OPEN ACCESS 90

ISSN 2356-5462
http://journals.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 11, No. 2, Dec 2025. pp. 90-110
doi:10.21108/ijoict.v11i2.9602

Prediction and Classification of Vehicle Traffic Congestion in Bandung City Using the Random Forest and K-Nearest Neighbors Algorithm

Muhammad Alauddin Angka Kurniawan[1], Sri Suryani Prasetiyowati[2], Yuliant Sibaroni[3*]

[1,2,3] *Informatics, Telkom University, Indonesia,*
Jl. Telekomunikasi No. 1 Terusan Buah Batu, Bandung, Jawa Barat, Indonesia, 40257

[*]yuliant@telkomunuversity.ac.id

Abstract

Traffic congestion remains one of the problems that continue to arise, especially in urban areas, one of which is Bandung City, when the causes of the problem are not managed properly. Continuous management of the causes of congestion problems will result in a controlled traffic system for the foreseeable future. This condition can be achieved if there is a congestion classification prediction system available. A reliable prediction and classification system can support the government in formulating data-based traffic management strategies. The Random Forest and K-Nearest Neighbor machine learning classification methods are strengthened with time-based feature expansion to capture traffic behavior in various time frames, so that the objectives can be achieved. The dataset obtained from Area Traffic Control System Bandung includes traffic flow recorded at 15-minute intervals at several intersections. Additional features such as red light duration, road width, and spatial proximity to residential and commercial areas are included to improve model performance. The results show that the Random Forest classifier with time-based feature expansion outperforms K-Nearest Neighbors, achieving the highest performance of 96%. These results show the potential contribution in short-term traffic prediction and its effectiveness in supporting urban traffic planning and congestion mitigation efforts in Bandung.

Keywords: Bandung, Congestion, Traffic, Kriging, Random Forest, K Nearest Neighbors.

I. INTRODUCTION

Traffic congestions are one of the most pressing challenges in urban transportation. This problem often occurs when the volume of vehicles on a particular road surpasses its intended capacity, resulting in a significant reduction in traffic speed, often approaching 0 km/h [1]. Congestions are even more common in major cities, including Bandung of which is among the largest urban areas in Indonesia and experiences high population growth [2]. In the year 2020, the population of Bandung City reached 2,510,103, reflecting an increase of 6,395 people from 2018 when it initially stood at 2,503,708 [3]. As the population continues to rise every year, road usage also escalates, driven by various purposes such as transportation, religious activities, sports, and etc. The number of personal vehicles in 2020 amounted in total to 1,538,788, which increased the likelihood of traffic volumes surpassing road capacity [4].

Received on 14 Jul 2025, Revised on 23 Jul 2025, Accepted and Published on 1 Dec 2025.

In order to resolve these issues, there are numerous different approaches that can be taken. One of the most effective modern approaches to address this issue is machine learning. Machine learning is a field of study focused on algorithms and statistical models that enable computers to perform tasks without explicit programming. By leveraging machine learning algorithms, it's possible to classify similar data and generate predictive insights from previously existing information. One of the core processes in machine learning is that each data needs to be grouped based on unique labels. This process is known as classification, where a machine learning model is trained to recognize and categorize data accordingly. By assessing existing data, the model learns patterns and characteristics to accurately classify new incoming data. To find the best solution for traffic congestion in Bandung, machine learning techniques can be utilized for prediction and classification, helping identify patterns and optimize decision-making.

Elaborating upon that topic, two of these machine learning techniques are Random Forest and K-Nearest Neighbors. Random Forest is a type of classifier that uses decision trees which works by creating many trees from different parts of data and features, then combining the results to make a final decision. A study shows how effective Random Forest is at making predictions, especially when dealing with large datasets that have various features [5]. K-Nearest Neighbors on the other hand is a classifier that groups data points based on how similar they are to other points in the area. It's effectiveness when less noise exists makes it useful for tasks like labeling traffic congestion or predicting future congestion levels [6].

The relevance of environmental attributes has also been highlighted in previous studies which uses Random Forest and other methods to model spatial traffic patterns. One study looked into how attributes such as road density, land-use diversity, bus route accessibility, and job-housing balance influence crash frequency across urban road networks [7]. Although the focus was on safety outcomes, the predictive modelling approach using Random Forest revealed that these built environment features are critical in understanding traffic dynamics, congestion included. The study highlighted that spatial urban environment attributes are effective features for machine learning models due to their influence on traffic-related phenomena.

Even though various studies have explored machine learning for traffic congestion prediction, inconsistencies are still present in algorithm performance, feature selection, and spatial relevance. Previous work often focuses only on temporal attributes without including contextual spatial features such as distance to use zones or timing characteristics. This study addresses these gaps by combining temporal feature expansion with spatial predictors including red light duration, road width, and distances to both activity centers and residential areas using ATCS data from Department of Communication and Information Technology, Bandung. Additionally, this study also seeks to assess both Random Forest and KNN classifiers across multiple time slices, analyzing performance and visualizing predictions to provide insights for city road planning and traffic policy.

II. LITERATURE REVIEW

A study established in Timika identified a range of contributing factors to urban traffic congestion. The factors mentioned include below-standard road infrastructure, poor traffic signal management, and driver behavior, emphasizing the varied essence of the issue [8]. Other than delays, traffic congestion can contribute to increased traffic accident risk and reduced road safety overall. Studies have shown that congestion, particularly in urban areas is linked to higher rates of traffic accidents especially when combined with poor city use planning and streetscape condition [9] [6] [10]. For example, a study found that congestion causally impacts accident occurrence in dense road networks [9]. Another study has demonstrated a U-shaped relationship between congestion and accidents across European cities [10]. Furthermore, urban infrastructure features such as intersections, road width, and land use diversity have also been found to affect both congestion intensity and accident frequency significantly [6] [11]. These results show the urgency of accurate congestion handling and the selection of features to reduce risk and support urban policy making.

In the hopes of finding the best method for addressing traffic congestion, multiple studies have been conducted. In one study discussing traffic congestion prediction, the findings resulted that the accuracy of the Support Vector Machine (SVM) classifier reached a percentage of 93%, while the Naïve Bayes classifier achieved only 90% accuracy using a dataset from Bandung provided by ATCS Bandung [12]. Meanwhile, another study conducted in the Amman region, specifically King Abdullah Road yielded significantly different results. This study employed various machine learning methods. In addition of using Random Forest and SVM, this study also includes Logistic Regression, and K-Nearest Neighbour (KNN) [13]. Two machine learning tools were used to support the findings. The tool developed by Waikato University, WEKA, and the tool created by the bioinformatics laboratory at the University of Ljubljana, Orange. Using the Orange Tool, the study found that the accuracy of the Random Forest and Logistic Regression classifiers reached 100%, while KNN and SVM achieved 99.8% and 99.1% accuracy, respectively. However, when using WEKA, the results differed, with SVM achieving 99.7% accuracy, while K-Nearest Neighbors, Logistic Regression, and Random Forest recorded accuracies of 98.7%, 97.6%, and 96.2%, respectively [13]. A comparative study was conducted on ensemble models for predicting traffic congestion in the year 2022. The study used a dataset from a highway in Gauteng Province, South Africa. The results were acquired using traditional machine learning methods such as Random Forest, Decision Tree, and Support Vector Machine (SVM), along with ensemble techniques like Bagging, AdaBoost, and Logistic Regression as a stacking method to combine Random Forest, Decision Tree, and Support Vector Machine (SVM). Attributes such as travel time, traffic volume, and average speed contributed to increasing prediction accuracy. Among all the methods, Decision Tree combined models gained the highest accuracy, with the traditional machine learning method reaching 98.3%, Bagging (DT) at 98.2%, and both AdaBoost (DT) and Logistic Regression (DT) achieving 99.7% [14].

Random Forest and K-Nearest Neighbors have both emerged as effective classifiers for traffic congestion prediction in past studies, often delivering high accuracy results. While Random Forest has been praised for its robustness and ability to handle high-dimensional, non-linear data, KNN offers simplicity and strong performance when feature distributions are well-separated. The previous findings suggest that while both methods are viable, their effectiveness is highly context-dependent prompting this study to compare their performances specifically on Bandung's sensor-based traffic data with enriched spatial-temporal features. This approach expands on previous studies by introducing a more contextualized feature set, including red light duration and proximity to land-use centers, which are often underexplored in comparative model evaluations.

In addition to the previous region-focused studies, one study also uses supervised machine learning models including Support Vector Machine, Classification Tree, and RUSBoosted Ensemble to predict congestion using environmental features from the Kaggle traffic volume dataset. Their study found that ensemble methods provided the highest accuracy. While their approach primarily relied on time-of-day and weather-related attributes, this study utilizes spatial elements such as red light duration, road width, and land use proximity, which have been shown to enhance prediction relevance in urban congestion contexts [15]. These findings support the use of spatial features to predict vehicle traffic congestions. A number of studies have also confirmed that land-use features especially the presence and distribution of residential, educational, and commercial zones play a crucial role in shaping traffic conditions in urban areas [16] [17] [18].

A study conducted in Xi'an, China, examined the spatial relationship between land use types and congestion sources, revealing that areas with high concentrations of educational and residential buildings frequently overlapped with congestion hotspots [16]. Another study focusing on a smaller urban context found that educational institutions and commercial facilities act as trip magnet during peak hours, adding to significant traffic buildup [17]. Even minor variations in land-use composition were shown to result in unstable high traffic volumes on surrounding roadways. A separate study centered in Kuala Lumpur demonstrated that commercial zones particularly those with high retail activity are consistently related with elevated congestion levels [18]. These findings shows the fact that traffic congestion is not solely influenced by vehicle counts, but also by how urban land is distributed and utilized. Therefore, the inclusion of distance to residential and activity centers as predictive features in traffic modeling is justified, as they represent real-world travel demand patterns.

III. RESEARCH METHOD

This study builds on existing traffic congestion prediction methods by integrating both spatial and temporal dimensions into the modeling process. Additional spatial features such as red light duration, proximity to residential areas, and proximity to activity centers are incorporated. These spatial features were supported by urban studies emphasizing the role of land-use distribution and traffic signal timing in congestion buildup [16] [18] [19]. Moreover, the approach this study uses introduces feature expansion over multiple time slices, allowing the model to learn short-term temporal patterns. By combining expanded temporal sequences with other spatial attributes, the study provides a more context-aware prediction of occupancy levels. These methodological modifications aim to improve accuracy, support urban traffic management, and provide interpretable information for local policy applications.

Fig. 1 presents the sequence of steps undertaken in this study, providing a clear and concise representation of the research process.

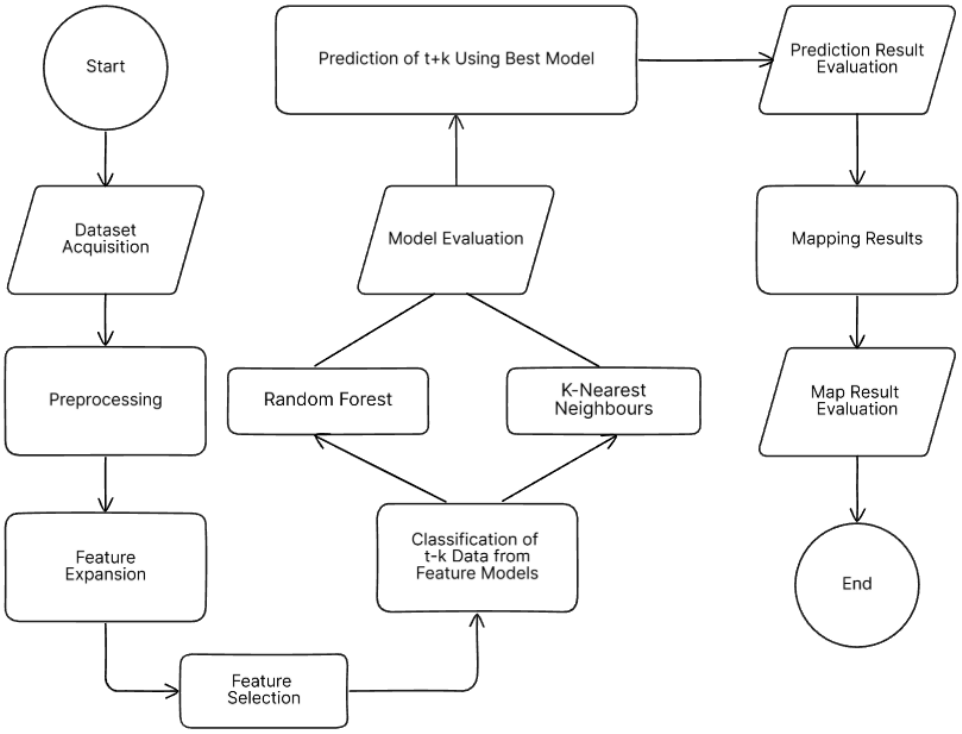


Fig. 1. Research Flowchart

A. Dataset Exploration

The dataset used in this study consists of traffic condition data from the Bandung City ATCS (Area Traffic Control System) for the year 2024, particularly the month of February. This dataset provides information on the state of vehicle traffic in Bandung at specific road points during particular time intervals. The description of the dataset attributes and their explanations are provided in TABLE I.

TABLE I
DATASET DESCRIPTION

Attribute	Description
Location	Name of Traffic Point Location.
Sensor	Label of Traffic Sensor. (According to Cardinal Directions)
Lane	Direction of Traffic Lane.
Time	Date and Time of Recorded Data. (DD-MM-YYYY hh:mm:ss)
Motorbikes	Number of motorbikes crossing the active sensor.
Cars	Number of cars crossing the active sensor.
Bus/Trucks	Number of bus and trucks crossing the active sensor.
Total Amount	Total amount of vehicles crossing the active sensor.
Headway (s)	Average time interval during which vehicles pass through the active sensor.
GAP (s)	Average distance of the back portion of vehicles with the front portion of vehicles behind it.
85p Speed	Average speed of the fastest 85% of vehicles on the road, uninfluenced by traffic conditions.
Average Speed	Average speed of vehicles.
Occupancy	Saturation value of the road.

Aside from the attributes in TABLE I, additional attributes related to the physical and spatial characteristics of each traffic point are required to better analyse observable patterns. The "Latitude" and "Longitude" attributes are added to indicate the precise location of each active sensor. This study also aims to investigate the relationship between Red Light Duration, and two spatial attributes which are Distance to Activity Centers, and Distance to Residential Areas in contributing to traffic congestion in Bandung.

The inclusion of red light durations are supported by recent studies, which show that the duration of red phases can significantly influence driver behavior and traffic flow. A study detailed that the timing of red lights affects drivers' route choices and decisions at intersections, which in turn alters congestion distribution across the network [19]. A localized case study on a signalized intersection in Bandung revealed that long queues up to 226 meters form during red phases, particularly at peak hours. This study's findings emphasize the importance of correct timing in shaping intersection performance and congestion buildup, reinforcing its value as a feature in traffic congestion modeling [20]. Previous studies have also found that areas near residential, commercial, to even educational land uses tend to experience higher traffic density due to increased travel demand and concentrated activity [9] [15] [16].

Attributes such as *Location*, *Lane*, *Sensor*, *Time*, *Latitude*, and *Longitude* will serve as ID labels for prediction results and are not included as features in the model. In accordance with the Indonesian Highway Capacity Manual (MKJI) 1997, the *Occupancy* attribute can be classified into several categories, each with its own label. The labels can be seen at TABLE II.

TABLE II
OCCUPANCY LABELS

Occupation Value	Congestion Level	Label
> 90	Highly Congested and Delayed Flow	5
>= 80 and < 90	Unstable Flow	4
>= 70 and < 80	Controlled Stable Flow	3
>= 60 and < 70	Stable Flow	2
< 60	Free Flow	1

The occupancy class is used to represent the level of vehicle density at a specific traffic light point at a given time, providing valuable insights for further analysis. This occupancy class attribute serves as the target variable

(Y) and is the main focus of this study. A more detailed explanation of all the attributes used in this study can be found in TABLE III, which presents comprehensive information about each attribute along with its description.

TABLE III
ATTRIBUTE DESCRIPTION

Attribute	Description
X_1	Red light duration
X_2	Distance to nearest activity point
X_3	Distance to nearest residential area
X_4	Total number of motorcycles that crossed the traffic point.
X_5	Total number of cars that crossed the traffic point.
X_6	Total number of buses and trucks that crossed the traffic point.
X_7	Total number of vehicles that crossed the traffic point.
X_8	Headway (Time difference between vehicles in a row when passing a certain point).
X_9	Gap.
X_{10}	85p Speed (Average speed of 85% of vehicles without concerning traffic).
X_{11}	Average speed.
Y_1	Occupancy class of related traffic point.

B. Dataset Acquisition

The dataset used in this study was obtained from active sensors installed at various traffic points. Data collection was carried out in collaboration with the Area Traffic Control System (ATCS) by scheduling a formal appointment to download the data directly to the ATCS server center. Currently, the database managed by the ATCS under the Department of Communication and Informatics (DISKOMINFO) of Bandung City does not provide an option to download one-month data ranges with fifteen-minute intervals per day. As a result, the data had to be downloaded manually by visiting the ATCS office and selecting data from each lane, sensor, and traffic point individually. This data collection process took approximately six weeks to complete.

The collected dataset includes vehicle traffic data from 00:00 to 23:45 with fifteen-minute intervals. To summarize the data and reduce computational load on the algorithm, each day's data was divided into four time classes for every measurement point, with six hours for each segment. This segmentation aims to simplify data representation and enable the model to obtain more contextualized patterns based on specific periods within a day. Each time class represents a time range with distinct traffic characteristics. A more detailed explanation of the definitions and intervals of each time class can be found in TABLE IV.

TABLE IV
TIME CLASS LABELS

Time	Label
00:15 – 06:00	I
06:15 – 12:00	II
12:15 – 18:00	III
18:15 – 00:00	IV

To gather data regarding the "Duration" attribute, direct field observations were conducted. The purpose of mentioned field observation was to determine the red light duration for each direction at the selected traffic light points. A single traffic light intersection can have varying durations for different directions. Consequently,

the red light durations were recorded separately for each direction. The collected data was then averaged to produce the final value for the "Duration" attribute. The complete values of the "Duration" attribute obtained through this process are presented in TABLE V.

TABLE V
TRAFFIC POINT RED LIGHT DURATION

Traffic Point Location	Red light Duration (s)
SP. Arcamanik	120
SP. Batununggal	180
SP. Buahbatu	150
SP. Gedebage	180
SP. Pasteur	300
SP. Samsat	360
SP. Moh Toha	150
SP. Ujungberung	150

The attributes "Distance to Activity Center" and "Distance to Residential Area" were obtained by calculating the proximity to the nearest activity centers and residential areas. In this study, activity centers include places such as universities, shopping malls, markets, or sports stadiums. The distances to the three closest activity centers or residential areas were averaged to generate the final attribute value. Once the distances were calculated and averaged, the resulting values were entered into the corresponding feature attributes. The averaged distance values for each point are presented in TABLE VI.

TABLE VI
TRAFFIC POINT DISTANCES

Traffic Point Location	Distance to Activity Centers (Km)	Distance to Residential Areas (Km)
SP. Arcamanik	0.560	0.600
SP. Batununggal	1.600	0.717
SP. Buahbatu	0.825	1.400
SP. Gedebage	0.675	0.717
SP. Pasteur	1.250	0.330
SP. Samsat	2.650	1.800
SP. Moh Toha	1.350	0.650
SP. Ujungberung	0.450	0.350

C. Data Preprocessing

Before the acquired dataset can be utilized, a comprehensive preprocessing phase is conducted to ensure optimal performance of the machine learning algorithms used. The details of this phase are described in the following subchapters.

1) Data Cleaning: The dataset must be cleaned to remove null values, as these may interfere with the algorithm's performance. Inconsistent rows or columns are excluded to enhance the algorithm's effectiveness in classification and prediction. During this preprocessing phase, the dataset is filtered to retain only the essential data.

2) Data Integration: After the cleaning process is done, the data is reintegrated into a single data frame. The merged data must be ensured that it remains coherent, accurate, and ready for use in analytical algorithms. A verification of data quality is then conducted to avoid inconsistencies or errors that may affect analysis.

3) Data Modelling: At this stage, the dataset is divided into eight distinct models. Seven models each representing separate portions of the data, and one additional model that combines all segments (a-g, and *gab*). The difference of features and data for each portion allows for more variation in data characteristics used for analysis. This approach is designed to explore differences in model performance based on the diversity of features and the methods used for feature extraction.

4) Data Transformation: At this stage, the data is normalized and generalized to ensure consistency and facilitate further analysis. The process is carried out to rescale different feature values into a uniform range, therefore improving the performance of the applied algorithm. This study applies the min-max normalization method, which transforms feature values into a desired range (0 to 1), in order to support the predictive model training process.

D. Feature Expansion

Feature expansion is performed to identify the most effective features for predicting traffic congestion occupancy at various traffic points based on the dataset. This process aims to enhance the information that can be obtained from the data, enabling the models to make more accurate predictions. In this study, feature expansion is carried out by incorporating data using t-k time slices from one to seven. Each time slice uses data and features from the previous slice respective to the value of k it has. Using the previous segmentation, each value of k represents 6 hours. These time slices will be used to predict future values, below is the formula to prepare such prediction:

$$Y_{t+k} = f(X_{t-1}, X_{t-2}, \dots, X_{t-k}) \quad (1)$$

E. Feature Combination

After feature expansion is performed, each model corresponding to each time step (t) is analyzed to determine the best feature combinations for prediction. In this study, the feature selection process involves selecting between three to the maximum number of top-performing features of each time slice.

F. Feature Selection

The expanded features are then selected using the SelectKBest method, applied to each possible combination for every model. Due to the feature expansion approach, the number of features generated per model ranges from eleven to seventy-seven. This results in the largest model having over 50 possible feature combinations, requiring careful computational analysis to identify the optimal set of features. Selecting only the most relevant features is essential not only for improving model accuracy, but also for minimizing noise and reducing overfitting. Methods like SelectKBest help prioritize input features that have the strongest statistical relationship with the target variable, improving the interpretability and efficiency of the predictive model. This approach is particularly important when dealing with time series data and high dimensional input spaces, where irrelevant or weakly correlated features can degrade model performance. [21]

G. Classification of T-k Data from Feature Models

At this stage, classification and prediction are performed using the Random Forest and K-Nearest Neighbors (KNN) classifiers. The prediction process involves training and testing each model based on its respective features and class labels.

H. Random Forest

At this stage, the prediction process is carried out using the Random Forest classifier. Random Forest is an ensemble learning algorithm that uses multiple decision trees and aggregates their outputs to determine the final classification result through majority voting. This classifier is particularly well suited for high-dimensional data and well known for its high accuracy and robustness [22]. According to a study, Random Forest builds several trees on randomly selected subsets of data and features, with the final prediction being determined by the majority vote across all trees, making it both stable and effective for traffic-related classification tasks. [23] The architecture of this classifier is visualized in Fig. 2.

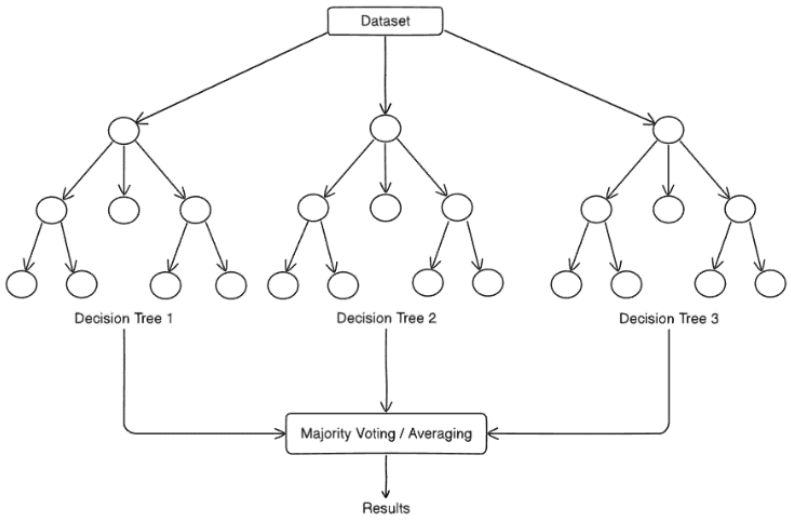


Fig. 2. Random Forest Architecture

The dataset that has been segmented and processed into various models is then used for training and testing in the prediction process. Each model is trained independently to evaluate prediction performance across different data partitions. The classifier is then tuned with several hyperparameters to prevent overfitting and ensure good generalization. These hyperparameters are the number of trees in the forest (n_estimators), the maximum depth of each tree (max_depth), the minimum number of samples at a leaf node (min_samples_leaf), the maximum number of features needed for splitting (max_features), and a random_state value to ensure reproducibility of the results.

The dataset segment for each model is then divided into two subsets, a training subset comprising 80% of the total data and a testing subset comprising the remaining 20%. This step ensures that the evaluation results reflect the model's ability to generalize to new data. After the training and testing processes are completed, the next step involves performing cross-validation to evaluate the model's performance more comprehensively.

Cross-validation is employed to calculate evaluation metrics that include accuracy, precision, F1-score, and recall. Those four metrics together provide a wider view of the model's effectiveness in predicting. This process ensures that the model is not overfitting, but also maintains a robust performance across different subsets of data. The selection of the best model for each time slice is based on the accuracy score achieved, while also taking into account the class distribution with the dataset. Class variation is an important consideration to ensure that the model does not exhibit bias and is capable of accurately crafting predictions.

I. K-Nearest Neighbors

At this stage, the prediction process is now carried out using K-Nearest Neighbors (KNN) classifier. Each model is trained independently to evaluate prediction performance on their respective data partitions. The KNN method classifies data based on the proximity to a defined number of nearest Neighbors, making the selection of appropriate parameters crucial to ensure prediction accuracy. It makes predictions by comparing the test data directly to the training instances. This property enables it to adapt flexibly to local structures in the data but also makes it sensitive to parameter selection and data scaling [22]. This classifier is further visualized in Fig. 3.

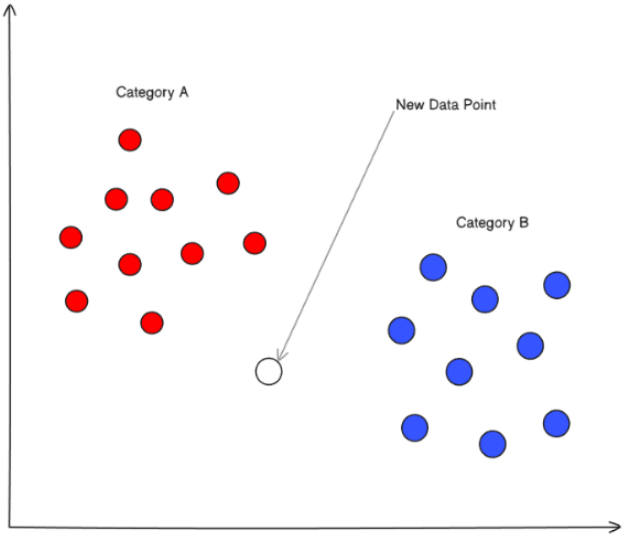


Fig. 3. KNN Architecture

To prevent overfitting and ensure good generalization, the KNN classifier is optimized using several hyperparameters. These hyperparameters are the number of Neighbors used in the prediction process (n_neighbors), the weighting scheme applied to each neighbour (weights), and the algorithm used to compute distances and identify Neighbors (algorithm). By tuning them, the KNN model can be adapted to handle varied characteristics of each data partition model, improving prediction accuracy and consistency.

As it has been done for the previous model, the dataset is split into two subsets, a training subset comprising 80% of the total data and a testing subset comprising the other 20%. Following the training and testing phases, cross-validation is again performed to further assess the obtained results. The best model for each time interval is selected based on accuracy while also considering the class distribution within the dataset.

J. Model Results Evaluation

Upon completion of training and prediction using both classifiers, the results are then compiled and compared to determine each classifier's characteristics that may be used.

K. Prediction Using Best Model

Once both classifier is properly trained, predictions are generated using each methods for each time intervals (t+1, t+2, t+k, and so on). The prediction results for each time step are systematically recorded in a new, dedicated dataframe designed to neatly store the output of the predictions. To offer practical insights for government policy making, the outcomes of this prediction will serve as the foundational framework for the map in the subsequent phase.

L. Mapping Using Ordinary Kriging Interpolation

At this stage, the prediction results are mapped using available mapping tools to provide clearer and modr informative visualization. In this study, ArcMap 10.8 and ArcGis Online was used to create layers and map the predicted occupancy class results generated by the models.

M. Results Evaluation

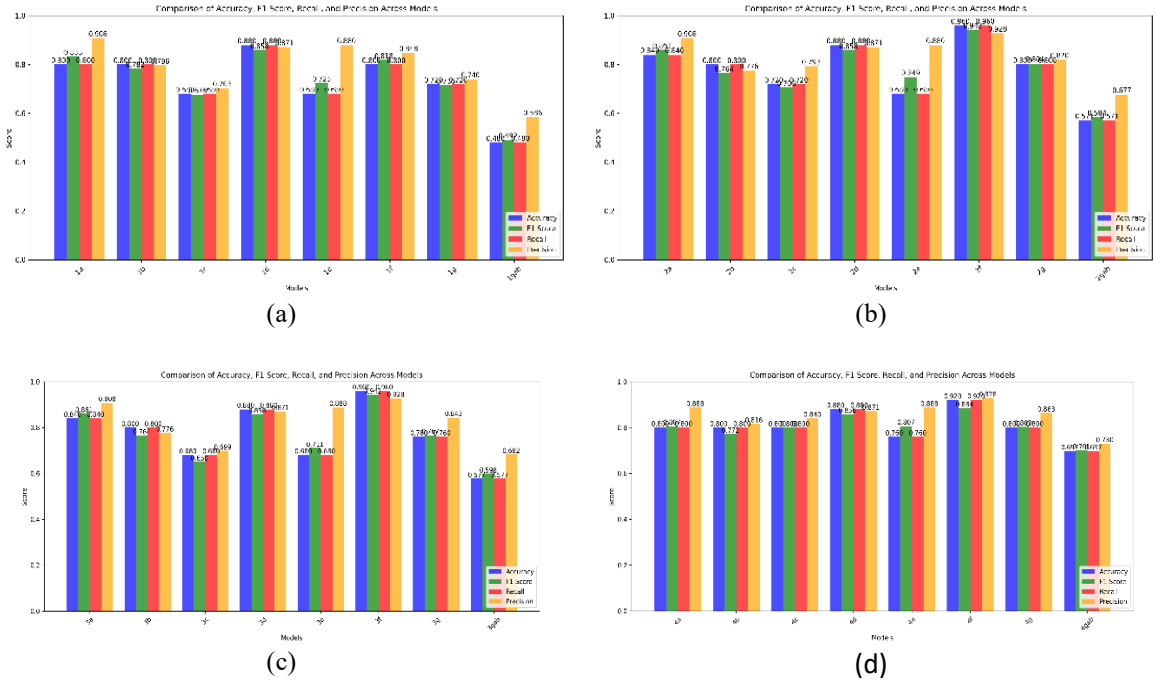
In the final stage of the study, the mapped occupancy class predictions were thoroughly re-evaluated to assess the accuracy and relevance. This evaluation involved analyzing traffic occupancy density patterns across various traffic points on the visualizations and across time slices. Based on this assessment, appropriate recommendations are formulated to support traffic management decisions. These recommendations are intended to assist authorities in designing more effective and efficient strategies to alleviate traffic congestion.

IV. RESULTS AND DISCUSSION

This study employs both Random Forest and K-Nearest Neighbors algorithms to identify which yields the highest accuracy, while also considering other performance metrics including variance, F1-score, recall, and precision.

A. Result of Random Forest Prediction Metrics

Following the selection of best features for each time model, these features were subsequently used in both the training and prediction phases of the algorithm. These results are presented in Fig. 4.



(a) (b)

(c) (d)

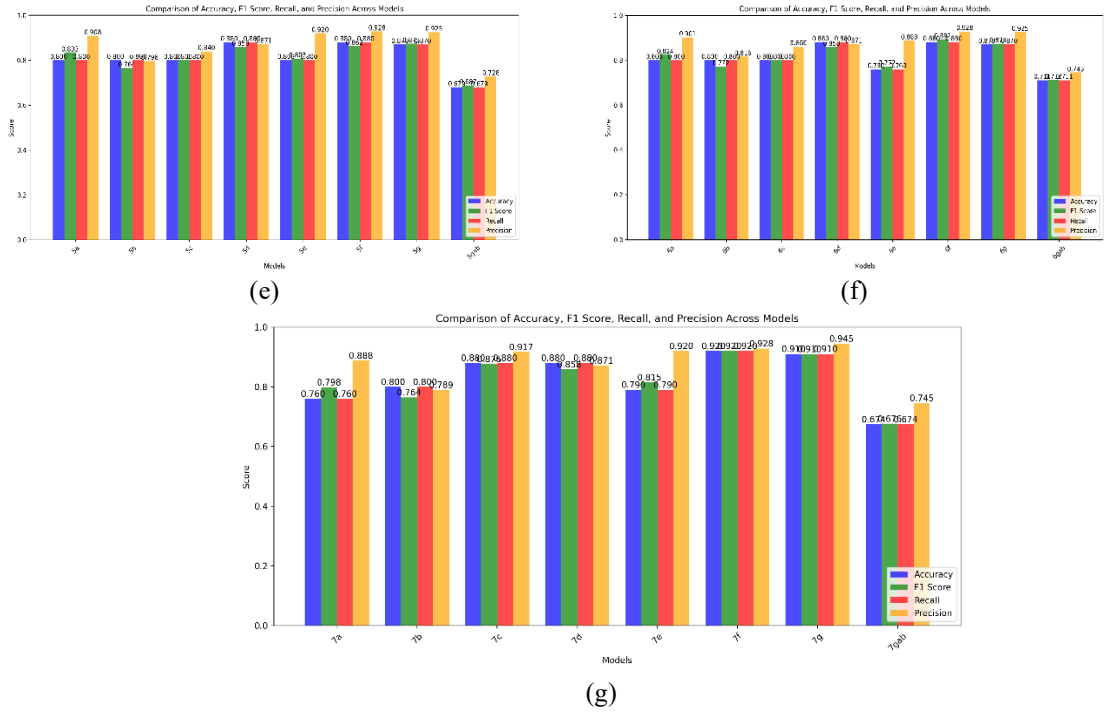


Fig. 4. Random Forest Metrics Results, (a) t-1, (b) t-2, (c) t-3, (d) t-4, (e) t-5, (f) t-6, (g) t-7 (BLUE) Accuracy, (GREEN) F1 Score, (RED) Recall, (YELLOW) Precision.

Each of the models generated a range of evaluation metrics, which varied depending on the specific features included. The results indicate that the Random Forest classifier in its best results achieved an accuracy rate of 96%, an F1-Score of 94%, a recall of 96%, and a precision of 91%. A more detailed breakdown of the model's performance across different time slices and feature sets is provided in TABLE VII, offering a comprehensive view of its predictive capabilities in various scenarios.

TABLE VII
BEST T-K RANDOM FOREST MODELS

t-k	Best Model	Number of Features	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)
t-1	1d	3	0.880	0.858	0.880	0.871000
t-2	2f	19	0.960	0.942	0.960	0.913704
t-3	3f	29	0.960	0.942	0.960	0.913704
t-4	4f	37	0.920	0.884	0.920	0.928000
t-5	5f	37	0.880	0.862	0.880	0.928000
t-6	6f	25	0.880	0.862	0.892	0.928000
t-7	7f	71	0.920	0.920	0.920	0.928000

B. Results of K-Nearest Neighbors Prediction Metrics

Enduring a similar process, after selecting the best features for each time model, these features were both used in the training and prediction phases of the algorithm. These results are presented in Fig. 5.

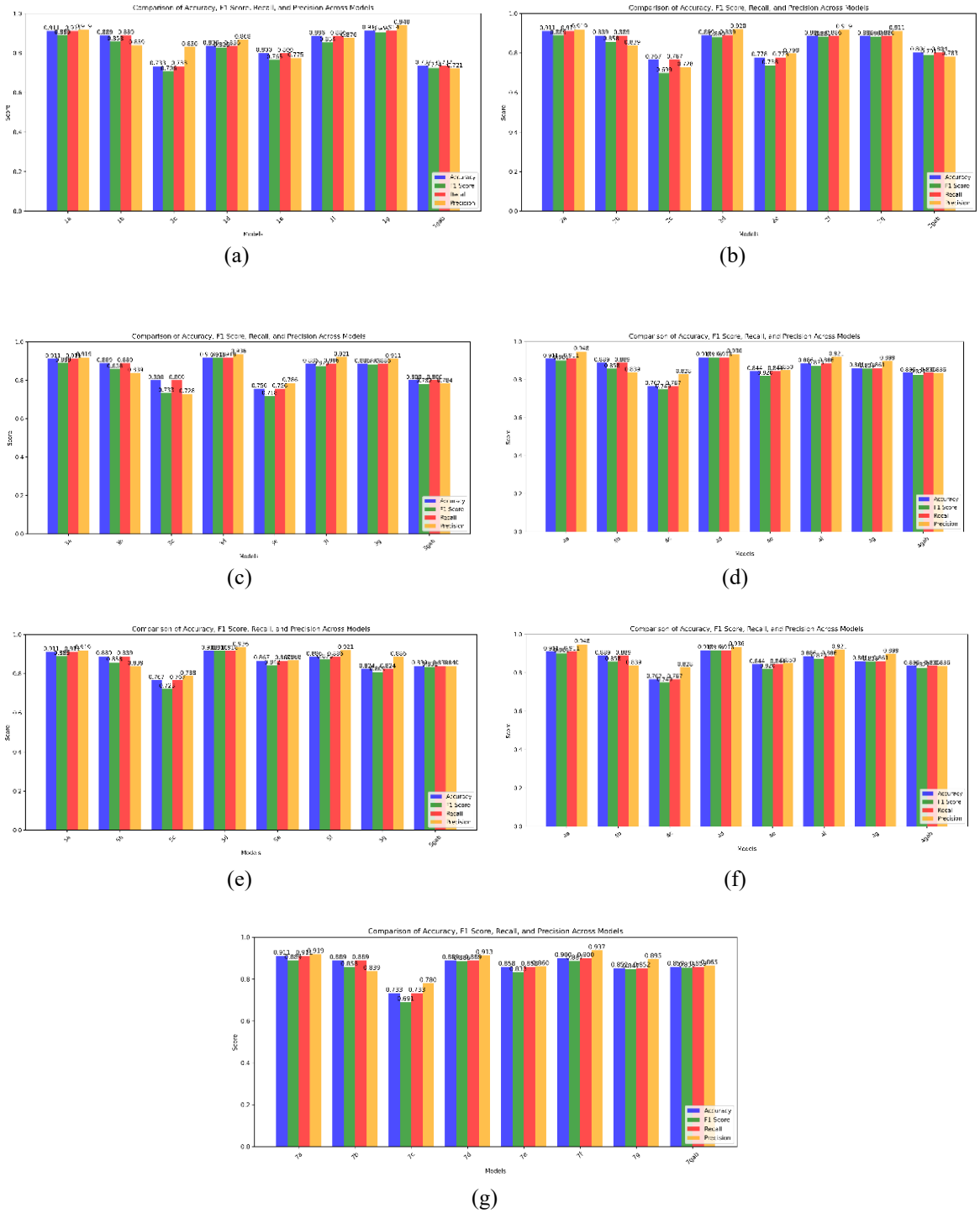


Fig. 5. KNN Metrics Results (a) t-1, (b) t-2, (c) t-3, (d) t-4, (e) t-5, (f) t-6, (g) t-7
(BLUE) Accuracy, (GREEN) F1 Score, (RED) Recall, (YELLOW) Precision.

The results indicate that the KNN classifier in the best results achieved an accuracy rate of 94%, an F1-Score of 94%, a recall of 84%, and a precision of 95%. Further detailed results are displayed at TABLE II.

TABLE II
BEST T-K KNN MODELS

t-k	Best Model	Number of Features	Accuracy (%)	F1-Score (%)	Recall (%)	Precision (%)
t-1	1a	5	0.911	0.889	0.911	0.919
t-2	2a	9	0.911	0.889	0.911	0.919
t-3	3d	4	0.918	0.916	0.918	0.936
t-4	4d	5	0.918	0.916	0.918	0.936
t-5	5d	6	0.918	0.916	0.918	0.936
t-6	6d	7	0.943	0.942	0.943	0.956
t-7	7d	9	0.914	0.912	0.914	0.933

C. Best Features Results

As previously explained, preprocessing of the dataset was conducted to gain reliable outcomes. Feature expansion techniques were used to determine the most relevant features for model construction. Specifically, the SelectKBest method was utilized to identify and select the optimal set of features. The features utilized in the Random Forest classifier can be seen at TABLE III, while the ones used in the K-Nearest Neighbors can be seen at TABLE IV.

TABLE III
BEST MODEL FEATURES OF RANDOM FOREST MODEL

t-k	Best Model	Features
t-1	1d	x1, x2, x5
t-2	2f	x1_1, x1_2, x1_4, x1_7, x1_8, x1_9, x1_10, x1_11, x2_1, x2_2, x2_3, x2_4, x2_5, x2_6, x2_7, x2_8, x2_9, x2_10, x2_11
t-3	3f	x1_1, x1_2, x1_4, x1_7, x1_8, x1_9, x1_10, x1_11, x2_1, x2_2, x2_3, x2_4, x2_5, x2_6, x2_7, x2_8, x2_9, x2_10, x2_11, x3_1, x3_2, x3_3, x3_4, x3_5, x3_6, x3_7, x3_8, x3_9, x3_10, x3_11
t-4	4f	x1_1, x1_2, x1_7, x1_8, x1_9, x1_10, x1_11, x2_1, x2_2, x2_3, x2_5, x2_6, x2_7, x2_8, x2_9, x2_10, x2_11, x3_1, x3_2, x3_3, x3_5, x3_6, x3_7, x3_8, x3_10, x3_11, x4_1, x4_2, x4_3, x4_4, x4_5, x4_6, x4_7, x4_8, x4_9, x4_10, x4_11
t-5	5f	x1_1, x1_2, x1_7, x1_8, x1_9, x1_10, x1_11, x2_1, x2_2, x2_3, x2_6, x2_8, x2_10, x2_11, x3_1, x3_2, x3_3, x3_5, x3_6, x3_7, x3_8, x3_10, x3_11, x4_1, x4_2, x4_3, x4_4, x4_5, x4_7, x4_10, x4_11, x5_1, x5_2, x5_3, x5_8, x5_10, x5_11
t-6	6f	x1_2, x1_9, x1_10, x1_11, x2_3, x2_10, x3_3, x3_5, x3_7, x3_8, x3_10, x3_11, x4_3, x4_5, x4_7, x4_10, x4_11, x5_3, x5_8, x5_10, x5_11, x6_3, x6_8, x6_10, x6_11
t-7	7f	x1_1, x1_2, x1_4, x1_5, x1_6, x1_7, x1_8, x1_9, x1_10, x1_11, x2_1, x2_2, x2_3, x2_4, x2_5, x2_6, x2_7, x2_8, x2_9, x2_10, x2_11, x3_1, x3_2, x3_3, x3_4, x3_5, x3_6, x3_7, x3_8, x3_9, x3_10, x3_11, x4_1, x4_2, x4_3, x4_4, x4_5, x4_6, x4_7, x4_8, x4_9, x4_10, x4_11, x5_1, x5_2, x5_3, x5_6, x5_8, x5_9, x5_10, x5_11, x6_1, x6_2, x6_3, x6_4, x6_5, x6_6, x6_7, x6_8, x6_9, x6_10, x6_11, x7_1, x7_2, x7_3, x7_4, x7_5, x7_6, x7_7, x7_10, x7_11

TABLE IV
BEST MODEL FEATURES OF KNN MODEL

t-k	Best Model	Features
t-1	1a	x1, x2, x3, x10, x11
t-2	2a	x1_1, x1_2, x1_3, x1_10, x1_11, x2_1, x2_2, x2_3, x2_10

t-3	3d	x1_1, x2_1, x3_1, x3_5
t-4	4d	x1_1, x2_1, x3_1, x3_5, x4_1
t-5	5d	x1_1, x2_1, x3_1, x3_5, x4_1, x5_1
t-6	6d	x1_1, x2_1, x3_1, x3_5, x4_1, x5_1, x6_1
t-7	7d	x1_1, x2_1, x3_1, x3_5, x4_1, x5_1, x6_1, x7_1, x7_5

D. Prediction Results

After gathering the prediction metrics of each classifier, both classifiers are utilized in the prediction algorithm using their optimal feature combinations and the most effective time slices identified during model evaluation. The results of the Random Forest classifier are presented at TABLE IV, while the results of the KNN classifier are presented at TABLE IV.

TABLE IV
RANDOM FOREST PREDICTION RESULTS

Location	Sensor	Lane	T+1	T+2	T+3	T+4	T+5	T+6	T+7
SP. Arcamanik	SOUTH	SOUTH	1	1	1	1	1	1	1
SP. Arcamanik	NORTH	NORTH	1	1	1	1	1	1	1
SP. Buahbatu	WEST	STRAIGHT RIGHT	1	1	1	1	1	2	2
SP. Gedebage	WEST	STRAIGHT RIGHT	2	2	2	2	1	1	1
SP. Gedebage	WEST	STRAIGHT	1	1	1	1	1	1	1
SP. Gedebage	SOUTH	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Gedebage	SOUTH	LEFT TURN	1	3	1	2	1	1	1
SP. Gedebage	NORTH	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Gedebage	NORTH	LEFT TURN	1	1	1	1	1	1	1
SP. Pasirkoja	WEST	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Pasirkoja	NORTH	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Pasteur	WEST	STRAIGHT	1	1	1	1	1	1	1
SP. Pasteur	EAST	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Pasteur	NORTH	STRAIGHT	2	1	1	1	1	1	1
SP. Samsat	WEST	RIGHT TURN	1	1	1	1	1	1	1
SP. Samsat	WEST	STRAIGHT RIGHT	1	3	1	3	1	1	1
SP. Samsat	NORTH	STRAIGHT RIGHT	1	1	1	1	1	1	3
SP. Moh Toha	WEST	STRAIGHT RIGHT	3	3	3	1	3	2	2
SP. Moh Toha	WEST	LEFT TURN	2	2	2	2	1	1	2
SP. Moh Toha	SOUTH	STRAIGHT RIGHT	1	2	2	1	1	1	1
SP. Moh Toha	SOUTH	LEFT TURN	1	1	1	1	1	1	1
SP. Moh Toha	EAST	STRAIGHT RIGHT	1	1	1	1	1	2	2
SP. Ujungberung	SOUTH	LEFT TURN	2	2	2	2	1	1	1
SP. Ujungberung	EAST	STRAIGHT LEFT	1	1	1	1	1	1	1

TABLE IV
KNN PREDICTION RESULTS

Location	Sensor	Lane	T+1	T+2	T+3	T+4	T+5	T+6	T+7
SP. Arcamanik	SOUTH	SOUTH	1	1	1	1	1	1	1
SP. Arcamanik	NORTH	NORTH	1	1	1	1	1	1	1
SP. Buahbatu	WEST	STRAIGHT RIGHT	1	1	1	3	2	1	3
SP. Gedebage	WEST	STRAIGHT RIGHT	1	1	1	1	1	1	3
SP. Gedebage	WEST	STRAIGHT	1	1	1	1	1	1	3
SP. Gedebage	SOUTH	STRAIGHT RIGHT	1	1	1	3	2	1	1
SP. Gedebage	SOUTH	LEFT TURN	1	1	1	1	1	1	1
SP. Gedebage	NORTH	STRAIGHT RIGHT	1	1	1	3	2	1	1
SP. Gedebage	NORTH	LEFT TURN	1	1	1	1	1	1	1
SP. Pasirkoja	WEST	STRAIGHT RIGHT	1	1	1	1	1	1	3
SP. Pasirkoja	NORTH	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Pasteur	WEST	STRAIGHT	1	1	1	1	1	1	1
SP. Pasteur	EAST	STRAIGHT RIGHT	1	1	1	1	1	1	1
SP. Pasteur	NORTH	STRAIGHT	3	1	1	1	1	1	1
SP. Samsat	WEST	RIGHT TURN	3	3	3	3	2	3	1
SP. Samsat	WEST	STRAIGHT RIGHT	3	1	1	1	1	1	1
SP. Samsat	NORTH	STRAIGHT RIGHT	3	3	3	3	2	3	1
SP. Moh Toha	WEST	STRAIGHT RIGHT	1	1	1	3	2	1	3
SP. Moh Toha	WEST	LEFT TURN	1	1	1	1	1	1	1
SP. Moh Toha	SOUTH	STRAIGHT RIGHT	1	1	1	3	2	1	3
SP. Moh Toha	SOUTH	LEFT TURN	1	1	1	1	2	1	1
SP. Moh Toha	EAST	STRAIGHT RIGHT	1	1	1	3	2	1	3
SP. Ujungberung	SOUTH	LEFT TURN	1	1	1	1	1	1	1
SP. Ujungberung	EAST	STRAIGHT LEFT	1	1	1	1	1	1	1

E. Mapping Interpolation Results

Mapping was conducted for each time slice generated by the prediction. The interpolation process was crafted using ArcMap's Geospatial analysis interpolation tool, specifically employing the ordinary kriging method, and ArcGis Online's sketching tool. Each occupancy class is assigned a unique color code to simplify interpretation. Detailed information regarding the color codes is provided in TABLE IV.

TABLE IV
COLOR CODES

Color	Class	Description
Black	1	Free flow
Orange	2	Stable flow
Red	3	Controlled Stable flow

The mapping results of the random forest classifier can be seen at Fig. 6, while the results of the KNN classifier can be seen at Fig. 7.

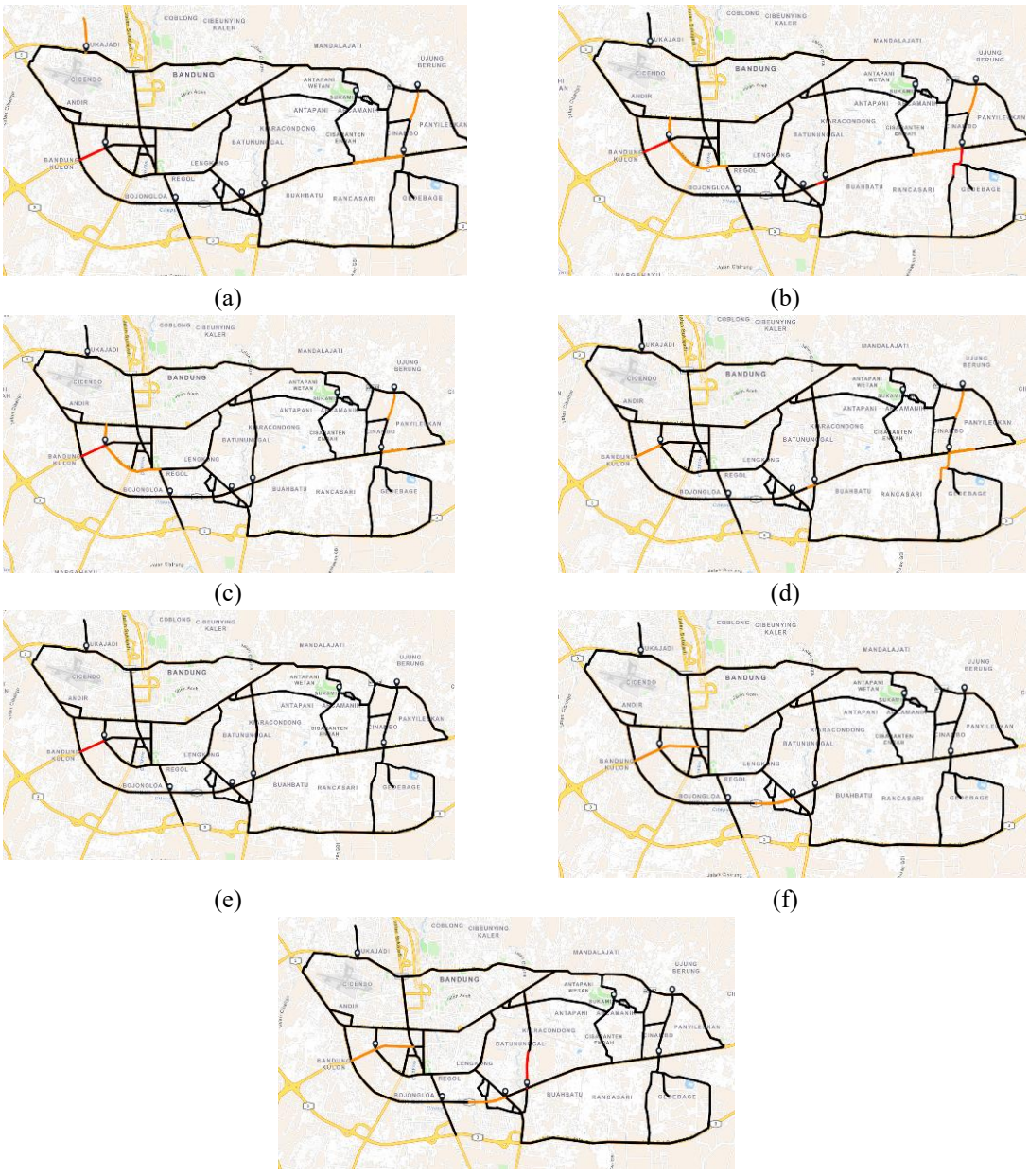


Fig. 6. Random Forest Prediction Mapping Results (a) t-1, (b) t-2, (c) t-3, (d) t-4, (e) t-5, (f) t-6, (g) t-7 (BLUE) Free flow, (ORANGE) Stable flow, (RED) Controlled Stable flow.

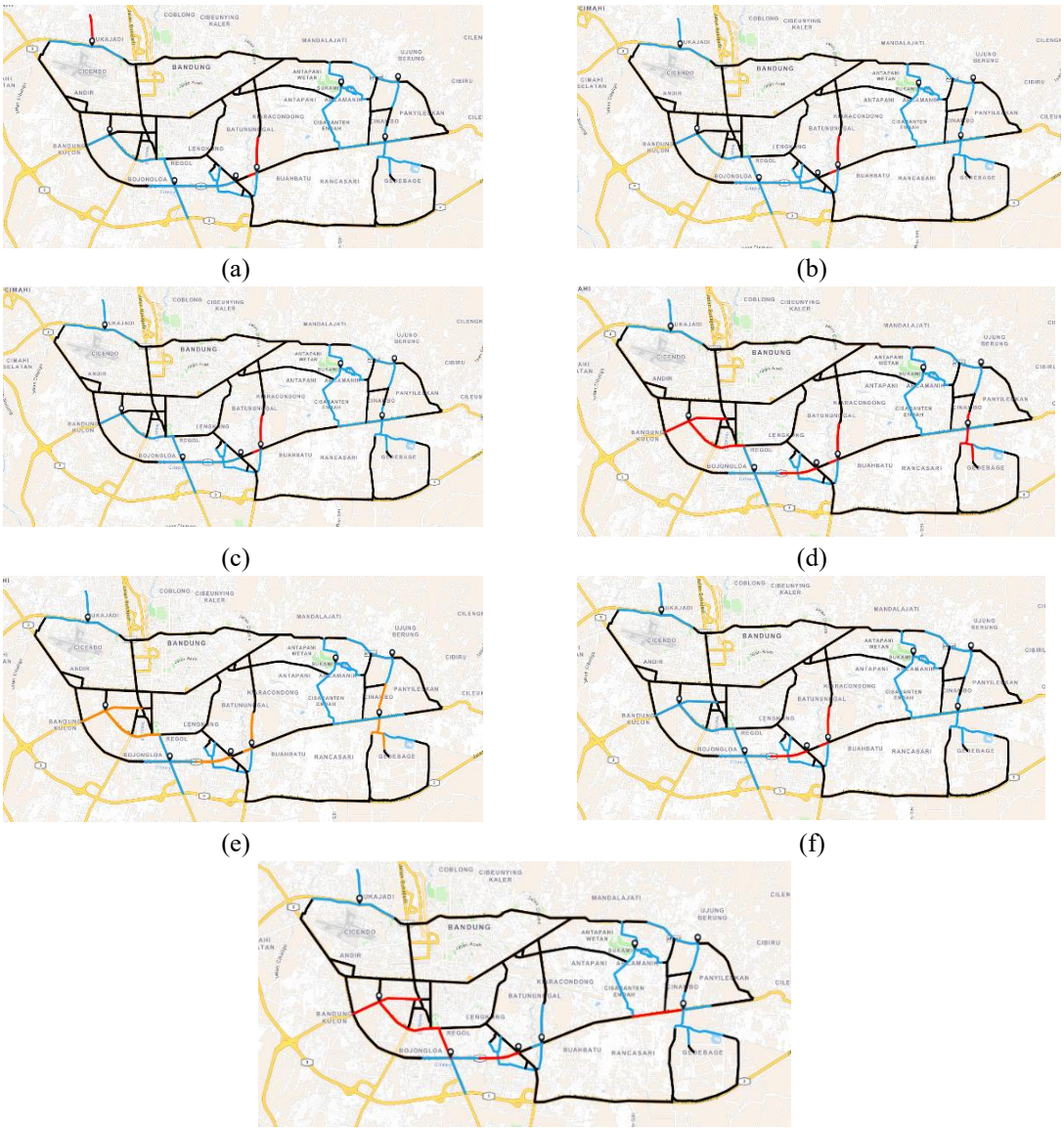


Fig. 7.
KNN Prediction Mapping Results (a) t-1, (b) t-2, (c) t-3, (d) t-4, (e) t-5, (f) t-6, (g) t-7
(BLUE) Free flow, (ORANGE) Stable flow, (RED) Controlled Stable flow.

F. Discussions

The results of this study demonstrate that both the Random Forest and K-Nearest Neighbors (KNN) classifiers are similarly effective in predicting traffic congestion occupancy across multiple time intervals, though Random Forest acquired higher percentage in overall metrics. By applying feature expansion, the models were able to capture traffic patters more accurately, which contributed to the improvements in predictive performance across different configurations. In its best-performing model, Random Forest achieved an accuracy of 96%, F1-Score of 94%, recall of 96%, and precision of 91%. Meanwhile, its lowest model could reach an accuracy of 48% ,an F1 Score of 49%, a recall of 48%, and precision of 58%, both shown in TABLE VII.

K-Nearest Neighbors (KNN) demonstrated competitive performance, achieving an accuracy of 94%, F1-Score of 94%, recall of 84%, and precision of 95%, in its best model. While the lowest model gained an accuracy of 73%, an F1 Score of 69%, recall of 73%, and precision of 78% as shown in TABLE II. These results suggest that while KNN may be more sensitive to feature composition, it can serve as a viable lightweight model for near-real-time applications or systems with limited computational resources.

Comparing the best features of each models with the other, it can be seen that Random Forest uses features even more extensively compared to KNN. The highest feature count a random forest model uses in this study reaches up to 71 features, while KNN only uses a maximum of 9 as seen at TABLE VII and TABLE II. It was found that features such as Red light duration, Distance to Activity Point, Distance to residential point, Total number of Vehicles, Headway, and 85p Speed influence the occupation prediction better than the rest. This was proven by these features consistently being used in best models of each time frame, especially both Distance to Activity Point and Distance to Residential Point where they appear in every time frame. Though the previous is true, time is also an influencing factor in this study. A few features can be considered only relevant because they are from a certain time frame, as the feature Distance to Residential Point is only consistently relevant from the t-2 frame and so on (x2_3) as shown in TABLE III.

In addition to predictive performance, the two classifiers exhibit distinct characteristics in terms of computational efficiency and deployment. In this study, the Random Forest algorithm required a longer runtime due to the complexity of training multiple decision trees and performing feature selections. Meanwhile, K-Nearest Neighbors demonstrated faster overall execution time, as it relies primarily on distance calculations during prediction. These distinctions highlight the practical trade-offs between both methods in terms of computation, transparency, and operational context.

Spatial mapping of the predicted congestion levels using ArcGIS online provided further insight. Visualizing the output highlighted congestion points across Bandung City area from free flow to controlled stable flow. These spatial patterns align with known traffic conditions in Bandung, enhancing confidence in the model's outputs and offering actionable information for urban planners. From the prediction results, it can be concluded that SP. Gedebage, SP. Samsat, and SP. Moh. Toha consistently shown higher predicted traffic levels across both classifiers, as shown in both Fig. 6 and Fig. 7. Notably, SP. Ujungberung was predicted to have high traffic only by the Random Forest classifier. These results suggest that features such as Distance to Activity Centers, Distance to Residential Areas, and Red Light Duration has a meaningful role in affecting traffic. For example, SP. Gedebage is located close to both activity centers and residential zones, which likely contributes to its traffic levels consistently being high. On the other hand, SP. Samsat is not located near significant residential or activity centers at the time of study, but has a long red light duration, which may have led to increased congestion predictions.

V. CONCLUSION

This study explored the use of machine learning algorithms Random Forest and K-Nearest Neighbors (KNN) to predict traffic occupancy levels using spatial data. Through a comprehensive preprocessing, feature expansion, training, and model evaluation using metrics such as accuracy, F1-Score, recall, and precision, the findings demonstrate that both algorithms show effectiveness in capturing certain traffic patterns. Random Forest achieved higher accuracy and overall metrics with an accuracy reaching 96% while KNN on the other hand obtained 94%, establishing the prior classifier as the more robust option for this study. The inclusion of additional spatial urban features, such as red light durations, distances to activity centers, and distances to residential areas further enhance performance. It could be concluded that features such as Red light duration, Distance to Activity Point, Distance to Residential Point, Total number of Vehicles, Headway, and 85p Speed influence the occupation prediction better than the rest. This conclusion is found because these features consistently become chosen features in the best model for each time frame, especially the Distance to Activity

Point as it appears in every time frame. Notably, the results revealed difference in predicted traffic levels across locations with varying values for these attributes. Additionally, traffic points such as SP. Gedebage, SP. Samsat, and SP. Moh Toha is predicted to have a higher traffic occupancy level compared to the rest.

THREATS TO VALIDITY

While this study uses feature expansion and spatial features for traffic congestion prediction, certain limitations should be acknowledged. First, the dataset is sourced exclusively from Department of Communication and Informatics (DISKOMINFO) of Bandung City's Area Traffic Control System's sensor database, which may impact the generalizability of the findings to other areas with different traffic infrastructure, urban plans, or vehicle behaviors. Additionally, while classifiers such as Random Forest and KNN have demonstrated high predictive performance in prior works [13] [14], their effectiveness may vary across datasets depending on feature usage and noise in each dataset. Moreover, the use of manually sourced features (e.g., red light duration and proximity to activity centers) introduces the potential for measurement error and subjectivity, which could affect internal validity The models also rely on historical data without using additional environmental factors such as weather or other events, potentially limiting responsiveness in live deployment contexts. Future enhancements may include integrating diverse data sources, testing models across different cities, and refining the usage of environmental features.

ACKNOWLEDGMENT

We would like to thank Telkom University, as their guidance and assistance supported this study from start to finish. In addition, we would like to give our largest appreciation to Area Traffic Control System Bandung and Department of Communication and Informatics (DISKOMINFO) of Bandung City for allowing us to collaborate with them in gathering their traffic counting data.

REFERENCES

[1] E. Harahap, Z. Aditya, F. Badruzzman, Y. Fajar, A. Bastia, S. Zein and A. Kudus, "Solusi Kemacetan Lalu Lintas Kota Bandung Melalui Pemerataan Arus Kendaraan", *Sains, Aplikasi, Komputasi dan Teknologi Informasi,* vol. 4, no. 1, pp. 27-36, 2022.

[2] F. Aditya, S. M. Nasution and A. Virgono, "Traffic Flow Prediction using SUMO Application with K-Nearest Neighbor (KNN) Method", *The International Journal of Integrated Engineering,* vol. 12, no. 7, pp. 98-103, 2020.

[3] Badan Pusat Statistik Kota Bandung, "Jumlah Penduduk (Jiwa), 2018-2020", 2020. [Online]. Available: https://bandungkota.bps.go.id/indicator/12/32/1/jumlah-penduduk.html. [Accessed 18 November 2024].

[4] Badan Pusat Statistik Kota Bandung, "Potensi Kendaraan Bermotor Per Jenis di Kota Bandung, 2020", 2021. [Online]. Available: https://bandungkota.bps.go.id/statictable/2021/03/05/1409/potensi-kendaraan-bermotor-per-jenis-di-kota-bandung-2020.html. [Accessed 18 November 2024].

[5] M. Akhtar and S. Moridpour, "A Review of Traffic Congestion Prediction Using Artificial Intelligence", *Journal of Advanced Transportation,* vol. 2021, no. 878011, pp. 1-18, 2021.

[6] S. A. M. Bagheri, B. Mojaradi, N. Kamboozia and M. Faizi, "Analyzing the effects of streetscape and land use on urban accidents and predicting future accidents by using machine learning algorithms (case study: Mashhad)", *Heliyon,* vol. 10, no. e33346, pp. 1-11, 2024.

[7] D. Xiao, H. Ding, N. N. Sze and N. Zheng, "Investigating built environment and traffic flow impact on crash frequency," *Accident Analysis & Prevention,* vol. 201, no. 107561 , pp. 1-12, 2024.

[8] T. Apriyono and D. P. Rumlus, "Analisis Faktor-Faktor Yang Mengakibatkan Kemacetan Lalu Lintas Pada Ruas Jalan Budi Utomo dan Jalan Hasannudin di Kota Timika", *Jurnal Kritis,* vol. 5, no. 2, pp. 96-114, 2021.

[9] D. Albalate and X. Fageda, "On the relationship between congestion and road safety in cities", *Transport Policy,* vol. 105, pp. 145-152, 2021.

[10] S. S. Gonzalez, F. Bedoya-Maya and A. Calatayud, "Understanding the Effect of Traffic Congestion on Accidents", *Sustainability,* vol. 13, no. 7500, pp. 1-19, 2021.

[11] Z. Zheng, Z. Wang, L. Zhu and H. Jiang, "Determinants of the Congestion Caused by a Traffic Accident in Urban Road", *Accident Analysis & Prevention,* vol. 136, no. 105327, pp. 1-9, 2020.

[12] W. Fazry, S. S. Prasetyowati and Y. Sibaroni, "Bandung City Traffic Classification Map with Machine Learning and Ordinary Kriging", *Jurnal Ilmiah Penelitian Pembelajaran Informatika (JIPI),* vol. 7, no. 4, pp. 1139-1148, 2022.

[13] A. M. Arabiat and M. Altayeeb, "Assessing the effectiveness of data mining tools in classifying and predicting road traffic congestion", *Indonesian Journal Electrical Engineering and Computer Science,* vol. 34, no. 2, pp. 1295-1303, 2024.

[14] T. Bokaba, W. Doorsamy and B. S. Paul, "A Comparative Study of Ensemble Models for Predicting Road Traffic Congestion", *Applied Sciences,* vol. 12, no. 3, p. 1337, 2021.

[15] E. Taiwo, G. O. Ogunsanwo and A. Olumuyiwa, "Traffic Congestion Prediction using Supervised Machine Learning Algorithms", *TASUED Journal of Pure and Applied Sciences,* vol. 2, no. 1, pp. 110-116, 2023.

[16] D. Wang, H. Chen, C. Li and E. Liu, "Exploring the Relationship between Land Use and Congestion", *Sustainability,* vol. 15, no. 9328, p. 1, 2023.

[17] Z. Bao, Y. Ou, S. Chen and T. Wang, "Land Use Impacts on Traffic Congestion Patterns: A Tale of a Northwestern Chinese City", *Land,* vol. 11, no. 2295, pp. 1-17, 2022.

[18] J. Y. Yap, N. Omar and I. Ismail, "A Study of Traffic Congestion Influenced by the Pattern of Land Use", *IOP Conference Series: Earth and Environmental Science,* vol. 1022, no. 012035, pp. 1-6, 2022.

[19] M. Scandella, A. Ghosh, M. Bin and T. Parsini, "Traffic-light control in urban environment exploiting drivers' reaction to the expected red lights duration", *Transportation Research Part C,* vol. 145, no. 103910, pp. 1-24, 2022.

[20] F. A. Ramadhan, F. Mulyawati and R. Gunawan, "Evaluasi Kinerja Suatu Simpang Bersinyal", *Jurnal Transportasi,* vol. 22, no. 3, pp. 249-254, 3 December 2022.

[21] R. E. Essa, S. S. Prasetyowati and Y. Sibaroni, "Performance of ANN and RNN in Predicting the Classification of Covid-19 Diseases based on Time Series Data", *Jurnal Riset Komputer,* vol. 10, no. 1, pp. 82-90, 2023.

[22] E. Y. Boateng, J. Otoo and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest Neighbor, Support Vector Machine, Random Forest, and Neural Network: A Review", *Journal of Data Analysis and Information Processing*, vol. 8, pp. 341-357, 2020.

[23] G. Meena, D. Sharma and M. Mahrishi, "Traffic Prediction for Intelligent Transportation System using Machine Learning", in 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), 2020.