X_t = \{X_{t-1}, X_{t-2}, \ldots X_{t-n}\}$ are recent time intervals for the prediction target of X_t, where X_t represents the passenger flow at time step *t*, X_{t-1} represents the passenger flow at time step *t-1* etc,. Here, the predictions are based on recent time intervals because number of passengers boarding in the past time periods has a certain impact on number of passengers boarding in the future time intervals.

Daily period bus passenger flow

Here, $X_{dt} = \{X_{(d-1)t}, X_{(d-2)t}, \ldots X_{(d-n)t}\}$ are the recent days same time period data for the prediction target X_{dt}, where X_{dt} represents the passenger flow for d^{th} day at time period *t*, $X_{(d-1)t}$ is the passenger flow for $(d-1)^{th}$ day at time period *t*, etc,. Here the predictions are based on the same hour for previous days, as there is a significant relation for the number of passengers boarding in the future days with respect to the number of passengers boarding at same time in the previous days. For example, the passenger flow of morning peak and evening peak will be larger than the passenger flow of the common peak.

Weekly period bus passenger flow

Here, $X_{wt} = \{X_{(w-1)t}, X_{(w-2)t}, \ldots X_{(w-n)t}\}$ are the recent week same time period data for the prediction target X_{wt}, where X_{wt} is the passenger flow for the w^{th} week at t^{th} time period,

$X_{(w-1)t}$ is the passenger flow for the $(w-1)^{th}$ week at time period *t-1* etc. The passenger flow pattern on weekends exhibits similarities to historical weekend patterns but differs significantly from the traffic pattern on Mondays. As a result, we develop a weekly-period component to capture the recurring weekly variations in public bus passenger flow.

Weather and recent period bus passenger flow

In order to evaluate weather parameter i.e, rainfall into the model, hourly variation along with rainfall has been taken into consideration. This is a multivariate time series problem with rainfall and hourly passenger demand data. Where, X_{wrt} = {W_{t-1}, W_{t-2},….W_{t-n}}*{ X_{t-1}, X_{t-2}, ….X_{t-n}} are the recent time period data along with the respective weather data for the prediction target X_{wrt}, where X_{wrt} is the passenger flow at the time step *t,* W_{t-1} represents the weather data at time step *t-1,* X_{t-1} represents the passenger flow at time step *t-1* etc,. The predictions are based on passenger flow and weather data of previous hours, as there is a small difference in passenger flow with respect to the weather change.

These temporal features on X_h, $X_{d,}$ $X_{w,}$ and X_{wr} are separately modeled with LSTM layers with previous time periods hourly, daily, weekly, and weatherly respectively.

Fusion

The flow of people using public transportation is constantly changing dynamically. These modifications will affect several external factors in addition to the existing passenger flow. For instance, when it's raining heavily, some individuals might decide to take a taxi rather than a bus. Fusion method for these external factors need to be designed.

The external data has been transformed to the needed structure that can be learned by the neural network. For the time of day, divided each day into 13 time periods each consisting 1 hour, base on operating conditions of the public bus. For day of week, divided then into seven categories. Similarly for the weather, divided into seven categories one for each day of the week.

The output (predictions) of hourly variation, daily periodicity, weekly trend and weather models are fed as input to the final model i.e., fused model. This fused model uses Long Short-term memory algorithm with multivariate time series analysis. The

output (predictions) of the fused model is then evaluated with the test data using RMSE, MAE, and MAPE.

3.3. Comparison of proposed model with other models

Comparing the proposed model with the state-of-the-art techniques is essential to know, how well the model is performing with reference to the base models. Compared proposed fused LSTM model with other algorithms, including Artificial neural networks (ANN), Simple Recurrent neural networks (SimpleRNN), Seasonal Auto Regressive Integrated Moving Average (SARIMA), and simple LSTM. Along with the base line models, to ensure the effect of each temporal characteristics and weather in the model, three other models have been compared they are: 1) **R**ecent time, daily **P**eriodicity and weekly **T**rend-LSTM (RPT-LSTM) without weather parameter, 2) **R**ecent time, daily **P**eriodicity and **W**eather-LSTM (RPW-LSTM) without weekly trend, and 3) **R**ecent time, weekly **T**rend and **W**eather-LSTM (RTW-LSTM) without daily periodicity. The selected models include a statistical method and deep learning method, which can ensure the fair comparison.

3.4. Validation of the proposed model

All the predictions for the adjacent time intervals, daily periodicity, weekly trend, and weather component are to be found and all these external factors are fused with the LSTM layers. In this way, the predicted target will be determined. After predicting, by including all the components, model is evaluated with the testing data with RMSE, MAE and MAPE. Model accuracy will be determined based on RMSE, MAE, and MAPE.

$$RMSE = \sqrt{\frac{1}{n}\Sigma(predicted - \ actual)^2}$$

$$MAE \ = \ \frac{\sum|predicted - actual|}{n}$$

$$MAPE \ = \ \frac{1}{n}\Sigma\left|\frac{predicted - \ actual}{actual}\right|$$

The accuracy metrics for accepting the model is given in the table 3.1.

Table 3.1: Accuracy metrics

Accuracy (in%)	**Error (MAPE in %)**	**Remarks**
>90%	<10%	Accepted and accurate
75-90%	10-25%	Accepted and satisfied
<75%	>25%	Rejected and unsatisfied

3.5. SUMMARY

A clear methodology has been proposed in this chapter which includes preparation of LSTM models for hourly, daily periodicity and weekly trend from Electronic Ticket Machine (ETM) data, developing a LSTM network for external factors including weather parameters like rainfall, fusing of temporal and weather characteristics for prediction of public bus passenger demand and also mentioned the models that are being compared with the proposed RPTW-LSTM.

In the next chapter, Study area, how the data is collected and processed are being explained.

CHAPTER 4

DATA COLLECTION AND PROCESSING

4.1. General

The weather of Udupi, which is on the west coast of Karnataka, varied greatly over the summer and winter. In this chapter we will go through the data collection and extraction required for the analysis of temporal and weather characteristics on bus passenger flow.

4.2. Study Area

Udupi is served by a dependable transportation system, with the KSRTC playing a vital role in providing bus services to the city and its environment. The numerous neighborhoods, economic districts, institutions of higher learning, and tourist attractions in the city are all connected by the intra-city bus services provided by KSRTC in Udupi. The buses have designated bus stops where passengers can board, disembark and the buses follow pre-planned routes.

Udupi is expanded over 68.23 km^2 has wide public transport facility since there is variation in the weather during every month there is change the passenger flow. The study takes into account 10 routes from the city of Udupi. The routes that were chosen pass via a hospital, educational facilities, public structures, tourist attractions, and residential neighborhoods.

4.3. Data Collection

Electronic ticket (issuing) machines (ETMs) are devices containing memory and processors that issue tickets for transportation. ETM data from Udupi KSRTC Depot from 2018 to 2022. The dataset includes the name of the boarding and alighting stop, journey time, the number of passengers, an anticipated ticket cost and passenger type description. The following years' meteorological information, including rainfall, temperature, and humidity, was obtained from the India Meteorological Department (IMD) Bangalore. GPS was used to record the stops' latitude and longitude and QGIS was used to display a map with different routes

Table 4.1: Sample Raw Data

ETD_WAY BILL_NO	ETD_DATE	ETD_ TD_TIME	ETD_CUR_ STOP_NAME	ETD_DST_ STOP_NAME	ETD_ KMS	ETD_ AMOUNT	ETD_ ADULTS	ETD_ CHILD	ETD_TICKET _TYPE_DESCR
135467	1/4/2022	29:09.0	SANTEKATTE	UDUPI	5	9	1	0	PASSENGER
135467	1/4/2022	39:52.0	PUTTUR	UDUPI	3	8	1	0	PASSENGER
135422	1/3/2022	42:49.0	UDUPI	BRAHMAVAR	10	14	1	0	PASSENGER
135422	1/3/2022	43:42.0	UDUPI	PUTTUR	3	8	1	0	PASSENGER
135422	1/3/2022	45:51.0	UDUPI	BRAHMAVAR	10	14	1	0	PASSENGER
135422	1/3/2022	55:18.0	SANTEKATTE	BRAHMAVAR	5	12	1	0	PASSENGER
135422	1/3/2022	11:14.0	BRAHMAVAR	MANDARTHI	10	15	1	0	PASSENGER
135422	1/4/2022	53:01.0	MANDARTHI	BRAHMAVAR	10	15	1	0	PASSENGER
135422	1/4/2022	00:13.0	NADUR	BARKUR	4	20	2	0	PASSENGER
135422	1/4/2022	02:31.0	RANGANKERE	BARKUR	2	8	1	0	PASSENGER
135422	1/4/2022	04:45.0	RANGANKERE	BRAHMAVAR	5	10	1	0	PASSENGER
135422	1/4/2022	22:42.0	BRAHMAVAR	UDUPI	10	28	2	0	PASSENGER
135422	1/4/2022	23:01.0	BRAHMAVAR	UDUPI	10	14	1	0	PASSENGER
135422	1/4/2022	23:39.0	BRAHMAVAR	UDUPI	10	14	1	0	PASSENGER
135422	1/4/2022	24:14.0	BRAHMAVAR	SANTEKATTE	5	12	1	0	PASSENGER
135267	1/1/2022	25:11.0	SANTEKATTE	UDUPI	5	7	1	0	SENIORCITIZ
135402	1/3/2022	45:21.0	SANTEKATTE	UDUPI	5	9	1	0	PASSENGER

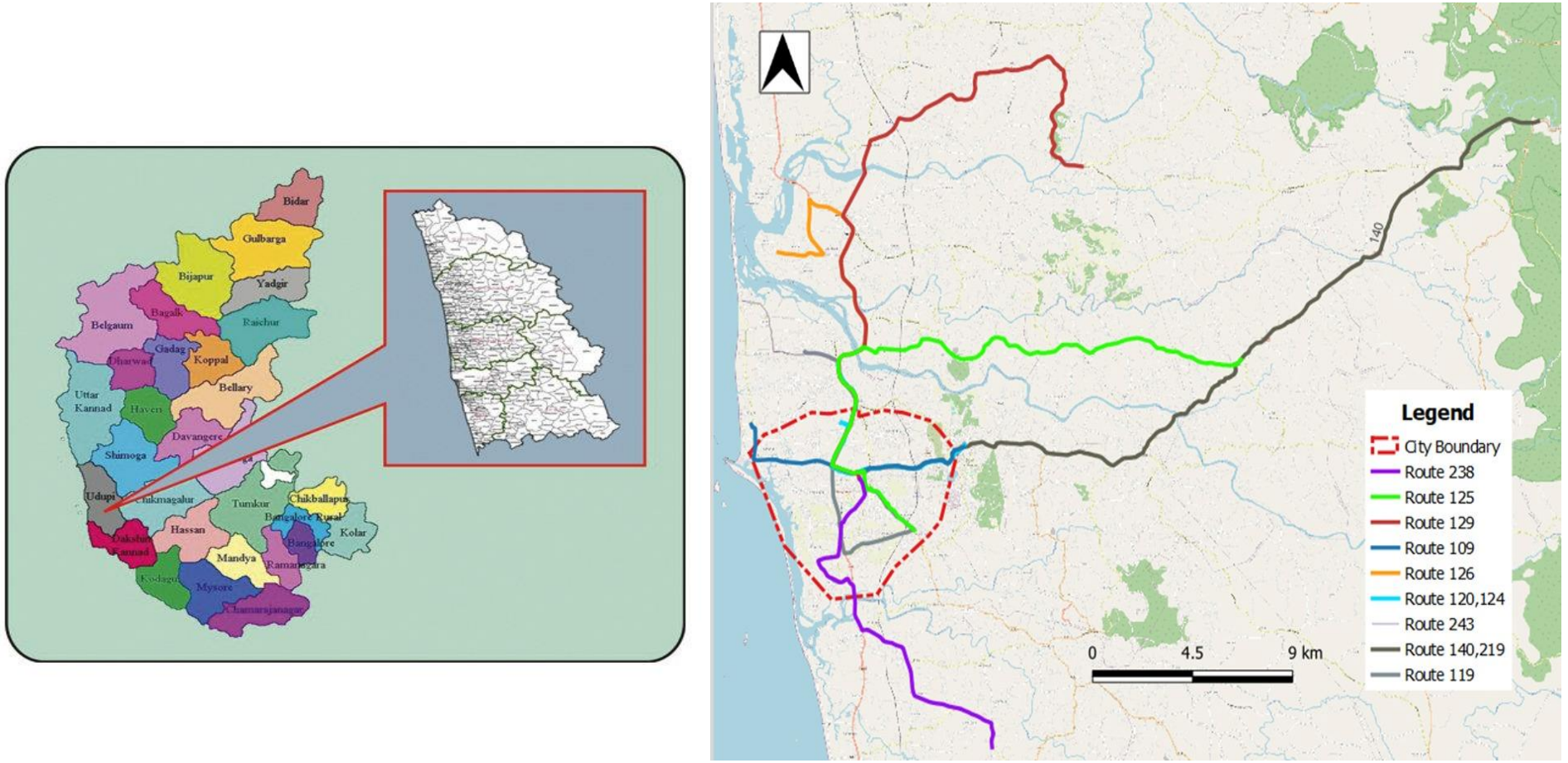


Fig 8: Study area and route map for Udupi

4.4. Data Processing

ETM (Electronic Ticket Machine) Data of Udupi intracity has been collected from the Udupi depot. The data consists a total of 10 routes namely Malpe, Honnala, Manchakal, Nellikatte, Perdur, Kokarne, Kelusanka, Alevoor, Kalyanpura and Hoode.

Weather data has been collected from Government of India, India Meteorological Department. Weather data consists of hourly temperature data, relative humidity and rainfall data. Weather data is used to check the correlation of temperature, relative humidity and rainfall with respect to demand to find out the impact of weather on bus passenger demand.

The raw ETM data mainly consists details of date, time, boarding station, departure station, amount, number of adults, number of children and passenger type description. Data mining (Data preprocessing) techniques like removing the unnecessary data like the amount of the ticket which is zero, the stops that are not present in any of the routes and normalizing the stop names (converting the different names of a single stop to one name) in both origin and destination columns. Identified the stops for all routes present in Udupi and segregated the data for each route by querying the data using pandas and panda SQL in python. The data for each route has been saved in parquet format which can store large amount of data with less space and the data types of all the columns will also be saved.

After cleaning the data by removing the unwanted stops and segregating routes, prediction has to be done for one route (considered Kelusanka route (route number 129), as it has more data compared to all other routes). Data is prepared for modelling by converting the data into three different formats hourly, daily and weekly. Firstly, the data has been converted to hourly data. It was observed that the data is consistent between morning 7 am and evening 7 pm. So, the remaining data which is before morning 7 am and after evening 8 pm has been removed from the dataset. Data consists of years 2018 (from march), 2021 and 2022. As there is an impact of covid for the years 2019, 2020, and 2021, so we considered 2022 for analysis. For some of the hours, the data has been missing for some days, in order to make the data more consistent the missing data has been replaced with the previous week same hour of that particular hour. For example, if the data for 8 am to 9 am is missing, first identified the dates that

are missing and replaced with previous week 8 am to 9 am data. The final data for hourly consists of date, time and demand with 4745 (365 days with each day having 13 hours from 7 am to 8 pm) rows. Using the hourly data, the data has been converted to daily demand by adding the demand column for the respective date and the final daily data consists of date and demand with 365 rows. Using the daily demand data, the data has been converted to weekly demand by aggregating the 7 consequent dates and adding these 7 days demand and the final weekly data consists of 52 rows (weeks). The sample of final data used for analysis is given in table 5.2.

Table 4.2: Final Kelusanka sample data

ETD_DATE	ETD_TD_TIME	DEMAND
1/1/2022	7	28
1/1/2022	8	70
1/1/2022	9	22
1/1/2022	10	14
1/1/2022	11	58
1/1/2022	12	46
1/1/2022	13	42
1/1/2022	14	15
1/1/2022	15	33
1/1/2022	16	53
1/1/2022	17	113
1/1/2022	18	96
1/1/2022	19	11
1/2/2022	7	20
1/2/2022	8	96
1/2/2022	9	82
1/2/2022	10	35
1/2/2022	11	66
1/2/2022	12	39
1/2/2022	13	28
1/2/2022	14	7
1/2/2022	15	17
1/2/2022	16	86
1/2/2022	17	91
1/2/2022	18	109
1/2/2022	19	11

Table 4.3: Sample daily demand data along with relative humidity and temperature

DATE	Passenger flow	RH	Max temp
1/1/2022	**20760**	60	33.9
1/2/2022	**19604**	58	34.3
1/3/2022	**26884**	67	34.1
1/4/2022	**26184**	66	33.8
1/5/2022	**25060**	67	33.0
1/6/2022	**24436**	83	32.3
1/7/2022	**26300**	85	32.7
1/8/2022	**9016**	81	32.7
1/9/2022	**4408**	71	31.4
1/10/2022	**28364**	75	31.9
1/11/2022	**25648**	72	32.1
1/12/2022	**25672**	70	30.6
1/13/2022	**27252**	79	30.1
1/14/2022	**27880**	68	31.0
1/15/2022	**13040**	72	31.3
1/16/2022	**6464**	75	32.0
1/17/2022	**28192**	74	32.4
1/18/2022	**23284**	68	32.1
1/19/2022	**25512**	61	33.8
1/20/2022	**25632**	72	33.6

4.5. SUMMARY

Udupi is chosen as the study area which includes 10 routes. Chosen, Kelusanka route for analysis. Electronic Ticket Machine (ETM) Data has been collected from Udupi KSRTC depot for the years 2018 to 2022 and weather data has been collected from Indian meteorological department, Bangalore which includes temperature, rainfall and relative humidity. Processing of data has been explained in this section.

In the next chapter the analysis of the data (correlation statistics, trend characteristics) and the results will be discussed.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1. General

The RPTW-LSTM approach for predicting the public bus passenger flow involves in fusing of external factor models. Passenger trend characteristics like hourly trend, daily trend and weekly trend has greater impact on future passenger flow. For example, if we see the Monday peak hour depends on previous hours, previous days same hour and also previous Monday's peak hours. In addition to temporal characteristics, weather also has a great impact on passenger flow and correlation of these weather parameters are studied in this project.

5.2. Passenger Trend characteristics

In order to forecast the bus passenger demand, it is essential to know previous trend of passenger demand for hourly, daily and weekly trend characteristics. Predictions are based on the previous patterns of passenger demand. Here the passenger trend for hourly, daily and weekly are done for the year 2022 and are shown in the figures 18, 19 and 20 respectively.

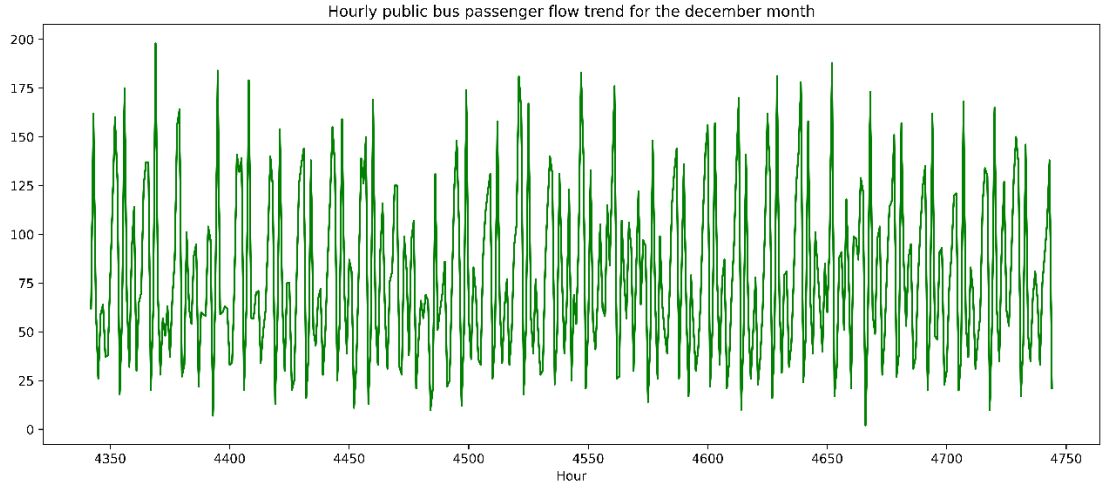


Fig 9: Hourly bus passenger demand trend

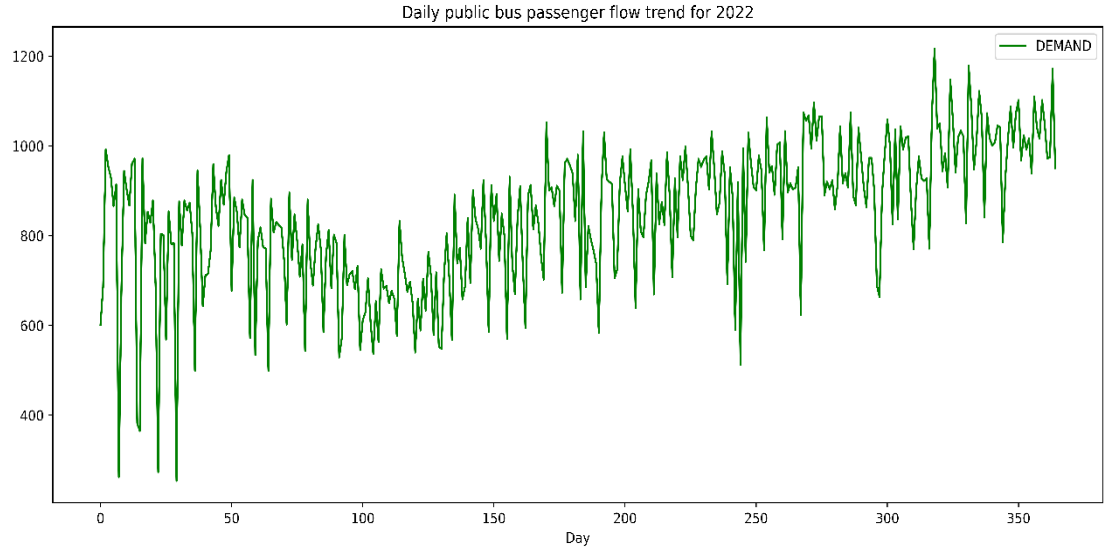


Fig 10: Daily bus passenger demand trend

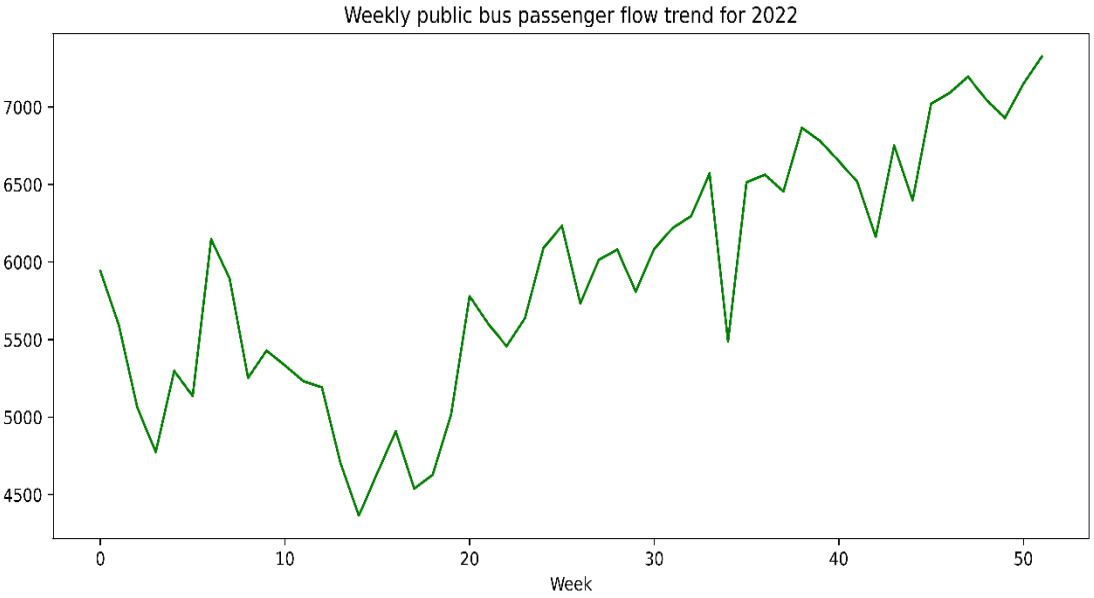


Fig 11: Weekly bus passenger demand trend

5.3. Determination of correlation between weather and passenger data

External factors like weather parameters influence a great deal on public bus passenger flow, hence it is important to know the correlation of these parameter with respect to bus passenger flow. For Rainfall we considered five months of data i.e., June, July, August, September, and October months as there is rainfall for these months. Regression plots are drawn and shown in the figures 21, 22 and 23.

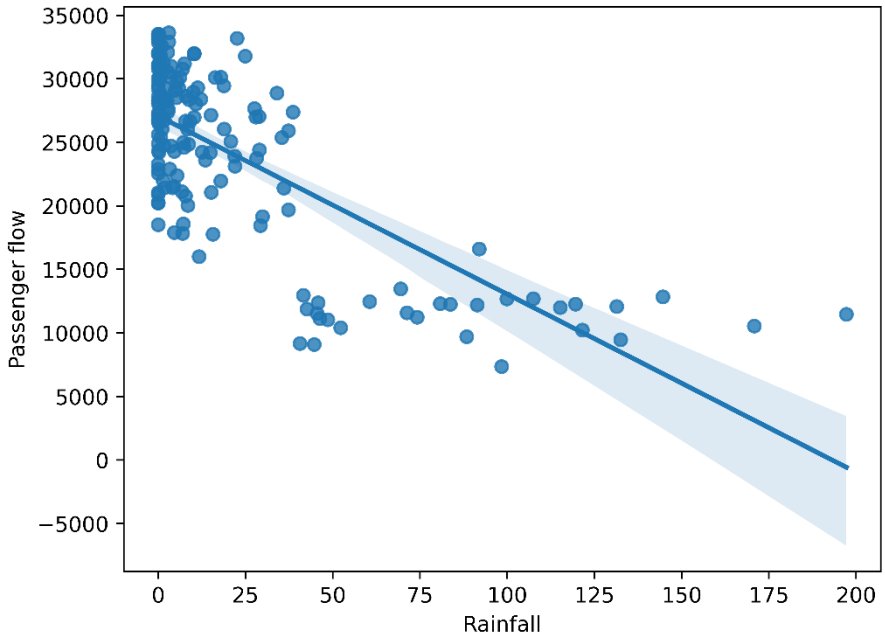


Fig 12: Regression plot of bus passenger flow with Rainfall

Regression plot shows how these weather parameters (rainfall, temperature, and relative humidity) are influencing the public bus passenger flow. Rainfall is highly negative correlated with passenger flow with a value of **-0.73496** i.e., with rainfall there is less likely for passengers to board on public busses. Relative humidity has a very less impact on passenger flow with a value of **-0.16096** i.e., with humidity increases passenger flow decreases. Temperature also has a less impact on passenger flow with a value of **0.171665** i.e., as temperature increases there is a small increase in passenger flow.

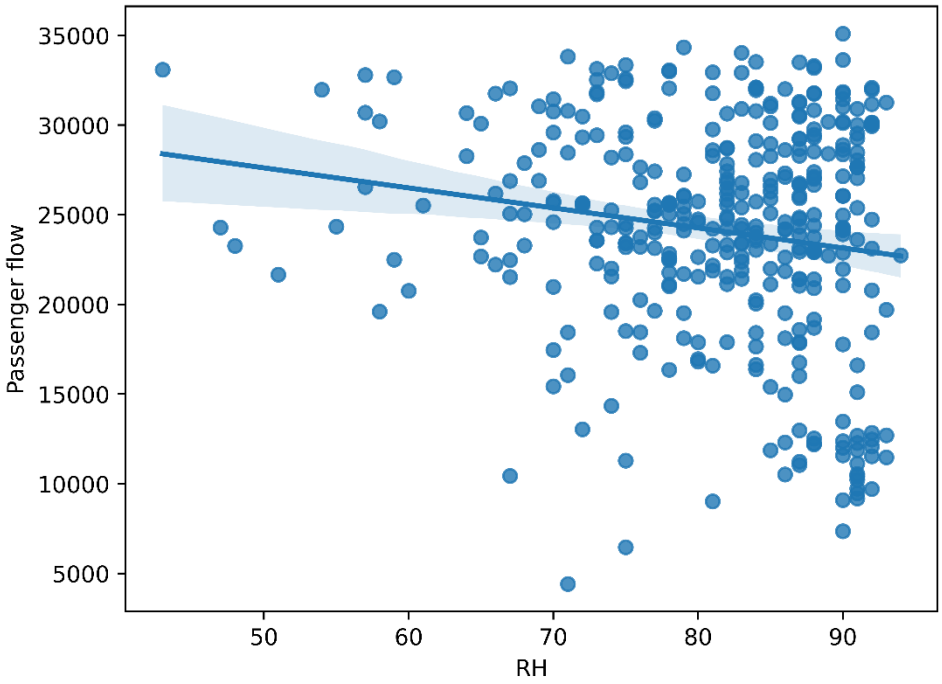


Fig 13: Regression plot of bus passenger flow with Relative Humidity

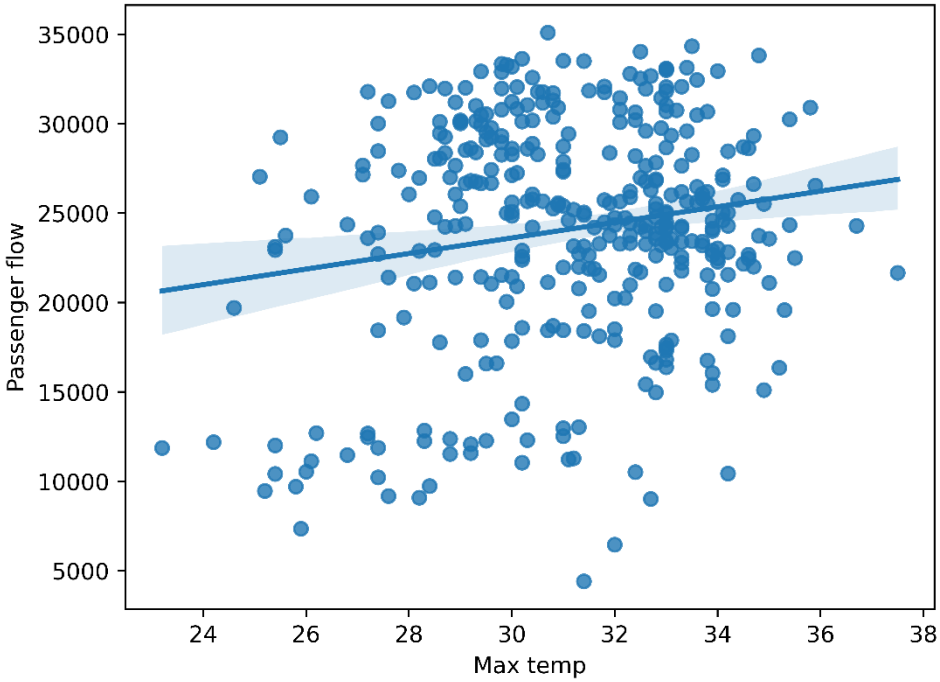


Fig 14: Regression plot of bus passenger flow with Maximum Temparature

After drawing the Regression plots of three weather variables rainfall, temperature and relative humidity with demand at daily level, it is found that there is significant impact of rainfall from June to October of 2022 on bus passenger flow. And also there is a very less correlation between temperature and relative humidity with respect to demand. Therefore for analysis (predicting) of the bus passenger flow we considered only rainfall and excluded relative humidity and temperature.

5.4. RPTW-LSTM

RPTW-LSTM (Recent, daily Periodicity, weekly Trend and Weather - Long Short-Term Memory) is the proposed model which considers temporal and weather characteristics. In order to reduce the overfitting problem in the model, used regularization techniques like dropout, L1 and L2 regularization. Later found that removing overfitting reduces the accuracy of the model, so found the best parameters using the hyper-parameter tuning. Explanation of these parameters are given below.

Recent period passenger flow forecasting

Input data taken is previous 10 hours passenger data to predict the next outcome. Previous 10 hours input has found to be better performance after trail and error method. Training is done using the sequential model with three LSTM layers with 128, 64 and 16 units and a DENSE layer of 1 unit for output. Return sequences is used for first two layers to ensure that the output of each layer is fed into the next layer in the sequence and ‘relu’ is used as the activation function. Optimizer used is adam and loss function used is mean squared error (mse) for compiling the model. After trial and error, found that 120 epochs with a batch size of 64 gives the best performance to fit the model. To check the performance after prediction, the forecasted and the actual values has been re-scaled to original data. Validation has been done using root mean squared error (rmse), mean absolute error (mae) and mean absolute percentage error (mape). Rmse=18.83821679, Mae = 13.677, and Mape = 22.09%. Predicted vs actual is shown in the figure 15.

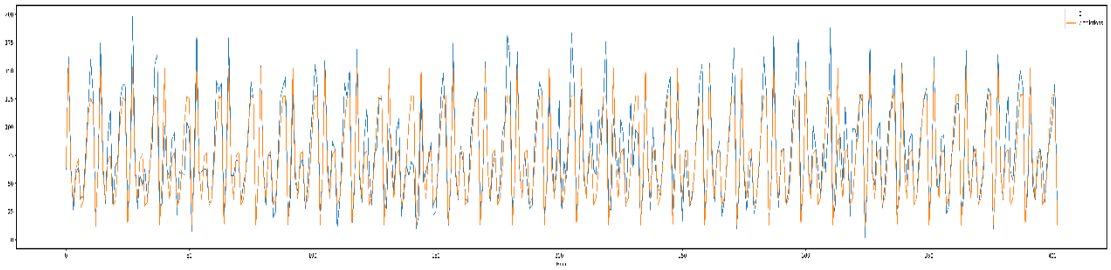


Fig 15: Actuals vs Predictions using recent time intervals

Passenger flow using daily periodicity

Input data taken is previous seven data points of same hour from the previous seven days to predict the next outcome. Previous seven data points has found to be better performance after some trial and error. Training is done using the sequential model with three LSTM layers of 128, 64 and 16 units respectively and a DENSE layer of one unit for output. Return sequences is used for first two layers to ensure that the output of each layer is fed into the next layer in the sequence. Activation function is relu, optimizer is adam and loss function is mse. By using trial and error, found that 230 epochs with batch size 100 gives the best performance to fit the model. The forecasted and the actual values has been re-scaled to the original range. Validation of daily is done by rmse, mae and mape. Rmse = 19.647, Mae = 14.689, and Mape = 22.89%. Predicted vs actual is shown in the figure 16.

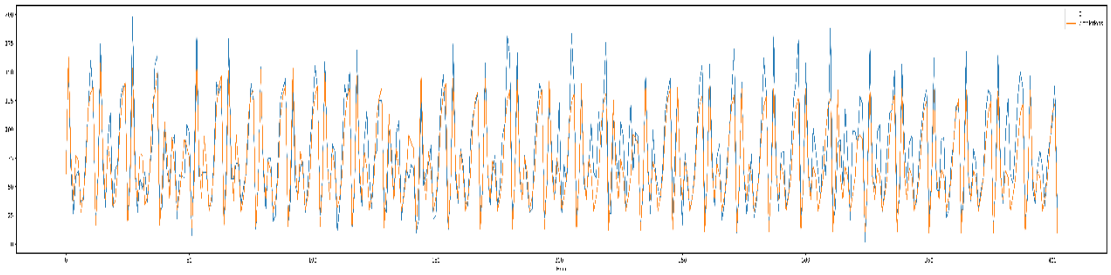


Fig 16: Actuals vs Predictions using daily periodicity

Passenger flow using weekly trend

Input data taken is eight data points for the same hour same day from the previous eight weeks to predict the next outcome. Input as eight data points has found to be better performance after some trial and error. Training is done using sequential model with three LSTM layers of 128, 64 and 16 units respectively and a DENSE layer of one unit for output. Return sequences is used for first two layers to ensure that the output of each

layer is fed into the next layer in the sequence. Activation function is relu, optimizer is adam and loss function is mse. By using trial and error, found that 150 epochs with batch size 64 gives the best performance to fit the model. The forecasted and the actual values has been re-scaled to the original range. Validation of weekly model is done by rmse, mae and mape. Rmse = 16.78, Mae = 12.479, and Mape = 19.52%. Predicted vs actual is shown in the figure 17.

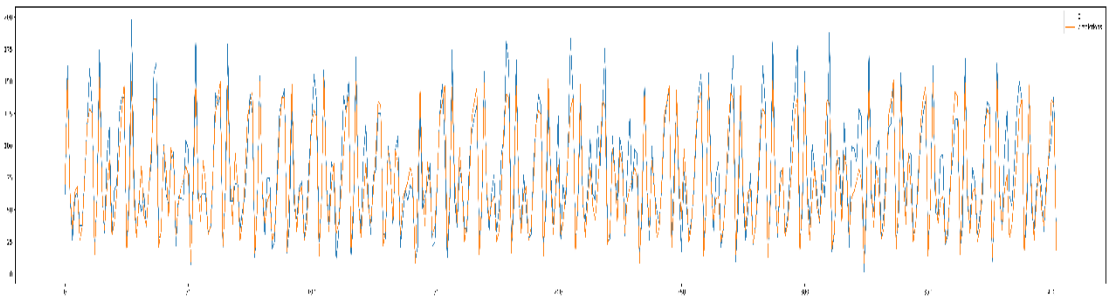


Fig 17: Actuals vs Predictions using weekly trend

Passenger flow using recent and rainfall

Input data has been taken as previous 10 hourly passenger and rainfall data to predict the passenger flow of next outcome. Input equal to 10 has found to be better performance by trial and error. Training is done using the sequential model with two LSTM layers of 128 and 64 units and a DENSE layer of one unit. Return sequences is used for the first LSTM layer to ensure that the output of the first layer is fed into the next layer in the sequence. Activation function is relu, optimizer is adam, and the loss function is mse. By using trial and error, found that 115 epochs with batch size 64 gives the best performance to fit the model. The forecasted and the actual values have been re-scaled to the original range. Validation of the weather model is done by rmse, mae, and mape. Rmse = 19.244, Mae = 14.143, and Mape = 21.98%. Predicted vs actual is shown in the figure 18.

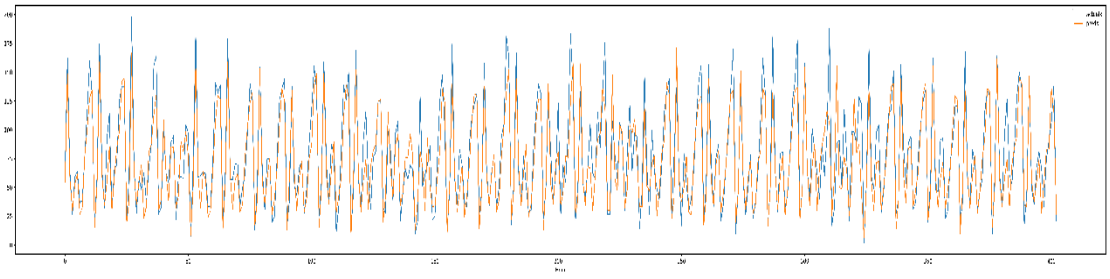


Fig 18: Actuals vs Predictions using recent time and rainfall

FUSION (RPTW-LSTM)

The output (predictions) of hourly, daily, weakly and weather models are fed as input to the final model i.e., fused model. Fused model is trained using sequential model with two LSTM layers of 128 and 16 units respectively and a DENSE layer of one unit for output. Return sequences is used for first LSTM layer to ensure that the output of each layer is fed into the next layer in the sequence. Activation function is relu, optimizer is adam, and loss function is mse. By trial and error, found that 60 epochs with batch size 10 gives the best performance to fit the model. The forecasted and the actual values have been re-scaled to the original range. Validation of the fused model is done by rmse, mae, and mape. Rmse = 9.47, Mae = 6.98, and Mape = 14.82%. Predicted vs actual is shown in the figure 19.

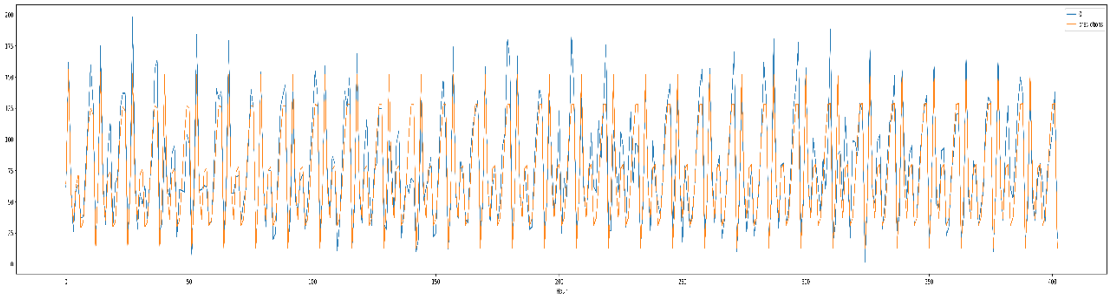


Fig 19: Actuals vs Predictions after fusion

When fused with four models it is found that the error percentage has decreased to 14.8% from 22.1%. The proposed model (RPTW-LSTM) has an accuracy of 85.2%, which is satisfactory and is acceptable for prediction of bus passenger flow.

5.5. Comparison of other models

Artificial Neural Networks (ANN) is a deep learning algorithm used for time series analysis, which is a feed forward neural network having an input layer and an output layer. Previous 10 hours are used as the input to predict the next outcome, to give the fair comparison. Build with the sequential model of one DENSE layer of 128 units for input with relu as activation function and one DENSE layer for output. Model is compiled with adam optimizer and loss function as mse. Epochs is set to 100 and a

batch size of 32 is set to fit the model. Predicted vs actuals of ANN model are shown in fig 20. The validation results are shown in the table 5.1.

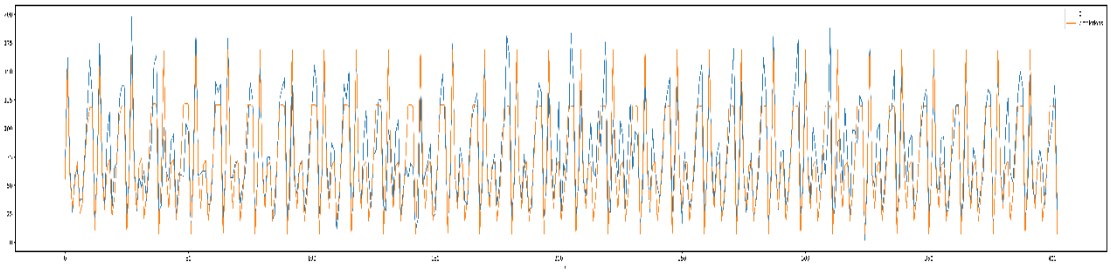


Fig 20: Actuals vs Predicted of ANN

Simple Recurrent Neural Network (RNN) is also a deep learning algorithm used for time series analysis and is considered to be better than Artificial Neural Networks. Previous 10 hours are used as the input to predict the next outcome, to give the fair comparison. Build with the sequential model of two SimpleRNN layers of 128, 64 units with relu activation function and a DENSE layer for output, return sequences is used in the first layer to ensure the output of the layer is fed into the next layer in the sequence. Model is compiled with adam optimizer, mse as loss function and with epochs 100 and batch size 64 model is fitted. Predicted vs actuals of RNN model are shown in fig 21. The validation results are shown in the table 5.1.

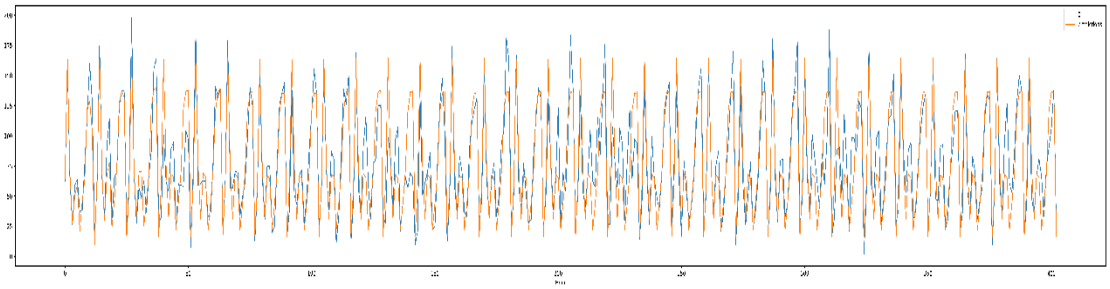


Fig 21: Actuals vs Predicted of RNN

Seasonal Auto regressive Integrated Moving Average (SARIMA) is used for time series forecasting and consists of three parts Auto Regressive (AR term i.e p), order of difference (d) and Moving average (MA term ie q) and apart from these seasonal parameter is also there in SARIMA unlike ARIMA, which is basically, how the seasonality is observed (ex: for hourly variation, 13 hours are repeated every day, for daily variation 7 days are repeated every week which implies 13 and 7 are the

seasonal order for respective variation). The selection of these parameters is critical for better performance. For hourly variation, we used p =1, d=1, q =2, and seasonal parameter = 13. Predicted vs actuals of ANN model are shown in fig 22. The validation results are shown in the table 5.1.

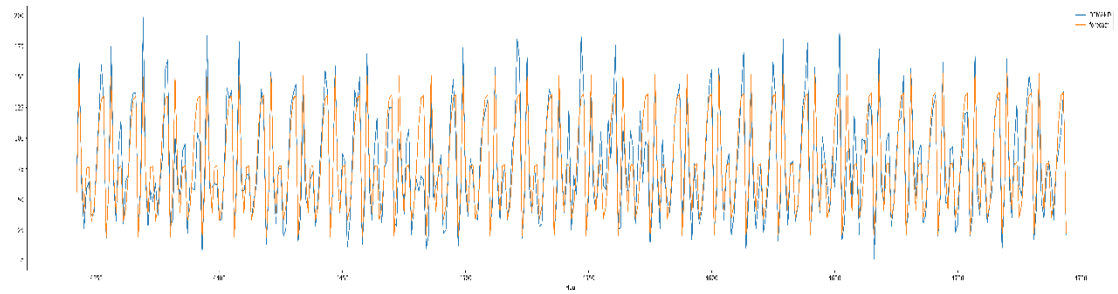


Fig 22: Actuals vs Predicted of SARIMA

There is a big question how each model is affecting in the accuracy of proposed model. To answer this question, developed three other models which are RPT-LSTM, RPW-LSTM, and RTW-LSTM.

RPT-LSTM algorithm uses recent period model, daily periodicity model, and weekly trend model to predict the passenger flow. This model doesn't consist of weather parameter to see weather impact on the proposed model. The output (predictions) of hourly, daily, and weakly models are fed as input to this model i.e., RPT-LSTM. Model is trained using sequential model with four LSTM layers of 256, 128, 64 and 16 units respectively and a DENSE layer of one unit for output. Return sequences is used for first three LSTM layers to ensure that the output of each layer is fed into the next layer in the sequence. Activation function is relu, optimizer is adam, and loss function is mse. 60 epochs with batch size 10 used to fit the model. The forecasted and the actual values have been re-scaled to the original range. Validation of the fused model is done by rmse, mae, and mape. The results are shown in the table 5.1. Predicted vs actual is shown in the figure 23.

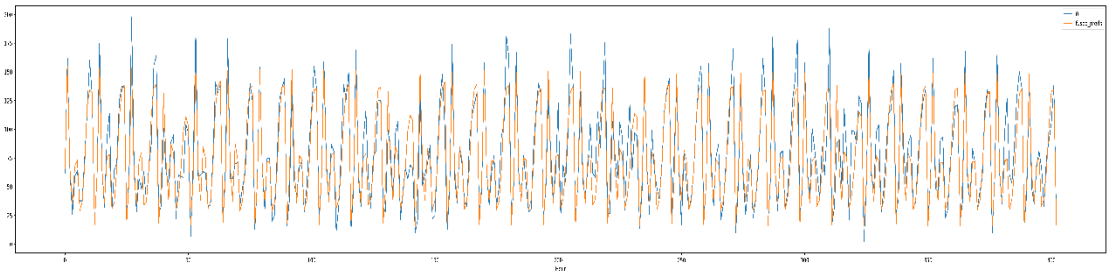


Fig 23: Actuals vs Predictions of RPT-LSTM

RPW-LSTM algorithm uses recent period model, daily periodicity model, and weather model to predict the passenger flow. This model doesn't consist of weekly trend to see weekly trend impact on the proposed model. The output (predictions) of hourly, daily, and weather models are fed as input to this model i.e., RPW-LSTM. Model is trained using sequential model with four LSTM layers of 256, 128, 64 and 16 units respectively and a DENSE layer of one unit for output. Return sequences is used for first three LSTM layers to ensure that the output of each layer is fed into the next layer in the sequence. Activation function is relu, optimizer is adam, and loss function is mse. 60 epochs with batch size 10 used to fit the model. The forecasted and the actual values have been re-scaled to the original range. Validation of the fused model is done by rmse, mae, and mape. The results are shown in the table 5.1. Predicted vs actual is shown in the figure 24.

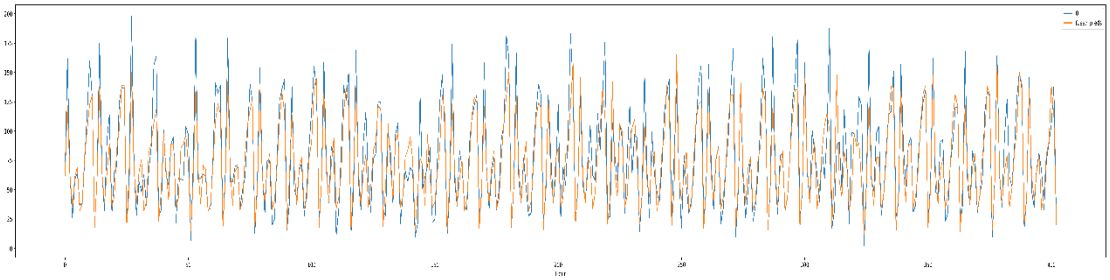


Fig 24: Actuals vs Predictions of RPW-LSTM

RTW-LSTM algorithm uses recent period model, weekly trend model, and weather model to predict the passenger flow. This model doesn't consist of daily periodicity to see daily periodicity impact on the proposed model. The output (predictions) of hourly, weekly, and weather models are fed as input to this model i.e., RTW-LSTM. Model is trained using sequential model with four LSTM layers of 256, 128, 64 and 16 units respectively and a DENSE layer of one unit for output. Return sequences is used for first three LSTM layers to ensure that the output of each layer is

fed into the next layer in the sequence. Activation function is relu, optimizer is adam, and loss function is mse. 60 epochs with batch size 10 used to fit the model. The forecasted and the actual values have been re-scaled to the original range. Validation of the fused model is done by rmse, mae, and mape. The results are shown in the table 5.1. Predicted vs actual is shown in the figure 25.

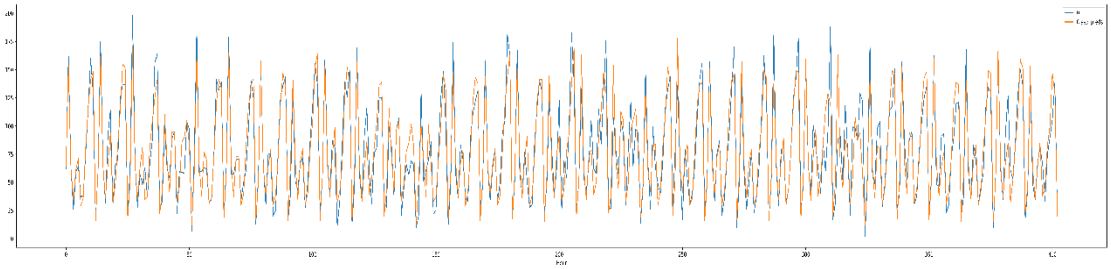


Fig 25: Actuals vs Predictions of RTW-LSTM

Table 5.1: Results of all the models

	MAPE	**RMSE**	**MAE**
ANN	25.98%	22.0221	16.315
RNN	25.45%	20.9984	16.3499
SARIMA	25.39%	16.6212	13.778
LSTM	22.09%	18.8382	13.677
RPW-LSTM	17.33%	10.82	7.785
RPT-LSTM	16.56%	9.508	7.002
RTW-LSTM	16.15%	9.443	7.017
RPTW-LSTM	14.83%	9.4729	6.985

From Table 5.1, although there is no much difference in the three models, we can say that RTW-LSTM has more accuracy than RPW-LSTM and RPT-LSTM, which means daily periodicity has more weightage, then followed by weekly trend and less weightage for weather. As the analysis is done for whole year, there is a very less impact if we considered rainfall in the analysis as there are only five months of rainfall when compared to a whole year.

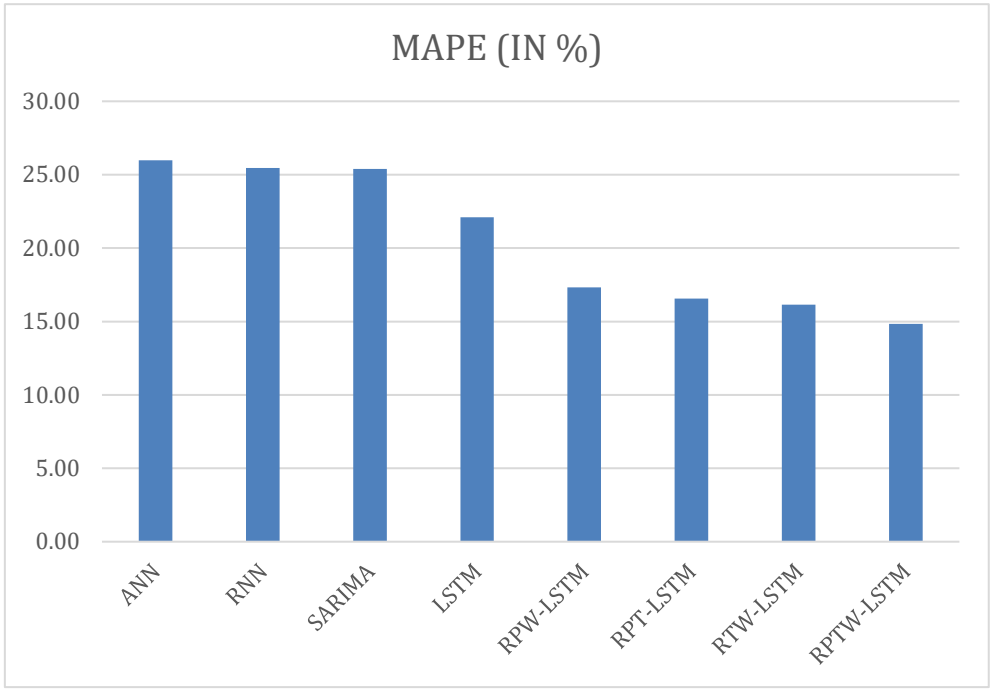


Fig 26: Comparison of MAPE of all algorithms

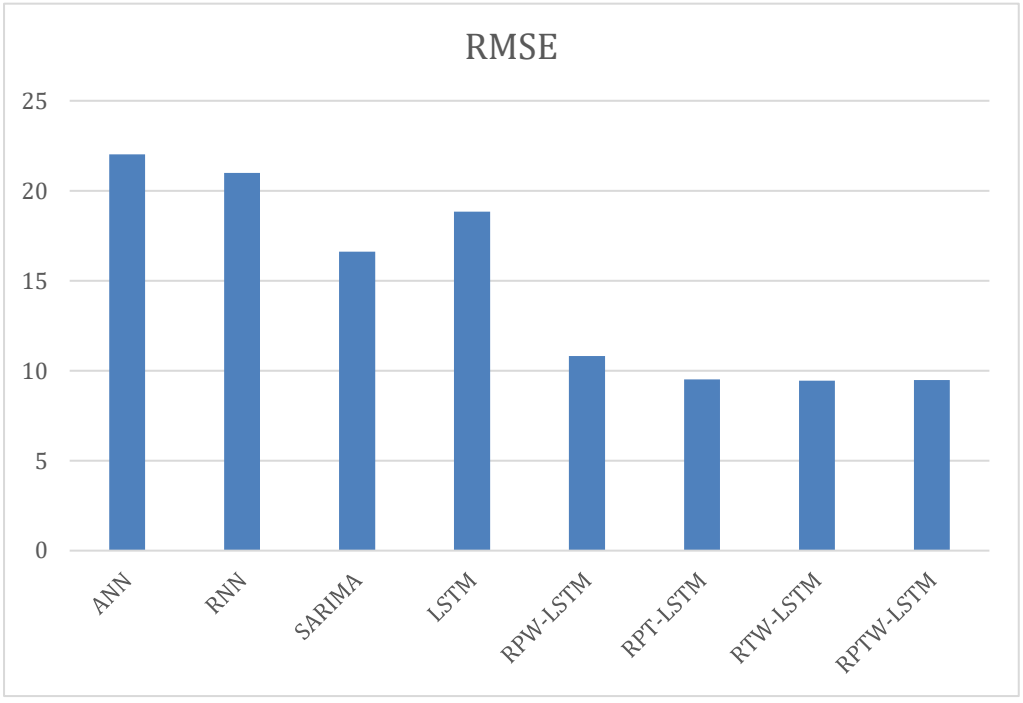


Fig 27: Comparison of RMSE of all algorithms

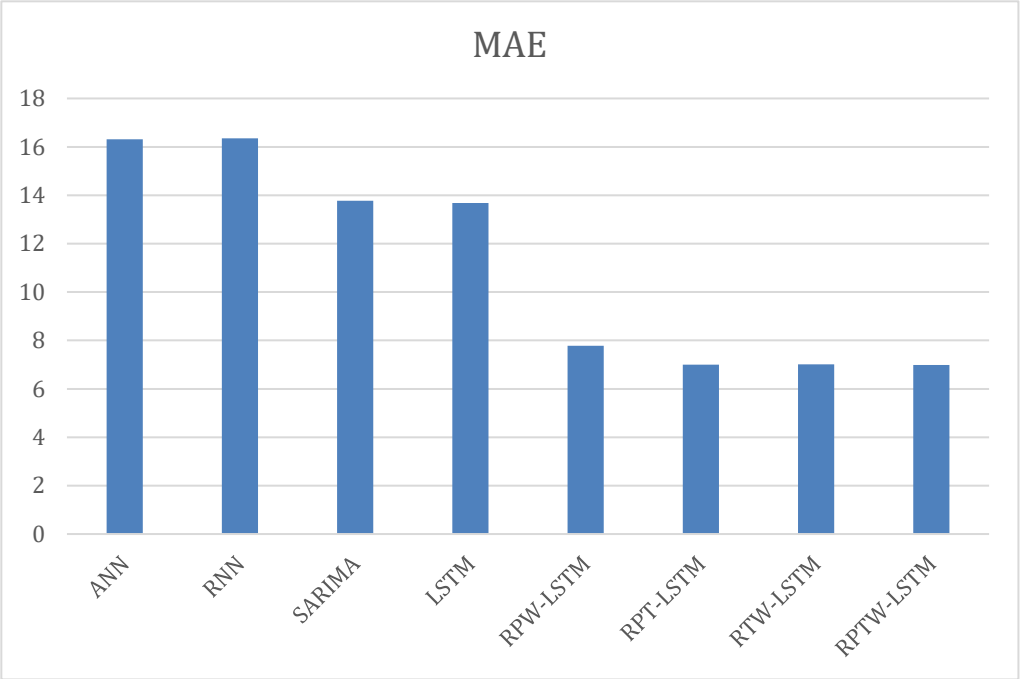


Fig 28: Comparison of and MAE of all algorithms

The percentage reduction in error for the proposed RPTW-LSTM model with respect to all the models in given in the table 5.2.

Table 5.2: Reduction in error of RPTW-LSTM w.r.t. all other models

	ANN	**RNN**	**SARIMA**	**LSTM**	**RPW-LSTM**	**RPT-LSTM**	**RTW-LSTM**
Reduction in error of RPTW-LSTM	43%	42%	42%	33%	14%	10%	8%

5.6. Summary

The proposed methodology was utilised in this chapter to predict the public bus passenger flow using the ETM data and weather data. Correlation of weather parameters with passenger flow has been found and rainfall affects the passenger flow the most and relative humidity, temperature has a least affect. Therefore, in the analysis only rainfall is considered. The proposed model was made by using hourly, daily,

weekly, and weather parameter (rainfall) variation and has been compared with other base line models. Evaluated the proposed model by comparing with other models like RPT-LSTM, RPW-LSTM, and RTW-LSTM which evaluates the performance of weather, weekly, and daily respectively. Figures 26, 27, and 28 are the error comparison graphs which gives an idea of how the proposed model is better than other models. Found proposed model (RPTW-LSTM) has better accuracy and minimal error than all other models.

CHAPTER 6

CONCLUSIONS

This study used deep learning LSTM algorithm to forecast the public bus passenger flow for Kelusanka route in Udupi city. The data was segregated in hourly basis before modelling. The prediction analysis and results have been described in earlier chapters. The main findings of the current study are summarized in this chapter.

Short term forecasting of the bus passenger flow is the major hotspot for making rationale bus scheduling decisions and to provide timetables to accommodate more passengers in peak hours. For the flow prediction problems, there are many similarities between classic transport models and feature engineering in machine learning. Therefore, the domain knowledge from existing transport models can be used to design deep learning models.

Short term passenger flow predictions highly depend on previous trends and seasonality i.e., hourly trend, daily trend, and weekly trend. Weather has a greater influence in passenger onboarding a public bus in an urban city. The ETM data and weather data has collected for a year 2022 and has a consistency from 7:00 am to 8:00 pm everyday. Correlation of passenger flow and weather variables rainfall, relative humidity, and temperature are found using pandas library, python. All the temporal characteristics (recent time intervals, daily periodicity, and weekly trend) and weather variable are modelled separately with LSTM neural networks and all the outputs of these models are fed into new LSTM neural network for final fusion of all models to predict the passenger flow. The fused model, RPTW-LSTM is then compared with various already present base line models and also with different combination of temporal and weather variable models to find the impact of each variable in the RPTW-LSTM. LSTM models are done in keras integrated TensorFlow version 2.11.0, various other machine learning libraries are used including scikit-learn

The conclusion from the study are as follows:

1. The correlation of weather parameters rainfall, relative humidity, and temperature with passenger flow gives rainfall has a greater influence, whereas relative humidity and temperature has less impact on passenger flow. Hence only rainfall has been considered for modelling.
2. RPTW-LSTM considers the temporal characteristics (recent time, daily periodicity, and weekly trend) and weather variable (rainfall) to predict the hourly bus passenger flow and has an error of 14.8%. This algorithm can be used for hourly bus passenger flow for any other routes which makes easy for transit controllers in bus scheduling and operating.
3. The proposed model RPTW-LSTM has better performance and more accurate when compared with baseline models like ANN, RNN, SARIMA, and LSTM. Also, when compared with different combinations of RPTW (RPT-LSTM, RPW-LSTM, and RTW-LSTM) with proposed model, found that daily periodicity has more weightage, then followed by weekly trend and less weightage for weather. As the analysis is done for whole year, there is a very less impact if we considered rainfall in the analysis as there are only three months of rainfall when compared to a whole year.

In conclusion, the model accuracy will be increased if the recent, periodicity, trend, and weather are included in the model and when compared to other baseline models the proposed model has a better performance in forecasting the bus passenger flow. Instead of whole route, this study can be further extended to modelling for stop wise which can give more information for each stop.

REFERENCES

Chan, Kit Yan, Tharam S. Dillon, Jaipal Singh, and Elizabeth Chang. "Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm." *IEEE Transactions on Intelligent Transportation Systems* 13, no. 2 (2011): 644-654.

Chen, Tao, Jie Fang, Mengyun Xu, Yingfang Tong, and Wentian Chen. "Prediction of Public Bus Passenger Flow Using Spatial–Temporal Hybrid Model of Deep Learning." *Journal of Transportation Engineering, Part A: Systems* 148, no. 4 (2022): 04022007.

Danfeng, Yan, and Wang Jing. "Subway passenger flow forecasting with multi-station and external factors." *IEEE Access* 7 (2019): 57415-57423.

Guo, Shengnan, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting." *IEEE Transactions on Knowledge and Data Engineering* (2021).

Hao, Siyu, Der-Horng Lee, and De Zhao. "Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system." *Transportation Research Part C: Emerging Technologies* 107 (2019): 287-300.

Hasnine, Md Sami, Jason Hawkins, and Khandker Nurul Habib. "Effects of built environment and weather on demands for transportation network company trips." *Transportation Research Part A: Policy and Practice* 150 (2021): 171-185.

Li, Can, Lei Bai, Wei Liu, Lina Yao, and S. Travis Waller. "Graph neural network for robust public transit demand prediction." *IEEE Transactions on Intelligent Transportation Systems* (2020).

Liu, Lijuan, Rung-Ching Chen, and Shunzhi Zhu. "Impacts of weather on short-term metro passenger flow forecasting using a deep LSTM neural network." *Applied Sciences* 10, no. 8 (2020): 2962.

Li, Yaguang, Rose Yu, Cyrus Shahabi, and Yan Liu. "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting." *arXiv preprint arXiv:1707.01926* (2017).

Li, Yang, Xudong Wang, Shuo Sun, Xiaolei Ma, and Guangquan Lu. "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks." *Transportation Research Part C: Emerging Technologies* 77 (2017): 306-328.

Liu, Yang, Zhiyuan Liu, and Ruo Jia. "DeepPF: A deep learning based architecture for metro passenger flow prediction." *Transportation Research Part C: Emerging Technologies* 101 (2019): 18-34.

Lin, Ciyun, Kang Wang, Dayong Wu, and Bowen Gong. "Passenger flow prediction based on land use around metro stations: a case study." *Sustainability* 12, no. 17 (2020): 6844.

Ma, Zhenliang, Jianping Xing, Mahmoud Mesbah, and Luis Ferreira. "Predicting short-term bus passenger demand using a pattern hybrid approach." *Transportation Research Part C: Emerging Technologies* 39 (2014): 148-163.

Mulerikkal, Jaison, Sajanraj Thandassery, Vinith Rejathalal, and Deepa Merlin Dixon Kunnamkody. "Performance improvement for metro passenger flow forecast using spatio-temporal deep neural network." *Neural Computing and Applications* 34, no. 2 (2022): 983-994.

Tang, Liyang, Yang Zhao, Javier Cabrera, Jian Ma, and Kwok Leung Tsui. "Forecasting short-term passenger flow: An empirical study on shenzhen metro." *IEEE Transactions on Intelligent Transportation Systems* 20, no. 10 (2018): 3613-3622.

Tao, Sui, Jonathan Corcoran, Francisco Rowe, and Mark Hickman. "To travel or not to travel: 'Weather'is the question. Modelling the effect of local weather conditions on bus ridership." *Transportation research part C: emerging technologies* 86 (2018): 147-167.

Wang, Peng, and Yuan Liu. "Network traffic prediction based on improved BP wavelet neural network." In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1-5. IEEE, 2008.

Wang, Xuemei, Ning Zhang, Ying Chen, and Yunlong Zhang. "Short-term forecasting of urban rail transit ridership based on ARIMA and wavelet decomposition." In *AIP Conference Proceedings*, vol. 1967, no. 1, p. 040025. AIP Publishing LLC, 2018.

Wang, Xuemei, Ning Zhang, Yunlong Zhang, and Zhuangbin Shi. "Forecasting of short-term metro ridership with support vector machine online model." *Journal of Advanced Transportation* 2018 (2018).

Wei, Ming. "How does the weather affect public transit ridership? A model with weather-passenger variations." *Journal of transport geography* 98 (2022): 103242.

Zhao, Feifei, Weiping Wang, Huijun Sun, Hongming Yang, and Jianjun Wu. "Station-level short-term demand forecast of carsharing system via station-embedding-based hybrid neural network." *Transportmetrica B: Transport Dynamics* 10, no. 1 (2022): 1-19.

Williams, Billy M., Priya K. Durvasula, and Donald E. Brown. "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models." *Transportation Research Record* 1644, no. 1 (1998): 132-141.

Zhang, Jinlei, Feng Chen, Zhiyong Cui, Yinan Guo, and Yadi Zhu. "Deep learning architecture for short-term passenger flow forecasting in urban rail transit." *IEEE Transactions on Intelligent Transportation Systems* 22, no. 11 (2020): 7004-7014.