IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 July 2024, accepted 24 July 2024, date of publication 1 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437172

RESEARCH ARTICLE

A Framework to Characterise, Estimate, and Predict Vehicle Class-Agnostic Traffic States and Class-Wise Speeds for Mixed Traffic Conditions

ABIRAMI KRISHNA ASHOK AND BHARGAVA RAMA CHILUKURI
Department of Civil Engineering, Indian Institute of Technology Madras, Chennai 600036, India
Corresponding author: Abirami Krishna Ashok (ce18d012@smail.iitm.ac.in)

This work was supported in part by Indian Institute of Technology Madras, and in part by the Ministry of Education, Government of India.

ABSTRACT For homogeneous traffic, where all vehicles are the same type, the traffic state is characterised by speed, flow, density, queue length, etc. In mixed traffic conditions, variations in static and kinematic characteristics among vehicles and the resulting asymmetric interactions that arise, these state variables are inadequate to represent class-wise behaviours. This paper proposes a novel framework for characterising mixed traffic conditions based on vehicle class-wise speeds rather than a single value of the aggregated stream speed. Also, it proposes an area occupancy-based approach to estimate class-wise speeds from class-agnostic disaggregated travel-time data. The empirical validation of the proposed traffic state definitions demonstrates their generalisability. Finally, parametric and non-parametric prediction models are also developed for state and class-wise speed predictions. The empirical results demonstrate that the joint prediction approach (simultaneous prediction of multiple classes using the proposed state definition) is more accurate, computationally effective, and more efficient for practical applications than the marginal predictions using class-wise speed predictions. Moreover, the order of the class-wise speeds is more robustly preserved in the former than in the latter. This research can open doors for a new family of class-wise speed-based traffic management strategies and applications for mixed traffic conditions.

INDEX TERMS Mixed traffic, speed-based state characterisation, class-wise speeds, travel-time data, traffic state prediction.

I. INTRODUCTION

Traffic state represents the traffic conditions on the road, generally defined using various metrics such as the number of vehicles, travel time, speed, traffic flow, density, queue length, etc. Traffic state estimation is important for transportation applications such as traffic control, traveller information, vehicle rerouting, etc. The above-mentioned state variables are generally used for homogeneous traffic conditions with one dominant vehicle class, car. However, mixed traffic conditions consist of a diverse range of vehicle classes with different static and kinematic characteristics. This inhomogeneity in the features of different vehicle classes would mean class-specific advantages and disadvantages during their manoeuvres; e.g., two-wheelers can seep through the traffic stream due to their smaller size, but bigger vehicle drivers can see farther due to their higher seating position. Similarly, varying acceleration, deceleration, and lateral movement are observed across vehicle classes due to intravehicular interactions and trade-offs between safety and efficiency under different congestion levels and traffic compositions. These driving behaviours manifest at the macroscopic level (e.g., class-wise dispersions along corridors) and the microscopic level (e.g., class-wise speed variations). As a result, different vehicle classes behave differently in mixed traffic conditions [1], [2], [3]. Due to the class-specific behaviours, traffic state characterisation is challenging in mixed traffic conditions.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

II. LITERATURE REVIEW

A. MIXED TRAFFIC CHARACTERISATION

Traditionally, researchers used Passenger Car Equivalency (PCE) to characterise the traffic state in mixed traffic conditions (e.g., [1], [4], [5], [6]). Philosophically, PCEs convert a mixed traffic stream into an equivalent car-only stream and use the traditional state variables of flow and density to characterise the mixed traffic state. However, the challenges associated with this approach are widely reported in the literature. The PCE values are dependent on various factors such as speed [7], composition [8], [9], facility type, number of lanes, and location [1], [4], [5], [6], etc. Therefore, this approach obscures the effects of intravehicular interactions and fails to represent class-wise behaviour.

Another approach taken by researchers to address heterogeneity is the concept of Area-Occupancy (AO) [10] which is used as an alternative for traffic density. In a mixed lane-free environment, all vehicles manoeuvre the entire road space, not in a single lane. Hence, the entire road width is considered while measuring the vehicle concentration in the area-occupancy approach. Several researchers applied the concept of area-occupancy to study mixed traffic conditions (e.g., [11], [12], [13], [14]). Specifically, some of these studies showed substantial noise in the Fundamental Diagrams (FDs), indicating asymmetric interactions across vehicles at the same density or AO but with varying composition [14].

PCE and Area Occupancy-based characterisation homogenize the traffic stream, thus resulting in only one state variable for the entire traffic stream. The major issue with this homogenization is that the resulting state variable does not have any physical meaning since it is neither observed nor any vehicle class represents this. In mixed traffic conditions, each vehicle class may travel at a different speed, and mapping from stream to individual classes and vice versa is challenging. Another limitation of these approaches is they fail to capture the traffic dynamics and vehicular interactions due to variations in the composition since they attempt to bypass the effect of composition. However, this leads to scatter observed in static and dynamic models, indicating that these approaches require two independent variables. (e.g., PCE density/area occupancy, composition)to uniquely define a state variable.

To address these challenges, multi-class models have been proposed as an alternative to aggregated approaches. In this approach, vehicle composition is retained, and class-wise interactions are studied under different traffic levels ([15], [16], [17], [18], [19], [20]). In multi-class models, the density range is divided into different regimes, such as free flow, semi-congested, congested, etc. Thus, it allows for different vehicle classes to have different speeds and unique relationships with other classes' behaviours (generally, speeds). However, most of the studies only focused on two vehicle classes (e.g., [15], [17], [21], [22], [23]). However, the limitation is that the approach is challenging to scale for multiple vehicle classes. It is difficult to calibrate since they require considerable class-specific data across a wide range of traffic conditions. Some multiclass models represent class-wise behaviour as a function of total density, but others represent them as a function of both class-wise densities and total density. While the former is relatively easier to estimate since the fundamental diagrams are monomials or binomials, the latter is difficult to estimate due to the polynomial nature of the speed function. Also, it is difficult to identify the number of regimes, regime boundaries, and the class-wise characteristics in each regime for the case of more than two vehicle classes, and it is not easily scalable. Therefore, there is a need to develop a traffic state characterisation methodology that preserves the class-wise vehicle states to characterise the mixed traffic conditions uniquely.

Even though the PCE and AO-based metrics are simpler than the multi-class models, they characterise traffic state using passenger cars or proportion of road space occupied, thus masking the effect of diverse vehicle interactions into a simplistic homogenisation metric for traffic characterisation. Even though multi-class models allow for interactions and behaviours at the class level, their development becomes complex with many vehicle classes, hindering scalability.

To overcome the limitations of the methods in the literature, this study:

- highlights the limitations of homogenisation of mixed traffic for traffic state characterisation.
- presents a framework to characterise mixed traffic using class-agnostic traffic states
- proposes a methodology to estimate class-wise speeds using class-agnostic travel time data.
- develops parametric and non-parametric models to jointly predict class-wise speeds based on the proposed traffic states.

Once the traffic state characterisation methodology is developed, one could uniquely characterise the traffic state for any traffic condition (generally called state estimation). However, anticipating future conditions requires traffic state prediction models, which will be discussed in the next section.

B. MIXED TRAFFIC STATE PREDICTION

In general, prediction methods are categorized as "Parametric approaches" and "Non-parametric approaches" [24]. In Parametric models, the parameters have to be specified before they can be used to make predictions, while non-parametric models do not rely on specific parameter settings. Statistical models [25], Kalman Filter [26], [27], [28], and regression models come under the Parametric approaches. Examples of non-parametric models include support vector machines [29], [30], [31], k-nearest neighbor [25] and deep learning methods such as Neural Networks [31]. Both parametric and non-parametric models work well with probe and location-based data. Even though parametric models can select from comprehensive options to represent the traffic system better, it is highly site-specific and not readily transferable. On the other hand, non-parametric models do not require any pre-selection of a model form. But they are

highly data-hungry [32]. Statistical or parametric approaches are the most commonly used prediction techniques, followed by neural network models and Ensemble techniques [33].

For mixed traffic conditions, studies explored different parametric and non-parametric methods for travel-time prediction of one particular vehicle class using probe data [25], [26], [28], [34]. Under mixed and less lane-disciplined traffic conditions, Kumar et al. [26] proposed a Kalman filtering method to predict stream travel time from bus travel times. However, it was found that only two-wheeler travel time was compared against bus travel time despite the study being based on mixed traffic conditions on arterial routes. Jairam et al. [25] evaluated the performance of bus travel-time predictions using Kalman filter, AutoRegressive Integrated Moving Average (ARIMA) and k-Nearest Neighbour (k-NN) classifier. But to arrive at a stream-level travel-time prediction, mapping predictions from individual vehicle classes to the traffic stream must be carried out. Prediction of the travel time for the concerned vehicle class through probe data is possible. Sihag et al. [34] predicted travel time using trajectory data using the historical average method, regression models and decision trees. The applicability of the aforementioned models in the context of travel-time prediction is well-explored. The major inference from the literature is that the studies in the literature generally dealt with predicting speed for one or two vehicle classes for heterogeneous traffic and one of the challenges of this approach is to map the probe travel-time prediction to other vehicle classes in the traffic stream. Another drawback of these studies is that the public transit frequency is generally much lower than the other vehicle classes, resulting in a small sample size (compared to other vehicle classes) for their prediction and extrapolation to the traffic stream. Banik et al. [35] investigated a panel modelling approach, a less data-intensive method to predict stream travel time. The developed model better captures bus and stream travel time spatiotemporal variability. However, an aggregate stream-level travel time is inappropriate for mixed traffic with varying vehicle dynamics.

Antoniou et al. [36] combined a data-driven approach with traffic flow theories. In this study, various traffic states were identified by clustering the observations, and a Markov process is used to estimate the transition process. Neural Networks were used to classify the new observations into appropriate clusters, and appropriate traffic models were used to predict the traffic speed. While this study combines the traffic flow theory and data-driven prediction models, this framework needs much traffic flow and density data. Also, it is valid only for homogeneous traffic conditions.

From the above literature review, the following gaps are identified. Passenger Car-equivalency-based characterisation homogenizes the traffic stream and fails to capture the traffic dynamics and vehicular interactions. Estimated PCEs are based only on the observed composition, making them static. Dynamic PCEs capture the vehicle dynamics but remain invariant to composition. Area occupancy overcomes the limitations of traditional density measurement for mixed traffic environments. But, the derived FDs represent the same speed for all vehicle classes. Class-specific area occupancy yields different speeds for different vehicle classes but is similar to dynamic PCE; this also remains composition-invariant. Composition-specific FDs result in the same speed for all the vehicle classes. Multi-class models consider almost two vehicle classes and assume the same speeds among the vehicle classes. They do not consider the vehicular interactions and differences in speeds of different vehicle classes. The mapping of individual vehicle classes to stream and vice versa is complex, and they do not consider vehicular interactions. Also, aggregated travel time may not be appropriate for a mixed traffic stream with highly varying vehicular dynamics.

For mixed traffic characterisation, it is impossible to use analytical methods such as multi-class models. Because when there are more vehicle classes, it is difficult to identify the number of regimes, regime boundaries, and class-wise characteristics in each regime. Hence, the traditional multiclass model approach is not easily scalable in the multi-class scenario. Most existing models assume class-specific speeds as a function of total density. However, a few models that proposed vehicle speeds as a function of class-specific densities are difficult to calibrate. Generally, these models are tedious to calibrate since they require a large amount of class-specific data under the full range of traffic conditions.

When the traffic evolves, each vehicle class impacts the other vehicle classes, and hence, the traffic state on the road is not an independent phenomenon; it is a joint distribution of individual vehicle classes' speed. Hence, characterising the traffic states as the joint distribution is most appropriate.

Traditional prediction methods characterise the traffic state by stream variables, which is inappropriate for mixed traffic conditions. Most of the existing prediction methods do not consider the traffic state progression into account. Limited studies have used data-driven approaches for traffic state prediction for homogeneous traffic but none for mixed traffic. Most data-driven modelling is based purely on data and does not include traffic flow theories. No prediction methods exist in the literature for class-wise behaviour-based traffic state characterisation.

To fill these gaps, this study develops a methodology for characterising mixed traffic conditions and models to predict traffic state in urban arterials. The study's objectives are: i) To characterise the mixed traffic conditions by vehicle-class-specific speeds incorporating the spatio-temporal data. ii) To validate the characterisation methodology using empirical data. and iii) To estimate class-wise speeds from class-agnostic disaggregated travel-time data iv) To devise a corridor-level traffic state prediction methodology for the proposed state characterisation and benchmark it with the state of the practice methods. Since many traffic states are possible, a data-driven approach is chosen to study the most commonly observed traffic states from the data. Also, the study combines traffic flow theory and a data-driven

approach for traffic characterisation, state estimation and state ordering, thus making it knowledge-guided data-driven modelling.

The contributions of this study are:

- A novel framework for mixed traffic characterisation based on individual vehicle class speeds rather than stream speeds.
- A flexible traffic states definition scheme based on the vehicle class-specific speeds in mixed traffic and local conditions.
- A framework to estimate class-wise speeds from class-agnostic disaggregated travel-time data.
- A location agnostic knowledge-guided state ordering methodology for mixed traffic.
- Demonstrated that the joint probability-based prediction outperforms the marginal probability-based prediction.

The rest of the paper is organised as follows. The ''Methodology'' section presents the overall methodology of the proposed characterisation and prediction framework. The next section presents the characterisation of mixed traffic conditions based on speeds. The ''Class-wise speed estimation'' section demonstrates the vehicle-classwise speed estimation scheme based on the proposed speed-based characterisation. Then, based on the defined traffic state, a state prediction methodology is proposed and demonstrated in the ''Prediction of traffic states and class-wise speeds'' section. Finally, the major conclusions are summarised, the significant contributions and directions for future work are proposed in the section ''Discussion and Conclusions''.

III. METHODOLOGY

The schematic diagram of the proposed speed-based characterisation and traffic state and speed prediction methodology is illustrated in figure 1. The speed of the vehicles was calculated from the travel time data collected through Wi-Fi Media access control Sensors (WMS). The travel time data of multiple vehicles collected during a five-minute period are pre-processed to transfer them into an equivalent binary form of speed-bin data. This binned data is used for data-driven state identification of mixed traffic state characterisation. Since many traffic states are theoretically possible due to various speed combinations of vehicle classes, a data-driven methodology is chosen to identify the most common traffic states. Unsupervised learning approaches are used on the dichotomous dataset (binary form of the speed bin data) to cluster the data. The distribution-based clustering method, the Gaussian Mixture Model algorithm [37], is chosen due to the nature of the input data. The clustering algorithm's efficiency and the clusters' homogeneity are analysed through 'Jaccard index' [38]. In unsupervised clustering results, the clusters with the highest observation counts are chosen to represent the dominant traffic states for characterisation. The observed traffic states from the study area are validated using data from another location (study area 2) that showed that the proposed characterisation is transferable to other locations.

To study the distribution of traffic states across the time of the day, each 5-minute traffic speed data is mapped to a corresponding traffic state and studied for possible trends in the traffic state dynamics and evolution within a day. The temporal evolution of traffic states is analysed for each day by grouping daily observations into distinct patterns representing a specific traffic scenario (i.e., morning peak, afternoon off peak, evening peak, off-peaking, etc.), using the k-means clustering algorithm [39]. In the pattern identification phase, a logistic regression model is employed to predict the corresponding pattern for each observation. Next, traffic states are predicted using lagged state information and pattern probabilities. The state predictions are then used to estimate the class-wise speeds simultaneously. This joint prediction capability is termed ''Joint Prediction'' in the study. Existing approaches to class-wise speed prediction are considered for benchmarking the proposed method. Vehicle-class-specific prediction models are developed using parametric and non-parametric methods (logistic regression and neural network), using lagged speeds as input features. Predicted speeds for all vehicle classes are combined to represent the predicted traffic state for each observation. The proposed joint and bench-marking marginal models were analysed for their performance by comparing the overall prediction and class-wise performance. Further, the speed prediction results are analysed to preserve the order or ranking, i.e., a rank is given to each vehicle class based on its speed.

IV. SPEED-BASED CHARACTERISATION

The proposed state estimation methodology comprised state definition and ordering of the defined traffic states based on class-wise speeds. An overview of the characterisation methodology is given in figure 2. Travel-time data are collected and clustered using unsupervised clustering techniques. Based on the cluster characterisation, the lower and upper-speed bounds are defined for traffic states. Considering the area occupancy, the defined traffic states are ordered from free-flow to congested conditions. Thus, the traffic state definition step implies the speed-based characterisation of mixed traffic. More details of the steps involved are explained in the following subsections.

A. DATA COLLECTION AND PRE-PROCESSING

The rapid advancements in computing technology have greatly improved automated vehicle detection and vehicle classification methods using sensors such as loop detectors, video cameras, infrared, RAdio Detection And Ranging (RADAR), LIght Detection And Ranging (LIDAR), Radio Frequency Identification (RFID), etc. (see [40], [41] for more details). The vehicle speeds used in this study were calculated from the travel time data collected through Wi-Fi Media access control Sensors (WMS) installed on the roadside. These sensors are installed on the roadside to passively capture Media Access Control (MAC) IDs, the signal strength, and the timestamps of Wi-Fi-enabled devices in the

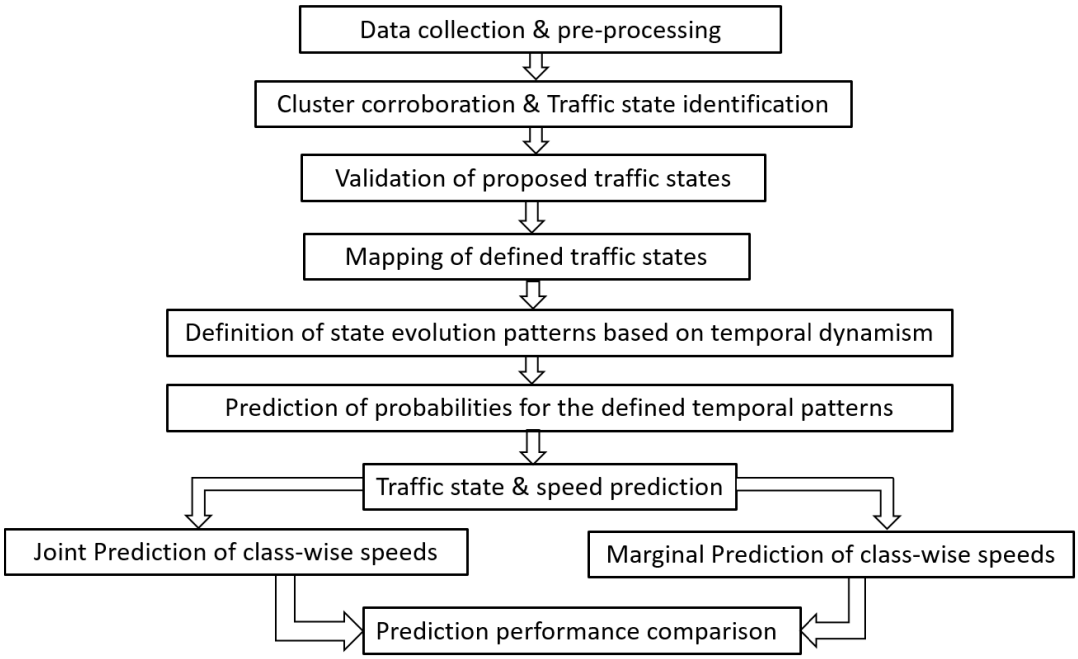


FIGURE 1. Proposed methodology.

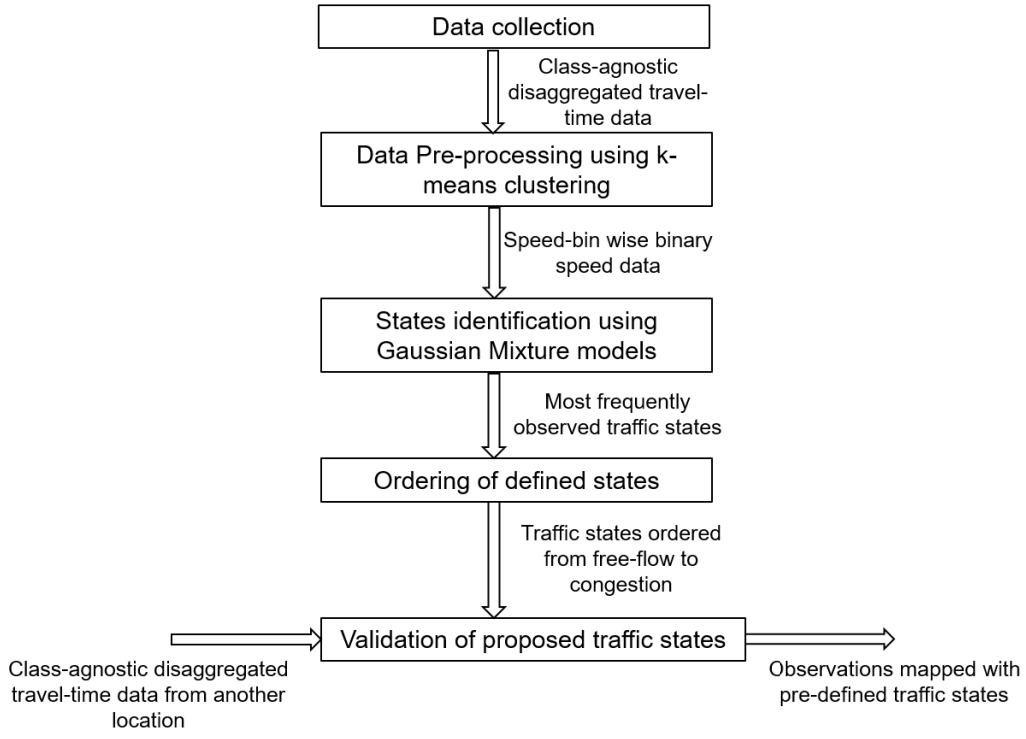


FIGURE 2. Proposed speed-based characterisation methodology.

vehicles crossing the sensor location. Travel time is obtained by matching the MAC addresses at both ends of the corridor. These sensors collect travel time data of the Wi-Fi-enabled devices in the vehicles. Therefore, the collected sample contains data from all vehicle classes. More details on the Wi-Fi sensors used in this study can be found in [42]. The travel time data is ''class-agnostic'', indicating that the travel time values are not marked with the corresponding class of

FIGURE 3. Data collection location (study area 1 and study area 2).

the vehicle from which it was collected. Several sensors, such as Bluetooth sensors [43], Wi-fi sensors [42], RFID sensors [44], etc., give such data. These devices reidentify the MAC address of the mobile devices in the vehicles or the tag id on the vehicle at different locations to extract the travel times and speeds without identifying the vehicle type, making the input data ''class-agnostic''. However, these sensors are popular since they help capture large amounts of data while capturing the variability in the speeds of various vehicle classes in the traffic stream. Note that the proposed speed-based characterisation methodology is sensor invariant and can use travel time or speed data from any sensors (e.g., Automatic Vehicle Location (AVL), RFID, Bluetooth, Global Positioning System (GPS), Radar sensor, Video camera, etc.) This study collects travel time data from two urban mid-block locations as shown in figure 3. The study areas are located between Madhya Kailash and Tidel Park intersections of Old Mahabalipuram Road (OMR), Chennai, India. The collected travel times are aggregated into five-minute periods, thus contributing to 288 periods over 24 hours. A sample five-minute travel time data is shown in Table 1. The database of study area 1 [(13.00645, 80.24422) to (13.00391,80.24750)] consists of 31 days of travel time observations on a midblock section of 700 m length and the database of study area 2 [(12.99518,80.24950) to (12.98792,80.25142)] consists of 48 days of travel time observations on a midblock section of 800 m length.

The input data is class agnostic and does not consider the vehicle composition. However, the smaller sample size of public transport is accounted for in two ways for the traffic state characterisation. First, the aggregation duration is set to 5 minutes to ensure that the duration is long enough to capture all the vehicle classes (note that the study area is a major corridor that serves multiple bus routes with small headways). Second, the 5-minute data is considered for analysis only if it has a sufficient sample size. Thus, the data from public transport is also included in the analysis.

The raw travel time observations aggregated for five-minute intervals are grouped into an optimal number of clusters, each representing a specific speed group. The optimal number of clusters is identified based on the Elbow method [45]. The pre-processing steps to identify the suitable speed bin are shown in table 1. First, the average travel time for each cluster is calculated, and then the average travel-time values are converted into cluster average speeds representing the space-mean speed values of each vehicle class. Note that space mean speed is commonly used in the traffic flow theory literature to represent traffic conditions. The traffic state definition based on Space Mean Speed (SMS) will be robust and relevant since SMS is consistent with the fundamental relationship of q=kv and more accurately represents the traffic conditions than the Time Mean Speed (TMS). According to Edie's generalised definitions, space mean speed can be obtained as $\frac{\Sigma\ Distance}{Average\ travel\ time}$ since all vehicles travel the whole distance between the sensors.

Since the range of speeds typically seen in urban areas is between 5 kmph to 65 kmph, five kmph is considered as the minimum speed (considering the pedestrian speed), and 65 kmph is taken as the highest speed (considering the posted speed limit). Based on the descriptive statistics of the given data and local conditions, speeds outside this range are considered to be outliers, thus yielding 12 bins ranging from 5 kmph to 65 kmph with 5 kmph intervals. Therefore, the space-mean speed values obtained are then associated to the respective 5 kmph speed bins. The equivalent binary form of the data from table 1 is shown as bold marked in table 2. This binned data is used for data-driven state identification and all further consequent steps for mixed traffic state characterisation.

B. STATE IDENTIFICATION

Traffic state in this study is defined as the combination of specific speeds of different vehicle classes. Each five-minute observation data has a unique combination of speed distributions. Since there are twelve-speed bins and in binary representation, 1111111111_2 is equivalent to 4095_{10}. Hence, there are 4095 combinations possible considering twelve-speed bins with binary data. Since there are theoretically many possible traffic states, a data-driven methodology is chosen to identify the most commonly occurring traffic states. Therefore, unsupervised learning is used to cluster the five-minute observation data. Since the dataset is dichotomous, distribution-based clustering methods are chosen. The given dataset is modelled as a mixture of different Gaussian distributions. Thus, the Gaussian Mixture Model (GMM) algorithm clusters the data using the Python package 'GaussianMixture' [37]. The number of clusters is decided based on the 'Silhouette score' [39], β, as defined in equation 1 below.

$$\beta = \frac{(b-a)}{max(a,b)} \tag{1}$$

where,

a = mean intra-cluster distance,

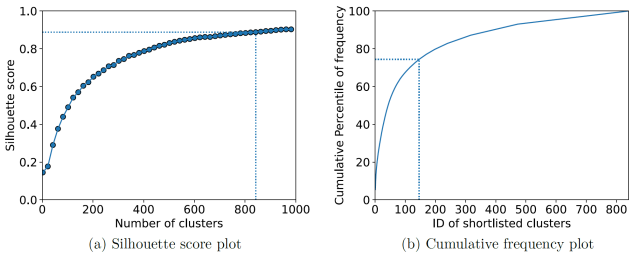
b = mean inter-cluster distance,

TABLE 1. Sample pre-processing of the data for a day in 5 minute aggregation period (09:15 to 09:20 am).

Travel time (s)	Cluster	Average travel time	Average speed (kmph)	Speed bin
125.05 132.91 126.72 118.74	A	125.85	20.02	20-25
95.00 89.85 108.46 83.26	B	94.14	26.77	25-30
179.98 165.74 173.39	C	173.04	14.56	10-15
145.05 154.82 160.42	D	153.43	16.42	15-20

TABLE 2. Binary form of the sample pre-processed data.

Time Period Speed bin	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65
08:10 to 08:15	0	0	0	0	0	1	0	0	0	0	0	0
09:15 to 09:20	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
12:15 to 12:20	0	0	0	0	0	0	0	0	0	1	1	1
12:25 to 12:30	0	0	0	0	0	0	1	1	0	0	0	0
17:15 to 17:20	0	0	1	1	1	1	1	0	0	0	0	0
18:15 to 18:20	1	1	1	1	1	1	0	0	0	0	0	0



(a) Silhouette score plot (b) Cumulative frequency plot

FIGURE 4. Data grouping using Gaussian Mixture Models (a) Identification of optimal number of clusters using Silhouette score plot (b) Cumulative frequency plot of top 840 clusters.

Note that the best value is 1, and the worst is −1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster. Clusters produced by a good clustering method will have high intra-class and low inter-class similarities.

To find out the optimal value of the number of clusters based on the Silhouette coefficient, GMM is fit by varying n=2 to 1000. It is found that the silhouette score stabilized with a marginal increase in the score after n=840, indicating a large number of traffic states. Furthermore, most states had very few instances of five-minute observation, suggesting that many of these states may be infrequent outliers.

Without sacrificing at the quality of the states, the cumulative frequency plot of the top 840 states is considered, shown in figure 4. From the plot, it can be inferred that about 70% of the data points (five-minute observation) are within the first 120 states, and the remaining 720 states account for only 30% of the data. Therefore, the first 120 more dominant and popular states are shortlisted for further study of the homogeneity of the observations within the state.

For a good clustering algorithm, all the observations within a cluster should be similar to each other. A clustering algorithm's efficiency and the clusters' homogeneity are analyzed through similarity measures. In this study, 'Jaccard index' [38] is taken as the metric to evaluate the performance of the Gaussian Mixture Models, the clustering algorithm used for data grouping. The Jaccard index or Jaccard similarity coefficient (J(A,B)) is defined as the ratio between $|A \cap B|$ and $|A \cup B|$ where $|A \cap B|$ gives the number of members shared between both sets and $|A \cup B|$ gives the total number of members in both sets (shared and un-shared). The Jaccard similarity will be 0 if the two sets don't share any values and 1 if the two sets are identical.

In the current context, the Jaccard similarity coefficient for the binary data can be written as follows:

$$J(A, B) = \frac{C_{11}}{C_{11} + C_{01} + C_{10}} \tag{2}$$

where

C_{11}: Number of speed bins where both A and B have the value 1.

C_{01}: Number of speed bins where observation A is 0 and B is 1.

C_{10}: Number of speed bins where observation A is 1 and B is 0.

For example, if two members are represented as A={1,0,1,1,0,1,1,1,1,1,0,0} and B={1,0,1,1,1,1,1,1,1,1,0,0}, the value of C_{11},C_{01},C_{10} are 8,1,0 respectively giving an J index of 0.89.

For the clustering results of the GMM, it is found that the Jaccard index ranged between 0.81 to 1 for the 120 clusters considered. Out of 120 clusters, 116 clusters showed Jaccard

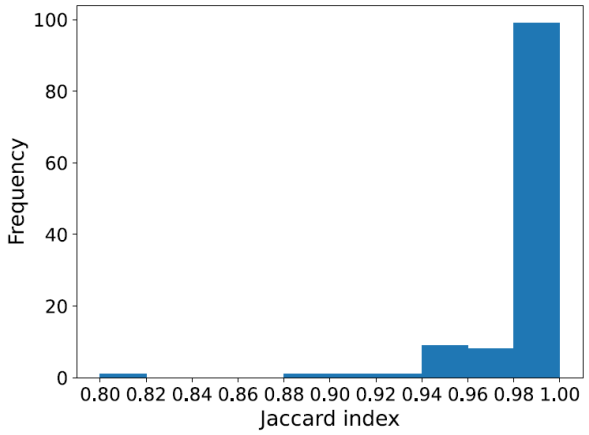


FIGURE 5. Evaluation of clustering efficiency using Jaccard index plot.

index greater than 0.95 (Figure 5). The remaining clusters showed Jaccard index in the range of 0.8 to 0.95. The Jaccard similarity scores show that the clusters have high levels of homogeneity, and the clustering using GMM was efficient.

The observations with continuous speed bins occupied by at most five vehicle classes are considered. From the unsupervised clustering results, considering the 12-speed bins of interest and most commonly observed patterns in the data, 50 traffic states are defined. [Figure 6]. These defined traffic states represent the traffic conditions. But, from one traffic state to another, it doesn't represent the traffic condition from congestion to free flow. Hence, there is a need for ordering traffic states based on some traffic flow principles.

C. STATE ORDERING

To order the proposed traffic states in ascending order from free flow to congestion, the concept of Area Occupancy (AO) has been considered as the measure. For each of the 50 defined states, AO has been calculated. Then, traffic states are ordered in the ascending order of AO so that state 1 denotes the least AO (free-flow condition) and state 50 denotes a high AO (highly congested conditions).

Mallikarjuna and Rao [10] have introduced AO to account for the width of the vehicle in the occupancy calculation. Thus, AO is calculated based on the entire road width irrespective of the number of lanes and AO associated with jam density equals one. The fundamental diagram parameters are calculated for each vehicle class considering the vehicular dimensions and the saturation headway values from [46] [Table 3].

Each traffic state is defined as a speed band with a lower and an upper-speed bin. There are 12 speed bins based on the speeds observed in the data ranging from 5-10 kmph to 60-65 kmph. The following assumptions are made to define the traffic states from the 120 observed clusters. A triangular fundamental diagram is used. 5 predominant vehicle classes observed in urban areas are considered, and their vehicular

TABLE 3. Fundamental diagram parameters of different vehicle classes.

Vehicle type	L (m)	h (sec)	w (kmph)	k_{cr} (veh/lane/km)	Q (veh/hr)	AO_{cr} (%)
2W	1.80	1.58	13	70	4571	12.5
3W	2.60	1.87	8	32	2080	32.5
CAR	4.20	1.73	14	30	1950	42.2
LCV	5.00	2.42	12	23	1488	44.8
HV	10.00	3.51	16	16	1027	61.2

dimensions are taken from [47]. They are two-wheeler (2W) (area 1.08 m^2), three-wheeler (3W) (3.64 m^2), car (CAR) (7.14 m^2), light commercial vehicle (LCV) (9.5 m^2), and heavy vehicle (HV) (25 m^2). Based on the assumption that all vehicle classes are observed during the 5-minute observation period, the ordering of states is invariant to vehicle composition.

It is to be noted that k_{cr} values are in the decreasing order while moving from 2W to HV whereas AO_{cr} is in the increasing order. From this, a major inference has been made that the dynamism exhibited by all these five vehicle classes is in the order of 2W,3W, CAR, LCV and HV. This order is reversed while considering k_{cr} and contradicts the order reported in literature [11]. The reason is that smaller vehicles will occupy less area with more vehicles. For example, let AO_{cr} of 12.5 % represent x number of two-wheelers. A similar number of heavy vehicles associated with AO_{cr} of 12.5 % be y. Due to the smaller vehicular area of 2W over HV, x is much larger than y. Therefore, based on table 3, it is clear that bigger vehicles have a smaller critical density, but higher AO_{cr} compared to that of the smaller vehicles. Moreover, it is observed that the AO_{cr} of 2Ws is much lower than that of the HVs. Therefore, 2W responds faster than other vehicle classes to any change in traffic conditions.

The present study uses the class-wise speeds corresponding to total Area Occupancy from the class-wise fundamental diagrams. This approach is taken since the composition information is not available from the Wi-Fi data and adequate composition-specific fundamental diagrams are not available in the literature.

Using the definition of AO, AO associated with jam density (AO_j) and critical density (AO_{cr}) are calculated [14]. From the relationship between flow (Q), density (k), and AO, the vehicle class-wise v vs AO have been estimated. Using the same, the AO values of all 12 speed bins of interest are calculated.

Each traffic state has been defined in such a way that, within the traffic state, each vehicle class belongs to a particular speed bin. For each traffic state, an area occupancy value is chosen so that it's the smallest possible AO that satisfies all the vehicle classes. Thus, each state has been assigned an AO value. Then, states were ordered in ascending order of AO so that state 1 has the lowest AO, which denotes the free-flow condition, and state 50 has the highest AO, which denotes the congested state. Figure 6 (c) shows the state-wise AO values. Due to the uniform slope for the line between state three and state 45, it is inferred that the states

are uniformly distributed. And beyond this range, the states are farther apart.

Figure 6 (d) shows the 50 defined states in the order of 1 being the free flow to 50 being the congested state based on AO. While comparing figure 6 (a) and 6 (d), it is observed that traffic states ordered based on AO clearly represent the traffic conditions in the order of free-flow to congestion.

As noted in [14], the wi-fi sensor has a bias in the data captured;it captures more slower vehicles than faster ones. Hence, to address this sensor bias, if observations have more than five continuous speed bins occupied, truncation is done on the lower speeds to account for the sensor bias. Thus, at the end of this step, the observations become labelled data with the state numbers as labels. This labelled data will be used in subsequent steps.

From figure 7 (a) and 8 (a), it is observed that all the defined traffic states are observed in the section of interest. Further, the traffic states are grouped based on their dispersion of speed bins. The states in which all the five vehicle classes have the same speed are considered in speed bin dispersion=1. Likewise, a dispersion plot is being tried to visualise the observed trend in the data [Figure 7 (b) and 8 (b)]. It is inferred that the section of interest had a reasonable number of observations in almost all possible types of dispersion, especially where the dispersion is >= 4. This justifies the need for vehicle-class-specific speed prediction. Because, in mixed traffic conditions, stream-level traffic prediction or average speed value for all the vehicle classes observed does not hold good.

The proposed state characterization defines traffic states based on multiple speed bins representing the variations of speeds across vehicle classes, unlike the traditional approaches that characterize traffic state using a single value of speed. Hence, the proposed state characterization is ''class-agnostic'' as it does not tag the vehicle class with each of the multiple speeds corresponding to a proposed traffic state. However, in section V of the paper, we relax this using area occupancy approach to map individual vehicle classes to each of the speed bins.

D. VALIDATION OF PROPOSED TRAFFIC STATES

To validate the characterised traffic states proposed in section 3.4, data from study area 2 is used. Figure 8 (a) shows the frequency plot of the predefined traffic states based on study area 1 against the data from study area 2. It can be inferred that most traffic states defined in section 3.4 are directly observed in study area 2.

Also, study area 2 had observations in all possible types of dispersion, indicating that the proposed speed-based traffic states may be suitable for other urban areas in mixed traffic conditions (Figure 7 (b)). However, the upper bin ranges may be extended to include new traffic states in addition to the proposed 50 states at higher speed locations.

Congested traffic states are observed more in both study area 1 and 2. The analysis also reveals that both the sections have a substantial number of observations across almost all possible types of dispersion. This significant presence of varied dispersion levels shows the necessity for vehicle-class-specific speed prediction models. In mixed traffic conditions, relying solely on stream-level traffic predictions or using an average speed for all vehicle classes proves inadequate. Each vehicle class may exhibit distinct speed patterns due to factors such as size, maneuverability, and driving behavior.

In this study, data from study area 1 is used for the traffic state definition of the characterisation framework and to validate the traffic states derived from the proposed speed-based characterisation, data from study area 2 is used. The validation using data from another location showed that the defined traffic states are reliable and transferable to similar heterogeneous traffic conditions.

V. CLASS-WISE SPEED ESTIMATION

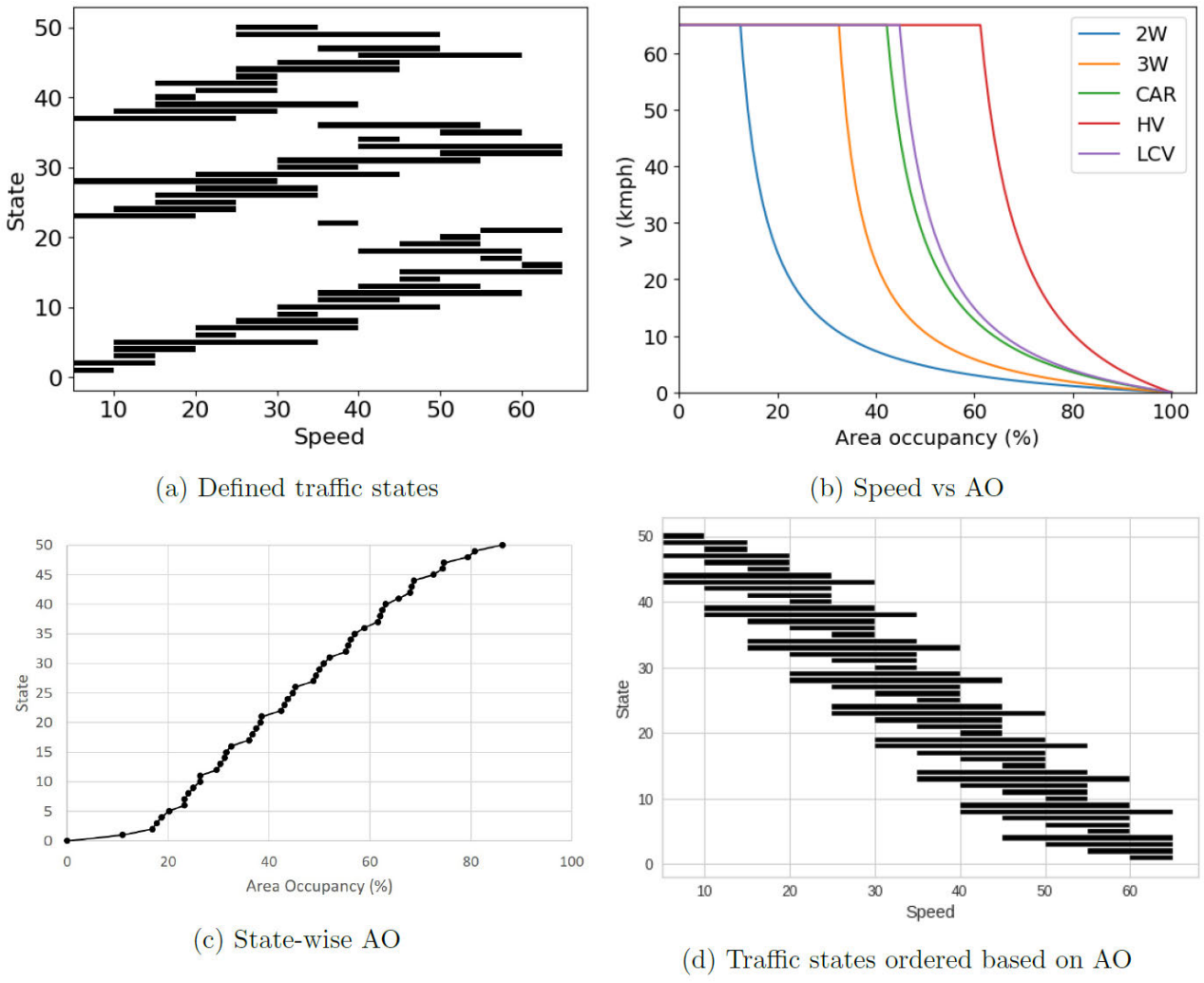
This section elaborates on the framework to arrive at the class-wise speeds from estimated traffic states. Section III shows that each defined traffic state has a specific range of speed bins occupied.

The present state definition allows for five different combinations as follows:

- All vehicle classes in the same speed-bin.
- There are Four vehicle classes in a speed bin and only one in the lower speed bin.
- Three vehicle classes in a speed bin, one in its immediate lower speed bin, and the other in one more lower speed bin.
- Two vehicle classes in a higher speed-bin and all the other three occupying the subsequent lower speed-bins.
- All the five vehicle classes occupy different continuous speed bins.

Figure 6 (b) shows the relationship between speed and AO for all the five classes considered. Until the critical AO, all vehicle classes remain at free-flow speed. After the critical AO, its speed starts reducing and reaches to zero at AO_{jam}. Hence, the ordering is influenced by the critical AO of vehicle classes. Based on the speed vs AO plot (see figure 6 (b)), it is inferred that the speed dynamism exhibited by different vehicle classes is in the order of 2W, 3W, CAR, LCV, and HV. Therefore, vehicle classes in the above-mentioned categories are allotted based on this exhibited speed dynamism. Figure 9 (a) shows the vehicle class-wise speeds for all the fifty predefined traffic states.

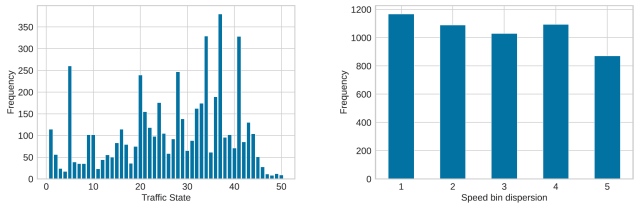
Each defined traffic state has a band of AO values. Defined traffic state one represents the free-flow scenario where all the vehicle classes travel at their free-flow speed. This traffic state corresponds to an AO value from zero to the AO_{cr} value of the 2W. The class-wise speeds for the first ten defined traffic states are detailed in figure 9 (b). The traffic state two represents the condition in which, except 2W, all other vehicle classes continue to be in their free-flow speed. 2W, being more dynamic than the other vehicle classes, slows down from 65 kmph to 60 kmph speed in traffic state two. But it is also to be noted that the traffic states with all the



(a) Defined traffic states (b) Speed vs AO

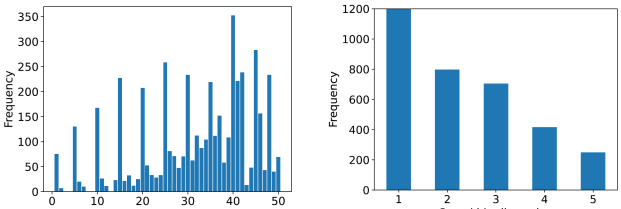
(c) State-wise AO (d) Traffic states ordered based on AO

FIGURE 6. Traffic State Definition (a) Defined traffic states (b) Vehicle class-wise speed vs Area occupancy (c) State-wise Area Occupancy (d) Defined traffic states ordered based on Area Occupancy.



(a) Frequency Plot (b) Dispersion Plot

FIGURE 7. (a) Frequency Distribution of Traffic States and (b) Speed Bin Dispersion.



(a) Frequency Plot (b) Dispersion Plot

FIGURE 8. Validation using data from Study area 2: (a) Traffic State Frequency and (b) Speed Bin Dispersion.

five vehicle classes occupying different speed bins represent a more congested scenario and are assigned to the state number eight and not five. Hence, it is inferred that the class-wise speeds are not monotonously decreasing and have notable differences in the order of states when all the five vehicle classes occupy different speed bins. Similarly, the defined fifty traffic states' last ten traffic states are represented in figure 9 (c). State fifty represents the most congested scenario, with all five vehicle classes traveling at 5 kmph speed. State 49 represents the traffic condition where all the other four vehicle classes travel at 10 kmph speed except 2W, whose speed is 5 kmph. Thus, the speed dynamism exhibited by figure 6 (b) is followed throughout the state definition and state ordering. Through figure 9 (b) and 9 (c), it is also inferred that traffic state 8 and 43 are the examples for traffic conditions where all the five vehicle classes are in different

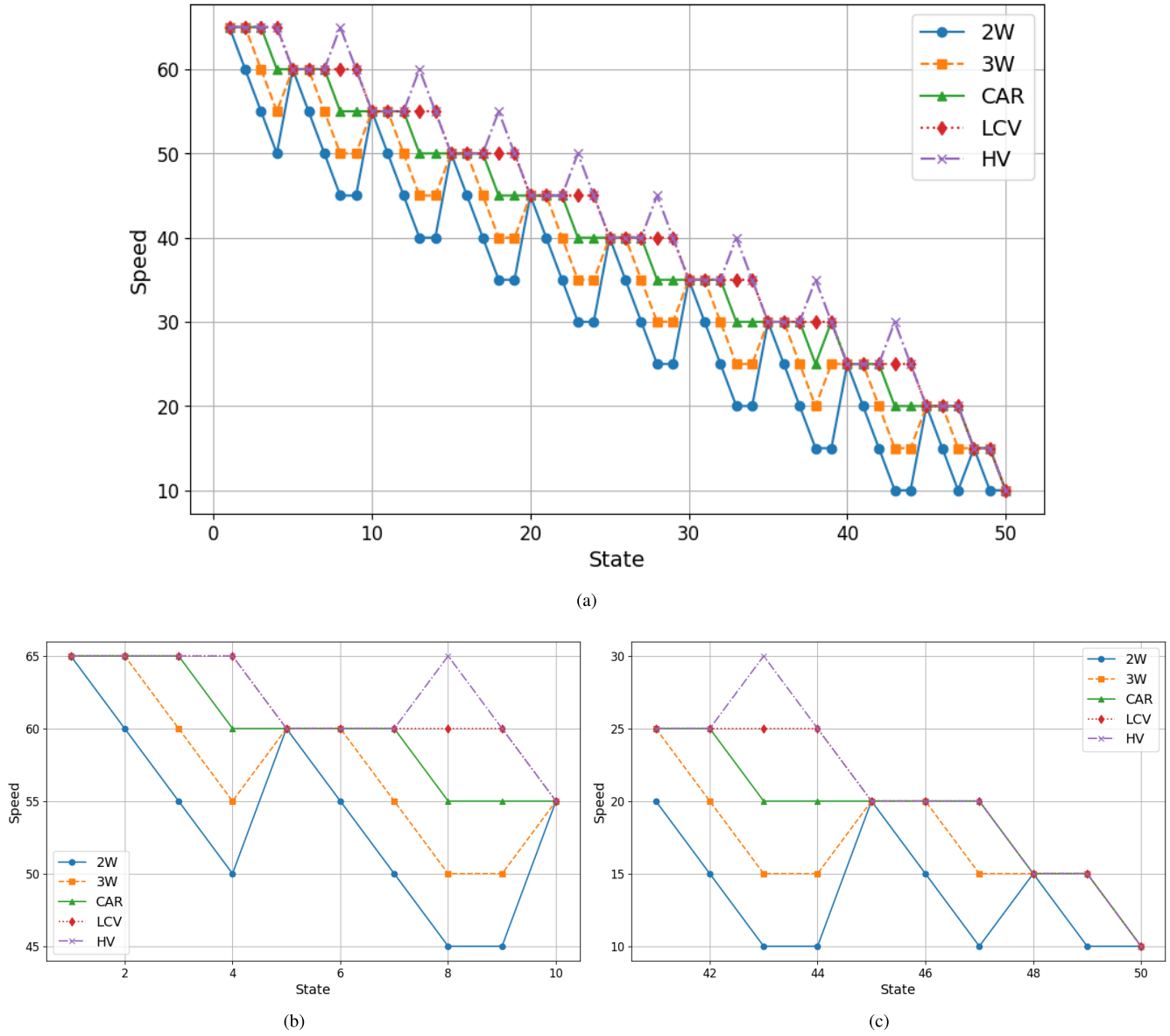


FIGURE 9. Details of Class-wise speeds; (a) 50 defined states (b) First 10 states (c) Last 10 states.

speed bins. Comparing the same with figure 7 and 8, it is evident that both the locations had significant number of observations in these traffic states where all the five vehicle classes are in different speed bins.

To estimate the class-wise speeds for class-agnostic observations, the observations are first mapped with the appropriate traffic states defined. Then, using figure 9, class-wise speeds can be estimated. Thus, this framework facilitates a framework to estimate class-wise speeds for class-agnostic disaggregated travel-time data.

VI. PREDICTION OF TRAFFIC STATES AND CLASS-WISE SPEEDS

In this study, the traffic state-based prediction methodology is proposed as a joint model and benchmarked with the existing speed-based prediction approaches, which are marginal models. The proposed methodology predicts the traffic state using lagged state information as input features in the prediction models. By using state-speed mapping, the vehicle class-wise traffic speeds are inferred. Thus, both state and speed prediction are possible using the proposed method. The labelled traffic states are mapped to the corresponding class-specific speeds to benchmark the proposed methodology. And vehicle-class wise prediction models are developed in which the lagged speeds are the input features. Once we predict the class-wise speeds, the predicted speeds of all vehicle classes are combined to represent the predicted traffic state for each observation. Figure 10 represents the proposed traffic state prediction methodology.

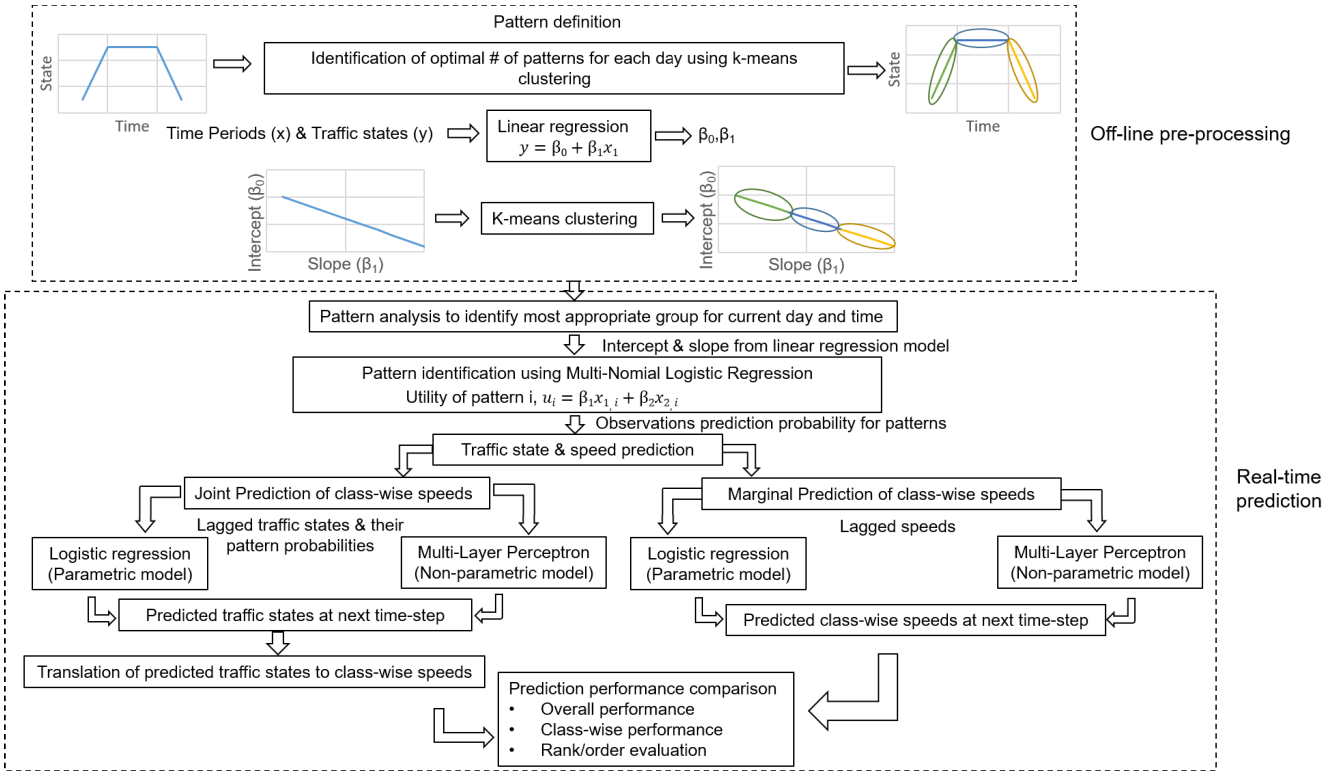


FIGURE 10. **Proposed state prediction methodology.**

As can be seen from figure 10, the pattern definition block of the pre-processing stage deals with pattern analysis of the historical data to identify their patterns. Initially, the temporal evolution of traffic states is analysed for each day by grouping daily observations into distinct patterns using the k-means clustering algorithm where each pattern represents a specific traffic scenario such as morning peak, afternoon off-peak, etc. Subsequently, a linear regression model is applied to observations within each pattern, with time period as the independent variable and traffic state as the dependent variable to obtain the intercepts and slopes for each day. These are then clustered using k-means clustering to define patterns based on specific ranges of intercept and slope values.

Once that data is available for all the historic days, the pattern analysis block of the real-time prediction stage deals with identifying the most appropriate group (3/4/5/6/7 patterns) for the current day (and time) based on analysis of historical data that captures the hourly, daily, weekly and monthly patterns, for state prediction.

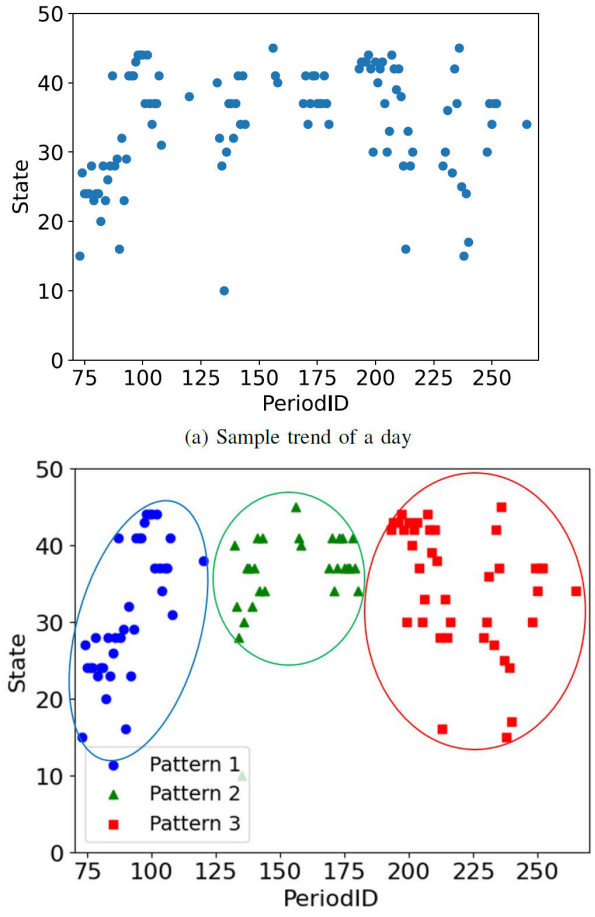
In the pattern identification phase, a logistic regression model is employed to predict the corresponding pattern for each observation. Since patterns are categorical, multinomial logistic regression is chosen, with intercept and slope as input features and pattern as the target variable.

This study employs both parametric (Logistic Regression) and non-parametric (Neural Network) models since the former provides theoretical insights and interpretability, while the latter enhances prediction accuracy, particularly for class-specific characteristics. In both models, lagged traffic states and pattern probabilities are used as input features for predicting the next traffic state.

The methodology allows for jointly predicting both traffic states and speeds by utilizing lagged state information and state-speed mapping. Thus, this prediction methodology is termed as ''Joint Prediction'' in the rest of the study. To benchmark the proposed joint prediction model, existing approaches for class-wise speed prediction are considered and referred to as 'Marginal prediction' in the paper. For each vehicle class separate prediction models are developed (one parametric and one non-parametric model), using lagged speeds as input features. Thus, we predict each class-wise speed from separate class-wise models to evaluate the performance of the joint prediction model. More detailed explanations of the proposed method and benchmarking approach are provided in sections VI-C and VI-D.

A. PATTERN DEFINITION

To study the distribution of traffic states across the time of the day, each 5-minute traffic speed data is mapped to a corresponding traffic state using the proposed methodology and visualized as scatter plots as shown in figure 11 (a). These visualizations show possible trends in the traffic state dynamics and evolution within a day. Each day's observations are grouped into m clusters using the k-means clustering algorithm; thus each pattern defines a particular traffic scenario. For instance, in the sample day shown in figure 11 (b), there are three patterns: pattern 1 represents



(a) Sample trend of a day

(b) Patterns identification of the day

FIGURE 11. (a)Trends in the traffic state dynamics within a sample day (b) Visualization of temporal patterns of the defined traffic states for a sample day.

the morning peaking, pattern 2 represents the afternoon peak period, and pattern 3 represents the evening off-peaking. The grouping analysis was done for 31 days of data to observe the trends and characteristics of state evolution. It may be noted that different days may have different numbers of patterns representing different traffic scenarios, such as 3 patterns (morning peaking, peak period, and evening off-peaking); 5 patterns (morning peaking, morning off-peaking, afternoon off-peak, evening peaking, and evening off-peaking); or even 7 patterns (morning peaking, morning peak period, morning off-peaking, afternoon off-peak, evening peaking, evening peak period, and evening off-peaking). However, the study area is a major corridor and exhibited a trend of three patterns in almost all of the 31 days of data used in this study. But other locations may exhibit five, six, seven, or more patterns. It should be noted that different days and locations may have different numbers of patterns representing the various traffic scenarios. Therefore, the application of this framework will require location-specific calibration of the parameters and pattern analysis prior to implementation.

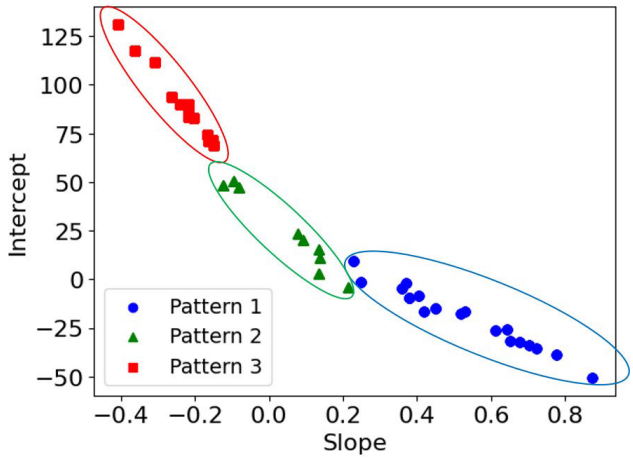


FIGURE 12. Definition of state evolution patterns based on temporal dynamism modelled using linear regression.

Each pattern has a specific state evolution; thus, different models are required for each pattern. One could employ various ways to model the patterns uniquely, but linear regression is a simple approach that characterises using only two variables. Hence, linear regression is performed on observation within each pattern of a day with time period as the explanatory variable and traffic state as the target variable; thus, intercepts and slopes are calculated. Figure 12 shows distinct and non-overlapping ranges of slope and intercept to define the pattern's characteristics uniquely.

The proposed method analyzes the temporal evolution of traffic states and models the relationship between the time of day and traffic states using historical data to better define traffic patterns. Therefore, for each day, a separate linear regression model is developed, and intercept and slope values are identified for each of the patterns within that day. The slope and intercept values for a pattern are aggregated from multiple days to characterize the pattern.

B. PATTERN IDENTIFICATION

Since the pattern prediction problem is categorical, 'Logistic regression' is chosen as the prediction model. The possible categories of logistic regression models are binary, multinomial, and ordered logistic regression. In our present prediction problem, since the patterns are nominal and discrete, the multinomial logistic regression model is chosen. It was found that the logistic regression can predict the patterns and the traffic states with good accuracy, possibly because the relationship between time-lagged features (State/Speed) and the target variable is simple and relatively less complex. Intercept and slope are the input features, and the pattern is the target variable. The model formulation is given below:

Utility equation is given by equation 3. The number of patterns for a given day was identified based on the traffic trend from the study location. The intercepts and slopes from the linear regression model were taken as

the input features. The probability function is given by equation 4. The probability of being in each pattern is calculated from the exponential function of the utility of each pattern. The developed multinomial logistic regression model exhibited very good predictions for the patterns with the best accuracy.

$$U_i = \beta_{1,i}x_{1,i} + \beta_{2,i}x_{2,i} \tag{3}$$

$$P_i = \frac{e^{U_i}}{\sum e^{U_i}} \tag{4}$$

where,

U_i: Utility of pattern i ($i = 1,2,\ldots$ n for 'n' patterns)

$\beta_{1,i}$: Regression coefficient of input feature 1 for pattern i

$x_{1,i}$: Input feature 1 (Intercept from linear regression model)

$x_{2,i}$:Input feature 2 (Slope from linear regression model)

P_i:Observation's prediction probability for pattern i

C. JOINT PREDICTION

This study considers a parametric model (Logistic Regression) to gain theoretical insights and interpretability and a non-parametric model (Neural Network) for better prediction accuracy of the future traffic states based on vehicle-class-specific characterisation. The logistic regression is trained using the multinomial option within the LogisticRegression function of the sci-kit library [48], employing the multiclass function to handle nominal patterns. The lagged traffic states and pattern probabilities are input features to predict the next traffic state. The model formulation of the logistic regression model is given as:

$$x_t = f(x_{t-1}, \ldots x_{t-n}, p(g^1_{(t-1)}), p(g^2_{(t-1)}), \ldots p(g^1_{(t-n)}), p(g^2_{(t-n)})) \tag{5}$$

where,

x_i: Traffic state at time i,

$p(g^j_{(t-k)})$: Probability of traffic state at time $t-k$ being in pattern j,

j:1,2,...p and k=1,2,...n,

p: total number of patterns in a day,

n: optimal number of lagged states considered for the model.

It can be noted from equation 5 that the traffic state at time t is a function of traffic states at previous n traffic states and their respective probability values for the corresponding patterns for the day. One pattern can be taken as the reference pattern, thus resulting in probability values of previous traffic states for the rest of the patterns.

Utility equation is given by equation 6. The input features are the probability values obtained from the multinomial logistic regression model from the previous step. Probability functions are given by equation 7. The probability of being in each traffic state is calculated from the exponential function of the utility of each traffic state.

$$U_i = \beta_{11}x_{1,i} + \beta_{2,i}x_{2,i} + \ldots\ldots + \beta_{n,i}x_{n,i} \tag{6}$$

$$P_i = \frac{e^{U_i}}{\sum e^{U_i}} \tag{7}$$

where,

U_i: Utility of traffic state i (i= 1,2,,...50)

β: Regression coefficients

$x_{k,i}$: Input feature k (Previous traffic state)

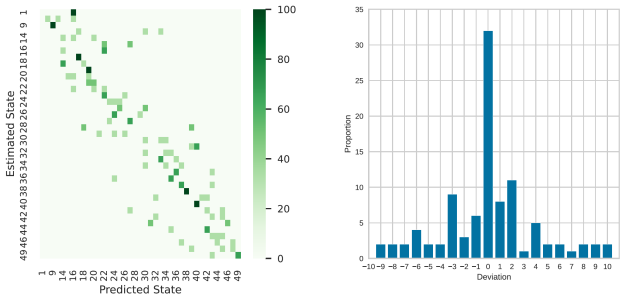
P_i:Observation's prediction probability for traffic state i

The Variance Inflation Factor (VIF) is used to identify the correlation between the independent variables. VIF value less than four is desirable and VIF greater than 5 represents a critical level of multicollinearity. Since the input features exhibit a correlation with VIF greater than 5, 'z standardization' was adopted, and VIF was brought to less than 2.

Logistic regression does not have an equivalent statistic to R^2. However, several pseudo R^2's have been developed [49] to evaluate the goodness-of-fit of logistic models. Even though pseudo R^2's cannot be interpreted independently or compared across datasets, they are valid and useful when evaluating multiple models predicting the same outcome.'Efron's R^2', 'McFadden's R^2', 'Nagelkerke R^2' and 'Cox and Snell's R^2' are the most commonly used pseudo R^2 indices for the logistic regression models. Among this,'McFadden's R^2' is direct and analogous to R squared in ordinary least square (OLS) estimation of linear regression. McFadden's adjusted R^2 mirrors the adjusted R-squared in OLS by penalizing a model for including too many predictors. Based on the McFadden's adjusted R^2 scores for n values varying from 1 to 6 for study area 1, considering elbow characteristics, the five lagged traffic states and their corresponding pattern probabilities are chosen as the potential optimal input features.

After careful consideration of Accuracy, F1 score, and Precision of the training data, $n = 5$ is chosen for modeling. It was observed that even in the testing data, $n = 5$ showed a superior performance for all three metrics. Ablation study is conducted on the input features to get insights into the relative importance of the individual features. Based on this study results, all the input features are retained since they all exhibited moderate importance. Based on the trend of training and testing loss, the optimal n is chosen as $n = 5$ to mitigate the overfitting issue. Note that the optimal number of lagged states for prediction should be calibrated based the location-specific traffic characteristics to capture daily and weekly patterns.

In this study, the hyperparameter optimisation is done using "RandamizedSearchCV" [50] with five-fold cross-validation. Also, the activation function is maintained same across all layers to simplify the model's architecture, to make it easier to debug and understand, also to reduce the complexity of hyper parameter tuning. The study uses "StratifiedShuffleSplit" [51] to split the dataset into training and testing datasets. Stratification based on 'state' ensures the training dataset contains samples from all the traffic states. Based on literature [52], [53], we chose



(a) Prediction results (b) Plot of predicted state deviation from actual

FIGURE 13. State prediction results of logistic regression model of the joint prediction based on state-based approach.

0.8-0.2 split up for training and testing in this study. However, other proportions could also be used for these tasks. Note that the hyperparameters for the models need to be carefully selected to prevent overfitting and/or reduce the complexity of the model (e.g., fewer neurons, fewer layers) to improve generalization, for different case studies.

If previous lagged input features are unavailable, suitable prediction techniques can be used to impute the missing data. These imputed lagged features can be used in the present framework as the input feature to predict the traffic states/speeds. However, note that the quality of the imputed data will influence the prediction's quality.

Figure 13 shows the state prediction results of the Logistic Regression (LR) model of the proposed state-based prediction approach. From figure 13, it is inferred that n=5 predicts about 70 per cent of the states with $\pm$ 3 states of the actual state. The deviation between actual and predicted states are plotted in figure 13. The skewness value of the deviation is 0.16. Lower variance for n=5 indicates better prediction results and lower positive skewness indicates higher symmetry in the prediction results. The model with five lagged states performed well, with about 32% of observations with zero deviation. Predictions of free-flow and congested traffic states are good. Only a few intermediate states have some dispersion.

After the state prediction, the predicted traffic states are mapped to the corresponding speed bands using the framework to estimate class-wise speeds. Figure 14 compares estimated and predicted class-wise speeds.

It can be observed from figure 14 that n=5 predict about 75 per cent of the speeds with + 5 kmph of the actual speed. The skewness values of the deviation are 0.24,−0.04,−0.19,−0.17 and-0.22 for n=5 for classes 1-5 respectively. Lower variance for n=5 than n=3 indicates better speed prediction results for all vehicle classes. By analysing the signs of the skewness, it is inferred that four of the five vehicle classes in n=5 had negative signs for skewness of deviation, indicating overprediction of the speed. Since all the reported skewness values are between −0.5 and 0.5, the deviation plot is approximately symmetric for all the five vehicle classes of both n=3 and n=5 models.

In this section, a neural network model has been developed using Multi-Layer Perceptron (MLP) Classifier, a simple and popular feed-forward neural network algorithm. Similar to the LR model, the input features are lagged traffic states and the corresponding pattern probabilities and the target variable is the traffic state. Solver (sgd,adam), learning rate (constant,invscaling, adaptive), hidden layer sizes ((100,), (256,100), (512,256,128)), batch size (8,16,32,64), alpha ([0.0001,0.05]), and activation (tanh,relu,logistic) are the hyperparameters used in this model and their parameter space are given in the parentheses. Hyper-parameter optimization is done by cross-validation using 'RandomizeddSearchCV' to avoid 'underfitting' or 'overfitting' of the model. Based on the cross-validation results, the optimized hyper-parameters used for the MLP Classifier are: solver (sgd), learning rate (constant), hidden layer sizes (512,256,128), batch size (32), alpha (0.0001) and activation (relu).

MLP models do not have the 'null model' as LR models, so pseudo R^2 can't be used to determine the optimal number of input features. Hence, the optimal n is determined as n=5 using 'F1 score', indicating the five lagged traffic states and their corresponding pattern probabilities as the potential input features.

Figure 15 shows the state prediction results of MLP models of the state-based prediction approach. From figure 15, it is inferred that the MLP model outperforms the logistic regression model, with the former showing less dispersion than the latter. The results show that the MLP model predicts both the free-flow and congested traffic states very well; only a few intermediate states appear to have some dispersion.

Then, the predicted traffic states are mapped to the corresponding speed bands to identify the class-wise speeds. The results are depicted in figure 16. The skewness values of the deviation of actual and predicted speeds are 0.09,0.04,−0.04,−0.06,0.24 for the neural network model. From these values, it can be inferred that the deviation distribution is almost symmetrical. Also, almost 60% of the observations' actual and predicted speeds are the same for all the vehicle classes. And more than 90% of the observations' predicted speeds coincide with the actual speeds with $\pm$ 5 kmph speed difference. Thus, the proposed MLP model performs well at free-flow and congested conditions and reasonably well at intermediate traffic states. The potential reason behind this is the underlying assumption of steady-state conditions. In mixed traffic conditions, the intermediate traffic states exhibit non-equilibrium conditions with few vehicle classes in steady state and a few in non-steady-state conditions.

The proposed methodology predicts the traffic state using lagged state information as input features in the prediction models. Then, the vehicle class-wise traffic speeds are inferred by using state-speed mapping. It should be noted that, in the proposed joint prediction model, the input features are the time-lagged traffic states, and the target variable is

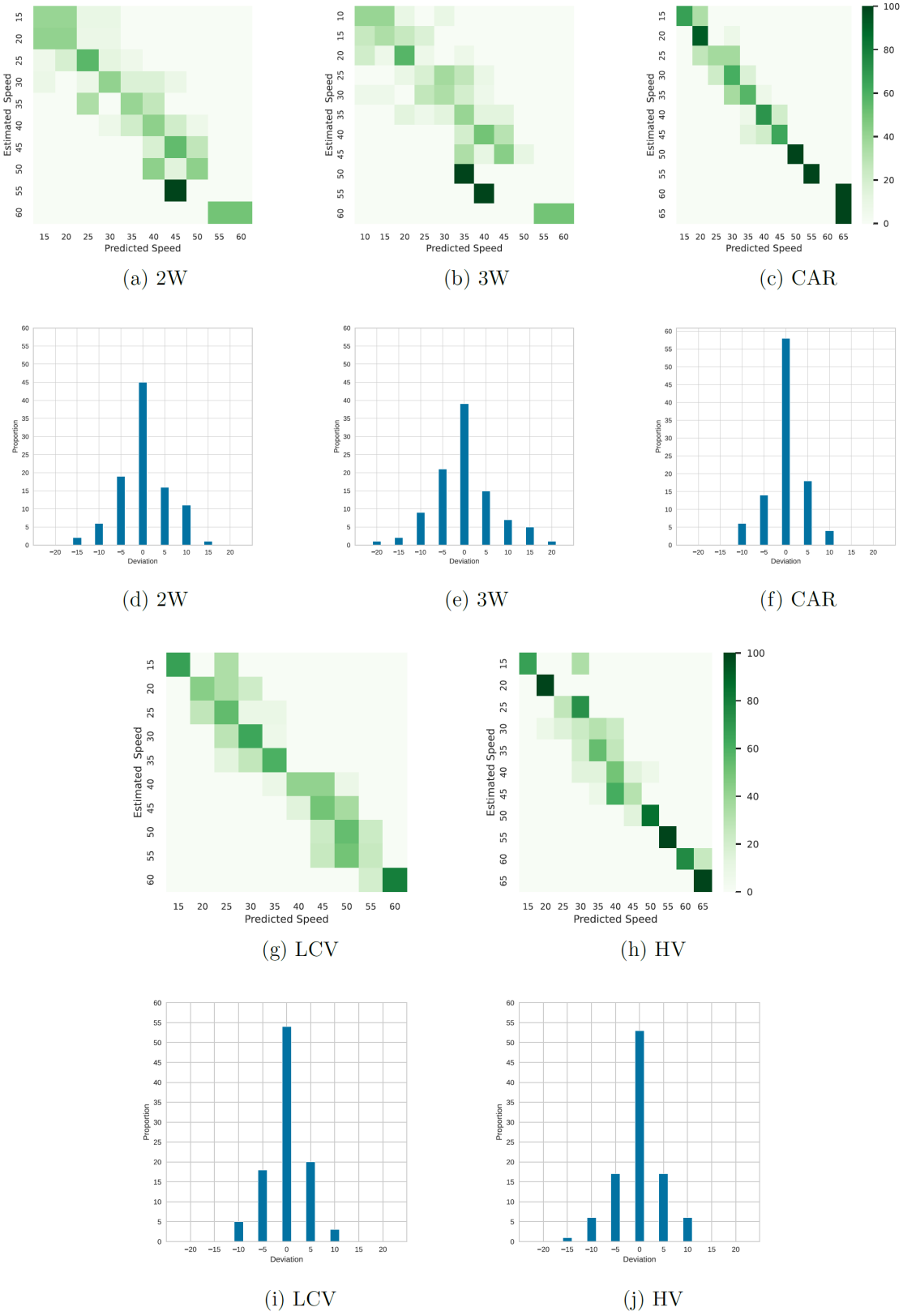
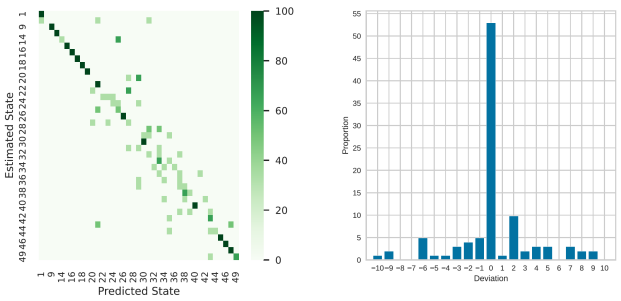


FIGURE 14. Speed prediction results of logistic regression model of the joint prediction based on state-based approach.



(a) Prediction results (b) Deviation Plot

FIGURE 15. State prediction results of MLP model of the Joint prediction based on state-based approach.

TABLE 4. Class-wise speed prediction performance of the proposed Joint model based on state-based prediction approach.

Class	Metric	LR model			MLP model		
		n=3	n=4	n=5	n=3	n=4	n=5
2W	Precision	0.41	0.39	0.39	0.50	0.56	0.70
	Recall	0.32	0.40	0.34	0.48	0.61	0.63
	F1 score	0.34	0.39	0.35	0.48	0.58	0.62
3W	Precision	0.35	0.32	0.32	0.51	0.54	0.70
	Recall	0.29	0.32	0.28	0.48	0.60	0.63
	F1 score	0.30	0.31	0.29	0.49	0.56	0.62
CAR	Precision	0.44	0.47	0.56	0.60	0.62	0.71
	Recall	0.39	0.47	0.48	0.59	0.66	0.68
	F1 score	0.40	0.46	0.50	0.59	0.63	0.67
LCV	Precision	0.48	0.51	0.51	0.66	0.70	0.72
	Recall	0.42	0.47	0.45	0.62	0.71	0.71
	F1 score	0.43	0.48	0.47	0.63	0.69	0.68
HV	Precision	0.43	0.48	0.53	0.65	0.66	0.67
	Recall	0.38	0.45	0.46	0.62	0.68	0.67
	F1 score	0.39	0.45	0.49	0.62	0.65	0.64

the traffic state. Once the traffic states are predicted using the proposed model, the class-wise speeds are inferred from the predicted states by using state-speed mapping and reported in figure 16. Thus, the proposed MLP model does not give a multi-output prediction.

While comparing the model performance across the vehicle classes, the model exhibited very good performance for CAR, LCV and HV (bigger vehicles) than 2W and 3W (smaller vehicles). The potential reason behind this is that the speed changes from one traffic state to the next state are more dynamic for smaller vehicle classes than for bigger vehicle classes. Due to the seepage behaviour, two-wheelers move within the gap of the larger vehicles. Hence, the smaller vehicles exhibit more dynamism than the bigger vehicles.

Since the prediction problem is formulated as categorical, Precision, Recall and F1 scores from the classification report are considered to evaluate different models.

From table 4, it is inferred that NN outperformed LR in almost all the cases. The prediction performance is very good for CAR, LCV and HV (bigger vehicles) than 2W and 3W (smaller vehicles). The logistic regression model with the three lagged traffic states as input features showed some dispersion in the predicted speeds, but the level of dispersion is reduced with the five lagged traffic states as input. However, the MLP model outperformed the logistic regression model with lesser dispersion. The optimal number of input features for the neural network models is also less than that of the logistic regression model for comparable performance.

D. MARGINAL PREDICTION

In this section, the proposed state prediction methodology is benchmarked with the existing practice of class-wise speed prediction approaches. For this, vehicle-class-wise prediction models are developed using popular parametric and non-parametric models (logistic regression and neural network approach) to perform speed predictions in which the lagged speeds are used as the input features. Since the VIF was found to be greater than 5, indicating the input features exhibit a correlation, 'z standardization' was adopted to reduce VIF to less than 2.

Similar to the proposed prediction method, the optimal number of lagged speeds to be considered as the input features for the benchmarking model was evaluated. It was found that n=7 is optimum for all five vehicle classes for the LR model.

The model coefficients are given in table 7 in the appendix.

Similarly, n=5 is the optimal value for the MLP model. Like the proposed method, hyper-parameter optimization is done by cross-validation using 'RandomizedSearchCV' [50] for the benchmarking model to avoid 'underfitting' or 'overfitting' of the model. The optimized parameter values of all the five vehicle class-wise models of the Speed-based prediction approach is shown in table 5.

The unique patterns observed in the predictions of the benchmarking model where all the five vehicle classes' predicted speeds differ from their actual are given by figure17. Figure 17 (a) and 17 (c) represent the consistent under and over-prediction of the speeds. Figure 17 (b) and 17 (d) are the case examples of over predictions for some vehicle classes and under predictions for some vehicle classes.

Proposed joint model and bench-marking marginal model were analysed for their performance. In figure 18,'0' corresponds to the percentage of observations where the predicted speeds coincide with the actual for all five vehicle classes. From figure 18, it is inferred that 42.20 % of the observations from the test data set coincided with the actual for the marginal model. The corresponding value for the joint model was 50.00 %.

For the percentage of observations where actual and predicted speeds differ, a detailed analysis has been carried out by examining the percentage of under-predictions and over-predictions vehicle class-wise (see figure 19).

Figure 19 represents the assessment of the proposed joint model and benchmarked marginal model for its class-wise performance. For 2W and 3W, the marginal model captured around 78% of the predictions matching the actual speeds. For CAR, and LCV, the marginal model showed around 70% of the observed predicted speeds match the actual speeds.

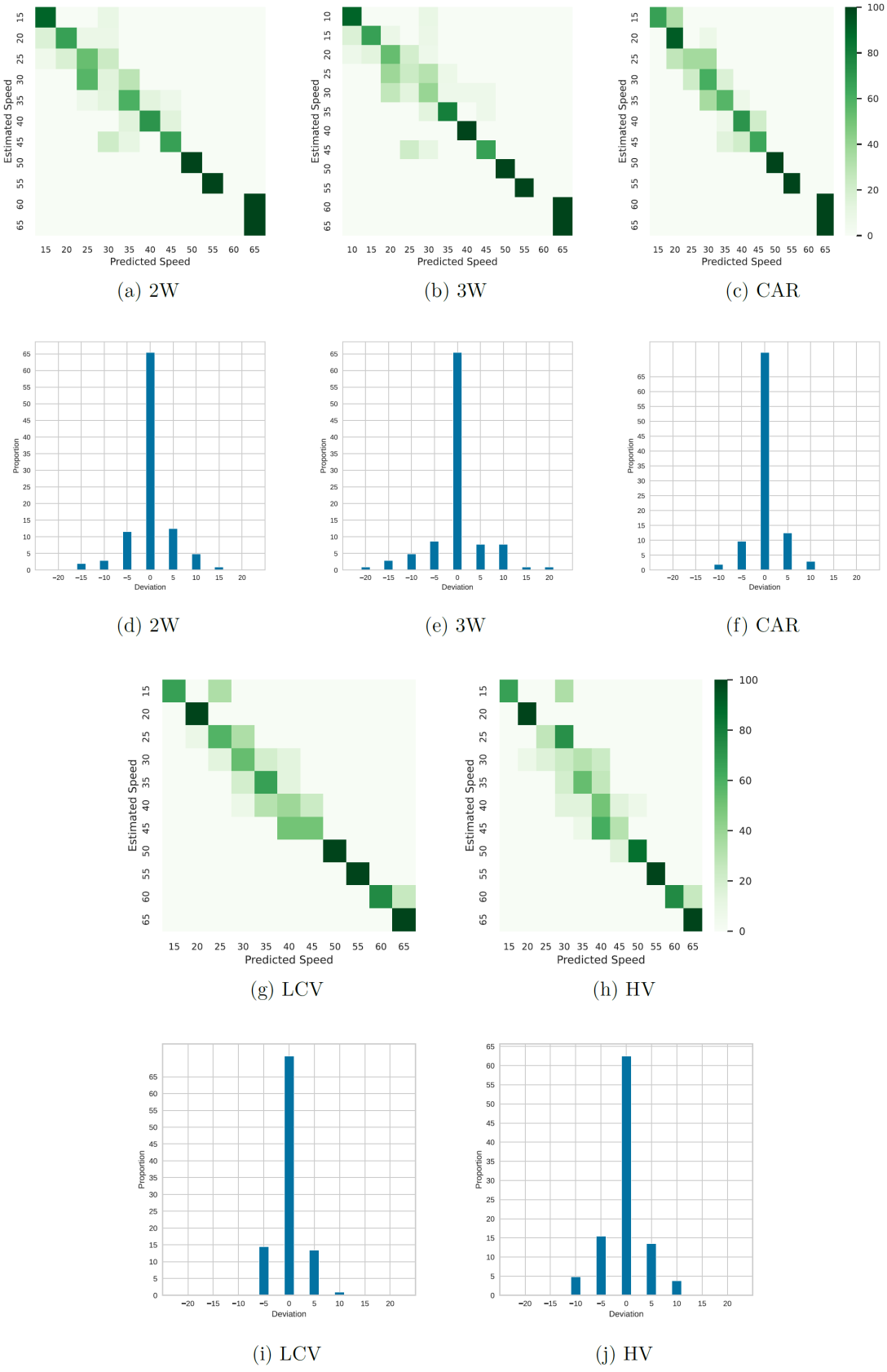


FIGURE 16. Speed prediction results of MLP model of the Joint prediction based on state-based approach.

TABLE 5. Optimized Hyper parameter values of the marginal model based on speed-based prediction approach.

Parameters	Optimized Value				
	2W	3W	CAR	LCV	HV
Solver	adam	adam	sgd	sgd	sgd
learning rate	adaptive	invscaling	constant	constant	constant
hidden layer sizes	(256, 100)	(512,256,128)	(256, 100)	(512,256,128)	(256, 100)
batch size	32	64	16	32	16
alpha	0.0001	0.05	0.0001	0.0001	0.0001
activation	relu	relu	tanh	relu	tanh

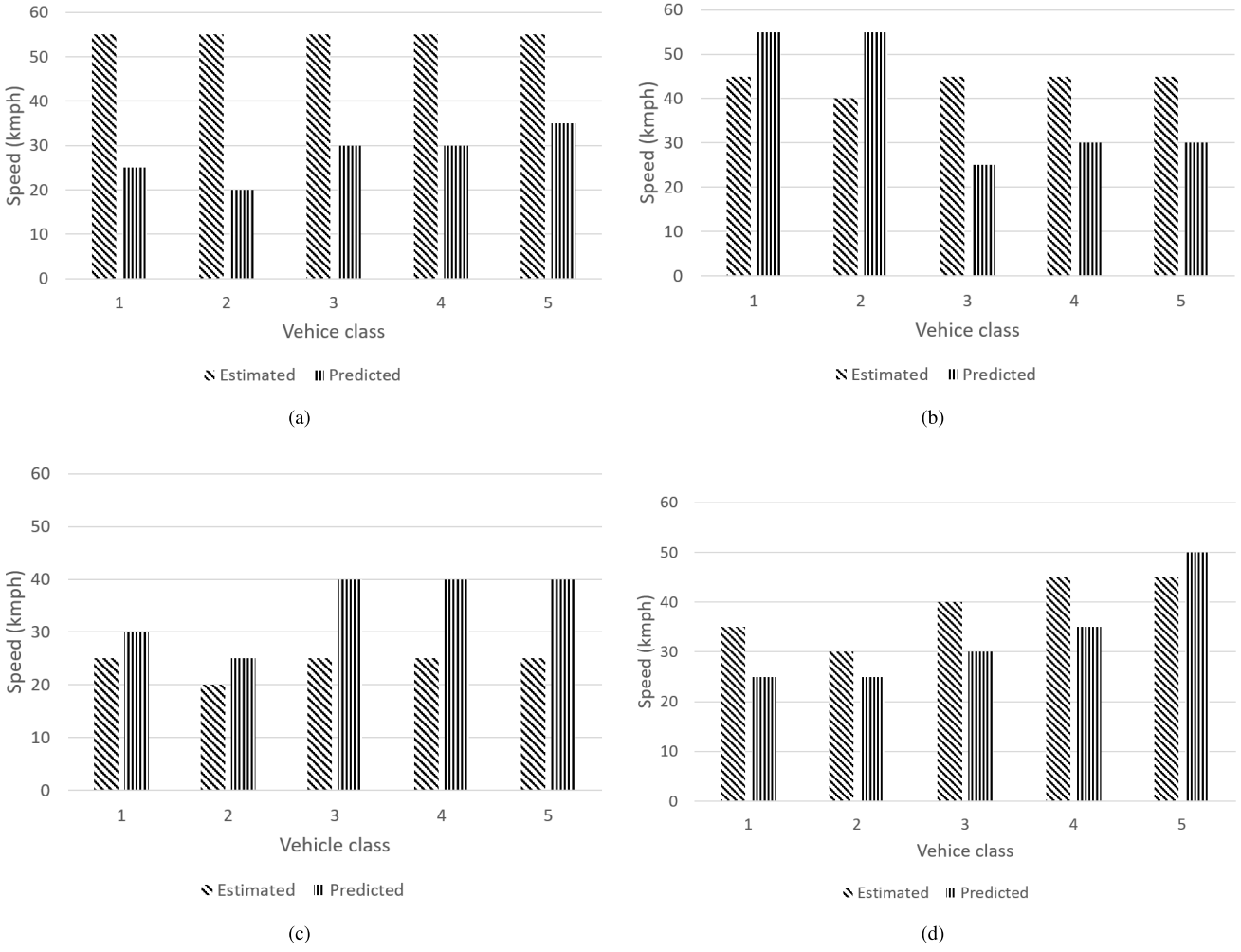


FIGURE 17. Analysis of observed prediction patterns.

The corresponding percentage for HV is 66%. For each vehicle class, the proportion of observation for which the predicted speeds do not match actual speeds is categorized as under-predicted and over-predicted (see figure 19). The proportion of under-prediction (on an average of 17%) is larger than that of over-prediction (on an average of 10%) for each of the five vehicle classes. In the joint model, the actual and predicted speeds match for about 67% of the time on average across all five vehicle classes. The joint model's performance was assessed as either under or over-predicting, similar to the marginal model. From figure 19 (b), it is evident that the proportion of under-prediction (average 17%) and over-predictions (average 16%) are almost equal for the proposed joint model. Hence, it is evident that the individual speed predictions are the strengths of the marginal models over joint models. However, the marginal model could not capture the combined effect or overall traffic state prediction. Even though, at the vehicle class level marginal model showed superior performance than the joint model (see figure 19 (a), from figure 18, it is evident that at the overall performance, the joint model outperformed the marginal model.

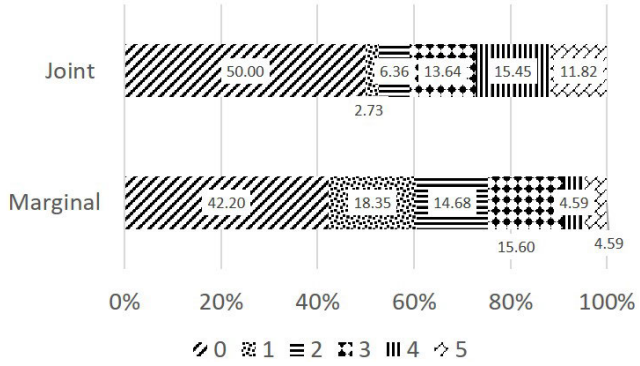


FIGURE 18. Performance evaluation of joint and marginal prediction models.

Further, the speed prediction results are analyzed for order or ranking. That is, each vehicle class of the observations of the test dataset was given a rank based on its speed value. The fastest vehicle class will receive the rank of 1, and the slowest vehicle class will be ranked 5. The results of the ranking analysis are shown in figure 20.

In figure 20 (a), '0' denotes all the vehicle classes' predicted order are as same as the actual. It is inferred from the figure that 46 % of the observations could receive the same order in the prediction for the marginal model and the respective value for the joint model is 58 %. The remaining % of observations were classified based on the number of vehicle classes that couldn't replicate the same order as in the actual. For the benchmarked marginal model the values are 16%,13%,10% and 16%, respectively, for two, three, four, and all five vehicle classes not replicated in actual order. The corresponding values for the proposed joint model are 12%,2%,27% and 3% respectively.

Furthermore, the ranking analysis was also done as vehicle classwise and the same is reported in figure 20 (b).

From classwise ranking analysis, it is evident that the joint model outperformed the marginal for all the five vehicle classes in terms of the percentage of observations for which the predicted model could capture the original order. Overall, the joint model showed around 11 % improvement over the marginal model. Further, vehicle class-wise, the joint model showed around 19% improvement over the marginal model for vehicle class 2W and CAR.

Speed-based prediction approach based on marginal model fails to capture the joint effects of multiple vehicle classes and performs poorly for speed prediction.

VII. DISCUSSION AND CONCLUSION

This research suggests a novel framework for mixed traffic characterisation based on vehicle class-specific speed and a real-time one-step ahead traffic state prediction mechanism. This study proposes a speed-based characterisation methodology to estimate class-wise speeds from class-agnostic disaggregated travel-time data. The travel time values are not marked with the corresponding class of the vehicle from which it is collected. However, the data inherently captures the kinematic variations by allowing them to occupy different speed bins. Travel-time data are collected and clustered using unsupervised clustering techniques. We ensured that all the periods used for state definition had at least five observations. Based on the cluster characterisation, the lower and upper-speed bounds are defined for traffic states. Considering the area occupancy, the defined traffic states are ordered from free-flow to congested conditions. Thus, the traffic state definition step implies the speed-based characterisation of mixed traffic. Also, while the traditional approaches characterise traffic states using a single value of speed, the proposed state characterisation defines traffic states based on a spectrum of speeds. Further, this study proposes the traffic state-based prediction methodology as a joint model and benchmarks it with the existing speed-based prediction approaches, which are marginal models.

In the proposed methodology, the traffic state is predicted by using lagged state information as input features in the prediction models. By using state-speed mapping, the vehicle class-wise traffic speeds are inferred. Thus, both state and speed prediction are possible using the proposed method. The labelled traffic states are mapped to the corresponding class-specific speeds to benchmark the proposed methodology. And vehicle-class wise prediction models are developed in which the lagged speeds are the input features. This paper's proposed characterisation framework is deployed over a Logistic regression model and an MLP model based on the literature [54], [55], [56]. However, it is possible to use other statistical and machine learning methods such as Long Short Term Memory (LSTM) and ensemble methods like Random Forest further to improve the performance in state and speed prediction [57], [58], [59], [60].

Since the described traffic conditions cover all potential traffic states of 12-speed bin ranges, the suggested characterisation methodology is applicable to any corridor. Even without considering vehicle composition, the technique performs well. Additionally, as the proposed methodology just needs information on travel time or spatial speed, it is sensor-independent. A data-driven strategy is adopted to examine the most frequently seen traffic states from the data due to the numerous possibilities for traffic states. To further make the model knowledge-guided, vehicle-class-specific area occupancy is added to organise the traffic states. This method is novel based on the approach since the prediction methodology depends on state progression rather than time. For mixed traffic conditions, speed predictions are typically made at the stream level; however, the current study investigates the possibility of making speed predictions for particular vehicle classes. The suggested approach is more effective than bench-marking because it only requires one model to forecast traffic conditions. In contrast, the bench-marking approach requires a different model for each vehicle class to forecast speed. Compared to the bench-marking approach, the proposed prediction approach is more

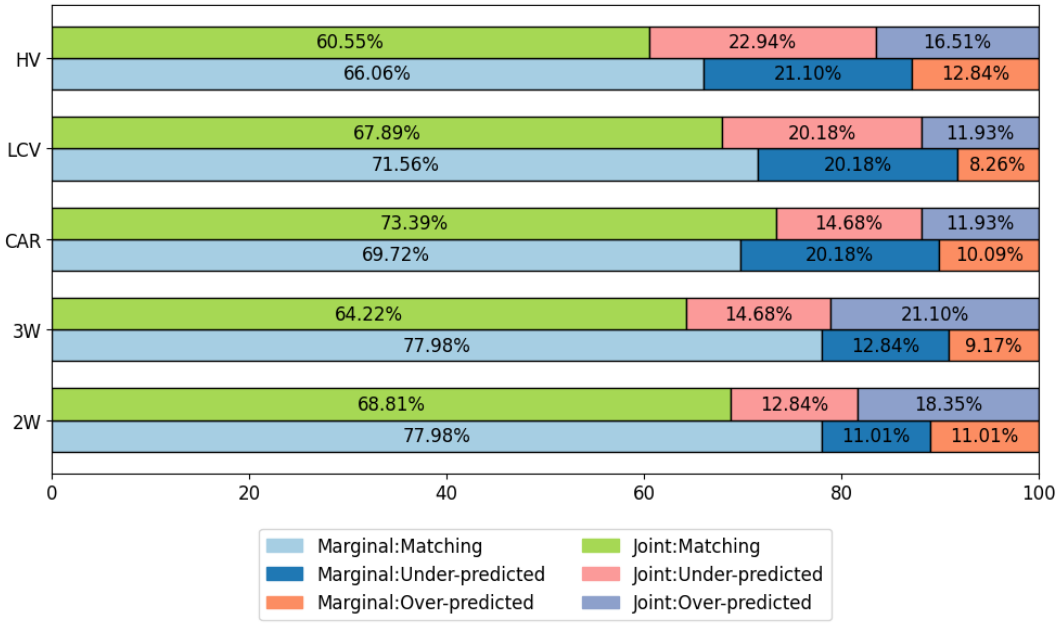


FIGURE 19. Class-wise analysis of Joint and Marginal models.

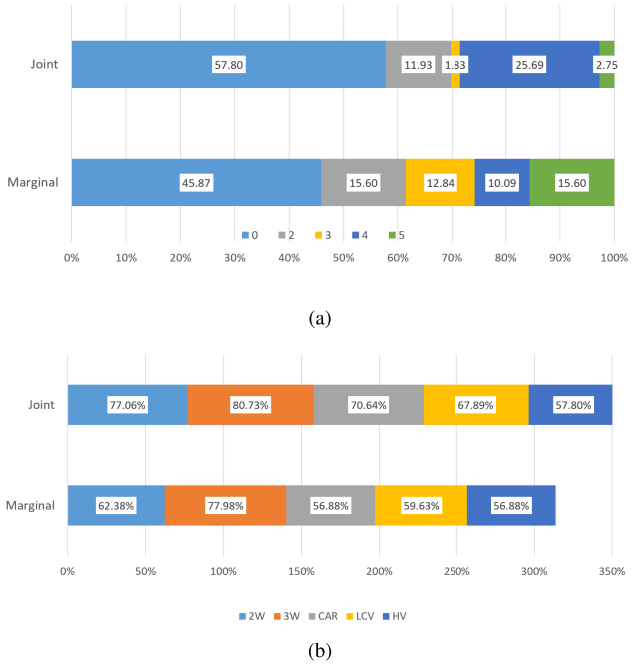


FIGURE 20. Prediction order Performance of benchmarking model (a) Overall (b) Class-wise.

straightforward and computationally effective. Furthermore, for both logistic regression and neural network models, the benchmarking model's optimal number of input features is greater than that of the suggested approach. The benchmarking approach has the following drawbacks: It needs more class-specific historical data, which is practically difficult, and a different mapping technique is required to translate class-specific speeds to traffic. The proposed methodology

TABLE 6. Sample results of traffic states Mapping based on the developed state mapping algorithm.

Rule	Observed	Mapped
1	111	111
	11101	11111
	111111	11111
2	1001	1
	11001	11
3	101111	1111
	11001001	11

characterises the traffic states as joint distribution which is most appropriate for mixed traffic conditions.

Concerning computational load, the proposed state-based prediction approach is more efficient than the Speed-based prediction approach since one model is enough to predict the traffic state in the State-based prediction approach. However, the speed-based prediction approach needs separate models for each vehicle class. The state prediction of this approach is complex since it involves mapping from speeds to states. Therefore, the limitation of the speed-based approach is that it needs more historical and class-wise data, which is practically difficult. A separate mapping algorithm is needed to convert the class-specific speeds to the traffic state. Hence, due to the requirement of class-specific prediction models and mapping algorithms, the computational load of the speed-based approach is higher than that of the state-based approach. One important insight is that since the prediction based on joint probabilities is better than the prediction based on marginal probabilities, it is evident that the class-wise speeds do not evolve independently and are interdependent. Therefore, it is better to predict states based on joint probabilities. The broader application areas of the

TABLE 7. Model coefficients of logistic regression model of the marginal prediction based on speed-based prediction approach with *n=5*.

Class	Feature	Speed-bin wise Co-efficients										
		15	20	25	30	35	40	45	50	55	60	65
1	Const		-0.25	1.37	1.41	1.38	0.58	0.74	-0.03	-1.42		-3.83
	1		-0.19	1.11	1.54	1.54	2.24	1.54	1.84	0.19		2.28
	2		0.15	-0.55	-0.39	-0.29	-0.29	-0.14	0.02	0.73		-1.67
	3		-0.36	0.06	0.18	0.36	0.26	0.55	-0.06	0.13		0.91
	4		0.30	0.05	0.20	0.19	1.03	0.32	-0.34	-0.02		1.52
	5		-0.32	0.80	1.25	1.06	0.46	1.57	2.25	-0.61		2.53
2	Const	-0.31	1.09	1.30	1.28	0.99	0.25	-0.38	-0.83	-2.11		-2.93
	1	-0.01	0.69	1.50	1.44	1.82	1.55	1.69	1.09	0.02		2.03
	2	-0.40	0.17	-0.35	0.04	0.01	-0.03	0.56	0.01	1.40		-1.10
	3	-0.80	0.08	0.19	0.34	0.21	0.82	-0.11	-1.03	-0.67		0.78
	4	0.06	0.35	0.11	0.18	0.78	0.48	-0.52	0.23	-0.15		1.20
	5	0.49	0.32	1.03	0.89	0.32	1.13	2.04	1.79	-0.90		2.03
3	Const		4.46	5.36	6.12	5.99	5.79	5.59	5.20	4.64		-0.02
	1		-2.03	-1.13	-1.05	-0.29	-0.50	-0.14	-0.39	-0.28		0.24
	2		-2.78	-2.46	-2.62	-2.90	-2.64	-2.38	-2.81	-1.74		-4.26
	3		1.20	0.87	1.11	1.19	1.65	1.12	1.95	1.61		2.26
	4		2.78	1.96	2.51	2.52	2.67	3.16	2.31	2.11		3.73
	5		1.76	3.11	3.21	3.46	3.27	3.37	4.16	3.22		5.37
4	Const		-0.32	3.69	4.06	4.23	4.02	4.16	3.80	3.19	-3.09	-5.00
	1		-1.42	-0.36	-0.52	0.08	0.41	0.30	0.33	0.53	1.48	1.839
	2		-2.80	-2.04	-1.88	-2.21	-2.06	-1.75	-1.90	-1.84	1.33	-3.23
	3		1.63	0.55	0.69	0.63	1.38	0.89	1.42	1.47	0.51	0.84
	4		4.18	2.24	2.65	2.87	2.74	2.97	3.21	2.98	1.57	4.42
	5		-3.52	0.75	0.96	1.16	0.98	1.46	1.57	1.12	2.54	4.52
5	Const		-0.39	3.32	3.91	3.95	3.85	3.97	3.66	3.13	-2.03	-14.8
	1		-0.71	0.08	-0.38	-0.10	0.17	0.84	0.09	1.25	1.04	5.55
	2		-2.42	-1.34	-1.30	-1.35	-1.47	-1.23	-1.22	-1.36	1.78	-4.33
	3		1.91	0.65	0.97	0.72	1.57	1.02	1.67	1.54	0.61	0.90
	4		4.39	2.52	2.88	2.95	2.88	3.06	3.39	3.10	2.36	5.79
	5		-5.08	-1.07	-0.62	-0.45	-0.34	-0.39	0.20	-0.41	0.66	4.95

proposed framework can be road network characterisation based on link speeds, state estimation of intersections and general prediction problems, including Natural Language Processing.

Following are the limitations and future directions for the study: The study presumed that all vehicle classes with sufficient sample size were present during the observation period since the study used data from Wi-Fi sensors that provide class-agnostic data. However, this assumption could be relaxed by using data from sensors such as cameras, radar, loop detectors, etc., that provide classification data to ensure sufficient samples for all vehicle classes to enhance the reliability of the speed-based characterisation. Also, since the data lacks vehicle composition, its impacts are not directly captured in the present framework but only through the area occupancy approach to obtain class-wise behaviour. While this approach may be acceptable without composition data and composition-specific fundamental diagrams, using the latter may better capture the vehicle interactions, resulting in improved estimation of class-wise speeds. Moreover, the prediction of traffic states and class-wise speeds used LR and MLP methods based on their superior performance in the previous studies from the literature. With new non-parametric methods such as LSTM, Random Forest, etc., gaining popularity [57], [58], [59], [60], one could use these methods for further improvement in the prediction accuracies. The proposed framework focuses on characterising the state of a roadway section based on a combination of class-wise speeds on that specific section. However, to characterise city-level states, the proposed framework needs to be extended to define states based on a combination of link-wise conditions such as speed, density, etc. Thus, while the speed-based characterisation proposed in this study is transferable for state characterisation on-road sections with variable speed bounds and geometric conditions, the proposed framework must be extended for city-scale characterisation.

One of the features of this method is that the states are composition agnostic. Under different compositions, different states arise, but all the predominant observable states are captured in the state definitions. However, the evolution of the traffic states is influenced by the composition and considering it as an explicit input variable will improve the performance of the prediction models. Furthermore, the standard deviation of speeds within and across clusters, cluster size and composition can also be used for more precise mapping.

In our study, we acknowledge the potential impact of rare events, such as extreme traffic jams, on our logistic regression model's learning effect and predictive accuracy. While these events may be infrequent in our traffic flow data, they can introduce data imbalance problems, leading to skewed class distributions and compromised model performance. Hence, we recognise the importance of future research endeavours aimed at collecting additional data on rare events,

exploring advanced modelling techniques, and addressing the underlying causes of extreme traffic conditions to enhance the reliability and applicability of our findings. In this paper, the proposed characterisation framework is deployed over a Logistic regression and MLP models. However, it is possible to use other statistical and machine-learning methods as well. Some of such methods include Long Short Term Memory (LSTM) and ensemble methods like Random Forest. Recent advancements in traffic prediction techniques, such as Graph Neural Networks (GNNs) and Long Short-Term Memory (LSTM) networks, are gaining popularity towards improved forecasts with higher accuracy. Some of the state-of-the-art prediction techniques integrate machine learning algorithms, deep learning architectures, and domain knowledge to capture the complex dynamics of traffic systems and provide timely and accurate forecasts for traffic management and planning purposes [57], [58], [59], [60].

The proposed speed-based characterisation framework relies on the speed bounds and number of vehicle classes to propose class-agnostic traffic states and state characterisation methodology. As validated in section IV-D of the manuscript, the proposed states and characterisation could be easily transferred to locations with similar characteristics. However, one could use the proposed framework to develop a location-specific set of traffic states for locations with a different set of characteristics. The class-based characterisation can help propose and evaluate various class-specific policy interventions such as vehicle segregation or class prohibition on specific links, bus rapid transit routes, class-specific rerouting, development of class-based advanced traveler information systems, etc., for improved equity, efficiency, safety, and sustainability on the roadway networks. Additionally, the proposed system can also be used to develop class-specific real-time traffic management strategies for effective utilization of transportation infrastructure. Based on the class-wise behavior (such as two wheelers, three wheelers, cars, trucks) and their respective speed profiles, traffic managers can design interventions that specifically target the needs and behaviors of each class such as dedicated lanes or scheduled restrictions during peak hours, to alleviate congestion caused by heavy vehicles.

APPENDIX

A. REAL-TIME IMPLEMENTATION

To implement the above method in real-time, the data requirement is the continuous travel time on a mid-block section. The travel time data will be converted to binary speed data (data pre-processing). The speed data will be labelled with appropriate state numbers (state mapping). Then, this labelled data will be used in a linear regression model to find the slope and intercept. Then, the slope and intercept values will be fed as input to an MNL model, and the group probabilities will be identified.

Based on historical data, the typical group (3/4/5/6/7 patterns) exhibited in a location can be predetermined. Once that data is available for all the historic days, one could identify the most appropriate group for the current day (and time) based on analysis of historical data that captures the hourly, daily, weekly and monthly patterns, for state prediction. After performing this pattern analysis, one could determine the probability of the pattern to which the current data point belongs for further analysis. Previous time steps' traffic state and the corresponding pattern probabilities are the inputs for the state prediction model (Logistic regression / Neural network). This model identifies the probabilities of all the defined traffic states. An inverse transform sampling will be done to identify which probabilities have to be chosen. Thus, the traffic states will be predicted. Once the traffic state is predicted, it will be mapped to the corresponding speed bands. Each traffic state has a defined speed for all five vehicle classes considered. Thus, the proposed methodology will predict vehicle class-specific speeds in real-time.

B. MAPPING OF OBSERVED TRAFFIC STATES TO DEFINED TRAFFIC STATES:

There are 4096 (2^{12}) traffic states possible with the 12-speed bins considered. To translate the predicted speeds of Method 1 to the most appropriate traffic state out of the predefined 50 traffic states, a separate mapping algorithm with three sets of rules was formulated and given below:

State Mapping Algorithm

```
String Manipulation
Manipulate String for each observation
Calculate String length (L) and count the number of zeros (C) in the string
Rule definitions
Rule 1: Chosen Pattern = Observed pattern
Rule 2: Chosen Pattern = Maximum consecutive 1's
Rule 3: Chosen Pattern = Max or 2nd Max consecutive 1 depending on the position
Rule Selection
    if L ≤ 5
        if C ≤ 1
            Apply Rule 1
    else
        Apply Rule 2
    if L > 5
        if C = 0
            Apply Rule 1
    else
        Apply Rule 3
```

Thus, at the end of this step, we have the labelled database in which all the observations are given a state number. Table 6 shows some sample observations and the traffic states to which it was mapped along with the rule applied.

ACKNOWLEDGMENT

The authors would like to thank the Emerging Mobility Technology (EmMo Tech) Project, IIT Madras and Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI) for supporting this research. They also like to thank Prof. Karthik K. Srinivasan for his suggestions of keywords for the title.

DATA AVAILABILITY STATEMENT

All data, models and code supporting this study's findings are available from the corresponding author upon reasonable request.

REFERENCES

[1] S. Chandra and U. Kumar, "Effect of lane width on capacity under mixed traffic conditions in India," *J. Transp. Eng.*, vol. 129, no. 2, pp. 155–160, Mar. 2003.

[2] J. Thomas, K. K. Srinivasan, and V. T. Arasan, "Vehicle class wise speed-volume models for heterogeneous traffic," *Transport*, vol. 27, no. 2, pp. 206–217, Jun. 2012.

[3] S. Basheer, K. K. Srinivasan, and R. Sivanandan, "Investigation of information quality and user response to real-time traffic information under heterogeneous traffic conditions," *Transp. Developing Economies*, vol. 4, no. 2, pp. 1–11, Oct. 2018.

[4] D. Basu, S. R. Maitra, and B. Maitra, "Modelling passenger car equivalency at an urban midblock using stream speed as measure of equivalence," *Eur. Transp.*, no. 34, pp. 75–87, 2006.

[5] C. Mallikarjuna and K. R. Rao, "Modelling of passenger car equivalency under heterogeneous traffic conditions," in *Proc. 22nd ARRB Conf.–Res. Pract., Canberra, Aust.*, 2006, pp. 1–13.

[6] S. Chandra, S. Gangopadhyay, S. Velmurugan, and K. Ravinder, "Indian highway capacity manual (Indo-HCM)," Council Sci. Ind. Res., 2017.

[7] S. Biswas, S. Chandra, and I. Ghosh, "An advanced approach for estimation of PCU values on undivided urban roads under heterogeneous traffic conditions," *Transp. Lett.*, vol. 12, no. 3, pp. 172–181, Mar. 2020.

[8] A. Maurya, S. Dey, and S. Das, "Speed and time headway distribution under mixed traffic condition," *J. Eastern Asia Soc. Transp. Stud.*, vol. 11, pp. 1774–1792, Oct. 2015.

[9] S. H. Demarchi and J. R. Setti, "Limitations of passenger-car equivalent derivation for traffic streams with more than one truck type," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1852, no. 1, pp. 96–104, Jan. 2003.

[10] C. Mallikarjuna and K. R. Rao, "Area occupancy characteristics of heterogeneous traffic," *Transportmetrica*, vol. 2, no. 3, pp. 223–236, Jan. 2006.

[11] R. Mohan and G. Ramadurai, "Heterogeneous traffic flow modelling using second-order macroscopic continuum model," *Phys. Lett. A*, vol. 381, no. 3, pp. 115–123, Jan. 2017.

[12] S. Singh, R. Vidya, B. K. Shukla, and S. M. Santhakumar, "Analysis of traffic flow characteristics based on area-occupancy concept on urban arterial roads under heterogeneous traffic Scenario—A case study of tiruchirappalli city," in *Lecture Notes in Civil Engineering*. Cham, Switzerland: Springer, 2021, pp. 69–84.

[13] R. Mohan, "On the modelling of speed–concentration curves for multi-class traffic lacking lane discipline using area occupancy," *Transp. Lett.*, vol. 14, no. 5, pp. 447–463, May 2022.

[14] N. Maiti and B. R. Chilukuri, "Empirical investigation of fundamental diagrams in mixed traffic," *IEEE Access*, vol. 11, pp. 13293–13308, 2023.

[15] C. F. Daganzo, "A continuum theory of traffic dynamics for freeways with special lanes," *Transp. Res. Part B, Methodol.*, vol. 31, no. 2, pp. 83–102, Apr. 1997.

[16] S. P. Hoogendoorn and P. H. L. Bovy, "Continuum modeling of multiclass traffic flow," *Transp. Res. Part B, Methodol.*, vol. 34, no. 2, pp. 123–146, Feb. 2000.

[17] G. C. K. Wong and S. C. Wong, "A multi-class traffic flow model—An extension of LWR model with heterogeneous drivers," *Transp. Res. Part A, Policy Pract.*, vol. 36, no. 9, pp. 827–841, Nov. 2002.

[18] S. Benzoni-Gavage and R. M. Colombo, "An-populations model for traffic flow," *Eur. J. Appl. Math.*, vol. 14, no. 5, pp. 587–612, 2003.

[19] D. Ngoduy and R. Liu, "Multiclass first-order simulation model to explain non-linear traffic phenomena," *Phys. A, Stat. Mech. Appl.*, vol. 385, no. 2, pp. 667–682, Nov. 2007.

[20] F. van Wageningen-Kessels, J. van Lint, C. Vuik, and S. Hoogendoorn, "Generic multi-class kinematic wave traffic flow modelling: Model development and analysis of its properties," *Transp. Traffic Theory*, vol. 64, p. 232, Sep. 2013.

[21] H. M. Zhang and W. L. Jin, "Kinematic wave traffic flow model for mixed traffic," *Transp. Res. Record: J. Transp. Res. Board*, vol. 1802, no. 1, pp. 197–204, Jan. 2002.

[22] S. Chanut and C. Buisson, "Macroscopic model and its numerical solution for two-flow mixed traffic with different speeds and lengths," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1852, no. 1, pp. 209–219, Jan. 2003.

[23] S. Logghe and L. Immers, "Heterogeneous traffic flow modelling with the IWR model using passenger-car equivalents," in *Proc. 10th World Congr.*, 2003, pp. 1–15.

[24] S. Oh, Y.-J. Byon, K. Jang, and H. Yeo, "Short-term travel-time prediction on highway: A review of the data-driven approach," *Transp. Rev.*, vol. 35, no. 1, pp. 4–32, Jan. 2015.

[25] J. R. B. A. Kumar, S. S. Arkatkar, and L. Vanajakshi, "Performance comparison of bus travel time prediction models across Indian cities," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2672, no. 31, pp. 87–98, Dec. 2018.

[26] S. V. Kumar, L. Vanajakshi, and S. C. Subramanian, "A model based approach to predict stream travel time using public transit as probes," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 101–106.

[27] B. A. Kumar, L. Vanajakshi, and S. Subramanian, "Pattern-based bus travel time prediction under heterogeneous traffic conditions," in *Transportation Research Board*. Washington, DC, USA: National Research Council, 2013.

[28] A. Achar, D. Bharathi, B. A. Kumar, and L. Vanajakshi, "Bus arrival time prediction: A spatial Kalman filter approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1298–1307, Mar. 2020.

[29] L. Vanajakshi and L. R. Rilett, "Support vector machine technique for the short term prediction of travel time," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2007, pp. 600–605.

[30] M. Yang, C. Chen, L. Wang, X. Yan, and L. Zhou, "Bus arrival time prediction using support vector machine with genetic algorithm," *Neural Netw. World*, vol. 26, no. 3, pp. 205–217, 2016.

[31] H. Xu and J. Ying, "Bus arrival time prediction with real-time and historic data," *Cluster Comput.*, vol. 20, no. 4, pp. 3099–3106, Dec. 2017.

[32] H. Maripini, A. Khadhir, and L. Vanajakshi, "Traffic state estimation near signalized intersections," *J. Transp. Eng., Part A, Syst.*, vol. 149, no. 5, May 2023, Art. no. 03123002.

[33] S. Suhas, V. V. Kalyan, M. Katti, B. V. Ajay Prakash, and C. Naveena, "A comprehensive review on traffic prediction for intelligent transport system," in *Proc. Int. Conf. Recent Adv. Electron. Commun. Technol. (ICRAECT)*, Mar. 2017, pp. 138–143.

[34] G. Sihag, M. Parida, and P. Kumar, "Travel time prediction for traveler information system in heterogeneous disordered traffic conditions using GPS trajectories," *Sustainability*, vol. 14, no. 16, p. 10070, Aug. 2022.

[35] S. Banik, L. Vanajakshi, and D. M. Bullock, "Mapping of bus travel time to traffic stream travel time using econometric modeling," *J. Intell. Transp. Syst.*, vol. 26, no. 2, pp. 235–251, Mar. 2022.

[36] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, "Dynamic data-driven local traffic state estimation and prediction," *Transp. Res. Part C, Emerg. Technol.*, vol. 34, pp. 89–107, Sep. 2013.

[37] *Gaussian Mixture Models—Scikit-learn.org*. Accessed: Jun. 2, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/mixture.html

[38] C. Apitzsch and J. Ryeng, "Cluster analysis of mixed data types in credit risk: A study of clustering algorithms to detect customer segments," M.S. thesis, UMEA Univ., Sweden, 2020.

[39] *Clustering—Scikit-Learn*. Accessed: Jun. 2, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html

[40] M. Won, "Intelligent traffic monitoring systems for vehicle classification: A survey," *IEEE Access*, vol. 8, pp. 73340–73358, 2020.

[41] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, "A review of vehicle detection techniques for intelligent vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 1–21, Sep. 2022.

[42] S. S. Patra, B. Muthurajan, B. R. Chilukuri, and L. Devi, "Development and evaluation of a low-cost WiFi media access control scanner as traffic sensor," in *Proc. 11th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2019, pp. 777–782.

[43] Z. Pu, Z. Cui, J. Tang, S. Wang, and Y. Wang, "Multimodal traffic speed monitoring: A real-time system based on passive Wi-Fi and Bluetooth sensing technology," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12413–12424, Jul. 2022.

[44] K. Mandal, A. Sen, A. Chakraborty, S. Roy, S. Batabyal, and S. Bandyopadhyay, "Road traffic congestion monitoring and measurement using active RFID and GSM technology," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1375–1379.

[45] *Elbow Method 2014; Yellowbrick V1.5 Documentation—Scikit-Yb*. Accessed: Jun. 2, 2024. [Online]. Available: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

[46] S. Radhakrishnan and G. Ramadurai, "Discharge headway model for heterogeneous traffic conditions," *Transp. Res. Proc.*, vol. 10, pp. 145–154, Jul. 2015.
[47] K. Venkatesan, A. Gowri, and R. Sivanandan, "Development of microscopic simulation model for heterogeneous traffic using object oriented approach," *Transportmetrica*, vol. 4, no. 3, pp. 227–247, Jan. 2008.
[48] *LogisticRegression—Scikit-learn.org*. Accessed: Jun. 2, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[49] *Psuedo R-squared for Logistic Regression 2014; Data Science Topics 0.0.1 Documentation—Datascience.oneoffcoder.com*. Accessed: Jun. 2, 2024. [Online]. Available: https://datascience.oneoffcoder.com/psuedo-r-squared-logistic-regression.html
[50] *RandomizedSearchCV—Scikit-Learn*. Accessed: Jun. 2, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
[51] *StratifiedShuffleSplit—Scikit-Learn*. Accessed: Jun. 2, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html
[52] D. D. Oliveira, M. Rampinelli, G. Z. Tozatto, R. V. Andreão, and S. M. T. Muller, "Forecasting vehicular traffic flow using MLP and LSTM," *Neural Comput. Appl.*, vol. 33, no. 24, pp. 17245–17256, Dec. 2021.
[53] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," *Int. J. Intell. Technol. Appl. Stat*, vol. 11, no. 2, pp. 105–111, 2018.
[54] T. S. Tamir, G. Xiong, Z. Li, H. Tao, Z. Shen, B. Hu, and H. M. Menkir, "Traffic congestion prediction using decision tree, logistic regression and neural networks," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 512–517, 2020.
[55] T. Lu, Z. Dunyao, Y. Lixin, and Z. Pan, "The traffic accident hotspot prediction: Based on the logistic regression method," in *Proc. Int. Conf. Transp. Inf. Saf. (ICTIS)*, Jun. 2015, pp. 107–110.
[56] I. Alam, D. M. Farid, and R. J. F. Rossetti, "The prediction of traffic flow with regression analysis," in *Emerging Technologies in Data Mining and Information Security* (Advances in Intelligent Systems and Computing), vol. 813. Cham, Switzerland: Springer, Sep. 2019, pp. 661–671.
[57] Y. A. Pan, J. Guo, Y. Chen, Q. Cheng, W. Li, and Y. Liu, "A fundamental diagram based hybrid framework for traffic flow estimation and prediction by combining a Markovian model with deep learning," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122219.
[58] Z. Peng, Z. Wang, and W. Zhang, "Short-term traffic flow prediction based on weather factors analysis and neural network," *J. Phys., Conf. Ser.*, vol. 2649, no. 1, Nov. 2023, Art. no. 012058.
[59] A. Rasaizadi, F. Hafizi, and S. Seyedabrishami, "The ensemble learning process for short-term prediction of traffic state on rural roads," in *Handbook on Artificial Intelligence and Transport*. Cheltenham, U.K.: Edward Elgar Publishing, 2023, pp. 102–123.
[60] J. Ke, H. Zheng, H. Yang, and X. Chen, "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach," *Transp. Res. Part C, Emerg. Technol.*, vol. 85, pp. 591–608, Dec. 2017.

ABIRAMI KRISHNA ASHOK received the B.E. degree in civil engineering from the Thiagarajar College of Engineering, Madurai, Tamil Nadu, India, in 2013, and the M.E. degree in transportation engineering from the College of Engineering, Guindy, Anna University, Chennai, India, in 2015. She is currently pursuing the Ph.D. degree with the Department of Civil Engineering, Indian Institute of Technology Madras, Chennai. Her research interests include mixed traffic characterisation, traffic state estimation and prediction, macroscopic modelling, and intelligent transportation systems (ITSs).

BHARGAVA RAMA CHILUKURI received the Ph.D. degree in civil and environmental engineering from Georgia Institute of Technology, Atlanta, GA, USA. He has several years of professional experience as a Traffic Engineer in multiple companies, USA. He is currently working as an Associate Professor with the Department of Civil Engineering, Indian Institute of Technology Madras. His research interests include traffic flow theory of homogenous and heterogeneous traffic, traffic operations, and networks.