POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

An Intelligent Network Traffic Prediction Model Based on Ensemble Learning for Vehicular Ad-hoc Networks

PARVIN AHMADI DOVAL AMIRI
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Août 2023

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

An Intelligent Network Traffic Prediction Model Based on Ensemble Learning for Vehicular Ad-hoc Networks

présentée par **Parvin AHMADI DOVAL AMIRI**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Soumaya CHERKAOUI, présidente
Samuel PIERRE, membre et directeur de recherche
Alejandro QUINTERO, membre
Wessam AJIB, membre externe

DEDICATION

To my mother Nasrin,
for all of her infinitive support, encouragement and self-sacrifices . . .

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my research director Professor Samuel Pierre, I could not have undertaken this Ph.D. study without his endless support, insightful suggestion and comments and constant support. I am grateful for providing me with the opportunity to conduct my research under your supervision in Mobile Computing and Network Research Laboratory (LARIM).

I would also like to especially thank Dre. Franjieh El Khoury for making me be in progress continuously. Thank you for guiding me with patience during my Ph.D. study.

I am grateful to Professor Alejandro Quintero, Professor Soumaya Cherkaoui, and Professor Wessam Ajib for kindly accepting to participate in this jury.

Words can not express my gratitude to my best friend, Frnoush, for her helpful comments throughout my Ph.D. research. She always challenged me to give a second thought to my ideas and solutions for my research.

I would like to extend my thanks to my colleagues and friends at LARIM, Eric, Amir, Sepehr, Lamia, Nasrin, Sanaz, Olson, Claudy, Jean, Loïc, and Marc for their helpful comments on my research and for making the past four years of my life much more enjoyable and memorable. I am grateful for being in a friendly environment where all of us help each other to grow. It has been an amazing experience working at the LARIM laboratory.

From the bottom of my heart, I am extremely grateful to my parents Emomgholi and Nasrin for their endless support, sacrifices, and guidance to follow the best future and be the best version of myself. This endeavour would not have been possible without my dearest brothers Fardin and Mohammad, despite the long distance between us, they provide me infinite support and peace of mind for me in all of the difficult situations in my life. Thank you, brothers, for being for helping me every step of the way in my future.

Last but not least, I am grateful to my dear husband Mehndi, for his understanding, support and encouragement during my Ph.D. study. Thank you for always believing in me.

RÉSUMÉ

Les réseaux véhiculaires Ad Hoc « Vehicular Ad-hoc networks » (VANET) constituent une partie importante du système de transport intelligent (STI). Les VANET ont le potentiel d'améliorer la sécurité routière et la gestion du trafic en fournissant des applications de sécurité ou autre. Cependant, l'évolution des VANET vers l'Internet des Véhicules (IoV) pose certains défis dans les services VANET. Les applications VANET utilisent des communications dédiées à courte portée (DSRC) pour la communication entre les véhicules et les unités routières, connues sous le nom de communications basiques, comme les communications véhicule-à-véhicule (V2V) et véhicule-à-unité-routière (V2R). De plus, le développement de VANET vers IoV apporte de nouvelles exigences telles que véhicule-à-tout (V2X), soit une variété de modes de communications entre les nœuds du réseau comme le véhicule-à-véhicule (V2V), véhicule-à-infrastructure (V2I), infrastructure-à-infrastructure (I2I) et de véhicule-à-piéton (V2P). Le défi principal des services VANET est lié à la large quantité de données générées par les usagers, entraînant un trafic sur le réseau et, par conséquent, la réduction de la qualité de service (QoS) pour les services VANET. Dans ce cas, lorsqu'il s'agit d'applications de sécurité, cela peut coûter une vie humaine.

Les techniques d'Intelligence Artificielle (IA) sont des solutions prometteuses pour résoudre les problèmes de trafic réseau dans les VANET. À cette fin, la prévision du trafic réseau est une tâche difficile qui peut aider les opérateurs de réseau à éviter les congestions et la réduction de la qualité de service dans les services de réseau véhiculaire. Ce domaine a attiré de nombreux chercheurs pour étudier la conception d'une solution d'IA pour les approches de prédiction du trafic réseau. Cependant, concevoir une méthode d'IA optimale capable d'obtenir une prédiction plus précise et stable reste un défi. Étant donné que chaque technique d'IA a ses propres limites et problèmes, si nous ne pouvons pas compter sur un seul modèle ML, cela imposera plus de problèmes. De plus, ces algorithmes doivent s'adapter aux nouvelles exigences de VANET comme la communication V2X. Par conséquent, il reste encore beaucoup de recherche inaboutie dans ce domaine.

Dans cette thèse, la première méthode proposée pour la prédiction du trafic dans VANET considère un apprentissage automatique (ML) basé sur les ensembles comme une sous-catégorie de méthodes d'IA. Ainsi, un modèle doté de l'IA peut obtenir de meilleures performances et une prédiction plus précise et plus stable qu'un modèle ML unique. Le problème de prédiction est défini comme un problème de classification. Un ensemble de données VANET réelles est utilisé comme entrée pour le modèle Ensemble Learning (EL). De plus, en ce qui concerne

l'importance de la qualité des données d'entrée, des méthodes de sélection de caractéristiques, notamment Boruta et LightGBM, sont utilisées pour extraire les attributs les plus utiles du V2V et du V2R en tant qu'ensemble de données fusionnées. La méthode proposée nommée STK-EBM est basée sur la stratégie d'empilement d'apprentissage d'ensembles qui comprend deux couches : la couche de base et la méta-couche. Dans la couche de base, l'algorithme Random Forest (RF), les modèles K-Nearest Neighbor (KNN) et XGBoost sont intégrés et sélectionnés en fonction de leur efficacité pour notre cible qui est la prévision du trafic dans le réseau. Les résultats de prédiction des modèles de couche de base sont agrégés par une régression logistique (LR) optimisée. L'analyse comparative des résultats de modèles ML uniques bien connus est effectuée dans le but d'indiquer pleinement l'avantage du modèle proposé qui apporte précision et stabilité de mode dans les résultats d'évaluation.

En ce qui concerne le VANET avec communication V2X, le modèle d'ensemble de vote souple est proposé dans notre deuxième modèle. Données simulées extraites des simulateurs Simulation of Urban Mobility (SUMO) et OMNet++. Les classes de trafic et de non-trafic sont définies à l'aide du Packer Delivery Ratio (PDR) qui est une métrique importante dans la prévision du trafic étudiée.

Il convient de noter que la réalisation de la meilleure stratégie appropriée pour intégrer les modèles d'IA et fournir des résultats de performance améliorés et équilibrés est un défi. Étant donné que la quantité de données collectées est presque doublée dans le deuxième modèle par rapport à la première méthode proposée, nous devons appliquer une stratégie simple qui fournit non seulement une performance améliorée par rapport aux modèles ML simples, mais fournit également une méthode qui n'augmente pas la complexité du modèle qui a entraîné plus de ressources de calcul et de consommation de temps. Le résultat montre que le modèle proposé est plus précis et stable avec moins de temps d'exécution que le modèle individuel. Cependant, les stratégies EL ont leurs propres coûts et avantages. Dans notre cas, la méthode STK-EBM atteint une meilleure stabilité que le deuxième modèle proposé au prix d'apporter plus de complexité.

Dans le troisième modèle de prédiction du trafic réseau, nous considérons différents problèmes pour obtenir un modèle EL plus généralisable, simple et précis. Les données VANET du monde réel avec communication de base et les données simulées de communication avancée sont toutes deux prises en compte dans la troisième méthode. Les données V2X extraites de l'architecture conçue sont l'intégration des technologies DSRC et cellulaires pour répondre à la fois à la couverture de communication à courte et longue portée. De plus, les ensembles de données considérés sont différents en taille et en fonctionnalités. De cette façon, la méthode peut montrer si elle est plus bénéfique et applicable pour la mise en œuvre réelle des applica-

tions VANET. Le réseau de neurones artificiels (ANN) est le modèle choisi pour la prévision du trafic dans la littérature. Cependant, les limites et les défis du modèle peuvent être résolus par un ensemble d'ANN et de Swarm Intelligence (SI). De cette manière, un modèle de prédiction intelligent efficace pouvant être appliqué aux données de communication de base et avancées est proposé. Les résultats obtenus indiquent que la méthode eSwaNN-NTP peut atteindre plus de précision, de stabilité et moins de temps que les simples ANN et DNN dans les deux ensembles de données. Enfin, les modèles proposés appliqués pourraient améliorer les performances du réseau de manière efficace.

ABSTRACT

Vehicular Ad-hoc networks (VANETs) consider an important part of the Intelligent Transportation System (ITS). VANETs have the potential to improve road safety and traffic management by providing safety and non-safety applications. However, the evolution of VANETs towards the Internet of Vehicle (IoV), bring some challenges in VANET services. VANET applications employ Dedicated Short-Range Communications (DSRC) for communication between vehicles and roadside units which are known as basic communications including Vehicle-to-Vehicle(V2V) and Vehicle-to-Roadside Unit (V2R). Moreover, the development of VANET to IoV brings new requirements such as Vehicle-to-everything (V2X), which means a variety type of communications among road entities consisting of Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Infrastructure-to-Infrastructure (I2I)and Vehicle-to-Pedestrian (V2P). The important challenge in VANET services is related to the enormous data generated by vehicular road users and resulting in traffic in the network and in turn the reduction of Quality of Service (QoS) for VANET services. In this case, when it comes to safety applications it may cost human life.

Artificial Intelligent (AI) techniques are promising solutions to address network traffic issues in VANETs. Toward this end, network traffic prediction is a challenging task that can help network operators to avoid traffic in the network and reduction of QoS in vehicular network services. This domain has attracted many researchers to investigate providing an AI solution for network traffic prediction approaches. However, finding an optimal AI method that can achieve more accurate and stable prediction still is a challenge. Since each AI techniques have its own limitations and problems, if we cannot rely on only one ML model will impose more problems, and it needed to adapt to the new requirements of VANET which is V2X communication as well. Therefore, there are a lot of rooms that still need to be considered in this domain. In this dissertation, the first proposed method for traffic prediction in VANET consider an ensemble-based Machine Learning (ML) as a subset of AI methods. In this way, AI empowered model can achieve better performance, and more accurate and stable prediction than a single ML model. The prediction problem is defined as a classification problem. A real-world VANET dataset is used as input for the Ensemble Learning (EL) model. Moreover, regarding the importance of the quality of input data, feature selection methods including Boruta and LightGBM are employed to extract the most effective attributes of the V2V and V2R as the merged dataset. The proposed method named (STK-EBM) is based on the stacking strategy of ensemble learning that includes two layers: the base layer and the meta layer. In the base layer Random Forest(RF), K-Nearest Neighbor

(KNN) and XGBoost models are integrated and are selected based on their effectiveness for our target which is traffic prediction in the network. The prediction results from the base layer models are aggregated by an optimized Logistic Regression (LR). The comparative analysis of the results of well-known single ML models is performed with the aim of full indication of the advantage of the proposed model that brings mode accuracy and stability in the evaluation results.

Regarding the VANET with V2X communication, the soft voting ensemble model is proposed in our second model. Simulated data extracted from Simulation of Urban Mobility (SUMO) and OMNet++ simulators. The traffic and non-traffic class are defined using Packer Delivery Ratio (PDR) which is an important metric in traffic prediction studied.

It should be noted that the realization of the best appropriate strategy to integrate AI models and provide improved and balanced performance results is a challenge. Since the amount of collected data is almost doubled in the second model compared to the first proposed method, we need to apply a simple strategy that not only provides an enhanced performance than single ML models but also provides a method that does not increase the complexity of the model that resulted in more computation resources and time consumption. The result shows that the proposed model is more accurate, and stable with less execution time than the individual model. However, EL strategies bring their own cost and benefits. In our case, the STK-EBM method achieves better stability than the second proposed model at the cost of bringing more complexity.

In the third network traffic prediction model, we consider different problems to achieve a more generalizable, simple, and accurate EL model. Real-world VANET data with basic communication and Simulated data from advanced communication are both considered in the third method. The V2X data extracted from the designed architecture is the integration of the DSRC and Cellular-based technologies to address both short-range and long-range communication coverage. In addition, the considered datasets are different in size and features. In this way, the method can show if it is more beneficial and applicable for real-world implementation of VANET applications. The Artificial Neural Network (ANN) is the well-chosen model for traffic prediction in the literature. However, the limitation and challenges of the model can be addressed by an ensemble of ANN and Swarm Intelligence (SI). In this way, an efficient intelligent prediction model that can be applied to both basic and advanced communication data is proposed. The obtained results indicate that the eSwaNN-NTP method can achieve more accuracy, stability, and less time consumption than simple ANN and DNN in both datasets. Finally, the applied proposed models could improve the network performance in an efficient way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

ACK	Acknowledgement
ACO	Ant Colony Optimization
AI	Artificial Intelligence
ANN	Artificial Neural Network
AU	Application Unite
AUC	Area Under Curve
CCH	Control Channel
CH	Cluster Head
CNN	Convolutional Neural Network
DAE	Deep autoencoder
DDR	Data Delivery Ratio
DNN	Deep Neural Network
DSD	Stochastic Diffusion Search
DSRC	Dedicated Short-Range Communication
DT	Decision Tree
EL	Ensemble Learning
FANET	Flying ad Hoc Network
FN	False Negative
FN	False Negative
FP	False Positive
FP	False Positive
GBM	Gradient Boosting Machine
GNS	Ground Station
GRU	Gated Recurrent Units
GVN	Green Vehicular Network
I2I	Infrastructure-to Infrastructure
IoV	Internet of Vehicle
ITS	Intelligent Transportation System
KNN	K Nearest Neighbor
LR	Logistic Regression
LSTM	Long-Short Term Memory
LTE	Long-Term Evolution

MAE	Mean Absolute Error
MANET	Mobile ad-hoc Network
ML	Machine Learning
MLP	Multi-Layered Perceptron
NB	Nive Bayse
OBU	Onboard Unit
OSM	OpenStreetMap
PDR	Packet Delay Ratio
PSO	Particle Swarm Optimization
QoS	Quality of Service
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RSU	Roadside Unite
SAE	Stacked Auto Encoder
SCH	Service Channel
SI	Swarm Intelligence
SL	Side Link
SUMO	Simulation of Urban Mobility
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UAVs	Unmanned Aerial Vehicles
V2B	Vehicle-to- Barrier
V2C	Vehicle-to-Cloud
V2CN	Vehicle-to- Cellular network infrastructure
V2I	Vehicle-to-Infrastructure
V2P	Vehicle-to-Pedestrian
V2PD	Vehicle-to-Personal Device communication
V2S	Vehicle-to- Sensors
V2U	Vehicle-to- UAV
V2V	Vehicle-to-Vehicle
VANET	Vehicular Ad hoc Network

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Intelligent Transportation Systems (ITS) make a significant contribution to the future of transportation systems in smart cities. Vehicular ad-hoc Networks(VANETs) are an essential part of ITS which can provide road safety and traffic management services to road users [3]. VANET takes advantage of the data produced by connected vehicles communicating with each other and other entities on the road to raise the drivers' awareness about the probability of an accident or inform them about road incidents and traffic jams in advance. However, maintaining the performance of the network while the vehicular network is experiencing enormous data generation by continuously increasing vehicular road users, is a not neglectable concern for researchers. Traffic prediction can be helpful in detecting the problem in the network before it causes a reduction in Quality of Service (QoS). Although Artificial Intelligence (AI) solutions can be applied to predict network traffic, still finding an optimal AI method in dynamic topology and the evolution of VANET toward IoV is a challenging task. Integration of VANET and efficient AI solution still needs lots of consideration [4].

This chapter is organized as follows. In Section 1, we introduce some basic concepts about VANETs communication, applications, and AI methods. We discuss open problems in Section 2. In Section 3, we explain the research questions and objectives. Then, the global methodology is discussed in Section 4 and research contributions are illustrated in Section 5. Finally, the outline of this dissertation is indicated in Section 6.

1.1 Definitions and Basic Concepts

Vehicular ad-hoc Network (VANET) is an important part of the Intelligent Transportation System(ITS) which can enhance the traffic-related problems on the road through various services and applications [3]. In the following sections, we will explain the basic concept related to VANET communication, VANET application and VANET integrated with Artificial Intelligence (AI) for traffic-related problems in VANETs.

1.1.1 VANET data transmission and communication

VANET consists of a set of mobile nodes which are moving vehicles on the roads, and fixed nodes which are equipment installed alongside the roads [5]. VANET include different type of communication, the basic communication including two main types of wireless communication between vehicles and roadside unit, known as Vehicle-to-Vehicle (V2V) and

Vehicle-to-Infrastructure (V2I) [3]. Data transmission in this basic communication is based on Dedicated Short-Range Communication (DSRC) which is based on IEEE 802.11 standard. DSRC is a common technology for sharing information between vehicles. However, it cannot be applied for large-scale vehicular communication and enormous data transmission [6].

On the other hand, with development of VANET and the emergence of Internet of Vehicles (IoV), impose new requirements and challenges including advanced communication known as Vehicle-to-everything (V2X) which is a promising solution for traffic problems [7]. V2X communication commonly consist of Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Infrastructure-to Infrastructure (I2I), and Vehicle-to-Pedestrian (V2P) [8]. Cellular-based technologies are another main consideration that can be used for data transmission in VANET communication in the case of long-range coverage and a large amount of data transmission in V2X communications [9], [10].

VANET enable connected vehicles to communicate together and roadside unit through the abovementioned basic and advanced wireless communication to exchange information such as location, speed, direction, road hazards, road traffic, and accident. In this way, it can provide emergency and warning messages to inform drivers about the expected incident on the road [5]. This field has attracted researchers to develop VANET applications for these kinds of services. As Fig.3.1 and Fig 1.2 show, the architecture of VANET with basic and advanced (V2X) communication.

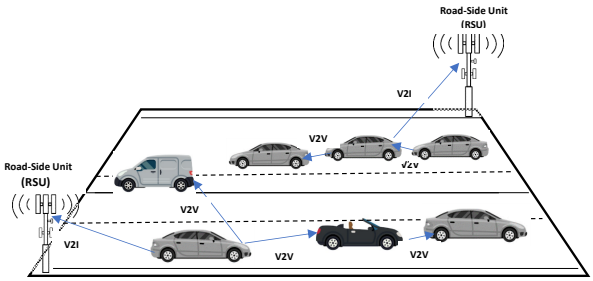

Figure 1.1 VANET basic and advanced (V2X) architecture

1.1.2 VANET applications

VANET applications can be categorized into two main classes including safety and non-safety applications. However, this application especially the safety application opened challenges regarding the performance of the network in the case of increasing data delivery through vehicular communication which causes network traffic [3], [11]- [12]. AI methods consider a potential solution that can employ an intelligent traffic prediction model to avoid a reduction

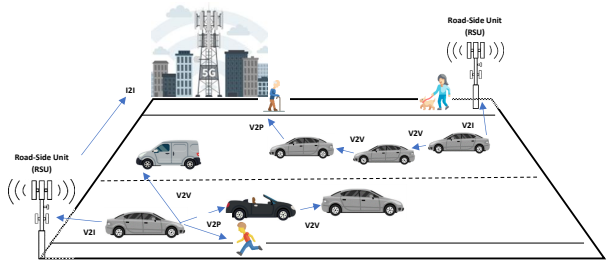

Figure 1.2 VANET basic and advanced (V2X) architecture

in Quality of Service (QoS) and a negative impact on network performance by detecting the traffic in the network before it causes failure or downgrade the provided services for road users' application [4].

1.1.3 VANET and AI

In this section, we explain different well-known Artificial intelligence (AI) methods. AI-techniques that combined with VANET to solve different issues can be categorized into Machine learning (ML), Deep learning(DL) and Swarm intelligence (SI) methods which are discussed in the following paragraphs.

ML methods can be divided into supervised, unsupervised, semi-supervised and reinforcement learning [13]. In this study, we consider supervised learning methods that can be matched with the traffic data which means it employs labeled data to train AI techniques to predict network traffic in VANET. Furthermore, supervised learning can be adapted to classifications and regression tasks [14]. There are different popular ML, DL and SI models that applied for the purpose of traffic prediction as explained in the next subsection.

1.1.4 AI methods for traffic prediction

Machine learning methods

Random Forest (RF) algorithm train decision trees on samples in parallel, then combine the output of the independent decision trees to gain a single result. It has key benefits including fast training, flexibility, generalizable outcome, managing overfitting problems and working with few input features. Therefore, this technique is a great solution for congestion prediction, control channel capacity and handover issues in multimedia data distribution [15].

A Decision Tree (DT) is a simple and limpid classification and prediction method. It works based on repeated data division into subsets that are selected regarding the most important

features of each node of the tree. This algorithm can predict the target by learning from decision rules of obtained features [4].

K-Nearest Neighbor (KNN) algorithm employs the whole dataset for the prediction model. This supervised learning method looking for k nearest points in the training data and employs their classes to predict the class of new data points. It uses the votes for the most repeated label in a classification problem [7], [16]- [17]. The highlighted benefit of the algorithm is related to its efficient classification based on votes [7].

Nive Bayse (NB) algorithm works based on classification of a set of observations of Baye's rules that are specified by itself. This method has significant advantages including robustness due to ignorance or elimination of unrelated features, simple implementation, and effortless training demand [7], [18].

Support Vector Machine (SVM) algorithm capable of learning from the past input data to make the prediction results. Therefore, it is best fitted to non-linear and dynamic traffic data [7], [4], [19], [20]. In addition, it can deal with many attributes but a low amount of input variable samples. Furthermore, this model consumes much more time for training [7].

Deep learning methods

DL algorithms as a subset of AI are more beneficial in the case of complex and big data [7], [21]. When it comes to traffic prediction, Artificial Neural Networks (ANN) as simple DL methods are a promising approach to enhance the accuracy [22]. The most popular NN model that can adapt better than other NN models with network characteristics and traffic data is MLP [23], [24]. The ANN algorithms include three layers: the input layer, the output layer and the hidden layer(s). The hidden layer can be one to many layers. ANN with more than one hidden layer(s) is known as Deep Neural Network (DNN). NNs bring their issues and challenges. The performance of NN-based techniques relies on proper network architecture determination, and they have a slow rate of learning that need to be improved by other approaches [25]. Fig.1.3 shows the architecture of Multi-Layered Perceptron (MLP).

There are other DL methods including Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU) that are modified Recurrent Neural Network (RNN) models to cover its issues and can be applied in the case of big data or video and image datasets [11].

Swarm intelligence methods

Swarm Intelligence (SI) as nature-inspired algorithms capable of performing autonomous and distributed approaches [7]. SI can be divided into four classes (i.e., biology-based, human

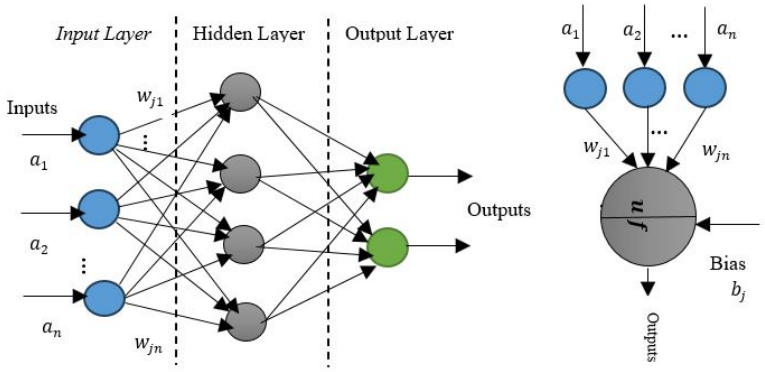

Figure 1.3 The general structure of a Multi-Layered Perceptron (MLP).

behaviour-based, evolution-based and physis-based) and each of mentioned classes have different SI algorithms [26]. When it comes to vehicular networks, SI defines as a population of vehicle entities that exchange information through their communication with other vehicles and entities on the road. There is no central controller, the vehicle follows basic regulations on the road including road structure, path, speed limit and traffic lights or road signs [7]. SI can be applied as a potential solution because of its powerful capability in modeling a population of agents with self-organization and in a cooperative manner [26]. Among all categories of SI methods, biology-based classes such as Particle Swarm Optimization (PSO) consider the trendy solution for prediction tasks [7]. It has the ability to assure each particle chooses one of the best positions that are experienced toward moving to new positions by applying the new best speed of the particles [27], the various possible solution by employing an initial speed as an input are plotted in this solution apace and the fitness value for particles moving over the space will be estimated by fitness value and gradually reached the optimal position toward those areas that show better fitness value, and finally the global optimum position will be met [7], [27], [28]. Different SI methods can be helpful to solve VANET challenges. However, SI being its own issues and challenges. PSO is determined by a variety of entities including vehicle velocity, vehicle position and the message produced by vehicles [29]. In addition, unreliable interference has an impact on the PSO algorithm due to its security issues [30]. The solution is an integration of PSO with other metaheuristic algorithms that is complex due to the extremely dynamic nature of VANET, or other SI algorithms such as Ant Colony Optimization (ACO) that also has problem-related to consuming much time for processing [31].

The important limitation of the PSO technique on VNAET is related to its need to cooperate with lots of vehicles and it leads to security issues in the vehicle communications and

vehicle verification problem [7]. However, the PSO technique is more beneficial than other SI methods such as Stochastic Diffusion Search (DSD) which require to be combined with other metaheuristic methods and is difficult due to the dynamic topology of VANET.

Ensemble Learning methods

Since each of the abovementioned AI-based techniques has its own advantages and disadvantages. Ensemble Learning (EL) techniques can help us to combine various ML/DL/SI methods together with proper integration strategy to provide a stable and strong, model that can address standalone AI-techniques problems [32], [33]. EL became a powerful method that has been applied in different research areas including health, finance and energy [34], [35].

EL can enhance the accuracy, stability robustness and generalization ability of individual AI methods. Therefore, it can be beneficial in the vehicular networks to provide a model with overall better performance in terms of accuracy, stability and time consumption especially in the case of we cannot rely on employing standalone AI methods due to their limitation and challenges

As Fig.1.4 depicted the AI methods integrated with the VANET application for traffic prediction.

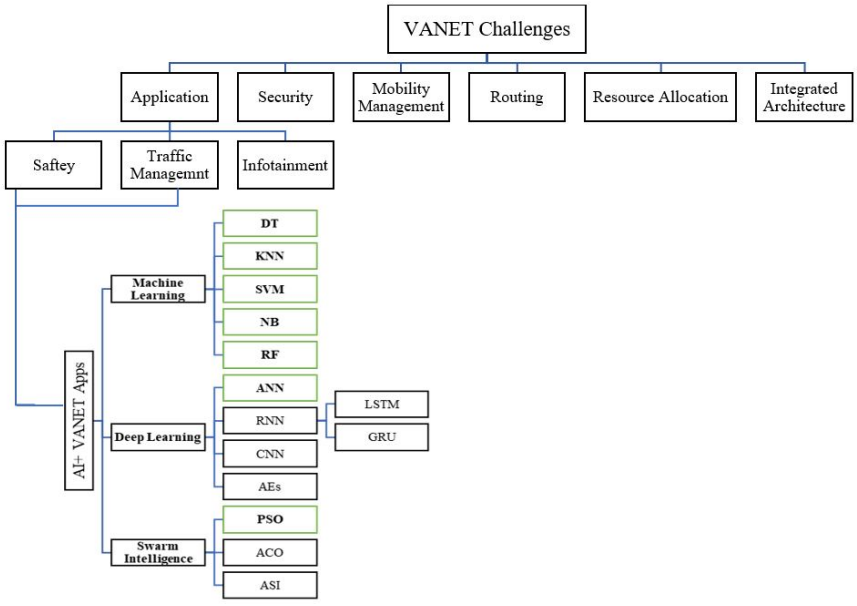

Figure 1.4 AI techniques in VANET applications

1.1.5 Traffic prediction in VANET

There are six specific domains that we can take advantage of AI solutions in VANET including application, routing, security, resource allocation and access technologies, mobility management and integrated architectures [7]. The deployment of AI techniques in each of these domains brings different issues and challenges. In this study, we focus on the utilization of AI in VANET applications. Applications in VANETs can be divided into three main classes including safety applications, traffic management applications and infotainment applications [7]. In the case of safety applications that can be used for broadcasting emergency messages and warning alerts to avoid accidents or reduce post-accident traffic congestion by using AI techniques for accident and traffic prediction or driver behavior prediction and motion detection. In the case of ML solutions for safety applications, RF is applicable to accident prediction, NB and DT [36] can be beneficial for reducing post-accident road congestion. In addition. Traffic management applications can address issues that cause problems on the road such as lane change, speed and traffic congestion. KNN [37] can predict vehicular speed and density. ANN [38] and CNN [39] can be used for traffic prediction to enhance road congestion. GRU [40] for traffic volume prediction, enhance traffic congestion with SI [41] and LSTM [42]. Regarding infotainment applications, there are problems such as multimedia propagation, video quality, and QoS support for real-time applications. In the case of QoS-aware video compression and multimedia data encoding, some DL methods including CNN [43], [43] can be applicable to solving challenges.

As noted in [21] and [44], applying complex and DL methods cannot be efficient in the case of small to medium size datasets. However, for large and complex data or image and video streaming processing, DL models bring great solutions than ML techniques. On the other hand, as noted in [2] and [45], EL techniques can be a great potential solution to generate a strong model by addressing the limitation and challenges related to individual AI models. However, identifying the best appropriate strategy to integrate different ML, DL and SI techniques together and achieve an efficient method with better performance is a challenging task that requires lots of consideration and a deep understanding of the advantages and disadvantages of each AI-based method and the proper strategy to combine them to reach the high potential effectiveness of single methods in an efficient way that not impose any new problem or computational complexity and much time processing to the final model. In this section, we summarize well-known AI-based techniques for traffic prediction in VANET applications, especially safety and traffic management applications and explain the limitation and strengths of each method.

1.2 Open Problems

In VANETs, the data related to each vehicle such as location, speed, direction and road conditions will be exchanged through their communications. Therefore, increasing the number of vehicles on the road cause increasing the number of sending and receiving packets through VANET communication which resulted in traffic in the vehicular network and its negative consequences on VANET safety applications [4]. There are several challenges in VANET services such as maintaining the performance of the network and providing an acceptable level of QoS for road users that are used VANET applications. Although traffic prediction can avoid network traffic, providing a highly accurate, stable and reliable prediction model is a challenging task.

AI techniques are a great potential solution for improving data-driven approaches and integration of AI and VANET, bring lots of opportunities and challenges and still lots of room for consideration [8]. Prediction of a network failure before it causes any problem such as delay or decline in provided services can optimize the entire driver-vehicle-road performance [46]. AI can help to predict network traffic that will cause the problem in the operation of the network [4]. ML, DL and SI are three major AI techniques for traffic prediction in VANET [7]. ML as a main part of AI, can be divided into three main learning models including supervised learning, unsupervised learning, and reinforcement learning [14]. Supervised learning is considered as an error correction method, that will be developed and completed through training set and experience despite unsupervised learning that learns and solves the problem without having error metrics to evaluate the solution. Therefore, supervised learning can be adapted to a non-linear dynamic problem such as prediction [47]. Moreover, the development of VANETs toward IoV, bring new requirement such as V2X communication [7]. Therefore, despite the different AI methods proposed by researchers for traffic prediction, the integration of AI and VANET still has lots of room for consideration and particularly when it comes to VNAET with V2X communication is a challenging task [7].

VNAET/V2X services have problems with the adoption of AI including the limitation of each access technology including DSRC and Cellular based technologies. DSRC has short-range coverage limitations and service degradation problems in traffic scenarios. On the other hand, Cellular-based challenges are related to its dependency on infrastructure, the higher price of network usage and end-to-end latency.

Accordingly, we require to optimize the network performance by considering basic and advanced communication in VANET with taking advantage of AI, which can be beneficial for traffic prediction. However, Each AI model brings its own limitations and problems and we

need to achieve an efficient prediction model that can address standalone AI methods challenges. Therefore, considering VANET adoption to AI techniques with basic and advanced types pf communications, and proposing intelligent traffic prediction models in an efficient way that can provide higher accuracy, and stability in evaluation performances is considered in this dissertation.

1.2.1 Research Questions

Considering the abovementioned open challenges, the following main research question may arise: Regarding the high potential capacity of ML methods to make an intelligent prediction method, and the advantages of applying DSRC and Cellular- based technologies in AVNET with V2X communication, how can we deal with network traffic problems and its adaption with AI technique and VANET with different communication in an efficient way?

More specifically:

- Regarding the highly dynamic nature of vehicular networks, how we can propose an algorithm which is able to predict traffic accurately using an AI-based network traffic prediction model?
- Considering the advantages of using V2X communication in traffic-related applications, how can we design an architecture, which intelligently and reliably predicts the network in VANET?
- Regarding the efficiency of the ensemble learning model, which integration strategy can help us to predict network traffic with more accuracy, stability, and reliability than standalone AI models?

1.3 Research Objectives

The main objective of this dissertation is to propose an efficient AI-based network traffic prediction methods that can be employed in VANET considering different types of communication and technologies. The three following sub-objectives are considered towards achieving the main objective of this dissertation:

1. Proposing an Ensemble Model for Intelligent Traffic Prediction in VANET using stacking ensemble strategy while considering basic communication that is a combined V2V and V2I data and extracting the most effective features using ensemble feature selection algorithm;

2. Proposing an Intelligent Network Traffic Prediction Model, using Soft Voting Ensemble Strategy Combination of Different ML Models to Address each Model Limitation and Provide Overall Better Performance Model in VANET/V2X;

3. Proposing an Ensemble Traffic Prediction in the Network, using Artificial Neural Networks (ANN) integrated with Swarm Intelligence(SI) and Designing a VANET with V2X Architecture that is more Applicable for Real-World implementation;

4. Evaluating the performance of the proposed methods.

1.4 Global Research Methodology

VANET applications impose some service requirement such as network performance that is highly important specifically for safety application that relates to human life [46]. When traffic occurs in the network, it causes a significant reduction in QoS and failure in the network. ML as a subset of AI can optimize the operation of the network [8]. Making an intelligent prediction of traffic in the network can help us to precisely identify the failure of the network and avoid service degradation, especially for services that are highly dependent on the performance of the network. However, each ML model brings its limitations and drawbacks for traffic prediction. In this case, ensemble learning can significantly improve the performance of machine learning in most problems [34]. Ensemble learning can achieve better performance with enhancing stability and generalization ability besides higher prediction accuracy and fast computation than a single ML [32]. We need to identify the best proper strategy to integrate ML models through EL. This motivated us to use EL in an efficient way for network traffic prediction to keep the QoS and performance of the network in the VANET application.

In this regard, VANET in basic architecture consists of V2V and V2I communication, However, toward the evolution of VANET to IoV, the new requirement such as V2X communication opened. Although it can enhance traffic issues in VNAET, it brings its own challenges such as using different access technologies in vehicular communication.

Regarding the main objective of this dissertation, an efficient AI-based solution for VANET with a basic and advanced type of communication is proposed with the aim of network traffic prediction to help network management and void service degradation for VANET applications. Network traffic prediction can be considered a regression and classification problem. We consider our problem as a classification task based on the features that we have in the datasets.

In the first objective of this dissertation, basic communication data in combined V2V and V2I datasets are considered for network traffic prediction based on the real-world VANET dataset. Based on the explanation in [8], [4]. when the number of sending and receiving packets through VANET communications increases (i.e., lots of vehicular users on the road), traffic occurs in the network. Therefore, for the classification task, we consider packet receiving as a network parameter to predict the network traffic. Then we labeled our data into two classes, class 1 for a case that packet is received, and we assume this is a non-traffic situation; and class 0 for not receiving a packet that is assumed as traffic in the network. Since, the input data in AI- models are highly important, the optimal and important features extracted from the VANET dataset by using LightGBM as an ensemble feature selection method [48]. The purpose of ensemble learning is to combine several simple models into a single strong learner to reduce errors or improve forecasting results. Ensemble learning can achieve higher accuracy and robustness, as well as a better generalization ability than a single model in most problems [22], [34]. The important issue is about selecting the best combination of popular ML algorithms when creating our EL model from scratch. This is one of the contributions of our research. In this context, we did an experimental analysis of the most mentioned and popular ML models for traffic prediction including RF, KNN, NB, DT, SVM, and MLP. These models are individually trained and the performance of each of them was evaluated. We find the best combination of these learners as the base learners by considering performance effectiveness. Base learners in our approach are diverse due to the point that each single ML model has a different view about solving the prediction task and bring its advantage and disadvantage for traffic prediction. Furthermore, we add a booster to the base learners to boost the prediction results. For this purpose, some ML algorithms, such as Xgboost [49], [50], can be employed. In these ways, we can bring more accuracy, adaptability, and stability to the dynamic nature of traffic.

In the first layer of the stacking ensemble of models, we selected RF, KNN and Xgboost. Each of them can help us to cover an issue. RF can solve the challenges related to scalability since it has the ability of training. models in parallel [51]. KNN was selected because it obtained better performance than the other ML models based on our dataset. Finally, Xgboost as an efficient and scalable implementation of the Gradient Boosting Machine (GBM) was selected to act as a booster to our base learners' results. Considering the distributed computing and parallel learning ability of this model, Xgboost will not impose extra time for the prediction result. However, it enables higher prediction accuracy and can make our model more precise. Moreover, this model can handle the challenge in DT, which is related to easily over-fitting. Eventually, Xgboost enhances the generalization ability [32], [49], [50].Therefore, we considered a model that can take advantage of all these points and use the best effective

combination of these models.

In the final step of the first objective, we aimed to simplify the interpretation of the base learner prediction results by using a simple meta-learner. Therefore, we employ an improved Logistic Regression (LR) as the meta-learner. It can find the optimal combination of the prediction results of all base learners. In other words, the meta-learner was trained based on the prediction made by previous learners. This helps improve the predictive performance of network traffic. We used grid search cross-validation [48], [52] to improve the accuracy of LR. It can tune the hyperparameter and result by using the best parameters of the algorithm.

To sum up, In the first place, we require to do an experimental analysis of the most popular ML models for traffic prediction including RF, KNN, NB, DT, SVM and MLP. In the second place, when we build an EL model from scratch the important issue is about selecting the best combination of ML algorithms. Therefore, in our EL approach which is composed of base learners and a meta learner, the base learners bring a different view about solving the prediction task. In this way, the model provides more accuracy, adaptability, and stability to the dynamic nature of traffic. The selected algorithms including RF, KNN and Xgboost can help us to cover an issue. RF can solve the challenges related to scalability, and KNN showed better performance results than the other ML models based on our dataset. Furthermore, Xgboost was selected as a booster to our base learners' performance. It also can enhance the generalization ability, but its parallel learning ability with distributed computation will not impose additional time. Finally, we aimed to simplify the interpretation of the base learner prediction results by using a simple meta-learner. Accordingly, we employ an improved Logistic Regression (LR) to find the optimal combination of the prediction results of all base learners. Eventually, we take advantage of the best effective combination of ML models to provide a stable prediction model with better overall performance in terms of accuracy, prediction error and execution time. We used Google Colab [53]], and Python programming for the implementation of AI models, we employ some libraries such as Scikit-Learn to scale and split input features into train and test sets. The parameter optimization which can help to enhance the performance of the model is performed by cross-validation as well [54], [55]. Finally, the performance of the proposed model is compared with the most common ML model in terms of classification evaluation metrics.

After considering VANET with basic communication, we need to apply V2X communication data which is more applicable in real-world applications. Therefore, V2X architecture is designed in the second and third objectives and data is generated using SUMO, OMNet++, INET, Veins, and SIMU5G simulators which are the most popular simulation tools for VANET. The simulated data is used packet delivery ratio to determine the traffic in

VANET with V2X communication because traffic is the reason for Packet Delay Ratio (PDR) and loss in the network. We labeled the target with PDR less than 0.3 as 1 which means traffic condition, and greater and equal to 0.3 as 0 which means a non-traffic condition in VANET.

In the second objective, we used simulation data to provide V2X communication considering V2V, V2R, R2R and V2P with DSRC technologies for their communication and we consider packet delivery ratio as a target to predict traffic in the network. We used SUMO and OMNet++ simulators. We applied an ensemble model from the integration of RF, KNN and DT to learn from training data and the soft voting strategy is applied to combine the result of the abovementioned algorithms and provide a final prediction. In this way, by using a simple integration strategy, we can balance the results of standalone algorithms without imposing more time in a parallel way. The same classification evaluation metrics and experimental environment like the first objective is considered to compare the result of the proposed model and four common ML models.

In the third objective, we design a VANET/V2X architecture including V2V, V2R, R2I and V2P, by integration of 5G as cellular-based technology and DSRC to address short- rang and long-range coverage in V2X communication. We provide an efficient data collection method by applying Cluster Head (CH) among the nearest vehicles and employ intra-cluster communication by DSRC as well. Then, we applied a combination of SI and ANN to cover the challenges of NN as a simple DL model and the most popular traffic prediction model and achieved an enhanced method with better performance. SI consider the best-matched method to integrate with NN, to improve the learning ability besides a more stable and accurate prediction model. In this objective, we consider not only standalone ANN and DNN with different hidden layers but also, we consider basic real-world VANET data to compare the proposed model effectively considering two different data sets in terms of the number of samples and the number of features.

To evaluate the performance of the proposed methods in VANET classification metrics including confusion matrix, classification report, ROC curve and CPU time. In classification metrics, accuracy represents the ratio of TP and TN overall number of samples. Sensitivity (recall) shows the ability of the classifier to identify all positive samples in the actual class. Precision indicates the accuracy of positive prediction. The F1 score is affected by precision and recall where the best score is 1.0 [56], [57]. We considered the Receiver Operating Characteristic (ROC) curve, which is a familiar tool to estimate the performance of binary classifiers and Area Under Curve (AUC) [58], to understand the stability of the model. Area Under Curve(AUC), For a predictor f, an unbiased estimator of its AUC: tests whether positives

are ranked higher than negatives. Finally, a comparative analysis of the proposed models in this dissertation and standalone ML/DL models are performed by employing the mentioned classification metrics.

1.5 Research Contributions

Regarding the realization of the full potential of AI techniques integrated with VANET, efficient intelligent network prediction methods are proposed in this desertion. Moreover, the contributions of this dissertation are as follows:

1. **Proposing an ensemble method to predict network traffic in VANET:** To realize the full potential of AI methods for traffic prediction, since the quality of input data is highly important before fed to the AI model and it can affect the performance of the model, we investigate on different feature selection techniques to realize the relevancy and the importance of the features with the target. Therefore, we employ the ensemble feature selection method to perceive the most efficient attributes based on combined datasets including V2V and V2I communication data in VANET. Then, we first select the most popular prediction model and make a comparative analysis among them based on the real-world VANET dataset including the fusion of V2V and V2I communication data. We employ a stacking ensemble learning strategy to integrate different ML models together in a two-layer structure to achieve a parallel computation and higher accuracy and stability in prediction results compared to individual ML models.

2. **Proposing a classification approach to predict network traffic in VANET with V2X communication:** Considering the evolution of VANET and its new requirement, which is V2X type of communication, which is more applicable in real-world VANET applications, we generate V2X dataset from the fusion of different communication in VANET by the cooperation of popular simulation tools for traffic modeling and wireless communication. Then, we employ a simple ensemble strategy based on a soft voting strategy to compare with standalone popular ML models. The ensemble model used a simple strategy; therefore, it will not bring more computation complexity while it capable of providing better overall performance in terms of time, accuracy, and stability than other considered individual models.

3. **Designing an efficient VANET with V2X communication and proposing NN-based ensemble model for traffic prediction in VANET:** WE designed VANET

with V2X architecture that collects data from various communication in an integrated architecture of DSRC and Cellular based technologies that can provide short-rang and long-range coverage. In addition, the data is collected in an efficient way by using grouping the nearest vehicles into a cluster and choosing the cluster head for communication among vehicles, roadside units, and pedestrians. Then, we apply both real-world datasets with V2V and V2X datasets which consider basic communication in VANET and the V2X dataset with double the size in the recording data and half in the input features than the real-world dataset. In this way, we can investigate the effect of the size of the dataset on ML and DL algorithms. Finally, the ensemble learning model based on the integration of the NN model and SI method is proposed to address the challenges related to ANN models which are mostly considered deep learning models for traffic prediction. Eventually, the performance of the proposed model was evaluated according to popular classification metrics and the obtained results were compared with standalone NN and DNN models considering two datasets.

1.6 Outline of Dissertation

This dissertation consists of seven chapters. Related works are reviewed in Chapter 2. In this chapter AI-based traffic prediction methods and their unsolved problems are presented and discussed.

In Chapter 3, an ensemble-based prediction method is proposed to predict the network state in the VANET. Indeed, the prediction model is considered a classification problem. A stacking ensemble strategy is applied to integrate different ML models in two layers and in parallel computation. Moreover, an ensemble feature selection technique is employed in which the best effective features based on the target can be applied.

In Chapter 4, the prediction of the network is performed considering a simple ensemble strategy of a combination of different ML models. A comparative analysis of the proposed model with the baseline ML model results shows the performance of the proposed model. In Chapter 5, a neural network method was combined with swarm intelligence and applied to predict the network traffic. Moreover, in Chapter 5, the V2X communication network concept in a highly dynamic environment of VANET is considered with the integrated architecture based on DSRC and 5G technology. a general discussion is presented about the proposed methods in this dissertation is presented in Chapter 6. Finally, in Chapter 7, the contributions, the limitations and future works are presented.

CHAPTER 2 LITERATURE REVIEW

Currently, the popularity of AI techniques can be seen in all research domains and VANET is not an exception. Although the integration of VANET and AI methods can provide an efficient solution to different VANET challenges such as VANET applications, still there are significant problems that need to be addressed [8]. In this chapter, we discuss the important challenges considering VANET architecture and provided applications and we explain the potential solution for these problems. More specifically, the VANET application and its requirement which is maintaining the performance of the network is discussed. This can be addressed by intelligent network traffic prediction models. However, the prediction methods need to be reliable, accurate and on time. Otherwise, it can affect vehicular users' life [11], [59]. In this matter, the limitation and advantages of AI techniques considering traffic prediction issues are discussed to pave the path toward the proposed methods in this dissertation.

2.1 Vehicular Ad-hoc Network Architecture, Applications, and Related Challenges

The communication-oriented architecture of VANET and intelligent traffic prediction models based on the data provided by these communications are the focus of this study. It can be applied to VANET applications. The number of vehicles will be reached 2 billion by 2035 [60], [1], increasing the number of vehicles on the road and consequently road congestion has become a serious issue that affects not only environmental domain such as air pollution but also human-related problems such as high fatality rate due to accident, wasting people time on traffic and fuel consumption and cost [61]. Although many actions have been taken to enhance road safety such as new traffic rules, 92 % of road accidents happen because of human-related issues including distraction, inadequate environment observation and awareness, poor decision-making, not maintaining a safe distance from other vehicles, not immediate reaction [60]. Furthermore, despite all the safety technologies and equipment inside the vehicles (e.g., seatbelts, airbags, and cameras), the recorded fatality rate of road accidents is not a neglectable issue [62].

The promising solution is to take advantage of technologies that can help to collect traffic data and share important information among drivers to avoid traffic issues [61]. Vehicular ad-hoc Networks (VANETs) is a subclass of Mobile Ad-hoc Network (MANET) and consider a significant element of Intelligent Transportation System (ITS) [1]. VANETs can take

advantage of wireless communication technologies to provide helpful applications for road users. In this way, VANETs can enhance road safety and manage traffic-related problems by broadcasting warning messages through safety applications to avoid accidents and congestion on the road [1], [61]. In the following subsections, we discuss the communication perspective of VANET architecture for traffic applications, its related challenges, and potential solutions.

2.1.1 VANET Basic Architecture and Related Problems

The basic architecture of VANETs includes two main types of communication known as Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I), where connected vehicles can communicate with each other and infrastructure [61]. The VANET basic communication has three primary elements including Roadside Unite (RSU), the Application Unite (AU) and Onboard Unit (OBU), which is placed in the vehicles to communicate with RSU and the other vehicles through V2V and V2I communication [1]. [61], [63].Moreover, IEEE 802.11p access technology with 5.9 GHz bandwidth and Dedicated Short-Range Communication (DSRC) are employed for V2V and V2I communication [61]. VNAET applications use the information collected from these communications. There are two basic categories of VANET applications including safety and non-safety services [61]. In the case of safety applications, the emergency messages need to be propagated without collision and corruption and this considers a major challenge that can affect the Quality of Service (QoS) for road users [5]. This means when traffic occurs on the road or after an accident on the road, VANET experiences a high density of nodes in the network that resulted in congestion of the channel and in turn the QoS for application of VANETs that is used by road users. Therefore, maintaining the QoS for safety applications in such a dense network is a critical issue that needs to be considered [61]. DSRC-based technology brings reliability and scalability challenges when it comes to large-scale congested vehicular networks [5]. This means vehicles employ DSRC- channel for data communication, in the case of network congestion which means all the vehicles require to transmit packets using the DSRC channel and which led to packet delay and downgrade of the throughput of the network and eventually, resulted in QoS degradation for road users [**?**]. Fig 2.1 shows that DSRC-based wireless technologies have seven channels, with 10MHZ including six service channels (SCH) and one control channel (CCH) that utilizes for safety and non-safety packet transmission and emergency messages respectively. Data communication among vehicles is through these channels and due to channel competition, the network may be congested that cause delays in packet transmitting, reduction in throughput and eventually degradation in QoS [64].

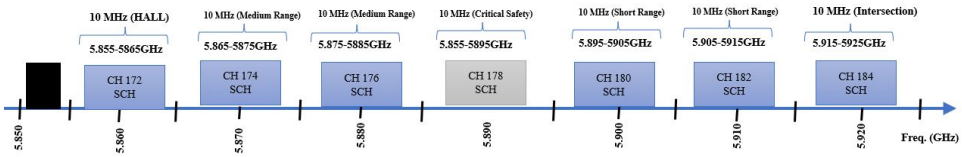

Figure 2.1 DSRC-based wireless technologies with seven channels [1].

2.1.2 VANET Advanced Architecture and Related Problems

VANETs include features such as high vehicle mobility, transportation infrastructure dependency, dynamic network changes and intermittent network connection that bring challenges in data delivery and communication reliability [5]. The evolution of VANET toward the Internet of Vehicles (IoV) leads to an increase in the success ratio of ITS services. However, there are important issues that need to be considered including data delivery delay, stable network performance, reliability, scalability, and flexibility. To achieve these requirements VANET needs to cooperate with other wireless technologies and network infrastructures such as cellular networks [1]. With the emergence of advanced technologies in both vehicles and wireless communication, the data can be exchanged not only between vehicles and RSUs but also among vehicles with pedestrians, bicycles, ground station (GNS) and Unmanned Aerial Vehicles (UAVs) which is called Vehicle- to-everything (V2X) [1]. Wireless communication provides different types of communication (V2X) that can be classified into Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Infrastructure -to- Infrastructure (I2I), Vehicle-to-Pedestrians (V2P), Vehicle-to- Barrier (V2B), Vehicle-to-Cloud (V2C), Vehicle-to- UAV (V2U), Vehicle-to- Sensors (V2S) [65]. In V2X, the data transmitted between vehicles and all road entities can address road-related issues such as accidents and traffic jams and offer a variety of information services as well [63]. Recently, researchers have been attracted to integrate VANET with other networks and infrastructures such as cellular networks [66], [67], WSN [1] and UAVs network [68], [69], [70], [71], [72] to deal with VANET issues such as reliable communication with low latency and enhance the delivery of data ratio with considering VANET cooperate with other networks [5]. 5G-based V2X is beneficial in VANET communication regarding lower latency, and higher network capacity [73]. However, in comparison with DSRC in V2V and V2I communication have its drawbacks and limitations. For instance, DSRC-based VANET needs cheaper infrastructure than a 5G-based network. In addition, user devices like a smartphone in cellular communication user share cellular frequency bandwidth. However, the DSRC frequency channel is dedicated completely to VANET users. Furthermore, V2V communication in DSRC is infrastructure-less and 5G-

based communication are depending on infrastructure [1]. Therefore, we need to consider these networks that can be merged with VANET and realize their limitations and challenges to develop vehicular networks with V2X communications that are applicable to real-world VANET applications.

2.1.3 Potential Solution for Challenges in VANET Applications

ITS services in smart cities can be applied by developing VANET applications and can be categorized into four different classes including safety-related applications, infotainment applications, transportation traffic management applications and driving system monitoring applications [1]. This study focuses on safety-related applications which mainly provide services to avoid accidents and post-accident traffic on the road. Drive assistance, safety information provision and driver warning applications are the subcategories of the safety-related services [1]. The vehicle can share its positions, velocities and directions to avoid crashes by alert messages in bad weather with not enough visibility [74]. Warning and emergency messages can also notify approaching vehicles about what happen on the road in advance including accidents, road hazards, and traffic to keep vehicles aware form incidents ahead of them [75]. Moreover, traffic management applications can be used to enhance road safety and traffic adapt traffic flow to control traffic jams. Intersection management, traffic congestion management and transportation information application are the three main subclasses of this type of application [1]. Before deploying VANET services in a real-time environment, we need to consider different challenges related to VANET applications. There are four significant challenges regarding the VANET application including computational complexity, prediction accuracy, packet storm problem and security [1]. VANET applications such as traffic management can take advantage of big amount of data produced by a variety of entities on the road (e.g., roadside unit, vehicle, pedestrian) that can be analyzed to gain useful information and knowledge for vehicular users to take optimal decision in traffic management [7]. However, providing an efficient solution to enhance the computational complexity and enable road traffic density prediction is a major challenge that needs to be considered in the development of traffic management applications [1]. Prediction accuracy is another highly important challenge that occurs due to the heterogeneous and high mobility topology of VANET [5]. Therefore, prediction of weather and traffic condition can avoid traffic-related issues. Traffic congestion can be reduced by the prediction of traffic flow. However, the Prediction of traffic flow is a critical task [76].

Another beneficial aspect of the VANET application is broadcasting warning messages to inform road users about congestion conditions. However, this may cause a significant amount

of warning messages transmitted via a vehicular network which resulted in network congestion and in turn reduction of QoS for road users [4]. In addition, regarding VANET data propagation there are some issues. Firstly, AVNET real-world establishments technologies are highly expensive and complicated. Therefore, especially, in big test and trial simulations tools can prevent three significant problems in terms of credibility and feasibility by accurate simulation parameters and standardization and can be certified by applying diverse techniques. Therefore, it can provide a complex vehicle mobility model that can simulate real traffic and behavior of road users. However, simulation scalability is a major challenge [1]. Secondly, data broadcasting between vehicles and other entities is a significant issue that needs to apply appropriate techniques to share critical information regarding safety alarms and traffic and weather update information among entities in VANET [1].

Finally, although traffic management applications can enhance traffic-related issues on the road, security is a critical challenge which has an impact on the entire transportation system in the case of false information by attackers [1]. The security-by-designer principal still has lots of room for consideration. However, it can help the developer of the VANET application to identify the security obligations while designing the applications, in terms of vulnerability detection, and security level estimation, and provide suggestions for security enhancement [77].

2.2 Intelligent Traffic Prediction Methods

AI-based algorithms are promising solutions to enhance the operation of the network by offering intelligent prediction models for traffic in the network, then, the network operator can detect and avoid failure of the network before it causes network traffic and QoS reduction in VANET applications [4]. ML and DL as subclass of AI can be beneficial for the purpose of designing an efficient traffic prediction model [8], [21]. ML can be divided into three main classes including supervised learning, unsupervised learning and reinforcement learning [8], [14]. Different supervised learning models are commonly applied for traffic prediction such as Random Forest (RF), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT) and Support Vector Machin (SVM) [4]. Although each of them being its limitation and challenges.

2.2.1 Machine Learning and Deep Learning Techniques

Previous related work proposed different intelligent traffic prediction methods using ML and DL that are discussed in the following paragraphs.

Meen *et al.* [78] Considered traffic data from V2V communication with key features (i.e., location, direction, and speed) to predict traffic flow in ITS. They applied three different ML algorithms (i.e., DT, RF and SVM) and evaluated them based on some popular classification metrics such as accuracy, precision, recall and time. The obtained results show the highest value of accuracy assigned to RF while the longest execution time is assigned to RF as well. and SVM consumed the lowest time compared to RF and DT.

The author in [19] considered traffic data from V2R communication in VANET. They compared five popular ML models for network traffic prediction including KNN, RF, SVM, NB and MLP. The simulation results evaluated on the basis of different classification metrics such as Receiver Operating Curve (RUC), Precision-Recall (PR) curve, accuracy, precision, recall, F1 score and time. In this study, RF showed better performance considering all metrics than other selected methods. The lowest execution time was assigned to NB while it shows the lowest values in all other metrics and KNN was the worst in terms of time consumption while it was the best model in respect of accuracy.

Stepanov *et al.* [54] applied three well-known ML methods (i.e., RF, Bagging and SVM) to predict LTE network traffic on cellular traffic dataset. The regression evaluation metrics (i.e., Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R2)) employed to compare the simulation results. The bagging algorithm outperformed the RF and SVM in terms of all regression metrics. However, it consumed much more time for execution as well. Considering time consumption, SVM performed well compared to RF and Bagging models.

The author in [23] conducted a comparative analysis on four popular ML methods (i.e., RF, GBR, KNN, and MLP) considering traffic data with trajectory features including vehicle ID, Vehicle position, vehicle lane, speed and time. The obtained results indicated GBR as the best and KNN as the worst prediction model.

The above-mentioned research studies applied single ML models based on the performance evaluation of non of the individual ML methods capable of providing efficient prediction that bring trade-offs between all considered metrics such as accuracy and time. In addition, each ML model has its own limitation and challenges that need to be addressed.

It should be noted that the existing studies focus on predictive models for network or road traffic using ML algorithms. Regarding the most important issue which is selecting the effective ML method, different popular ML models including RF [4], [54], [78], KNN [23], NB [78], DT [78], SVM [46], [54], [78], [79], [80], and MLP [23], [], [22], [24], [80] are commonly considered for designing predictive models in mobile network and VANET applications. Although each of them brings its own weaknesses and problems. Some of the most important ad-

vantages and limitations of well-known ML and DL models are mentioned in the following paragraphs.

The SVM method is mostly considered for traffic prediction due to its adaptability to the dynamic and non-linear characteristics of traffic data. it also can handle a big number of features and a small number of input variables. However, optimizing kennel choice is challenging in this algorithm [7]. KNN is efficient classification algorithms rely on votes. However, finding the optimum K values in this model is difficult, and it needs much time for processing [7]. NB is a robust method since irrelevant features are ignored or eliminated. This algorithm is easy to implement. The limitation of NB is related to its advantage, since it handles attributes independently, it cannot extract valuate hints and suggestions [7]. DT is transparent and simple to use. However, it needs big storage beside it became complex if many DTs are needed. Therefore, it cannot handle overfitting and instability [7]. However, DT is faster than KNN and more flexible than NB. RF can solve the challenges related to scalability since it has the ability of training models in parallel. It can address the overfitting problem and only require a small number of input values. This algorithm is based on many DTs, therefore, in the case of time-sensitive applications and when the test data is large, it cannot be effective [7]. However, RF is more robust and accurate than DT and can solve the overfitting issue of the DT algorithm, and still, it consumes much more time than DT [**?**].

When it comes to DL models, NN like MLP [21], [46], [81] is a simple DL model that is mentioned as a well-chosen problem solver in the case of increasing the accuracy of the traffic prediction model. However, it has challenges related to dependency of its performance to choose proper network architecture and it has a slow rate of learning ability as well [25]. It should be noted that DL models are commonly applied to large and complex traffic data to be more efficient in terms of strong learning abilities, provide higher accuracy and better adaption to the complex and non-linear large dataset [24], [80], [82].

Another type of AI method used in VANET is Swarm Intelligence (SI) [7]. This type of algorithm is known as a nature-inspired algorithm capable of performing autonomous and distributed approaches [7]. PSO is a more efficient SI method than some other popular methods such as SDS, ACO and ASI for traffic data because no hypothesis is required concerning optimizing the problem. In addition, there are no neglectable limitations regarding the other mentioned methods. SDS requires to integrate with other metaheuristic algorithms and the highly dynamic nature of VANET can affect it. ACO requires high processing time for the purpose of generating multiple solutions and ASI has computation delay issues [7].

2.2.2 Ensemble Learning Methods

Therefore, based on the discussion in the previous section, each ML, DL and SI model that is applied in VANET being its own advantages and drawbacks for the prediction model. Recently, Ensemble Learning (EL) methods have attracted researchers in different domains (i.e., health, Finance, Energy) [34]. EL models have the potential to develop an efficient prediction model because of integrating different ML or DL models together and providing a strong model that can cover standalone ML and DL model limitations. In addition, it can be adapted to the dynamic topology of vehicular networks [83]. EL can provide more acceptable performance in terms of accuracy and computation time and improve stability and generalization ability than individual models [32]. Zheo *et al.* [84] considered a big traffic dataset (500,000 recorded samples) for road traffic predictions in IoV. They applied an ensemble model combining two DL models including Long-Short Term Memory (LSTM) and Stacked Auto Encoder (SAE) named EnLSTM-WAPO. The simulation results indicated that the proposed model outperformed the standalone DL models such as ARIMA, GRU, DBF, LSTM and SAE in terms of error metrics. The author in [81], proposed a novel ensemble model by merging LSTM, Deep autoencoder (DAE) and CNN for short-term traffic flow prediction. They considered two real-world traffic datasets. The simulation results showed the ensemble model outperformed than standalone DL model and two other EL models.

The author in [85], proposed a hybrid model by integration of RF and GRU named RF-GRU-NTP for network traffic prediction in VANET. Moreover, they employ V2V and V2R communication data. they considered both road and network parameters. The simulation results showed the hybrid model outperforms standalone ML and DL methods with better accuracy and execution time. The author in [48], proposed an ensemble learning model named STK-EBM, that employs a stacking learning strategy to combine different ML models in two layers structure including RF, KNN, Xgboost in the first layer and Logistic Regressor (LR) in the second layer in order to obtain more accurate and stable network traffic prediction model in VANET. They considered V2V and V2R communication datasets. They employ Boruta and Light GBM as the ensemble feature selection method to extract more efficient features from the real-world merged dataset. In this way, they can maintain the quality of data that is fed as input to the EL model which can affect the performance of the model. They compared the obtained results with the most popular individual ML models and the proposed model showed stable and balanced results in all considered metrics including (i.e., Roc curve, confusion matrix, precision, recall, F1-score, accuracy, and time). The STK-EBM outperformed standalone ML methods. The author in [80], presented an ensemble model by combing MLP and Self-Adaptive Support Vector Regression (SSVR) to predict mobile

network traffic. The evaluation results indicated more accuracy and stability than different single Dl models including DNN, LSTM and CNN.

The author in [45], proposed an ensemble of three various ML models using soft voting strategy including RF, DT and KNN for forecasting network traffic. They employed V2X communication data including V2V, V2R, R2R and V2P. The simulation results showed better performance of the proposed model than five standalone popular ML models (i.e., RF, KNN, NB, DT, SVM and MLP) in all classification metrics inclusion precision, recall, F1-score, accuracy, ROC curve and execution time.

In summary, an efficient intelligent network traffic prediction model can help to avoid network failure which can affect network performance and in turn the QoS for VANET application. There are key important points that need to be taken into account to propose an efficient model. The first important consideration for AI-based model is the data that need to be high quality and effective before importing to AI models [54]. In addition, the integration of AI and vehicular network bring some challenges from not only keeping the quality of input data, but also proper integration strategy and effective evaluation of the prediction model considering computation complexity and time, accuracy, stability and generalizability. In this way, we can propose an efficient traffic prediction model that can be beneficial for network management and optimization especially when it comes to safety applications [80], [82].

In summary, an efficient AI-based prediction model relies on the size of the dataset, quality of data, selected features and type of problem. Subsequently, we need to identify the best-matched iteration strategy and AI models to achieve the best and most efficient prediction results. Therefore, designing a network traffic prediction model that is intelligent and efficient for VANET applications considering the dynamic, non-linear and complex nature of VANET topology is a critical task. Eventually, we can take advantage of the best proper EL strategy and efficient ML, DL and SI model selection to provide a stable prediction model with better overall performance in terms of accuracy and computation time.

Ensemble Learning (EL) techniques are presented as the base study by Dasharathi and Sheela in 1979 [86], they applied a technique to divide the feature space based on various basic classifiers. Basically, EL methods can be categorized into two types consist of parallel and sequential ensemble methods. As depicted in Figures 1 and 2, in the parallel ensemble methods, various base learners are trained independently and in parallel. Then, the prediction results of all the base learners are merged by applying an integration strategy to make the final prediction. However, in sequential ensemble methods, the base learners are trained in a correlative way which allows the correction of the error made by the previous base learner in each iteration [87]. Figs and show the block diagram of sequential and parallel ensemble

learning strategies.

The base learners can be homogeneous or heterogeneous, in the case of heterogeneous base learners which means employing various ML methods to cooperate with each, we need to select the most effective techniques to integrate that prediction of base learners and achieve a better performance [88], A key important consideration in a combination of base learners to create an ensemble model is the integration strategy [87]. There are three main categories of ensemble methods including boosting, bagging, and stacking [89]. In addition, the most popular integration method of the base models is majority voting which can be applied to classification and regression problems [90]. It should be noted that selecting an appropriate subset of the base models and ensemble strategy is a crucial consideration in ensemble learning methods [87].

The stacking ensemble learning was initiated in 1992 [91]. This framework can be applied to enhance the generalization error in ML issues. It can be beneficial in the case of multiple choice of suitable ML models on a specific problem. The stacking strategy employs a separate ML model to learn the prediction from the various models [92]. It includes two levels consisting of level 0 which is the base model and level 1 which is the meta-model. The based models in level 0, utilize heterogeneous ML models that are trained on the same dataset. The prediction results of the base models combine by meta-model in level 1, to build a new strong model that provides the best prediction results of the base models [87]. It should be noted that, unlike stacking, bagging ensemble models that are trained on the subset of the input individually to fit many decision trees and use averaging the predictions result. A boosting ensemble model works based on adding ensemble models in a sequential way, which means each model learns from the prediction mistake of the previous model to produce a better prediction and the result is a weighted average of the predictions [87].

2.2.3 Taxonomy of the Proposed Network Traffic Prediction Methods in VANET

AI solutions have attracted researchers to apply them to different approaches and domains because of their ability to improve data-driven techniques [7]. On the other hand, with the emergence of wireless technologies, intelligent transportation systems and novel designs in automotive technology the industry, VANET has been developed toward IoV, which brings new requirements that need to be considered. For instance, V2X communication opened up new challenges because of various types of communication between vehicles and all other surrounding entities including Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Infrastructure -to- Infrastructure (I2I), Vehicle-to-Personal Device communication (V2PD) and Vehicle-to- cellular network infrastructure (V2CN) [7]. 5G network has a novel creative char-

acteristic known as the New Radio V2X (NR-V2X) standard that integrates with Side Link (SL) communication to facilitate direct radio communication between vehicles and personal devices of the pedestrians for traffic data transmission without the demand for RSUs [7]. There are a variety of VNAET applications and services that use these communication data, especially with the aim of safety applications including accident avoidance and broadcasting warning messages in the case of any incident on the roads [93]. However, these services impose challenges related to security, privacy, performance and QoS [7]. AI-solution (i.e., ML, DL and SI) integrated with the VANET application can solve different issues in VANET. It should be noted that the Integration of AI techniques to vehicular networks has lots of room for consideration [94], [95].

The next three chapters contain the published and submitted papers which are the main parts of this dissertation. The main objective of this dissertation is to propose an efficient network traffic prediction model based on ensemble learning methods in VANETs. Therefore, in Chapter 1, considering the main objective, we defined three sub-objectives in this dissertation. In each sub-objective, we employ an AI method to solve a classification problem along with applying different ensemble strategies and considering the basic architecture of VANET with the aim of predicting traffic in the network.

In Chapter 3, the article entitled " An Ensemble-Based Machine Learning Model for Forecasting Network Traffic in VANET " has been published as a journal paper in "IEEE Access" journal. This article considers the network prediction problem as a classification problem. In this article, a stacking ensemble strategy method is proposed to predict the network traffic in VANET. Moreover, regarding the significance of input features in the performance of the ML models, the ensemble algorithms including Boruta and LightGBM are employed to find the best effective attributes in the merged V2V and V2I datasets. The proposed method predicts traffic and non-traffic situations in the network.

In Chapter 4, considers the VANET with advanced communications known as V2X with the aim of network prediction problem as a classification problem. A soft voting classification method is proposed to show the network performance in terms of the simple ensemble of heterogeneous ML models including RF, DT and KNN in the VANET. This part of the dissertation was published as a conference paper with the title of " A Soft Voting Classification Model for Network Traffic Prediction in VANET/V2X " by IEEE publisher.

The article presented in Chapter 5, an eSwaNN-NTP method is proposed to predict the network traffic in VANET. In addition, a simple DL model that is ANN is applied. In this approach a VANET architecture is designed that integrates DSRC and 5G technologies to provide both short-term and long-term coverage for V2X communication in vehicular

networks. This chapter of a dissertation is submitted to "IEEE Access" journal with the title of " Swarm-Based Ensemble Model for Network Traffic Prediction Considering Basic and V2X Communication in VANET".

In summary, the main objective and the three sub-objectives of this dissertation are met and will be extensively explained in more detail in the following three chapters. Predicting network traffic in terms of packet receiving and data delivery ratio in the network is important and beneficial in order to make appropriate decisions and avoid service degradation in VANET applications. Therefore, the contributions of this dissertation help the optimization of the network performance, QoS and user satisfaction of VANET services. Moreover, the presented methods help us towards forming efficient AI-based models which are more applicable in the real-world implementation of VANET applications.

CHAPTER 3 ARTICLE 1: AN ENSEMBLE-BASED MACHINE LEARNING MODEL FOR FORECASTING NETWORK TRAFFIC IN VANET

Authors: Parvin Ahmadi Doval Amiri, Samuel Pierre

Status: published in *IEEE Access*, March 2nd, 2023.

Abstract Vehicular Ad-hoc Networks (VANETs), as the most significant element of the Intelligent Transportation Systems (ITS), have the potential to enhance traffic efficiency and road safety by making the transportation system smarter and are still at the initial point of development. In this paper, we propose an ensemble-based machine learning model for network traffic prediction in VANET. We take advantage of Ensemble Learning (EL), which combines different Machine Learning (ML) models to achieve better performance and improve accuracy. We consider the most informative attributes of the VANET dataset using Boruta and LightGBM as ensemble feature selection methods. Our proposed model is based on Stacking Ensemble Learning with Booster Model (STK–EBM) designed with a stacking ensemble of heterogeneous ML models. The framework of the proposed model consists of two layers, including a base layer and a meta layer. The first layer integrates Random Forest (RF), K-Nearest Neighbor (KNN) and XGBoost as a booster of the base learners. An optimized Logistic Regression (LR) employs as our meta learner in the second layer. We evaluate the performance of our model considering classification metrics and then compare it with the most popular traffic predictive models. Simulation results show that the STK–EBM model gives a more stable prediction than the single algorithm, as well as better overall performance in terms of prediction accuracy and execution time.

Keywords: Vehicular ad-hoc network, network traffic prediction, classification, machine learning, deep learning, ensemble learning.

3.1 Introduction

Currently, people have become more dependent on transportation, and the number of road users has significantly increased. This growth leads to multiple problems in terms of air pollution, fuel consumption and costs by wasting the time of drivers in traffic and various losses due to road accidents. An efficient way to enhance road safety and tackle these various losses is by taking advantage of technology to raise the awareness of drivers, especially about

the probability of an accident or congestion on the road. Intelligent Transportation System (ITS) uses information, communication, and control technology to manage transportation networks. Vehicular Ad-hoc Networks (VANETs) are considered the most significant elements of ITS to enhance traffic efficiency and road safety [3]. The basic architecture of VANETs includes Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication [5]. These communications in VANETs employ the Dedicated Short-Range Communication (DSRC) and Wireless Access in Vehicular Environment (WAVE) (or IEEE 802.11p) standards [96]. VANET application uses the data collected from the mentioned communication that can produce valuable and shareable knowledge among vehicular users on the road to raise the awareness of drivers to make an informed decision, especially to prevent accidents and traffic congestion. Therefore, it brings safety, efficiency, reliability, comfort and convenience [11]. However, it imposes various service requirements, such as network performance, which is highly important with regard to VANET safety applications.

In the case of a time-sensitive application that relates to human life, not only time but also data accuracy and reliability are critical [11], [59]. Consequently, the existence of traffic data and easily reachable tools for analyzing these data and extracting knowledge pave the way for researchers to enhance the requirements of networks connected to road users' lives. Ultimately, researchers can make the transportation system smarter. They enhance road safety and decrease crashes and other losses related to traffic by developing new driving rules, policies and predictive models based on Artificial Intelligence (AI) that optimizes traditional data-driven approaches. Therefore, vehicular networks are in the early stages of challenges related to the exploitation and adaptation of AI tools [8]. Moreover, the implementation of Machine Learning (ML) techniques as a subset of AI could optimize the operation of the networks in predicting failure before it causes a significant reduction in the Quality of Service (QoS) [4]. In VANETs, information about road conditions and other vehicles will be exchanged among communications when the number of sending and receiving packets through these communications increases (i.e., many vehicular users on the road) and traffic occurs in the network, which will cause a delay or decline in important services. Accordingly, an efficient prediction of traffic in the network can help enhance the QoS, accuracy, reliability and time for road users. It can improve whole driver-vehicle-road performance [46]. The question is how can we propose an efficient network traffic prediction model using AI?

ML, as a major part of AI, can be categorized into three parts: supervised learning, unsupervised learning and reinforcement learning. Moreover, transfer learning, online learning and Q-learning can be considered subclasses of these three main learning models [8], [14]. DL is closely related to the three mentioned classes of the ML model. It is a deeper network of neurons in multiple layers that is used for traffic prediction in a large and complex

dataset [8]. In this study, we focused on supervised learning, in which training data are based on labeled data. Moreover, supervised learning can be designed as a classification and regression task [14]. We considered our problem as a classification task. Furthermore, supervised ML algorithms, including Random Forest(RF), K-Nearest Neighbor(KNN), Naive Bayes (NB), Decision Tree(DT) and Support Vector Machines (SVM), are commonly considered for designing predictive models in traffic [4]. Among them, Support Vector Machines (SVM), which can adapt to the dynamic and nonlinear nature of traffic data, have problems with selecting the kernel type and resolving this issue. The optimized SVM is an adaptive model for forecasting traffic and fitting times [46]. Moreover, Neural Network(NN) is a simple DL model mentioned as a well-chosen problem solution in increasing the accuracy of traffic flow prediction [22]. However, due to the limitations and drawbacks of single ML models in traffic prediction, Ensemble Learning(EL) has become popular and used in various domains, including health, finance, and energy [34]. The abovementioned studies on VANET application issues, AI limitations and advantages can work together to match the traffic problem in VANETs around the best AI solution. In summary, VANET applications impose some service requirements, such as network performance, that are highly important specifically for safety applications that relate to human life [46]. When traffic occurs in the network, it causes a significant reduction in QoS and failure in the network. ML, as a subset of AI, can optimize the operation of networks [4]. Making an intelligent prediction of traffic in the network can help us to precisely identify the failure of the network and avoid service degradation, especially for services that are highly dependent on the performance of the network. However, each ML model has limitations and drawbacks for traffic prediction. In this case, ensemble learning can significantly improve the performance of machine learning in most problems [34]. Ensemble learning can achieve better performance with enhanced stability and generalization ability in addition to higher prediction accuracy and fast computation than a single ML [32]. We need to identify the best proper strategy to integrate ML models through EL. This motivated us to use EL in an efficient way for network traffic prediction to maintain the QoS and performance of the network in VANET applications.

In this paper, we propose a network traffic prediction model using ensemble learning and integrating various ML approaches. We identify the best strategy to integrate ML methods through ensemble learning in an efficient way with the aim of providing a balanced result in addition to improving the overall performance.

The originality of our model lies in proposing an efficient ensemble of ML models to predict traffic with the aim of obtaining more stable and accurate prediction results. We consider VANET communication from the integration of V2V and V2I data. In addition, the most informative features are built from the extracted datasets using LightGBM and Boruta as

the feature selection approaches. Using this approach, we can maintain the quality of input data that will be highly important for efficient prediction. Therefore, the main objective of our proposed model is to design an efficient network traffic prediction model for VANETs. The detailed contents of this paper can be summarized as follows:

- Compare the top popular ML models for traffic prediction, including RF, KNN, NB, DT, SVM and MLP, as simple Deep Learning (DL) models to identify which one results in better performance and to realize the effective incorporation of ML models that bring more accuracy and adaptability to the dynamic nature of traffic, then use them as the base learners in the first layer of our proposed model.
- Propose a hybrid stacking ensemble model to predict traffic to obtain a more stable prediction and overcome the inherent weakness of every single model. The proposed model has two layers, including the base layer and meta layer, in which the most effective combination of ML models was selected. The first layer integrates RF, KNN and XGBoost as a booster of base learners, and an optimized Logistic Regression (LR) is employed as our meta-learner in the second layer.
- Evaluate the performance of the proposed model according to the classification metrics that can effectively assess the prediction results.

Eventually, we develop an efficient AI solution for network traffic prediction to prevent service degradation in VANET applications. The proposed model considers highly important points for being a more beneficial and powerful model, such as maintaining the quality of data, a heterogeneous integration strategy with reducing complexity and covering single ML model problems, in addition to taking advantage of the most effective models. Therefore, it can provide a more adaptable and stable prediction model with better overall performance in terms of accuracy, prediction error and execution time.

The rest of this paper is organized as follows. Section II discusses the related work. Section III describes the proposed methodology based on ensemble learning for network traffic prediction, and our results are presented in Section IV. Section V concludes the paper.

3.2 Background and Related Work

Currently, the onward movement in the field of communication and computing systems gives researchers the opportunity of a new way of solving problems related to intelligent traffic to enhance traffic efficiency and road safety. Several researchers have used AI to optimize

traditional data-driven approaches. The various branches of AI will be able to bring out an optimized solution that will not cause or generate more problems. Vehicular networks are in the early stages of challenges relevant to the exploitation and adaptation of AI tools [8]. Sultan *et al.* [5] indicated VANET architecture and applications. They discussed V2V and V2I communication that make possible the advancement of many applications. Faezipour *et al.* [97] discussed this communication and related challenges, as well as available solutions in intelligent Vehicle Area Networks as a future transportation system. Therefore, VANETs offer several applications, including safety and no-safety, using technologies to provide operative traffic management in vehicular networks. However, these services have diverse requirements in VANETs, such as network performance, which plays a significant role, notably for VANET safety applications. In this matter, the outcome data must be reliable, accurate and timely due to their effects on the vehicular road user's life [11], [59].

ML and DL, as a subset of AI, can be utilized for developing effective models and provide a better and higher rate of prediction accuracy [8], [21]. Although ML approaches provide better performance than the traditional model, each of them has challenges and issues. The way to cover a single ML model's problem to be more beneficial and powerful is to combine different ML models, which is called EL [34]. Previous related works presented different ML, DL, EL and optimized approaches to enhance the performance of their model for traffic prediction on a road or in other networks or other domains. Most of them focus on improving accuracy. In this study, we try to take advantage of EL and focus on providing a model that improves the accuracy with an efficient prediction time and keeps the overall performance better and more stable. Moreover, the proposed model is designed explicitly for VANETs considering the effective features of both V2V and V2I datasets to maintain the quality of input data, which is challenging in intelligent vehicle area networks. The authors in [78] indicated that the existing traffic flow prediction in ITS has problems adapting to real-world applications. They worked on traffic data obtained from V2V with important features such as location, direction and speed. They applied three ML models including DT, SVM and RF. The obtained results were assessed based on classification metrics (i.e., accuracy, precision, recall and time) and showed a higher value of accuracy for the RF and consumed longer time than other models. The minimum time of prediction was assigned to SVM. The study needs to consider more metrics like the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) metrics to better evaluate the models, and none of these models performed well in all performance metrics (i.e., accuracy and time). In urban traffic prediction, Lee *et al.* [23] compared four different and popular ML models consisting of RF, Gradient Boosting Regression (GBR), K-Nearest Neighbor (KNN) and Multi-Layer Perceptron (MLP). The trajectory data with features including Vehicle ID, Vehicle position

and lane, time and speed were considered. The evaluated results showed GBR and KNN performed best and worst, respectively, among the methods. Tong *et al.* [79] discussed traffic flow prediction in VANETs. They used an improved particle swarm optimization (PSO) algorithm to enhance support vector regression (SVR) parameters. The presented algorithm performed the best compared to DT and SVR with grid search optimization.

The hybrid LSTM-SAEs model was proposed in [84] for urban road traffic prediction on the Internet of Vehicles (IoV). The authors considered Long Short-term Memory (LSTM) as a developed structure of Recurrent Neural Network (RNN) and a strong model for prediction of time-series traffic data, which requires historical data. Moreover, they used the influential DL model stacked Auto Encoder (SAE) that can learn automatically from input data to have a pinpoint feature description. This research work took advantage of these dominant DL methods by merging them into one model called the EL model [98], [99]. Of note, considering the nonlinearity of traffic data, EL models have been famous these days. In other words, using different models together, we can build a stronger new model than each individual model and cover their flaws [83]. In addition, this EL model used feature engineering to improve accuracy. They considered big traffic data of 500,000 recorded samples from 51 road sections collected every 5 minutes. The approach achieves less prediction error than the base models, such as the autoregressive integrated moving average (ARIMA), Gated Recurrent Unit model (GRU), Deep Belief Network (DBN), LSTM and SAE.

Zheng *et al.* [81] presented a new EL model named EM for short-term traffic flow prediction. They combined three DL algorithms comprised LSTM, deep autoencoder (DAE) and Convolutional Neural Network (CNN), where CNN and LSTM allow to consider both temporal-spatial traffic features. They employed two real-world traffic datasets to validate the model performance. Their approach includes hidden and softmax layers for final prediction. The output of each model was individually used as input for the EL hidden layer to ensure that each individual output came up with an equal quantity of features for the softmax layer before the final prediction. The obtained result shows a higher accuracy value compared with every single model besides the other two EL models named DA and CNN-LSTM (CLTFP). Moreover, they mentioned that their approach was robust in the case of high variance.

Stepanov *et al.* [54] emphasized the point that ML and DL models can optimize network traffic prediction. They collected cellular traffic data and applied three ML algorithms consisting of RF and bagging. They indicated the advantage of bagging that each tree can learn freely from realizing the results in another tree. In contrast, RF obtained results for each object based on the output from each tree. The evaluation of the results by RMSE, MAE and coefficient of determination (R2) shows that bagging performed well in all metrics. However,

it consumed more learning time than the others. In the case of learning time, SVM is the best. This research work mentioned some interesting and helpful points related to the use of ML models for network traffic prediction, such as keeping the quality of feed data to ML by preprocessing, evaluating the importance of the features, and tuning the hyperparameters of ML models that can provide better prediction results.

A novel stacking ensemble learning approach aimed at mobile traffic prediction was presented in [80]. The proposed EL-MS model merges two ML algorithms named MLP as a base learner and the self-adaptive support vector regression model (SSVR) as a meta learner. The obtained result was evaluated by MSE to assess the difference between the actual and predicted values. The model shows stable and more accurate results than some of the other ML models (i.e., RNN, LSTM and CNN) and some ensemble models (i.e., DBN-SVM). Of note, MLP, as a simple DL model and well-known NN, can adapt to network properties and traffic patterns [80], [82], [24]. The EL model for VANETs is also a promising solution for designing efficient predictive models [7]. Designing an efficient and reliable prediction model is essential for network traffic management and optimization [100]. EL methods are applicable for minimizing bias, enhancing predictions and being robust to overfitting [7]. EL methods can achieve better results than standalone ML models in traffic prediction [49]. However, time consumption is one of their main problems [7].

In summary, although VANET services will result in safety and comfort for road users, inaccurate and false predictions, especially for safety applications, may affect the life of vehicular users. To our knowledge, designing an efficient intelligent model with the integration of AI and vehicular networks still needs much consideration. This motivated us to investigate highlighted points of related works from the input data to the model selection and effective evaluation of the model based on important factors for predictions. The existing research works indicate their approaches using a subset or combination of ML for traffic prediction on roads or for different types of networks. Some of them considered the quality of input data, and others just considered optimization on one ML algorithm. The lack of intelligent traffic prediction specifically for VANETs with both V2V and V2I data and covering the overall performance leads us to take advantage of EL and try to design an efficient network traffic prediction model with real data that can provide better overall performance and balance between accuracy and consumption time, which is highly important, especially for safety applications in VANETs.

3.3 Methodology

In this section, we propose an optimized and efficient stacking ensemble learning model by taking advantage of EL models. This model is applied to VANETs to predict network traffic. Figure 1 shows that the basic architecture of VANETs is composed of V2V and V2I communication.

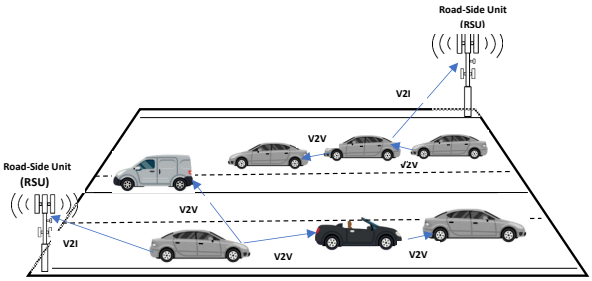

Figure 3.1 The basic architecture of VANET.

Vehicular networks are in the early stages of challenges relevant to the exploitation and adaptation of AI tools. ML is a subset of AI used for accurate analysis and prediction models [8]. However, each ML model suffers from its weaknesses and drawbacks. The way to cover a single ML model problem to be more beneficial and powerful is to combine different ML models, which is called EL [34]. Of note, considering the nonlinearity of traffic data, EL models have attracted considerable attention. The purpose of EL is to build a new strong model using several simple ML models together that can face the limitation of every single model in addition to taking advantage of their different views in solving the prediction task [83]. It also helps with reducing errors, achieving higher accuracy and robustness, fast computation and a better generalization ability than a single model in most problems [34], [32]. Figure 2 illustrates the workflow of our research work. We divided our model into three parts: data collection and preprocessing, model building and analysis of the result. This classification aims to differentiate the contribution of each part and then explains the procedure of the model in detail.

3.3.1 Data Preproccessing

The benefits of data preprocessing and feature selection from input data before feeding into ML models are highlighted in several studies. Keeping the quality of the data with some preprocessing, such as cleansing the missing data, normalization and choosing the more relevant important features, can play a significant role in increasing the efficiency of the

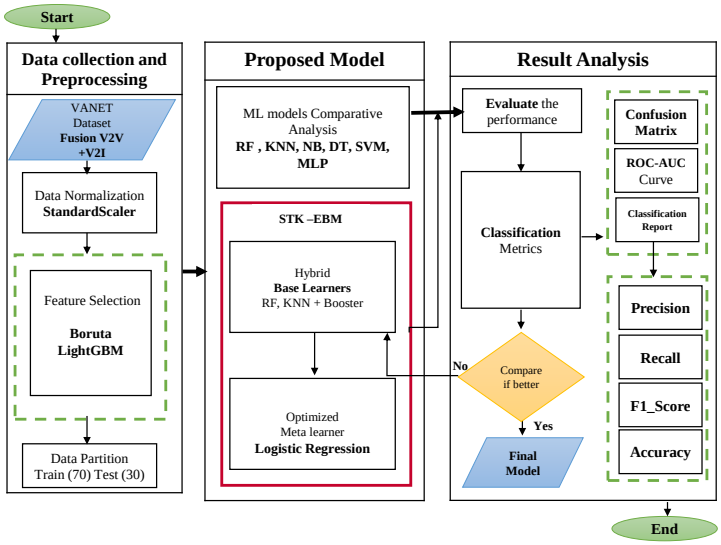

Figure 3.2 The workflow of this study.

model [50], [48] [52]. Therefore, in the first part of our model, we considered normalization and feature selection methods to achieve a high-performance predictive model.

Normalization

In this phase, we first removed some redundant, missing, and meaningless values in the raw dataset. Then, we normalized our data using StandardScaler normalization [52], which scaled all feature values in the range of [0,1], and in this way, we performed simple preprocessing. The formula is defined as:

$$Z_{scaled} = \frac{(X - \mu)}{\sigma}, \tag{3.1}$$

where X = input variable, μ = Mean and σ = Standard Deviation.

After normalization, we selected the important features that will be discussed in Section 2. Finally, we separated our dataset into training and testing sets. We assumed variable X with (i) the number of selected features as the input data. Accordingly, we labeled our target variable (y), which is the prediction of the traffic, in two classes of traffic (1) and no traffic (0).

Feature Importance and Selection

In this section, we present some research and several experimental analyses on various techniques that help us gain insight regarding the relevancy and importance of the features with the target variable. Each method gives a different perspective about how the variable can be useful depending on how the algorithms learn the target. Ching *et al.* [50] performed a comparative analysis of feature selection methods by considering different types of datasets. They focused on feature selection importance with the aim of data classification using machine learning algorithms. Inspired by this research work, we considered Boruta [50], which is an RF-based feature ranking technique. This algorithm determines the importance of variables with statistical judgment and detects all significant relevant features. Then, we used LightGBM as an ensemble feature selection method [48] to find optimal features in our dataset while considering V2V and V2I datasets separately and together. Finally, as one of the contributions of our research work, we come up with the best and most important, efficient and confirmed subset of the selective features by LightGBM and Boruta methods. Therefore, we can optimize network predictive model performance in our real dataset.

3.3.2 Overview of the Proposed STK-EBM Model Architecture

In the first section, we collected and merged data from V2V and V2I in VANETs, and then we preprocessed the data. We considered the most informative attributes of the data using LightGBM and Boruta approaches. At the end of the first section, we divided our dataset into training and testing to feed to the ML models for prediction. In the following subsections, we describe the components of the presented model: stacking heterogeneous ensemble model structure, base learner element selection and meta learner.

Stacking ensemble learning, because of a heterogeneous integration strategy, has the ability to increase the generalization of the model. Strong model can be generated by combining several models, and the structure of stacked ensemble learning is composed of two layers [101], [91]: base learner and meta learner. The reason can be justifiable in the real world when an important decision needs to be made. Several experts in the related field have provided a consensus opinion and achieved one strong professional decision. We built our proposed model named the stacking optimized heterogeneous ensemble model for the network traffic prediction problem (STK– EBM). It is composed of two layers. The first layer is constructed from selective ML algorithms, which are called base learners. The combination of the base learners is based on a comparative performance analysis of the most popular ML models used in previous studies for traffic prediction. The second layer considers one algorithm called the meta learner and is responsible for the final prediction of the whole model. The proposed

model focused on enhancing the overall performance in classification evaluations. In addition, there are some considerations and optimization in each layer that will be discussed in the following subsections. The global architecture of the proposed model is presented in Figure 3.

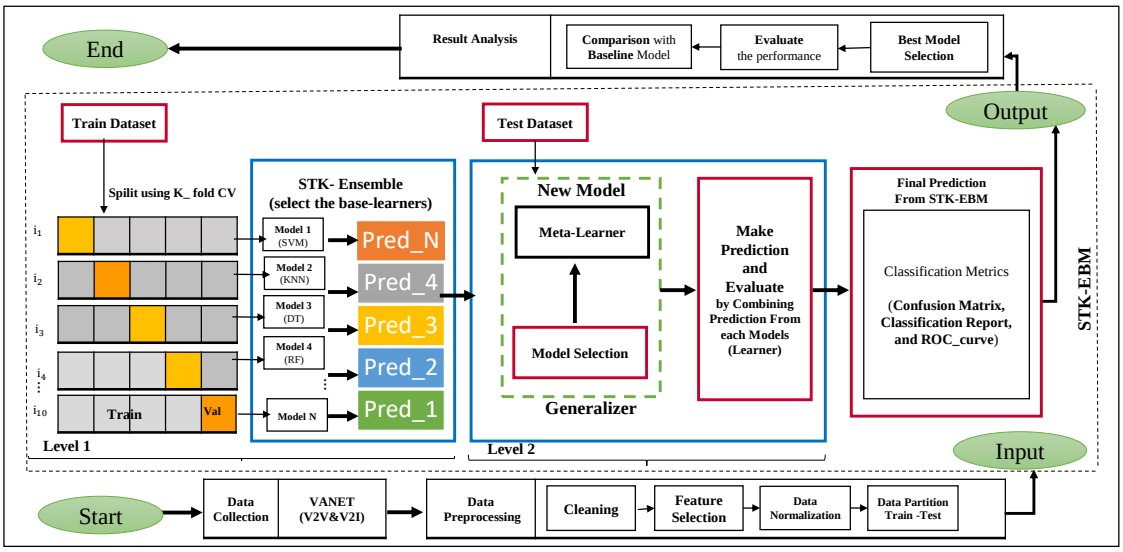

Figure 3.3 Architecture of the proposed stacking ensemble-based machine learning model for traffic prediction in VANETs.

Stacking Heterogeneous Ensemble Model Structure

"Stacked generalization is a generic term referring to any scheme for feeding information from one set of generalizers to another before forming the final guess" [91]. The deployment procedures for stacking an ensemble of models with the aim of network traffic forecasting are described as follows.

Step 1) The input variable is input into x to represent $x = x_1, x_2, ..., x_n$ that each x_i R^d (which is an attribute vector with d dimensions), and we put the target variable in (Y) that is labeled 0 for no traffic and 1 for the existence of traffic in the network. We split our dataset (Ds) into training and testing sets.

Step 2) The training set is divided into (k) equal-size subsets: $Ds = D_1, D_2, ..., D_K$. Therefore, the input of our model is the training set (train on k-1 one of these subsets), and our model is trained on the training set. The model evaluation is performed on the last subset as a validation set. Therefore, the validation is separated from the training set and is used

to validate our model performance during training. In this way, we ensure that the same data point is not present in both testing and training. This helps to prevent the model from overfitting. The output is the final prediction from the meta learner of STK–EBM.

Step 3) In layer one, the base learners (b) learn from the training set, where $b = b_1, b_2, ..., b_m$ are the (m) base learners and form 1 to *k-1* fold and continue the learning process and the prediction results on the last fold (k). Then, all base learners predict in k repetition, where $p = p_1, p_2, ..., p_k$.

Step 4) The new dataset will be generated for the meta learner $D_s^{'}(x_i^{'}, y_i)$, where $x_i \in D_k$. It is designated by:

$$x_i^{'} = \left\{ b_{k_1(x_i)}, b_{k_2(x_i)}, .., b_{k_m(x_i)} \right\}$$

Step 5) In the second layer of our model, the meta learner learns from a newly generated dataset $(p^{'})$

Step 6) The final output is the combination of base learner prediction in layer one by meta-learner as follows.

$$P_{(x)} = p^{'}_{(b_1(x), b_2(x), ..., b_T(x))}$$

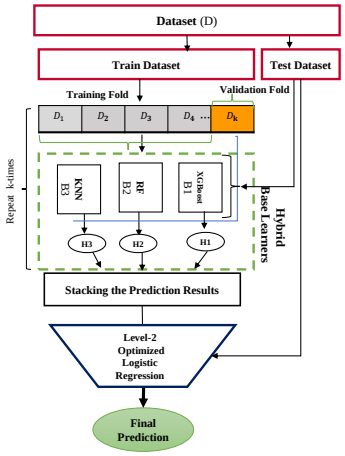

Figure 3.4 Framework of our proposed STK-EBM model.

In summary, we proposed a hybrid stacking ensemble model. The model is composed of two layers. We stacked a set of high-performance heterogeneous base learners in the first layer. These base learners were selected according to the performance analysis of the most popular ML prediction models. In this way, we reduced the complexity of stacking the useless and unsuitable models. The second layer is the meta learner, which combines the results of base

learners and provides a final prediction. The framework of the stacking ensemble model with the steps mentioned above is depicted in Figure 4.

Base Learner Element Selection

In this section, several popular ML algorithms used for traffic prediction are taken into account. The important issue is selecting the best combination of these algorithms when creating our model from scratch. This is another contribution of our research. In this context, we performed an experimental analysis of the most mentioned and popular ML models for traffic prediction, including RF [4], [78], [54], KNN [23], NB [78], DT [78], SVM [46], [78], [79], [80], [54] and MLP [22], [23], [80], [82], [24]. These models are individually trained, and the performance of each of them was evaluated. We find the best combination of these learners as the base learners. Base learners in our approach are diverse because each single ML model has a different view about solving the prediction task and brings its advantage and disadvantage for traffic prediction. Furthermore, we add a booster to the base learners to boost the prediction results. For this purpose, some ML algorithms, such as XGBoost [**?**], [56], and AdaBoost [102], can be employed. In these ways, we are able to bring more accuracy, adaptability and stability to the dynamic nature of traffic. Because our dataset is not sufficiently large to try on different DL models, we decided to perform MLP as a simple DL model, which is also common for the traffic prediction domain [22]. The MLP model, as the classical type of feed-forward neural network [23], [24], consists of three layers: the input layer, the output layer and the hidden layer, while the number of output nodes is based on the machine learning task. The classification task in our problem includes two output nodes. Regression tasks commonly consist of one node [24]. The general structure of MLP is depicted in Figure 5. The related formula is designated by [80], [82], [101].

$$X = (\sum_{i=1}^{n} w_{ij} a_i) + b_j, \tag{3.2}$$

where a_i = input variable, n is the number of inputs, w_j= the connection weight, w_{ij}= input into the summing junction, b_j=the bias of neuron employed for summation.

$$F(X)_{=} u_j = F[(\sum_{i=1}^{n} w_{ij} a_i) + b_j], \tag{3.3}$$

where X generates the output through the transfer function F and u_j is the summing junction.

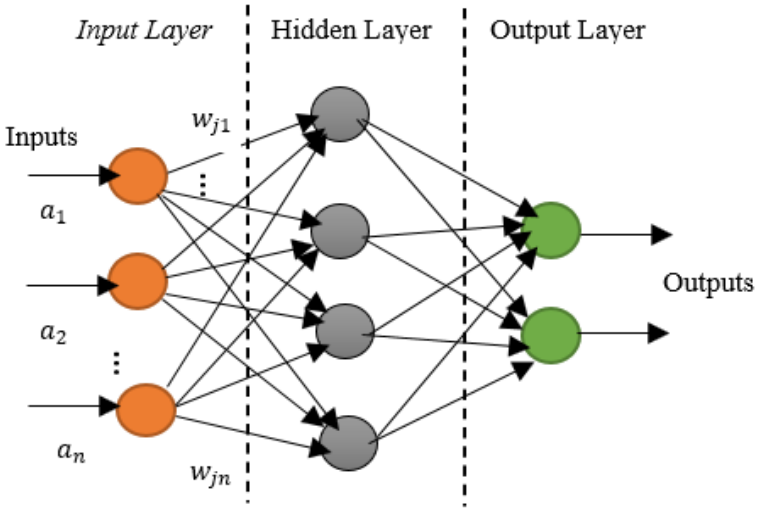

Figure 3.5 The general structure of a Multi-Layered Perceptron (MLP).

$$F(X) = \frac{1}{1 + e^{-x}}, \tag{3.4}$$

where the sigmoid activation function which is the connection weight from the i the input to the j th hidden neuron.

Meta Learner

In the final step of this section, we aimed to simplify the interpretation of the base learner prediction results using a simple meta learner. Therefore, we employ an improved Logistic Regression (LR) as the meta learner. It can find the optimal combination of the prediction results of all base learners. In other words, the meta learner was trained based on the prediction made by previous learners. This helps improve the predictive performance of network traffic. We used grid search cross-validation [103], [104] to improve the accuracy of LR. It can tune the hyperparameter and result using the best parameters of the algorithm [105]. Therefore, we can obtain more accurate results:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{3.5}$$

$$f\,(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{\beta_i X_i}\right)}}, \tag{3.6}$$

where X= input variable, β = predicted weights and z = predicted output.

input variable (X) with 1 to k as the number of features are merged linearly with predicted weights (β) to predict an binary value as output (z). The weight indicates the variable impact on prediction. $\beta = \beta_1, \beta_2, ..., \beta_k$ in which β_1 *to* β_k used for assessing weight of input variable and β_0 for assigning the bias value. Moreover, The probability of existence traffic in the VANET represent by $f(z)$. It called transformation function with a range between one and zero ($0 \leq f(z) \leq 1$). This function transforms probabilities into a binary value. where, $z < 0.5$ *output* $\rightarrow 0$ *(no-traffic)*, else ($z \geq 0.5$) *output* $\rightarrow 1$ *(traffic)*.

In summary, we first perform an experimental analysis of the most popular ML models for traffic prediction, including RF, KNN, NB, DT, SVM and MLP. Second, when we build an EL model from scratch, the important issue is to select the best combination of ML algorithms. Therefore, in our EL approach, which is composed of base learners and a meta learner, the base learners bring a different view about solving the prediction task. In this way, the model provides more accuracy, adaptability, and stability to the dynamic nature of traffic. The selected algorithms, including RF, KNN and XGBoost, can help us to address some issues. RF can solve the challenges related to scalability, and KNN showed better performance results than the other ML models based on our dataset. Furthermore, XGBoost was selected as a booster of the base learners' performance. It can also enhance the generalization ability, but its parallel learning ability with distributed computation will not impose additional time. Finally, we aimed to simplify the interpretation of the base learner prediction results using a simple meta learner. Accordingly, we employ an improved Logistic Regression (LR) to find the optimal combination of the prediction results of all base learners. Eventually, we take advantage of the most effective combination of ML models to provide a stable prediction model with better overall performance in terms of accuracy, prediction error and execution time.

3.3.3 Evaluation and Analysis Metrics

We used the confusion matrix, classification report and CPU time as the most common classification evaluation metrics. Furthermore, we considered the Receiver Operating Characteristic (ROC) curve, which is a familiar tool to estimate the performance of binary classifiers and Area Under Curve (AUC) [106] to understand the stability of the model. Table 3.1 indicates the relationship between the actual and predicted classes. Accuracy represents the ratio of TP and TN to the overall number of samples. Sensitivity (recall) shows the ability of the classifier to identify all positive samples in the actual class. Precision indicates the accuracy of positive prediction. The F1 score is affected by precision and recall, where the

best score is 1.0.

Table 3.1 Relationship between actual and predicted classes.

		Predicted Class	
		Positive Class	Negative Class
Actual Class	Positive Class	TP	FP
	Negative Class	FN	TN

where TP = True Positive, FP = False Positive[1]
FN = False Negative and TN = True Negative[2]

$$Accuracy = \frac{(TP + TN)}{TP + TN + FN + FP} \tag{3.7}$$

$$Sensitivity(Recall) = \frac{(TP)}{TP + FN} \tag{3.8}$$

$$Precision = \frac{(TP)}{TP + FP} \tag{3.9}$$

$$F1score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{3.10}$$

Area Under Curve(AUC), For a predictor *f*, an unbiased estimator of its AUC: tests whether positives are ranked higher than negatives

$$AUC(f) = \frac{\sum t_0 \quad \in D^0 \sum t_1 \quad \in D^1 \; 1\left[f\left(t_0\right) < f(t_1)\right]}{\left|D^0\right| . \left|D^1\right|}, \tag{3.11}$$

where $1[f(t_0) < f(t_1)]$ notes an indicator function which returns 1 if $f(t_0) < f(t_1)$ otherwise return 0, D^0 is the set of negative examples, and D^1 is the set of positive examples.

3.4 Experimental Results and Performance Evaluation

3.4.1 Dataset

We used a real VANET dataset with DSRC-based communications between vehicles and between vehicles and roadside units in a realistic highway scenario [107].The experiments were performed in the northwest sector of Atlanta, GA along I-75 between Exit 250 and Exit 255. The selected area of the highway has five regular lanes and one High Occupancy Vehicle (HOV) lane that has been monitored during the day between 2 pm and 5 pm. This can be

representative of most roads in the U.S. cities [108]. The data were acquired from GPS in 822.11 ad hoc networks. The GPS reported features such as location, longitude, latitude, speed and heading of the vehicles every two seconds. The accuracy of the location information recorded by interpolation was approximately five to seven meters. Moreover, IPerf was employed cooperatively with GPS reading network parameters. The V2V communication was measured based on the following vehicles, and both the sender and receiver were placed in vehicles that were moving in the same lane. The V2R communication was measured for moving vehicles, and the RSU station was the receiver, which was located on an elevated bridge with different heights. The sender was placed in the vehicle, and it broadcasted the packets while moving in the rightmost lane. The number of packets in V2R communication was 1470 bytes, which were broadcasted by the senders at an approximate rate of 150 packets/s [108]. All communication features, such as "log time", "location information of both sender and receiver", "velocity", "packet sent/received", and "signal quality", were associated and parsed together and were recorded. When the number of sending and receiving packets through VANET communications increases (i.e., many vehicular users on the road), traffic occurs in the network [8], [4]. We obtained 39,998 records of VANET communication data that combined V2V and V2I datasets. We solve our problem, which is intelligent network traffic prediction as a classification task. We take full advantage of all effective features of the real VANET dataset for traffic prediction using the LightGBM and Boruta methods. Therefore, the target is network traffic prediction, and we consider packet receiving as a network parameter to predict the network traffic. The target corresponds to the binary class (0: no traffic, 1: traffic).

3.4.2 Experimental Details

For the implementation of the model proposed in this paper, the following modeling environments were used. Jupyter Notebooks is an open-source and browser-based tool. It can work both locally and on the cloud [109]. Google Collaboratory is "a product from Google Research" [53], which is hosted on the Google Cloud Platform and is based on Jupyter Notebook. It is appropriate for ML and data analysis by providing fundamental AI libraries such as TensorFlow, Matplotlib, and Keres. It allows to write, execute and share python code via browser with others [110]. We used Google Colab to implement the model, and the programming language was Python version (3.7.13). Since the values of the dataset vary in unit and range, we normalized the data with the aim of bringing them into the same range for an accurate prediction model. Furthermore, for data visualization and analysis, we employ some well-known libraries, such as NumPy (fundamental computation), Pandas (data analysis), Scikit-learn [111], for scaling the features and data partitioning, and Matplotlib

(visualization).

3.4.3 Performance Evaluation of the Proposed Model

There are three steps from the loaded dataset to model validation that will be described as follows.

1. In the first step, after importing the data into Google Colab [53] and reading the data, we preprocessed the data (i.e., cleansing redundant data and removing space) and checked for missing values in variables. Then, we put the feature variables to X and the target variable to y. Next, we scale the features using Scikit-Learn libraries. At the end of this step, the data were split using Scikit-Learn [111], in which 70% of the data were considered for training, while the remaining 30% were used as test data. The 10-fold cross-validation method was used in our model for parameter optimization by tuning the hyperparameters and configuration of the model, which eventually led to a boost in the performance of the model [54], [55].

2. In the second step, considering most related features in our dataset, several popular ML algorithms, which have been used for traffic prediction, were taken into account including RF, KNN, NB, DT, SVM and MLP.

3. In the last step, we trained all abovementioned models, and then we evaluated and analyzed the performance of each prediction model from the literature and our proposed model. We compared them in terms of classification metrics, which will be analyzed in the following section. Furthermore, we highlighted the comparative analysis results of popular ML models and our proposed model.

Feature Selection Results

In this section, we extracted the importance of features, their relevancy and how they affected the prediction of the target (i.e., network traffic prediction). We considered the V2V and V2I datasets separately and together to determine which variable needs to be kept in our dataset. First, it is worth mentioning that in all individual and merged datasets, time remained a highly important feature. Second, an interesting correlation was found in V2V data, where sender speed and receiver speed were placed in the almost same degree of importance variable. The significance of sender speed was much higher in the V2I dataset and ranked third after time and sender location. Turning now to the V2V and V2I as combined data, the highest

value of importance belongs to the sender and receiver location and then the sender and receiver speed.

We plotted the feature importance for the V2V and V2I datasets separately and together using lightGBM, as shown in Figure 6. The abovementioned analysis, the provided plots, and eventually the features confirmed by Boruta (i.e., relevant features with a ranking of one) were taken into account. In conclusion, we selected as many variables as possible that are important, efficient and confirmed by these methods including time, sender location, sender and receiver speed, and receiver location.

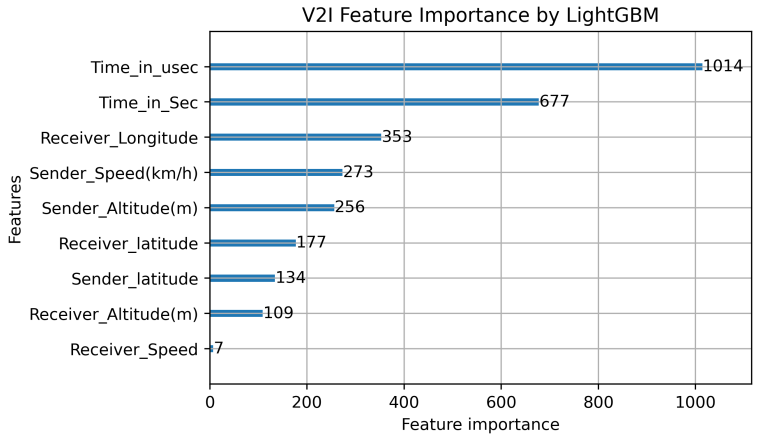

Figure 3.6 Feature importance results. (a) V2I dataset using LightGBM.

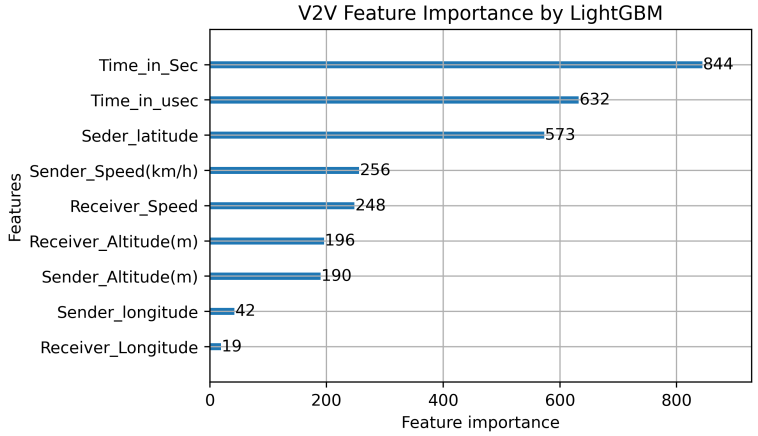

Figure 3.7 Feature importance results. (b) V2V datasets using LightGBM.

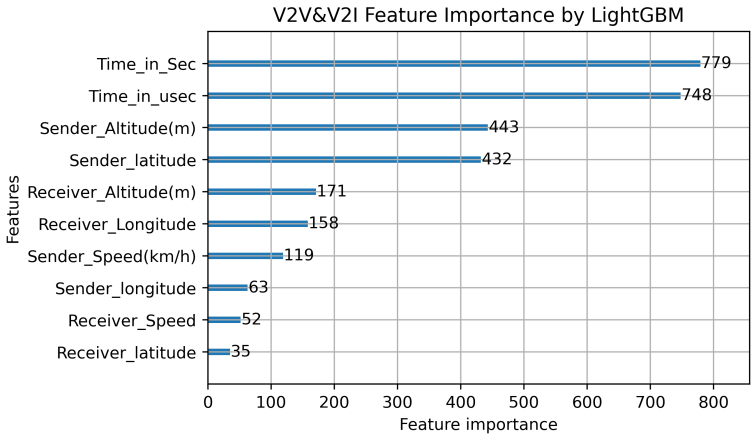

Figure 3.8 Feature importance results. (C) V2I and V2V datasets using LightGBM.

Classification Results

Based on the literature review, some ML models may be a good fit for a particular problem. However, each model can fail in different ways. Therefore, we find the best model that fits our dataset to solve the problem of network traffic prediction. We used the confusion matrix, classification report, ROC curve and CPU time as the classification metrics [**?**] to evaluate the performance of the most commonly used models for traffic prediction. We first applied a "dummy" classifier, which is a simple ML model that randomly makes predictions by considering the class distribution of the training set [112]. We obtained 0.668 accuracy, which confirms the importance of considering the feature variable as our input. In addition, it improves our prediction results. In the following paragraphs, we compared well-known popular ML models with our proposed models using the abovementioned metrics. Regarding Table 1, the relationship between the actual and predicted classes is presented [58]. There are four states in the confusion matrix. Therefore, the confusion matrix can give us insight into the probability that a model is confused when it produces the prediction results. The classification reports consist of precision, recall F1 score and accuracy that complement each other to evaluate the classification models. Precision is related to True Positive (TP) and False Positive (FP) states, which measure only positive prediction, and in our case, they are related to correctly predicted existence of traffic or incorrectly predicted traffic situations.

Accordingly, this metric ignores the negative states and must be coupled with recall, which considers True Positive (TP) and False Negative (FN) in which we receive the wrong alarm about the existence of traffic in the network and bring us unexpected decision results. In this paper, FN means no traffic as a prediction result while the network is in a traffic situation,

and FP gives us traffic as a prediction result while the network is in a no-traffic situation. Both FN and FP cause error and incorrect traffic prediction, and the F1 score is a weighted average score of recall and precision. In our analysis result, the predictive model with a high recall value means that the majority of the existence of traffic in the network is predicted correctly, and there is a slight probability of incorrect prediction. A model that can give us good accuracy while consuming more time for prediction results cannot be applicable in the case of time-sensitive tasks such as traffic prediction. Therefore, we tried to maintain a balance between accuracy and time consumption in our model.

Furthermore, the ROC for estimating the performance of binary classifiers and Area Under Curve (AUC) [58] for evaluating the stability of the model can be considered. When the ROC curve is distant from the middle-dotted line and converges to the upper left side and the AUC value is close to one, this implies an ideal model. Ultimately, all discussed issues of the results are provided in Figures 6-8, and Table 3 will be analyzed.

Figure 7 provides the results obtained from the confusion matrix performance of NB, RF, DT, KNN, SVM and MLP. Considering the abovementioned points, an unpleasant outcome that came from the incorrect prediction of traffic with both FN and FP resulted in a negative impact on vehicular network users on the road and a decline in the quality of service.

Regarding measuring the error rate from the confusion matrix, the best FN belonging to RF was 0.02%. However, the worst value of FP is also associated with this model (9.03%). On the other hand, KNN is best in FP, which was 4.15%, but it was not sufficiently good for TP. Comparing the different model results clearly shows that the proposed model maintained a good balance between these factors, which are 4.88% for FP and 1.00% assigned to FN, as well as a better rate of correct prediction (TP) than KNN.

Based on previous studies, SVM and MLP are mostly considered for traffic prediction. SVM can adapt to the dynamic characteristics of traffic. Moreover, the important drawback of this model is related to selecting an adequate kernel and parameter. Thus, in Table 2, we made a comparison of the SVM performance with different types of kernels, and the obtained results confirmed that SVM with the polynomial kernel shows more accuracy than the other types with slightly lower CPU time. Therefore, we considered the best kernel and optimized SVM. Finally, we compared the accuracy of this optimized SVM model with our proposed model.

Additionally, the MLP model is the classical type of feed-forward neural network [23], [80]. It consists of three types of layers: the input layer, the output layer and the hidden layer. Since our dataset is not sufficiently large to try on the different DL models, we decided to perform the simple neural network model, which is popular for the traffic prediction domain [22]. Figure 7 shows that the percentages of FN and FP for the MLP classifier were not noticeable

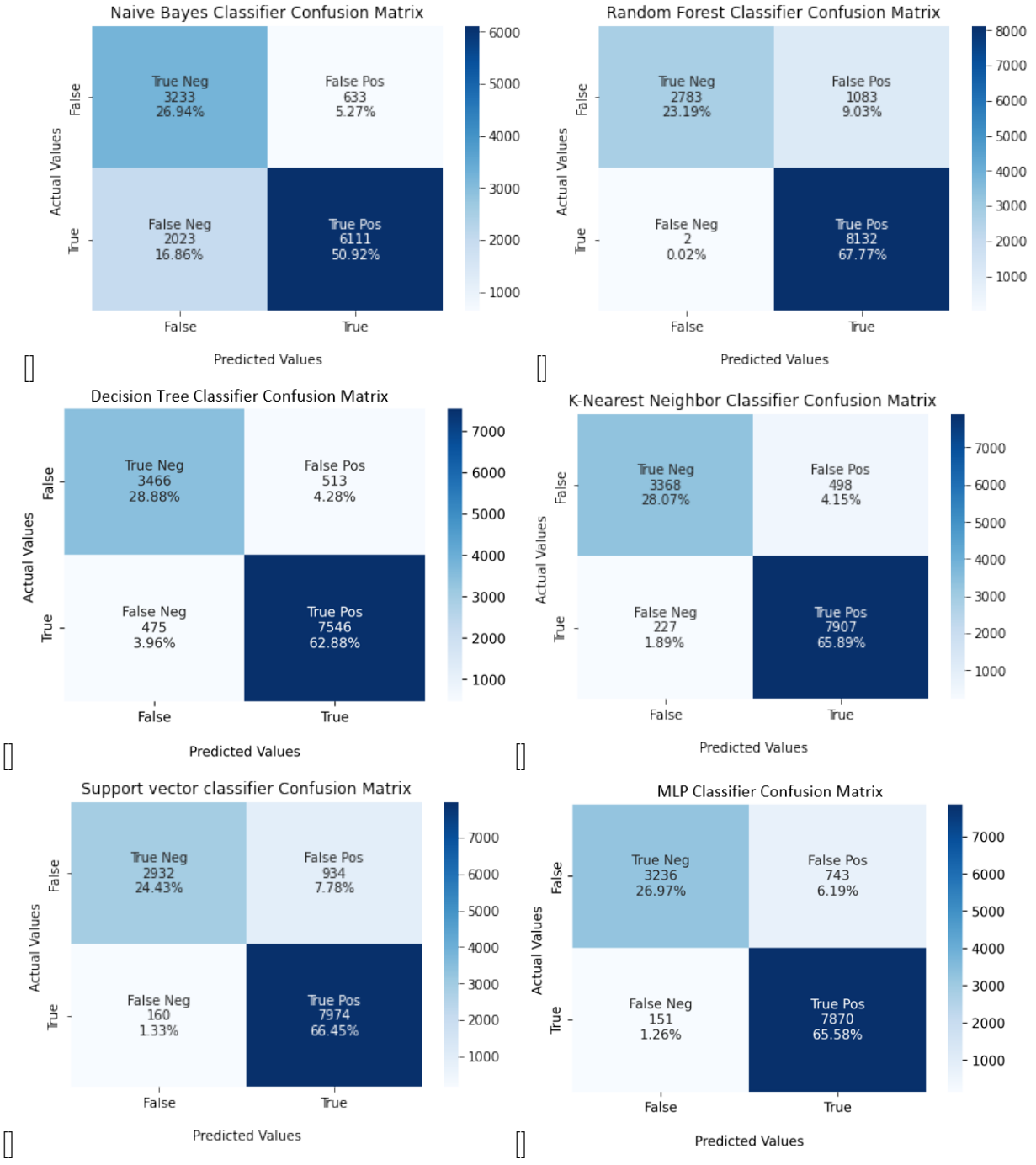

Figure 3.9 Confusion matrix of classification performance by (a) Naive Bayes Classifier, (b) Random Forest Classifier, (c) Decision Tree Classifier, (d) K-Nearest Neighbor Classifier, (e) Support Vector Classifier, and (f) MLP Classifier.

Table 3.2 Comparison of the performance of SVM with different types of kernels

Support vector machine classifier	**Accuracy**	Time(ms)
SVM(Linear Kernel)	0.907	8.11
SVM(Sigmoid Kernel)	0.851	7.63
SVM(RBF Kernel)	0.582	5.72
SVM(Polynomial Kernel)	0.910	6.68

among the other models and our proposed model.

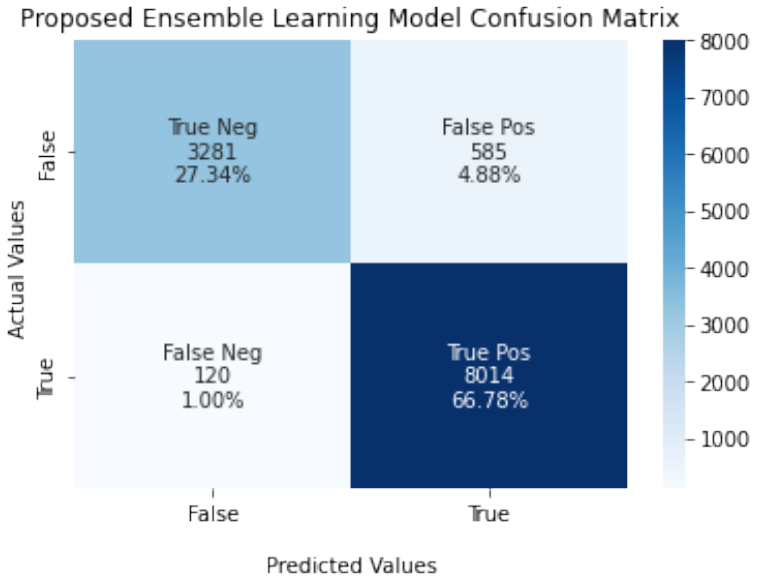

Figure 3.10 Confusion matrix of classification performance by the proposed ensemble learning model.

As we explained in the methodology section, our model consists of two layers: base learners and meta-learners. In the first layer of the stacking ensemble of models, we selected RF, KNN and XGBoost. Each of them can help us to cover an issue. For example, RF can solve the challenges related to scalability since it has the ability to train models in parallel [51]. KNN was selected to obtain better performance than the other ML models. Finally, XGBoost, as an efficient and scalable implementation of the Gradient Boosting Machine (GBM) was selected to act as a booster to our base learner results. Considering the distributed computing and parallel learning ability of this model, XGBoost will not impose extra time for the prediction result. However, it enables higher prediction accuracy and can make our model more precise. Moreover, this model can handle the challenge in DT, which is related to easy overfitting.

Eventually, XGBoost enhances the generalization ability [113]. Therefore, we considered a model that can take advantage of all these points and use the best effective combination of these models. In the second layer, which is the meta learner, we used logistic regression. In addition, we made some improvements at this layer, such as hyperparameter tuning, and we built our meta learner using the grid search algorithm. It helps us to select the best configuration of model parameters, which leads to maximizing the model performance.

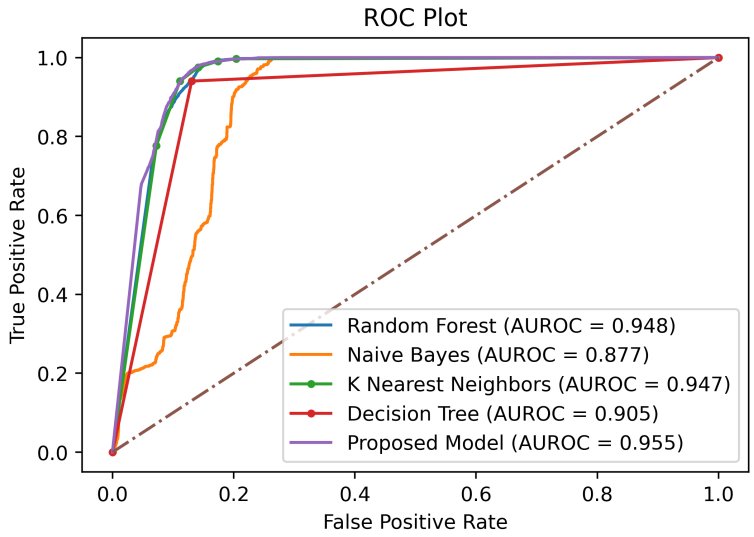

Figure 3.11 Comparison the ROC Curve with different ML models and the baseline models and our proposed model.

In this section, the confusion matrix of our proposed model is presented in Figure 8. In Figure 9, the results obtained by the ROC curve and AUC values were compared, which can help us to understand the stability of a classification model. When the ROC curve is distant from the middle-dotted line and converges to the upper left side and the AUC value is close to one, this implies an ideal model [58]. Considering this analysis, the proposed model showed better performance and stability than the other ML models in distinguishing between positive and negative classes. The AUROC value for our proposed model was 0.955, which is the highest among the other models.

Furthermore, we added a booster in the first layer of our stacking ensemble model. The comparison results of the AUROC value and curve with and without the booster are presented in Figure 10. The computed AUC value is 0.948 without using XGBoost as a booster and 0.955 with considering a booster. It confirmed obtaining better results with the booster.

XGBoost is known as a high-power predictive model for increasing the efficiency of the model.

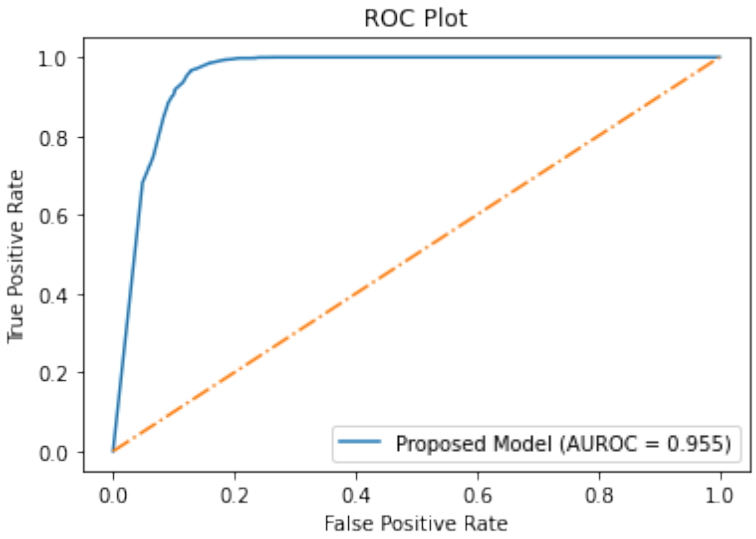

Figure 3.12 ROC curve of the proposed STK–HEM model (a) Using a booster.

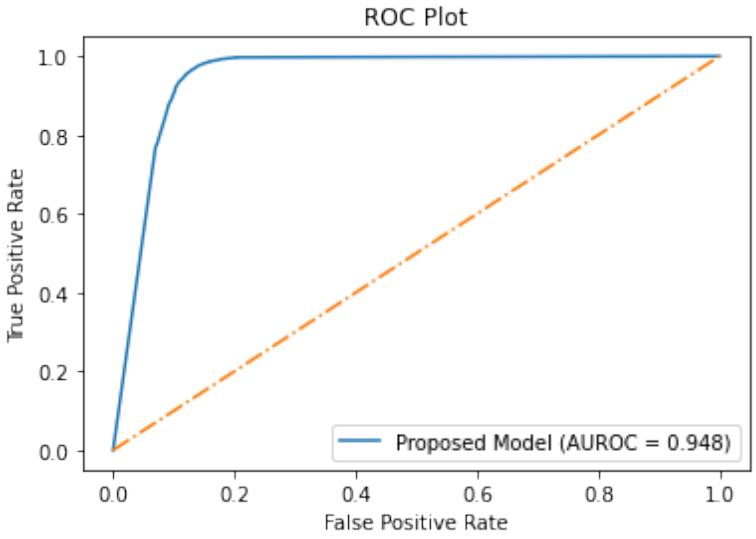

Figure 3.13 ROC curve of the proposed STK–HEM Model (b) Without using a booster.

In Figure 11, the training and testing accuracies of XGBoost as our booster algorithm are shown. The training accuracy is 0.9426, and the test accuracy is 0.9405, which are close.

In the second layer, logistic regression is employed as our meta learner. We performed parameter optimization using a grid search algorithm combined with cross-validation (CV) called Grid Search CV [54], [55], in which grid search was used for parameter tuning. It

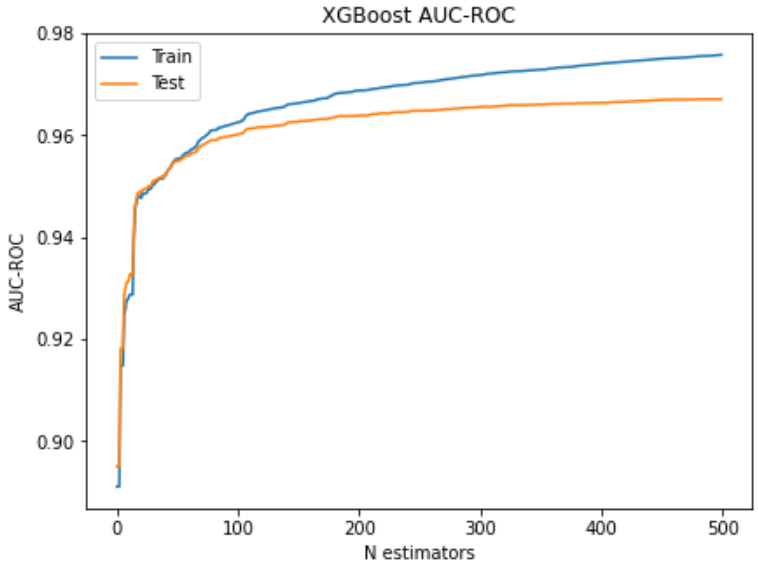

Figure 3.14 XGBoost as a booster in the first level of the proposed model. (a) AUC-ROC curve.

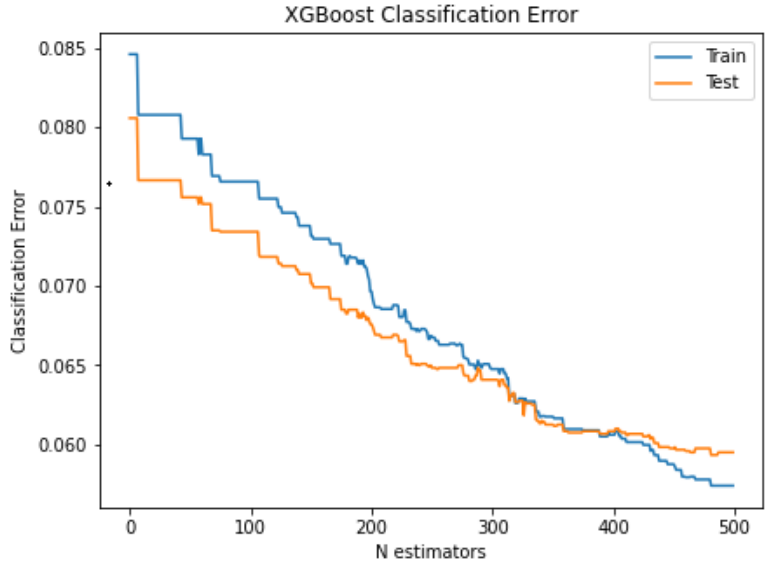

Figure 3.15 XGBoost as a booster in the first level of the proposed model(b) classification error.

considers each fusion of algorithm parameters specified in a grid. It can help to boost the performance of the model. Additionally, because CV performs oversampling, there is a special algorithm known as group k-fold cross-validation. It ensures that the same data point is not present in both testing and training. This helps us avoid overfitting.

In addition, we plotted the learning curve of the meta learner, as shown in Figure 12. The learner has mastered the learning task with a high validation loss at the beginning. However, after 10,000 training samples were trained and validated, the data converged together and stayed close to each other with a minimal gap until the end. This means that our meta-model was well-fitted. Furthermore, the learning curve can be applied as a mechanism to diagnose the machine learning model bias-variance problem, and it is possible to see the trade-off between bias-variance with our chosen super estimator model.

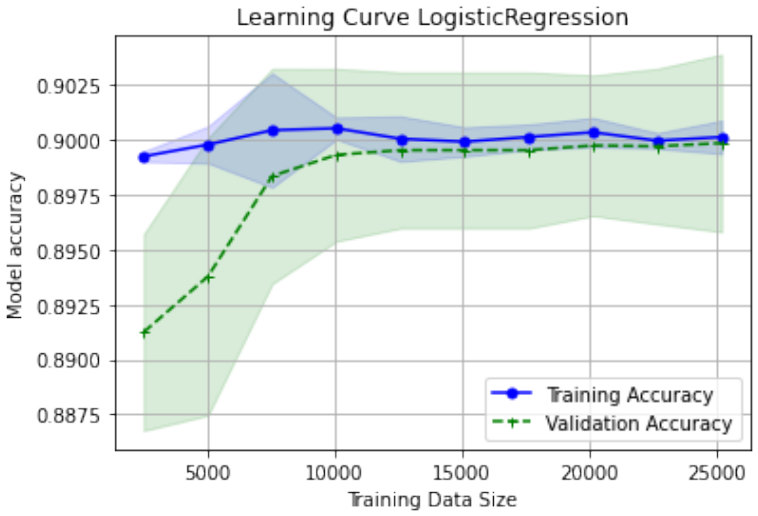

Figure 3.16 Learning curve super learner.

In summary, we built a stacking ensemble model with a booster for network traffic prediction problems. We focus on increasing overall efficiency. Moreover, the main part of our research work was selecting the best combination of the algorithm and model when constructing our model from scratch. We made many considerations to ensure that our model works well for each dataset (V2V)(V2I) and for the combination of them. Of note, because the time may vary by each execution, we considered an average value of 10 runs of the model. Moreover, the model gave us even better results for the vehicle-to-vehicle dataset, which means that when the sender and receiver were both on the move, the predictive model was stable and performed well. Finally, the proposed experimental analysis indicated that our proposed model was the winner considering every aspect. Additionally, the CPU time for training the model was similar to NB as the faster learner classifier in the baseline model. The performance analysis of the different prediction models and our proposed model with classification metrics for V2V and V2I are represented in Table 3 and Figure 13.

In summary, the efficiency of ML models relies on the size of datasets, the selected features

Table 3.3 Comparison of the performance of SVM with different types of kernels

Prediction Models	**Precision**	**Recall**	**F1_Score**	**Accuracy**
Random Forest	0.94	0.86	0.89	0.909
K-Nearest Neighbor	0.94	0.92	0.93	0.936
Naive Bayes	0.76	0.79	0.77	0.777
Decision Tree	0.91	0.91	0.91	0.917
Support Vector Machines	0.92	0.87	0.89	0.908
Multilayer Perceptron (MLP)	0.94	0.90	0.92	0.925
Proposed Model	0.94	0.98	0.95	0.941

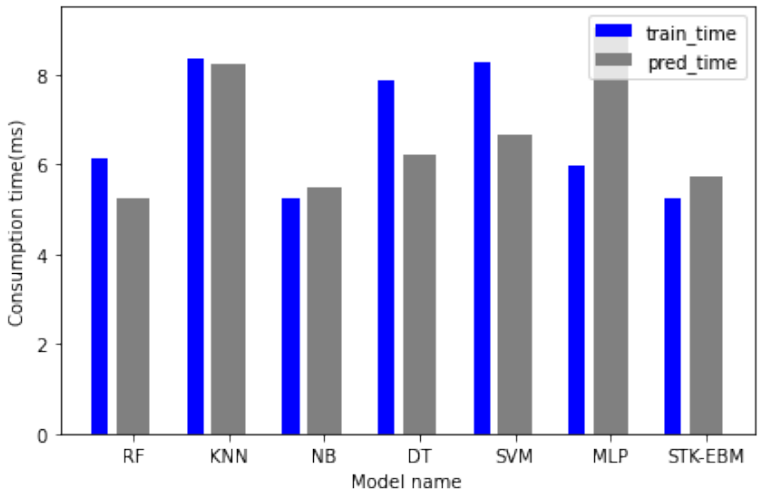

Figure 3.17 Comparison of the time consumption of different ML models and our proposed model.

and the type of problems. Subsequently, we identified the best-fitted ensemble model for network traffic prediction in VANETs. Our proposed model not only obtained a balance considering several aspects (including accuracy, error and time) but also showed improvement and better results.

Comparative analysis of popular ML models for traffic prediction with the proposed model

In the experimental results, most of the common machine learning models are tested and compared, which can fully demonstrate the advantages of the proposed model. In this section, we highlight the results of the comparative analysis by considering all evaluated metrics. Regarding the confusion matrix, the proposed model maintained a good balance between

both FN and FP that caused the error and incorrect traffic prediction. The other ML models showed best, worst, or not sufficiently good results in one of these factors. The ROC curve and AUC values can be considered to understand the stability of a classification model. Our model showed better performance and stability than the other ML models in distinguishing between positive and negative classes. Moreover, the AUROC value was the highest among the other models, which indicates that our model was an ideal model and winner in this metric. The accuracy of the proposed model was higher than that of the different tested ML models, which is highly important in our problem. The incorrect prediction of traffic in a network causes an error and undesirable decision in a real application. Our model also provided the highest recall value among other models, which means that the majority of the existence of traffic in the network was predicted correctly, and there is a slight probability of incorrect prediction. Furthermore, the F1 score as a weighted average score of the recall and precision also in our model was higher than that in the best standalone ML models (e.g., KNN, MLP). However, a model that can give us good accuracy while consuming more time for prediction results cannot be applicable to time-sensitive tasks such as traffic prediction. Therefore, we also consider the consumption time for prediction, and the result of the proposed model was almost near the best one, which is NB. Finally, the proposed model considers a trade-off between all popular ML models. It enhances the overall performance by increasing accuracy with minimum time and providing stability in the results.

Ultimately, a more accurate and stable prediction of traffic in the network can help to identify the failure of the network and its dependent services. It can effectively help us to predict traffic in the network and mitigate it before declining the quality of services for the users. Specifically, vehicular networks that are related to important services such as preventing congestion and accidents for road users will be more essential. Ensemble learning techniques have been investigated for decades but still attract all domain researchers with valuable advantages in different learning tasks. Of note, some challenges must be taken into consideration. For instance, identifying the best set of base learners to integrate for a given problem without much trial and error is time-consuming. However, EL provides an efficient tool to extract highly accurate and robust models, especially from dynamic, noisy, and heterogeneous data, bringing notable benefits to real-world applications. For example, traffic prediction.

3.5 Conclusion

In this paper, we proposed an ensemble-based ML model for forecasting network traffic in VANETs. We compared various most commonly used ML models that were suitable for forecasting traffic and used them in our proposed model. Moreover, to effectively evaluate

the performance, we considered different classification metrics including confusion matrix, ROC-AUC curve, accuracy, precision, recall, F1 score, and time. Then, we discussed how we used the stacking ensemble strategy for building a best-fitted model for designing an efficient network prediction model. The proposed stacking ensemble boosted model (STK–EBM), enhances the overall efficiency in all metrics and obtains stable prediction using the integration of RF, KNN, XGBoost, and LR.

The limitation of the presented model is related to DSRC access technology, which is just for short-range coverage. Additionally, there is just a basic type of communication between vehicles and roadside units. However, in a practical application, we will have various types of communication, such as vehicle to pedestrian and vehicle to a cellular network that is called vehicle to everything (V2X). Therefore, we need to provide more communication types with different access technologies, such as LTE/5G, that provide large-range coverage. In this way, we can obtain a better perception of the dynamic nature of traffic for such a prediction model in practical applications. In future work, we plan to work on different technologies, such as vehicular communication in cellular networks and V2X communications in VANETs, to design an efficient prediction model.

CHAPTER 4 ARTICLE 2: A SOFT VOTING CLASSIFICATION MODEL FOR NETWORK TRAFFIC PREDICTION IN VANET/V2X

Authors: Parvin Ahmadi Doval Amiri, Samuel Pierre

Status: register submit in International Conference on Wireless Mobile Computing, Networking and Communications(WiMob), April 28, 2023.

Abstract

Vehicular network services in the smart cities generate enormous data by vehicular road users, which is a critical challenge. Network traffic leads to a negative impact on safety applications. AI techniques are a promising solution to address network traffic in VANETs with V2X data. In this paper, we propose a soft voting classification model, which consists of hybrid supervised machine learning algorithms to predict traffic in the network. We evaluate the prediction performance of five well-known machine learning models and the proposed model based on various classification evaluation metrics. The simulation results show that the proposed network traffic prediction model performs better than other considered machine learning models in terms of accuracy (0.94%), time consumption (12.25 seconds) and AUROC (0.907) that proves its stability.

Keywords: Vehicular ad-hoc networks, artificial intelligence, ensemble learning, network traffic prediction, machine learning, intelligent transportation system.

4.1 Introduction

Intelligent Transportation Systems (ITS), which are considered as future of transportation systems in smart cities and Vehicular Ad-hoc Networks (VANETs), play an important role in such systems [3]. In a vehicular network, data is produced by vehicles and can be transmitted through basic communications [5]: Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I). VANETs provides various services that take advantage of the data collected from the mentioned communication, and the typical use of services is related to traffic management [11]. The evolution of VANETs to IoV opens new opportunities and challenges in this domain, such as advanced communication known as Vehicle-to-everything (V2X) that can address traffic challenges with the aim of Artificial Intelligence (AI) techniques. V2X is composed of a variety of communication, such as Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Infrastructure-to-Infrastructure (I2I) and Vehicle-to-Pedestrian (V2P)

[8]. Furthermore, Dedicated Short-Range Communication (DSRC), which is based on IEEE 802.11p standard, is a long-lasting and continued candidate for V2X communication [114]. However, the integration of VANETs with AI methods and its adoption within V2X, is a challenging task [7]. Several studies applied AI techniques in vehicular networks to provide a more accurate and efficient intelligent traffic prediction model. This is related to a critical issue in network traffic, especially in safety vehicular network applications [46].

In this paper, we propose a soft voting strategy that can enhance the performance of the single Machine Learning (ML) models. We consider V2X simulation data for network traffic prediction in VANETs. Moreover, we implement five well-known ML models consisting of Random Forest (RF), k-Nearest Neighbor (KNN), Naive Bayes (NB), Support Vector Machine (SVM) and Decision Three (DT), which are supervised ML models and commonly used for network traffic prediction [4]. Additionally, we compare the obtained results of the proposed model with the abovementioned popular models based on different classification metrics (i.e., precision, recall, F1 score, accuracy and time).

The main contribution of this paper is to show how soft voting method can improve the performance of ML techniques applied to vehicular networks regarding accuracy and time. Furthermore, different from the existing research, the traffic data is generated from the fusion of various communications (V2X) in VANETs by employing the cooperation of simulation tools for traffic modelling and wireless communications [115], [116], [117], [118]. Finally, the performance comparison of five commonly used ML models and the proposed model is presented.

This paper is organized as follows. Section II presents a brief background and related work. Section III explains the methodology and prediction method. The simulation scenario and evaluation results are discussed in Section IV. Finally, Section V concludes the paper.

4.2 Background and Related Work

Network traffic prediction as a significant challenge in the vehicular network can affect the operation of the network, which is highly critical especially when it comes to VANETs safety application. Several authors from the literature have applied the most popular ML models for network traffic prediction.

Meena *et al.* [78] used V2V traffic data to predict traffic flows. They applied RT, DT and SVM models and evaluated their performance based on classification metrics (i.e., precision, recall. F1 score, accuracy and time). The simulation results show a higher accuracy dedicated to RF while consuming a long time for execution.

Sepasgozar *et al.* [19] used V2R communication data in VANETs for network traffic prediction by employing five well-known AI models consisting of KNN, RF, SVM, NB and MLP. The performance of these models is evaluated based on classification metrics including Receiver Operating Curve (ROC), Precision-Recall (PR) curve, accuracy, precision, recall, F1 score and time. The comparative results show that RF outperformed the other models in all considered metrics with acceptable results. However, KNN is the most performed one with a high accuracy of 97 % and a long-consumed execution time.

Stepanov *et al.* [54] presented an LTE network traffic prediction method by using three ML techniques including SVM, RF and bagging. They considered the public cellular traffic dataset. They employed error metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of determination (R2) for evaluation. The results indicate that SVM performed better in terms of training time than RF and bagging. However, the bagging algorithm showed the lowest error with the highest training time. Moreover, Deep Learning (DL) models are applied in the case of large and complex traffic data in several related works. Ramakrishnan *et al.* [119] presented a network traffic prediction model by considering DL techniques including RNN, LASTM and GRU. They considered error evaluation metrics for analyzing the performance of the results, such as Mean Square Error (MSE), and the simulation results show less error when using single Dl model. In [83], the authors considered big traffic data with 500,000 recorded samples and proposed a hybrid DL model. This proposed model shows less error in prediction results than the well-known single DL models, such as ARIMA, LSTM, GRU and SAE.

Therefore, network traffic prediction is a major issue in vehicular networks because of non-linear traffic data and dynamic topology in VANETs. However, V2X communication data can help us covering traffic challenges by integrating it with AI-techniques to find an efficient AI method.

4.3 Methodology and Prediction Techniques

Vehicular communications in VANETs play a critical role in safety applications. In this paper, we use the Simulation of Urban Mobility (SUMO) and OMNet++ simulation tools [116], [117], [118], which have been widely used by researchers, to simulate the traffic data and validate the proposed model for network traffic prediction. We collect the fusion of data from Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Infrastructure-to-Infrastructure (I2I), and Vehicle-to-Pedestrian (V2P) communication, and Roadside Unit (RSU) to support the Vehicle-to-Infrastructure (V2I) communication and Infrastructure-to-Infrastructure (I2I). We consider the packet delivery ratio as a target, which is an important metric in network traf-

fic prediction and control studies [19], [21]. We perform some preprocessing steps, such as cleaning the missing data and data normalization, to provide high-quality input to feed to ML models. Finally, we divide our dataset into training and test sets.

The output is the probability with a range between one and zero ($0 \leq f(z) \leq 1$). This function transforms probabilities into a binary values. where, $z < 0.5$ *output* $\rightarrow 0$ *(no-traffic)*, else ($z \geq 0.5$) *output* $\rightarrow 1$ *(traffic)*.

4.3.1 Selecting Machine Learning Techniques

In this section, we discuss the five well-known ML techniques to choose the most performed one in terms of accuracy and execution time.

- Random Forest (RF): It is a supervised learning model that can be used in both classification and regression problems. This model trains decision trees on different samples in parallel. Moreover, it can increase the accuracy and decrease the overfitting in decision tree model [78], [19], [54], [120].
- K-Nearest Neighbors (KNN): This supervised learning model searches for nearest neighbors, then defines the classes used by those neighbors, and finally, the distance metric selects the k nearest neighbors [16], [23].
- Decision Tree (DT): This learning algorithm creates a model that predicts values of the target variable by learning simple decision rules acquired from features [4].
- Support Vector Machines (SVM): This model can predict the output by learning from the past input. It can efficiently handle non-linear data and can adapt to dynamic traffic data. However, it has long training time [7], [46], [4], [19], [23].
- Naive Bayes (NB): This model is a simple probabilistic classifier uses the Bayes theorem and considers features with naive correlation [20]. It can provide reliable and stable prediction and consume less time for execution than other models. However, it has a long training time [19], [21].

We selected RF, KNN and Logistic Regression (LR) since each of them can bring advantages to the final prediction model in an efficient way. RF can deal with overfitting, which is the problem of DT, besides its ability to train models in parallel and provide scalability and robustness [78], [19], [54], [120]. KNN performs better than SVM which has the problem of selecting the kernel type and long training time [7], [46], [4], [19], [16], [23]. Moreover, LR can

help to find the optimal prediction outcome of merging the different models without imposing more computation time and complexity because of its efficient computation [104]. Finally, we use a soft voting strategy as the classification model to enhance the performance of the proposed prediction model [121]. Fig. 1 depicts the proposed methodology, which consists of four modules. In the following paragraphs, we briefly describe the important consideration in each module.

In the first module, we obtain the V2X communication data from the integration of several types of communication including Vehicle-to-Vehicle (V2V), Vehicle-to-Roadside Unit (V2R), Vehicle-to-Pedestrian (V2P), and Roadside Unit (RSU) to support the Vehicle-to-Infrastructure (V2I) communication. This is called V2X communication [8]. In the second module, we preprocess the data to feed into the third module related to the proposed model as described in the following four steps.

Step 1) The input variable is input into x to represent $x = x_1, x_2, ..., x_k$, is the set of the k input features, where $x_k \in [0, 1]$, and the target variable in (Y) that is labeled 0 for no traffic and 1 for traffic. We split our dataset (D) into training and test sets.

Step 2) In the proposed model, there are three combined classifiers consisting of RF, KNN and LR that learn from the training set, where $c = c_1, c_2, ..., c_m$ is the set of the (m) classifier models.

Step 3) In this step, all base-classifiers predict the result based on generating class probability output, where $p = p_1, p_2, ..., p_n$ is the prediction output of the individual models.

Step 4) The final output is the combination of classifier predictions estimated by the soft voting classifier is defined [**?**] as follows.

$$Y = (argmax \sum_{j=1}^{m} w_j p_{ij}), \tag{4.1}$$

where m is the number of classifier, w_j= the assigned weight for jth classifier prediction, p_{ij}= prediction input into the summing function.

Furthermore, we consider five well-known ML techniques including RF, KNN, NB, DT and SVM to compare their performance in the third module. Finally, in the result analysis module, the classification evaluation metrics are applied to each ML model and the model that can achieve acceptable results in all considered metrics (i.e., precision, recall, F1 score, accuracy and time) will be selected as the best model. We propose a soft voting ensemble model that can balance the individual learning model by combining the prediction results of multiple models in a simultaneous way.

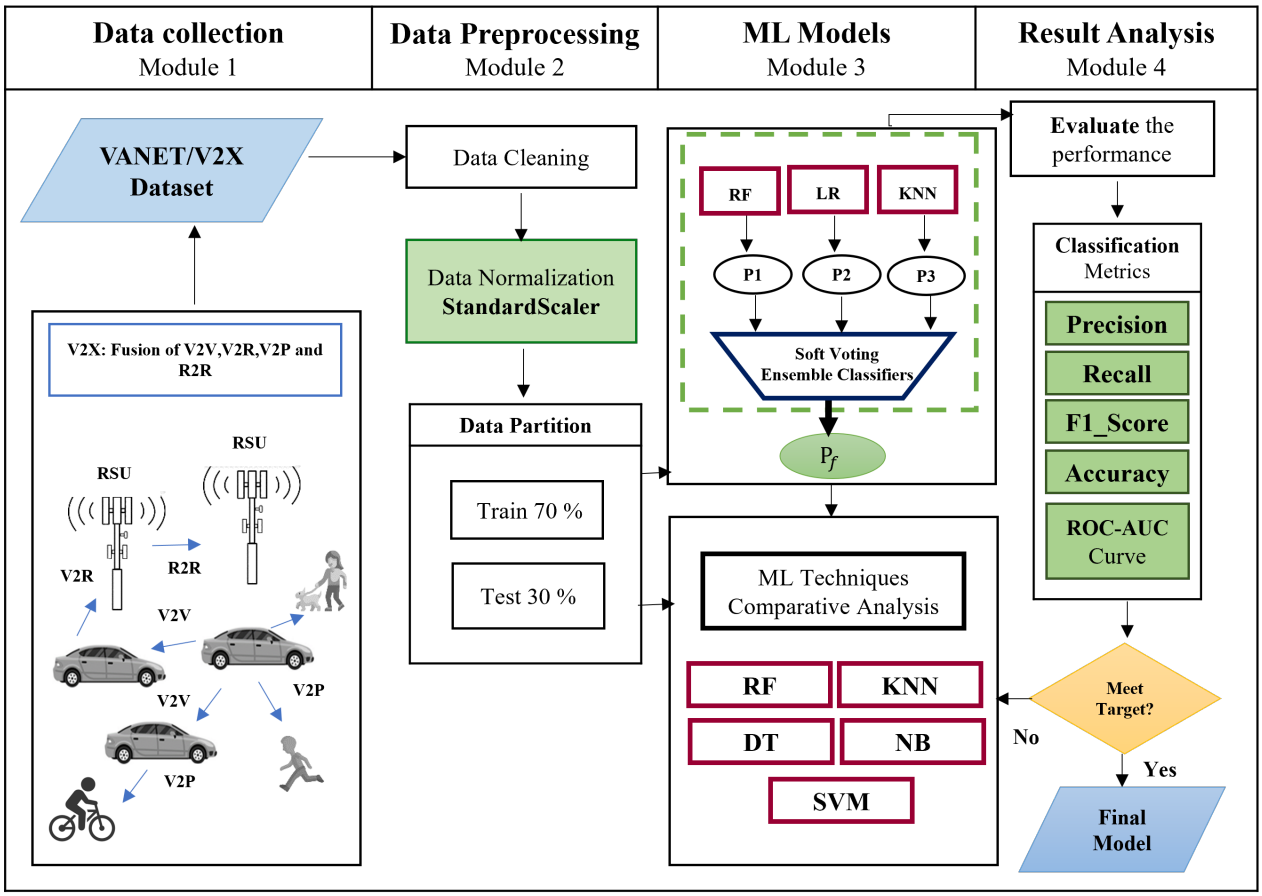

Figure 4.1 Overview of the proposed methodology for traffic prediction in VANET.

4.3.2 Evaluation and Analysis Metrics

We used four classification metrics to evaluate the efficiency of the proposed model: the accuracy, the classification report, ROC curve and AUC value and the consumption time.

Table I indicates the relationship between the actual and predicted classes. In case our target is in a traffic situation, but the predicted result incorrectly indicates a non-traffic situation, which is called a False Negative (FN), this leads to undesirable consequences for vehicular road users. Additionally, if the target is in a non-traffic state but the predicted result shows a traffic state, which is called a False Positive (FP), this can be a fault in the performance of the proposed model. However, it will not bring a negative impact on vehicular road users. Moreover, Receiver Operating Characteristic (ROC) curve is considered to determine the performance of the binary classification, and the stability of the model is estimated by Area Under Curve (AUC) [58].

Table 4.1 Relationship between actual and predicted classes.

		Predicted Class	
		Positive Class	Negative Class
Actual Class	Positive Class	TP	FP
	Negative Class	FN	TN

where TP = True Positive, FP = False Positive[1]
FN = False Negative and TN = True Negative[2]

$$Accuracy = \frac{(TP + TN)}{TP + TN + FN + FP} \tag{4.2}$$

$$Sensitivity(Recall) = \frac{(TP)}{TP + FN} \tag{4.3}$$

$$Precision = \frac{(TP)}{TP + FP} \tag{4.4}$$

$$F1score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{4.5}$$

Area Under Curve(AUC), For a predictor f, an unbiased estimator of its AUC: tests whether positives are ranked higher than negatives

$$AUC(f) = \frac{\sum t_0 \quad \in D^0 \sum t_1 \quad \in D^1 \; 1\left[f\left(t_0\right) < f(t_1)\right]}{|D^0| \, . \, |D^1|}, \tag{4.6}$$

where $1[f(t_0) < f(t_1)]$ notes an indicator function which returns 1 if $f(t_0) < f(t_1)$ otherwise return 0, D^0 is the set of negative examples, and D^1 is the set of positive examples.

Finally, the classification report consists of four evaluation metrics: (1) The accuracy represents the ratio of True Positive (TP) and True Negative (TN) to the overall number of samples; (2) the recall indicates the ability of the classifier to detect all positive samples in the actual class; (3) the precision provides the accuracy of the positive prediction; and (4) the F1 score is calculated by precision and recall while the best score is 1.0.

4.4 Simulation Scenario and Performance Evaluation

4.4.1 Simulation Scenario

We use the cooperation of simulation framework for traffic modeling and simulating wireless communications between vehicles and the other entities in VANETs. Firstly, we extract a part of the downtown map data of Montreal city in Canada from " OpenStreetMap" [115] as a ".osm" file, which is used for a more realistic simulation of vehicular traffic in SUMO [116]. Secondly, we adopt SUMO (1.7.0) to generate urban vehicular traffic. Thirdly, we use the network simulator OMNet++ (5.6.2) with the cooperation of SUMO, Veins (5.2)/INET(4.4.0) [117], [118] to simulate the wireless communication technologies and generate the traffic data. Fourthly, we consider DSRC for exchanging information among vehicles, vehicles and RSUs, and vehicles to pedestrians. Finally, we define an accident to happen at a specific time of running the simulation scenario to generate much load of data in the vehicular network. We perform a 1000 s duration for each run of the simulation. We calculate the Packet Delivery Ratio (PDR) values while the number of vehicles has increased and we assume the values of PDR min as 0.4 and PDR max as 0.6. [122]. The transmitted data within 1000 s of the simulation scenario is considered to measure PDR. Based on the PDR value, each data record with PDR less than 0.3 is labeled by 1 as a traffic condition in the network, and 0 as a non-traffic condition in the network. Table II presents the attribute and parameter values considered in each of the 260 running simulation scenarios.

4.4.2 Experimental Details

In this section, we indicate the implementation environment of the proposed method. We use Google Colab, which is based on Jupyter notebook and hosted on the Google Cloud Platform [53] to implement ML models. We use Python (3.7.13) as our programming language and fundamental AI libraries [106], such as TensorFlow and Keras, for our ML model. Moreover, we employ some other well-known libraries including NumPy for fundamental computation, Pandas for data analysis, Scikit-learn for scaling the various features in the same range and dividing the dataset into training and test sets, and Matplotlib for data visualization [106], [111].

4.4.3 Performance Evaluation of the Popular ML Techniques

The classification reports consist of precision, recall, F1 score and accuracy that complement each other to evaluate the classification models. Precision is related to True Positive (TP) (i.e., prediction of traffic correctly) and False Positive (FP) (i.e., prediction of traffic wrongly)

Table 4.2 Parameters and Assumed Values in the Simulation

Parameter	**Value**
Size of simulated Area	1000 m × 1000 m
Number of Lanes	4 (two in each direction)
Number of Vehicles	100
Number of Pedestrians	10
Number of RSU	4
Bandwidth (IEEE 802.11P)	10 MHz
Minimum transmission power (IEEE 802.11p)	-20 dBm
Maximum transmission power (IEEE 802.11p)	32 dBm
Transmission rate (IEEE 802.11p)	6-27 Mbps
Spectrum band	5.895-5.925 GHz
Maximum Transmission range (IEEE 802.11p)	1000 m
Message Size	400 Bytes
Message generation rate	10 Hz
Vehicle Speed	0-40 km/h
Propagation model	Nakagami (m=3)
Simulation time	1000 s
Simulation runs	260s

states. However, this metric should be coupled with Recall that considers True Positive (TP) and False Negative (FN) to take into account wrong negative prediction. Moreover, FN is predicted incorrectly, and FP gives traffic as a prediction result, but actually the network is in a non-traffic condition. Both FN and FP lead to incorrect traffic prediction and the F1 score is a weighted average score of the recall and precision. In our research work, correct prediction is highly important. Therefore, the model needs to provide the best values in all considered metrics in the classification report. Table III shows the performance value of each metric for individual models and the proposed model. As you can see, our proposed model provides a higher value in each metric. Considering the base models, RF provides a higher accuracy. In contrast, NB provides less accuracy than all models. However, the proposed model gives better accuracy than RF and higher precision, recall and F1 score than all other ML models.

In Fig. 2, the result obtained by the ROC curve and AUC values is compared with the proposed model to determine the stability of a classification model [58]. When the ROC curve is far from the dash-dotted line in the middle and the tendency to the upper left side while providing an AUC value closer to one, that means the better model at distinguishing between positive and negative classes and stability. Considering The designation of the acronym (AUROC), the proposed model proves better performance and stability than the

baseline models.

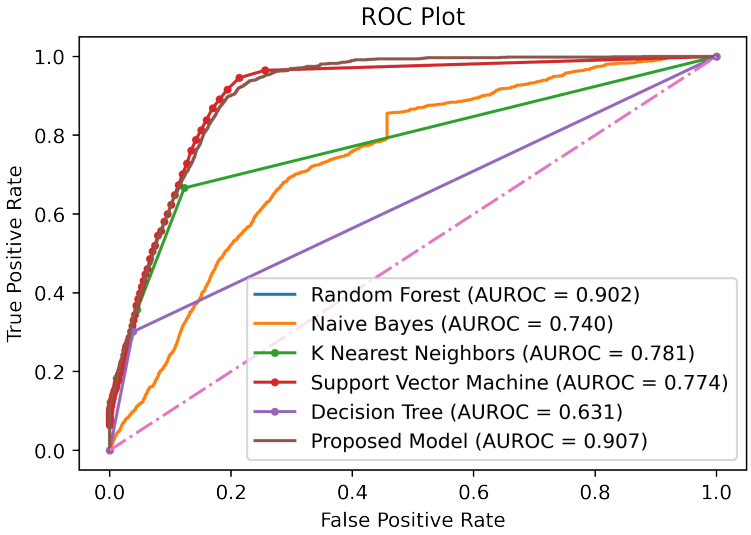

Figure 4.2 Comparison of the ROC curve of considered popular ML models and the proposed model.

Table 4.3 Performance analysis of the different ML prediction models and our proposed model with classification metrics

Prediction Models	**Precision**	**Recall**	**F1_Score**	**Accuracy**
Random Forest [78]- [54]	0.934	0.949	0.938	0.939
K-Nearest Neighbor [19]	0.932	0.942	0.939	0.938
Naive Bayes [19]	0.862	0.890	0.871	0.887
Decision Tree [78], [19]	0.931	0.929	0.930	0.928
Support Vector Machines [78]– [54]	0.905	0.951	0.928	0.923
Proposed Model	0.952	0.956	0.939	0.941

Regarding the execution time, as shown in Fig. 3, the most and the best time consumption are assigned to SVM and NB, respectively. The proposed model keeps the balance in time consumption and performs better than SVM and slightly better than RF.

It is interesting to note that identifying the best proper strategy to combine heterogeneous ML models and achieve enhanced performance is a challenging task. This paper differentiates itself from our previous work [2], by employing V2X communication which is more applicable to real-world traffic management applications. In addition, in our previous paper, we applied an advanced stacking-based ensemble learning strategy (STK-EBM) and this paper considers

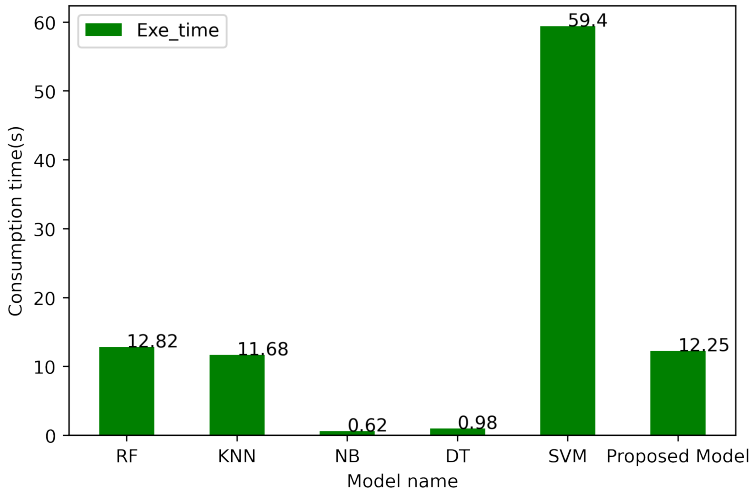

Figure 4.3 Comparison of the time consumption of different ML models and the proposed model

a basic ensemble learning strategy based on the soft voting ensemble method. The simple ensemble strategy in this work not only reduced the complexity but also improved the predictive model's performance even with the doubled size of the recorded samples. Therefore, it will not cause an increase in the computation resources. However, the STK-EBM model achieves much better stability in AUROC and recall metrics which are 0.955 and 0.98 respectively compared to the soft voting ensemble model in this paper with 0.907 for AUROC and 0.956 for recall. Finally, the ensemble learning (EL) model can help us to provide a stronger prediction model and improve the performance of the model, especially, if we cannot rely on using single ML models because of their limitations and challenges. Furthermore, different EL strategies can bring their own costs and benefits. Therefore, we need to realize the best proper ensemble learning strategy which can address challenges related to each standalone popular AI technique based on the type of problem, size of the dataset, and the number of features without imposing more problems such as time and complexity.

4.5 Conclusion

In this paper, we proposed a soft voting ensemble model for network traffic prediction in VANET/V2X. We consider our problem as a classification task and the target variable is packet delivery ratio, which is labeled 0 for non-traffic and 1 for traffic conditions in the network. We evaluated the obtained results based on classification metrics (i.e., Roc curve, precision, recall, F1 score and time). We compared the results of the proposed model with five

well-known machine learning models including RF, NB, KNN, DT and SVM. The simulation results proved that the proposed model outperforms the other considered model in terms of accuracy and time. As future directions, we will consider DL models for traffic prediction. Moreover, we will integrate the DSRC access technology with cellular-based technology like 5G, which can provide short and long-range coverage communication for exchanging information between infrastructure in V2X.

Acknowledgment

The authors would like to thank Dr. Franjieh El Khoury for the quite helpful comments and proofreading of this paper.

CHAPTER 5 ARTICLE 3: SWARM-BASED ENSEMBLE MODEL FOR NETWORK TRAFFIC PREDICTION CONSIDERING BASIC AND V2X COMMUNICATION IN VANET

Authors: Parvin Ahmadi Doval Amiri, Samuel Pierre

Status: submitted in *IEEE ACCESS*, May 5, 2023

Abstract

Recently, with the advent of the Internet of Vehicles (IoV), and the evolution of vehicular networks, VANET applications face new challenges. These applications are based on collected data from communication in vehicular networks. Therefore, network traffic prediction became a challenging task for the network operators to avoid degradation of the quality of service, which might adversely affect the services provided by the applications. Toward this end, several studies have investigated approaches related to integrating Artificial Intelligence (AI) and vehicular networks to provide intelligent, accurate and stable traffic prediction methods, in order to save human life. VANET with Vehicle-to-everything (V2X) communication is more practical in a real-world implementation. Furthermore, to achieve an efficient intelligent prediction model, we need to enhance individual AI challenges and limitations that can be done by ensemble learning. In this paper, we propose a network traffic prediction model named (eSwaNN-NTP) using a combination of Artificial Neural Network (ANN) and Swarm Intelligence (PSO) methods. Lastly, we analyze the performance of the proposed model based on different classification metrics (i.e., precision, recall, F1 score, accuracy, and time) and compare the results with standalone ANN and DNN with a different number of hidden layers. Furthermore, to validate the effectiveness of the proposed model, we employ two different datasets with essential communication (Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I)) and advanced communication (V2X) data in VANET. The simulation results show the proposed model outperforms the single Neural Network (NN) models in both datasets in terms of accuracy (95.83%), stability and time.

Keywords: Artificial intelligence, artificial neural network, ensemble learning, internet of vehicles, network traffic prediction, swarm intelligence, vehicular ad-hoc networks, Vehicle-to-everything.

5.1 Introduction

Today, we are closer to the future of transportation systems known as Intelligence Transportation Systems (ITS). The significant part of ITS is Vehicular Ad-hoc Networks (VANETs) which enable connected vehicles to communicate with each other on the road by exchanging driving information (e.g., location, speed, direction, road hazards, road traffic, accident, etc.) to enhance the perception of drivers and avoid serious road-related issues [3]. The basic architecture of VANETs consists of two main types of communications, Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) that use Dedicated Short-Range Communication (DSRC) access technology with (IEEE 802.11 p) standard to exchange information among their communications [5], [96]. VANETs offer various services by taking advantage of technology and data-driven methods to extract and share valuable knowledge among road users. The typical use case and VANET services are traffic management that bring safety, efficiency and comfort for vehicular road users [7], [11].

However, these kinds of services open further challenges related to maintaining the performance of the network and Quality of the Service (QoS), which can be critical when it comes to safety applications that might cost human life [11]. The performance of the network is compromised when VANETs experience an increasing number of sending and receiving packets through its communications and traffic occurs. It is mainly because of increasing the number of vehicles on the roads, which subsequently causes a reduction in the data delivery ratio and QoS [3], [123], [12]. Traffic prediction in such a dense network can help to avoid network traffic. However, an efficient and highly accurate traffic prediction model is a major challenge in this domain. AI- methods are considered as great potential solutions for traffic prediction before they cause a reduction in QoS and performance of the network [4]. On the other hand, the evolution of VANETs to the Internet of Vehicle (IoV) bring new problems and requirements, such as advanced communications known as Vehicle-to-everything (V2X). It can be considered as a solution for traffic issues [7]. However, the integration of VANETs with AI-methods still has lots of room for consideration [8], more specifically in V2X communications [7].

Despite the variety of AI-based techniques for traffic prediction, some limitations that must be taken into consideration. Machine Learning (ML), Deep Learning (DL) and Swarm Intelligence (SI) are three main AI techniques that have been applied for traffic prediction in VANETs [7]. Furthermore, ML is a subset of AI that can be classified into three classes: supervised learning, unsupervised learning, and reinforcement learning [8], [14]. DL methods as another subset of AI have been widely used in the case of large and complex traffic datasets [8]. Ensemble Learning (EL) methods by combining different ML and DL tech-

niques provide a stronger model that can face the limitations and problems of standalone AI techniques and achieve higher accuracy, stability, and generalization ability than a single model [32], [33]. EL became a trendy technique and attract many researchers in most domains, such as health, finance, vehicular network, and energy [34]. Considering ML-based models, different supervised algorithms including Randon Forest (RF), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machines (SVM), have been commonly applied for traffic prediction [4]. However, the best-matched characteristics of Neural Network (NN) models as a simple DL model for traffic prediction lead to a well-chosen solution in this scope [22]. Although there are some problems and limitations in the structure of the NN-based model that need to be solved. In this paper, the ensemble of SI algorithms with NN covers its limitation and provides an accurate, stable, and time-efficient model. The paper distinguishes itself from the abovementioned challenges and limitations in three key considerations and makes main contributions as follows:

- Design VANETs with V2X communication architecture for data collection from different communications while we go beyond a simple data collection, and we generate the data that comes from V2X communication through integrated DSRC and 5G-based technologies to cover both short-range and long-range areas and using grouping nearest vehicles for V2V communication. Additionally, we employ a real-world VANET dataset with integration of V2V and V2I data as a basic communication.
- Propose a swarm-based ensemble model (eSwaNN-NTP) to predict network traffic to optimize the NN limitation and provide a highly accurate and stable predictive model.
- Evaluate the performance of the proposed model according to the classification metrics and comparative analysis of the evaluation results with real-world VANETs dataset considering basic architecture and generated V2X dataset, in addition to standalone NN models.

In summary, despite all the research have been done for traffic prediction, there are still important challenges that must be considered. Most of the literature considered the basic type of communication for predicting network traffic in VANETs while the V2X communication is trendy and more practical in the future of vehicular networks and real-world implementation of VANETs. Considering the AI-methods integrated with VANETs still we require an efficient and less complex model as a prediction model in the dynamic and non-linear nature of VANETs. This motivated us to propose an intelligent network traffic prediction model, which can be best matched with traffic data and cover the limitation and challenges of NN

as a well-known model. Eventually, the experimental results will be analyzed with different classification metrics and will be compared not only with the basic VANETs communication dataset, but also with the single NN and Deep Neural Network (DNN) models to validate and highlight the higher accuracy, stability, and time efficiency of the proposed model.

5.2 Background and Related Work

Recently, a dramatic increase in the number of vehicles cause serious challenges ranging from environmental issues to human-related matters (i.e., air pollution, fuel cost, traffic congestion, time and accident) which have become inevitable challenges to be solved. ITS is a part of the smart city that is the future of transportation systems, and VANET is a substantial part of ITS that takes advantage of technology to provide a range of applications for solving the abovementioned issues [3]. Traffic management is a typical service of VANETs, in the case of safety applications, such as sending an alarm to prevent a car crash, traffic congestion, and immediate health help, accuracy and time are critical. Therefore, this kind of service must be done within an appropriate time and precisely, otherwise, it causes significant losses such as human life [11], [21]. The importance of maintaining the performance of such a network, leads to additional challenges (e.g., network traffic prediction), because traffic can negatively affect data transmission in the network, network performance and in turn the QoS for road users, respectively [96]. Although AI-techniques are considered a promising solution for developing an accurate traffic prediction model, there are noticeable challenges in this domain.

In VANET, data is transmitted through communication and the basic types of communication are V2V and V2I [5], [96]. However, the development of VANETs toward IoV, and the appearance of V2X as a new type of communication, lead to new requirements and problems [7]. Especially, there are significant challenges for VANETs integration with AI-methods to adapt to V2X communication [7]. DSRC and Cellular-based networks are two important technologies in VANETs communication to provide acceptable results for real-world implementation of V2X communication and consider a very active research domain in vehicular networks [124]. However, each of these technologies tackles its limitations. DSRC as an efficient technology in both safety and non-safety VANETs application has significant benefits such as real-time information exchange among vehicles and infrastructure-free communications which are the key to ensuring safety in vehicular communications. Nevertheless, DSRC was designed only for short-range communication coverage [8], [124]. On the other hand, although cellular-based network such as 5G has coverage for long-range communication, it has challenges in terms of dependency on connectivity with infrastructure, higher price of

network usage, and end-to-end latency [124], [125].

In the present day, researchers believe that the most functional implementation of V2X is in the integration of network technologies [114]. Therefore, DSRC and cellular-based should cooperate to provide an efficient solution for their weaknesses. Moreover, when the number of vehicles on the road increased, we need to employ a method that can effectively collect the data between moving vehicles, and roadside units [126]. Finally, an efficient way to collect the data in VNETs through V2X communication is mainly based on sharing of the information that comes from Vehicle-to-Vehicle (V2V), Vehicle- to- Infrastructure (V2I), and Vehicle-to-Pedestrian (V2P) communications in V2X, which can provide timely and valuable traffic flow information [8]. Traffic prediction is a critical challenge and V2X communication data can address traffic issues of safety applications by taking advantage of AI-methods [124], [114].

Concerning the adoption of VAET/ V2X services with AI-techniques, there are important challenges [11]. Therefore, we need to provide an efficient intelligent method that can be beneficial in vehicular networks for traffic prediction. ML and DL are subdivisions of AI-techniques. Based on the literature, ML models as a supervised learning task have been commonly considered for traffic prediction in vehicular networks. The following paragraphs discuss some of the recent literature on vehicular networks and VANETs for traffic forecasting.

The author in [78], applied three popular ML methods (i.e., DT, RF and SVM) for traffic flow prediction. They consider V2V communication data with some significant attributes of traffic dataset among vehicles including location, direction and speed. They evaluate the experimental results on the basis of important classification metrics (i.e., accuracy, precision, recall and time). The simulation results show RF and SVM as a highest accurate and lowest execution time model, respectively. Additionally, the more consumption time belongs to the highest accurate one, which is RF.

A comparative analysis of five popular supervised ML models including KNN, RF, SVM, NB and MLP presented in [19] with the purpose of network traffic forecasting in VANET. The traffic dataset was based on V2R communication. The performance of the selected ML models evaluated based on multiple classification metrics consist of Receiver Operating Curve (ROC), precision-Recall (PR) curve, accuracy, precision, recall, F1-score and time. This study considered traffic in the network when the packets in the V2R communication do not receive to the destination. Regarding the obtained analysis results, although the highest accuracy assigned to KNN, it consumed much more time than the others. In addition, the fastest algorithm was NB in terms of time with the lowest values in the other evaluation metrics. Finally, RF presented as the best balanced and acceptable model in terms of all metrics.

Lee *et al.* [23] compared four well-known ML techniques including RF, GBR, KNN and MLP for urban traffic prediction. They considered trajectory data with features (i.e., vehicle ID, vehicle position, vehicle lane, speed, and time). The prediction results showed GBR outperformed the other methods while KNN was the worst method.

A novel technique for LTE network traffic prediction is proposed in [54]. They applied three popular ML techniques on cellular traffic datasets consisting of RF, Bagging and SVM. They emphasized the benefits of ML models for optimizing network traffic. They considered the regression evaluation metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of determination (R2). The simulation results showed bagging algorithm perform well in all considered error metrics while the worst learning time consumption belong to bagging as well. Moreover, SVM was the lowest time consumption. This study indicated the quality of imported data to ML methods and the selected hyperparameters of ML models can strongly affect the prediction results.

Moreover, DL models are employed for large and complex traffic data. However, Artificial Neural Network (ANN) is a simple DL model is considered a well-chosen method for traffic prediction due to its characteristics including strong learning ability, providing higher accuracy, and adoption of complex non-linear traffic data [80]- [24]. However, there are challenges in the structure of NN-based models that affect their performance such as the dependency of their performance on selecting the right network architecture and slow rate of learning ability [25].

Of note, EL methods have been attracted researchers due to their ability to develop better overall performance by combining different ML and DL methods into strong models that can cover single ML flaws. Moreover, it can be matched better with the non-linear and complex nature of vehicular network traffic data [84].

Zhao *et al.* [84] employed big traffic data with 500,000 recorded samples that were generated every 5 minutes from 51 road sections. They proposed EnLSTM-WAPO, an ensemble model that provides less error prediction that well-known DL models, such as ARIMA, GRU, DBF, LSTM and SAE.

Sepasgozar *et al.* [85] considered a merged dataset from V2V and V2R communication data in VANETs and applied a hybrid model named RF-GRU-NTP by combining RF as an ML model and GRU as a DL model for network traffic prediction considering road and network parameters. The evaluation results show a better performance than a single model in terms of accuracy and time consumption.

A new ensemble model was proposed with the aim of mobile traffic prediction in [80]. They

combined two ML models including MLP and Self-Adaptive Support Vector Regression (SSVR). The obtained results were evaluated based on Mean Square Error (MSE) and it showed the ensemble model was stable and more accurate than other models (i.e., RNN, LASTM, and CNN).

Ahmadi *et al.* [2] employed an integration of V2V and V2I communication data in VANET with the most practical features extracted by the ensemble feature selection model (Boruta and LightGBM). They proposed Stacking Ensemble learning with a Booster model (STK-EBM), for network traffic prediction. They applied an ensemble of ML models including RF, KNN, Xgboost and logistic regression using a stacking strategy to achieve a stable prediction model. Moreover, they compared their proposed model with the most popular ML models to show its effectiveness and advantages. The experimental results proved that the proposed model provides a trade-off among the results of all considered models and obtains a more accurate and stable prediction.

Of note, the number of research studies on network traffic prediction in VANETs considering V2X communication is very low. Moreover, EL is considered a promising solution for developing an efficient prediction model in vehicular networks, which is necessary for network traffic management [7], [100].

Therefore, the popularity of the NN-based model and its performance problems such as selecting an appropriate network architecture, slow learning ability [25], and on the other hand, efficiency of ensemble models, lead to use as an ensemble model. The best combination matched for covering the problem of NN methods is swarm intelligence, which is another popular subset of AI techniques for VANET applications [7]. SI can be categorized into four classes including biology-based, human behavior-based, evolution-based, and physics-based with a variety of algorithms in each of these classes. SI has been widely used for solving problems because of its ability to model a population of agents that are capable of self-organization and interaction with each other [26]. Among them, Partial Optimization Swarm (PSO) as a nature-inspired algorithm in biology-based classes is considered as the most popular solution for prediction tasks [7], [26]. It provides self-organization ability, and parallel computation in a distributed manner besides flexibility and robustness [7], [26].

In summary, EL techniques can help us to cover a single ML model limitation by combining different methods together and make a new strong model [32], [33]. EL can achieve enhanced performance results by reducing errors, increasing accuracy and robustness, fast computation, and better generalization ability than a single model [34]. Therefore, EL can be a solution for taking advantage of NN-based model while covering their problems. For this purpose, we can develop an efficient intelligent model by cooperation of NN and SI.

5.3 Methodology

In this section, we illustrate the proposed model named (eSwaNN-NTP). This model takes advantage of EL models and is applied to VANET/V2X to predict network traffic. In the basic architecture of VANETs, data was collected from communication between moving vehicles and roadside units, as shown in Fig. 1 However, VANET with V2X data by considering communications among vehicles and any other entities on the road is a significant promise for the future of real-world VANET applications. Fig. 2 shows the architecture of VANET/V2X environment.

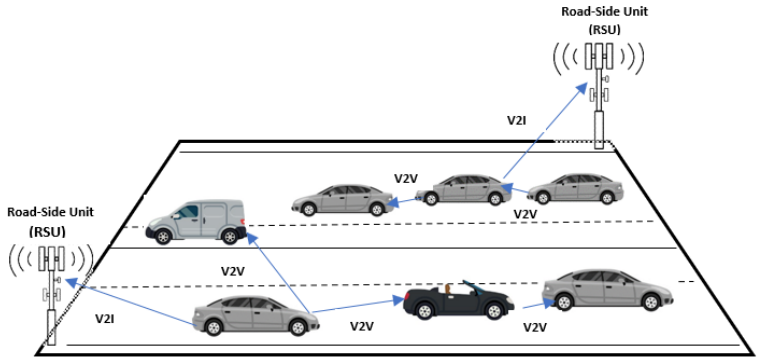

Figure 5.1 The basic architecture of VANETs [2].

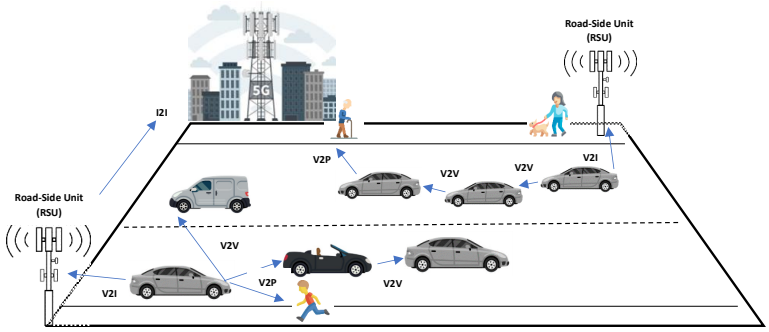

Figure 5.2 The advanced architecture of VANETs.

Moreover, the problem is when the number of sending and receiving packets through VANETs communication increase (i.e., many vehicular road users on the road), traffic occurs in the network [126], [127]. Network traffic prediction in the VANET environment is a complicated and challenging task due to the highly dynamic nature of such networks. To maintain the

performance of the network, machine learning, deep learning and on top of them EL methods, can achieve higher accuracy, reliability and fast computation of prediction models, even in complex and non-linear traffic data [32], [22], [126].

The proposed model in this study takes VANET with both basic and advanced communication data into account and provides an efficient ensemble model to predict network traffic. Fig. 3 depicts the workflow of our research work. We divided our model into three main sections: data collection and preprocessing, the framework of the proposed model and analysis of the evaluation results. This classification aims to specify the contribution of each part and then explain the procedure of the proposed method in detail.

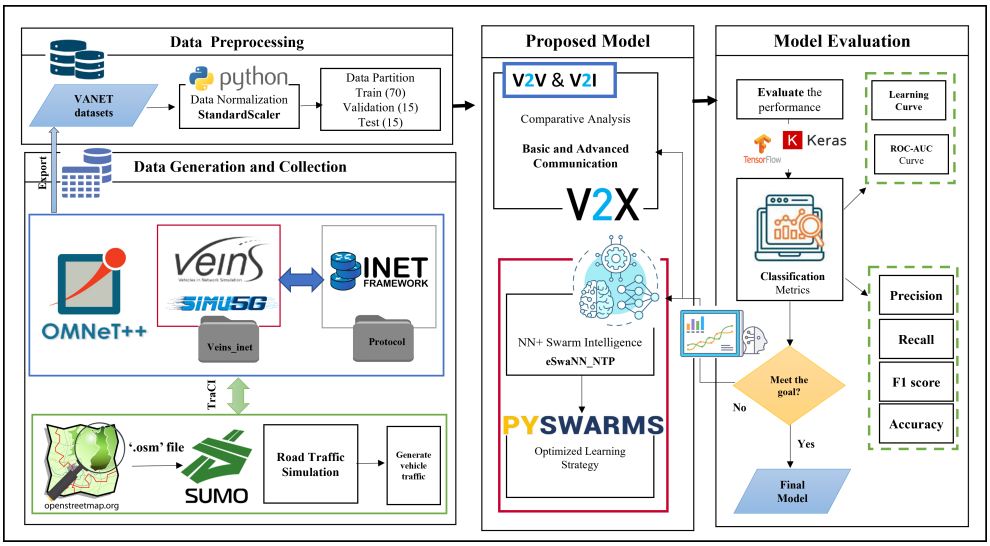

Figure 5.3 Workflow of the Swarm-Based Ensemble Model for Network Traffic Prediction with Basic and V2X Communication in VANET.

5.3.1 Data Collection and Preprocessing

The lack of a dataset containing V2X communication traffic information was the reason why we generated a dataset of traffic scenarios in VANETs considering V2X communication. Therefore, we collected the V2X communication data from V2V, V2R, V2P and R2I [8]. Moreover, we consider an efficient way to collect the traffic data. First, we collected information about all vehicles, and based on the information we make a cluster between vehicles. Then, we select Cluster Head (CH) vehicles and perform the intra-cluster communication process by using DSRC. In this way, we collect efficient information exchange among vehicles without the need for roadside communication infrastructure [124], [126], when Roadside Unit (RSU) is not directly in the range of vehicles. After that, we perform communication between

CH of vehicles and pedestrians. Moreover, we employ a 5G base station that can cooperate with RSUs to share updated information about the VANET environment. Fig. 4 shows the data collection from the V2X communication in VANETs in this paper.

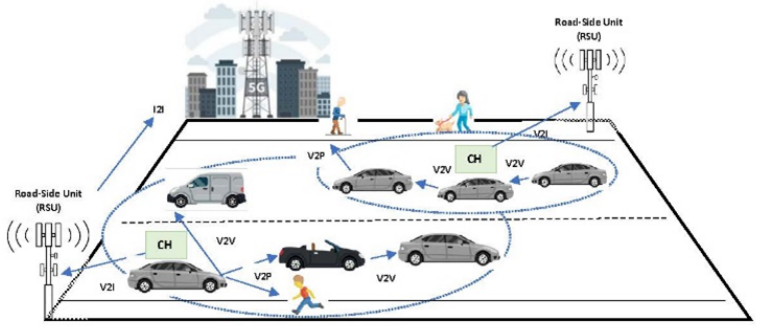

Figure 5.4 Data collection from the V2X communication in VANETs.

Finally, we perform packet transmission and measure packet delivery ratio while VANET experiences an increasing number of vehicles in traffic scenarios and dense network communication. Moreover, data preprocessing can help us to keep the quality of input data before feeding it into AI models which plays an essential role in increasing the efficiency of the model [50], [52]. Therefore, in the preprocessing procedure, we first eliminated some redundant, missing, and meaningless values in our dataset. Then, we employ StandardScaler normalization [52] that helps scaling the values of the dataset in the range of [0,1]. At the end of this section, we separated the dataset into training and test sets. The StandardScaler is denoted by:

$$Z_{scaled} = \frac{(X - \mu)}{\sigma}, \tag{5.1}$$

where X = input variable, μ = Mean and σ = Standard Deviation.

We consider our problem as a binary classification task. Therefore, we assumed variable$X$ with (i), the number of input features and accordingly, we labeled our target variable (y), which represents the prediction value of traffic in the VANET/V2X dataset. We define two classes of traffic (1) and non-traffic (0) in this study.

5.3.2 Overview of the Proposed eSwaNN-NTP Model Architecture

In this section, we demonstrate the framework of the proposed model in two subsections including neural network and the proposed (eSwaNN-NTP) model that is based on the en-

semble of NN and SI.

Neural Network

NN as a simple DL model is considered the best solution for increasing the accuracy of traffic flow prediction [22]. MLP is a well-known NN adaptable with network properties and traffic patterns than other NN. MLP model as the classical type of feed-forward neural network [23], [24] consists of three layers: the input layer, the output layer, and the hidden layer, while the number of the output nodes is based on the machine learning task. For classification task in our case include two output nodes. The General structure of Artificial Neural Network (ANN) is depicted in Fig. 5 The hidden layer can be one to many layers.

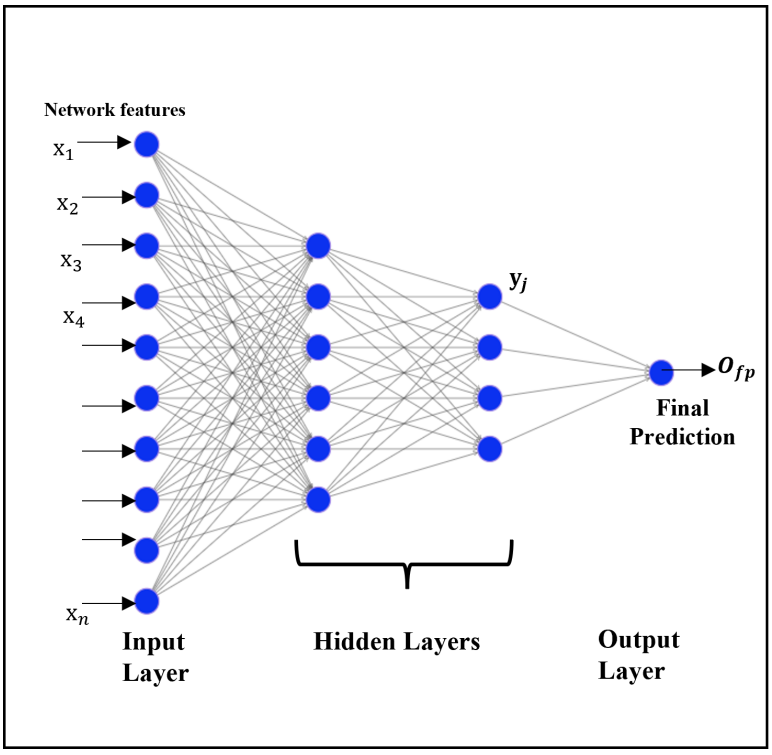

Figure 5.5 The general architecture of Neural Network (NN) model.

The related formula is designated by [80], [82], [101]. The output of all hidden layers is calculated by:

$$y_j = f(\sum_i w_{ji} x_i) + b_j)(i = 1, ..., n), \tag{5.2}$$

where x_i is the input variable, i represent the number of inputs, w_{ji} represent the connection

weight from input node i to hidden node j, b_j stands for the bias of the neuron j, employed for summation, y_j stands for the output of hidden layer node j and f is the activation function of a node, which is the connection weight from the ith input to the jth hidden neuron. The sigmoid function is the most used nonlinear transfer function.

The output of neural network is designated by:

$$O_fp = f(\sum_j w_{oj} y_j) + b_0)(j = 1, ..., n), \tag{5.3}$$

where w_{oj} represent the connection weight from node j to output node o, b_0 stands for the bias of the neuron,O_fp stands for the output data of network and f is the activation function of output layer node.

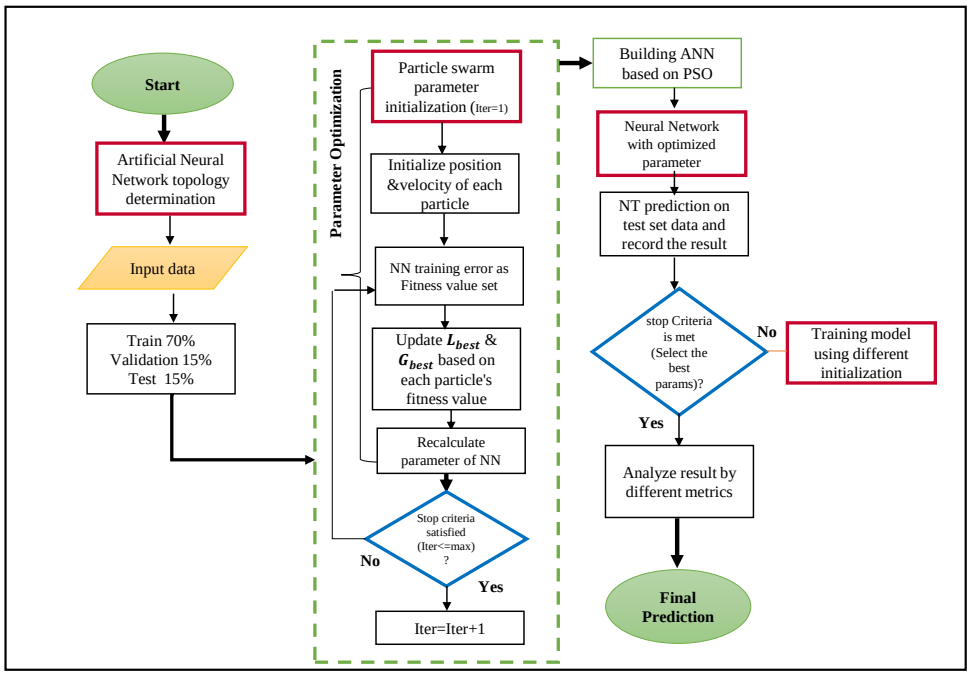

Figure 5.6 Framework of the proposed eSwaNN-NTP model.

However, there are challenges related to these popular NN-based models, such as dependency of their performance on selecting the right network architecture, and they need appropriate training algorithms to increase their learning rate ability [25]. ANN challenges besides being a well-chosen model for traffic prediction task, motivated us to find a solution to tackle NN-based model problems and limitations. SI as a best-fitted intelligent optimization algorithm can help us to address the NN problems [7], [26]. Furthermore, EL methods are become

popular due to their ability to take advantage of each individual model by combining them and make a strong and more accurate predictive model [32], [33], [34]. Therefore, we used ensemble of ANN and PSO as one of the trendy SI methods. In this way, we can take advantage of EL model to achieve fast convergence learning ability while avoiding local optimum problem and obtain more stable and accurate prediction model.

The Proposed eSwaNN-NTP Model

The term "Swarm Intelligence" was officially introduced by Beni and Wang in a cellular robotic system domain of research [128]. In recent studies, swarm intelligence has been widely used as a problem solution due to its ability to model a population of agents that are capable of self-organization and interaction with each other [26]. SI can be categorized into four classes including biology-based, human behavior-based, evolution-based, and physics-based and each of these classes consist of a variety of SI algorithms [26]. However, among these nature-inspired algorithms in biology-based classes, PSO is considered as the most popular and trendy solution for prediction tasks [7]. PSO stimulates the birds' folds behavior in nature. This method acts like birds that are flying together in multi-dimensional space. They continuously change their movement behavior and their distances for searching and finding the optimal place. They cooperate in an energy-saving manner with parallel operation [26], [129]. PSO includes N particles in a swarm and these particles are like birds and they initialized randomly, each particle flying in the search space and moving toward the global optimum solution. They provide parallel computation in a distributed manner besides being flexible and robust [26], [130]. Therefore, we can reduce ANN model complexity related to numerous hyperparameters, connection weights between different pairs of neurons and bias value [25], which can significantly affect the performance of the predictive models. The framework of eSwaNN-NTP with the aim of network traffic prediction is described as follows.

Step 1) The train dataset feed to the (eSwaNN-NTP) model where the swarm has N particles and *i-th* particle at *t-th* iteration are generated and their initial positions and velocities are set randomly. The related formula is inspired by [**?**], which is denoted by:

$$x_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{iD}(t)), \tag{5.4}$$

where $x_i(t)$ is the current position of the particle and searching space in the range $x_i(t) \in [-x_{max}, x_{max}]$.

$$V_i(t) = (V_{i1}(t), V_{i2}(t), \ldots, V_{iD}(t)), \tag{5.5}$$

where $V_i(t)$ is the current velocity of the particle and $V_i(t)$ limited to certain range $V_i(t) \in [-V_{max}, V_{max}]$.

Step 2) ANN learning error which is defined as MSE considered as the fitness value of each particle. The MSE will be evaluated to compare the actual and predicted values in each iteration and designated by:

$$MSE = 1/M \sum_{j=1}^{M} \left(NT_{j,actual}, NT_{j,predicted}\right), \tag{5.6}$$

where M is the number of actual values and $NT_{j,actual}$,$NT_{j,predicted}$ represents measured and predicted NT, respectively.

Step 3) The process is repeated until the stop criteria are met by PSO method.

$$P_i(t) = (P_{i1}(t), P_{i2}(t), \ldots, P_{iD}(t)), \tag{5.7}$$

where $P_i(t)$ is the local best position of each particle.

Step 4) The particle position and velocity will be updated. The updating position and velocity formula is denoted by:

$$\begin{aligned} V_{id}(t+1) = \omega V_{id}(t) + c_1 r_1(t) \left[P_{id}(t) - X_{id}(t)\right] \\ + c_2 r_2 \left[G_{id}(t) - X_{id}(t)\right] \\ X_{id}(t+1) = V_{id}(t) + V_{id}(t+1), \end{aligned} \tag{5.8}$$

Step 5) Based on the updated positions and velocities of each particle this process will be continued until the particle with the lowest learning rate can be discovered, which is the termination condition.

Step 6) The generated best results will be determined and if so, we executed Step 7 which means, ANN will be initialized by these optimum parameters. Otherwise, we start over with Step 3.

Step 7) Finally, the ANN model used PSO to find its best network configuration and increase its learning ability. Then, test dataset will be feed to the model for traffic prediction and evaluated based on the classification metrics that are defined in the next section. Therefore, EL can achieve excellent performance than standalone AI models [32], [33], [34]. We proposed an eSwaNN-NTP model, which is an ensemble of SI with NN to achieve a high-performance prediction model. The framework of our proposed model depicted in Fig. 6.

5.3.3 Evaluation and Analysis Metrics

In this paper, to evaluate the performance of the proposed model and analyze the efficiency of the model compared to other baseline models, we consider important classification metrics including precision, recall, F1 score, accuracy, the ROC curve and AUC value, the classification report, and the consumption time. The correlation between the actual and predicted classes shows in Table. I. Considering our target that is traffic prediction in VANET, False Negative (FN) in Table. I, means the prediction results indicate a non-traffic condition. In the case of traffic in the network, this impact negatively on VANET applications, especially safety applications that might be at the cost of human life and is critically important in traffic prediction. Moreover, False Positive (FP), illustrated in Table. I means non-traffic situation predicted as traffic situation. In addition, it can be a failure in the performance of the prediction models. However, it will not affect vehicular users and is more tolerable than FN. Accuracy, recall, precision and F1 score, are factors that can help evaluating in an efficient way the performance of the model.

In this way, the accuracy represents the ratio of TP and TN to the overall number of samples. Recall indicates the ability of the classifier to detect all positive samples in the actual class. Precision provides the accuracy of the positive prediction and F1 score is calculated by precision and recall, while the best score is 1.0. Finally, Receiver Operating Characteristic (ROC) curve, will be performed to determine the performance of binary classification and Area Under Curve (AUC), estimates the stability of the model [106].

Table 5.1 The relationship between the actual and predicted classes

Network Traffic		Actual Classes	
		Positive	Negative
Predicted Classes	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

$$Accuracy = \frac{(TP + TN)}{TP + TN + FN + FP} \tag{5.9}$$

$$Sensitivity(Recall) = \frac{(TP)}{TP + FN} \tag{5.10}$$

$$Precision = \frac{(TP)}{TP + FP} \tag{5.11}$$

$$F1score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{5.12}$$

Area Under Curve(AUC), For a predictor f, an unbiased estimator of its AUC: tests whether positives are ranked higher than negatives

$$AUC(f) = \frac{\sum t_0 \quad \in D^0 \sum t_1 \quad \in D^1 \; 1\left[f\left(t_0\right) < f(t_1)\right]}{|D^0| \, . \, |D^1|}, \tag{5.13}$$

where $1[f(t_0) < f(t_1)]$ notes an indicator function which returns 1 if $f(t_0) < f(t_1)$ otherwise return 0, D^0 is the set of negative examples, and D^1 is the set of positive examples.

5.4 Data Collection and Performance Evaluation

5.4.1 Data Collection and Simulation

In this section, we explain the details of two datasets we used including the simulated VNATE/V2X dataset and real-world VANET/V2V andV2I dataset, respectively.

GPS data-based techniques provide high accuracy, scalability, and high privacy in traffic prediction methods [118]. Vehicular communications in VANETs play a critical role in safety applications. Simulation tools such as Simulation of Urban Mobility (SUMO) and OMNet++ have been wildly used by researchers to validate and evaluate their proposed methods. This is due to the high cost of the physical resources needed such as vehicles equipped with OBU, RSUs and Base stations (5G), to implement vehicular network tests in real-world [118]. Moreover, the lack of a VNAET/V2X dataset containing different access technologies designed for traffic prediction approaches was the reason why we design a V2X communication architecture with integration technologies to generate traffic and collect recorded data from V2X communications. However, we considered real-world VAMETs dataset with basic communication inclusive fusion of V2V and V2I to evaluate the proposed model.

For VANET /V2X dataset, we consider the integration of the most popular simulation framework for traffic modeling and wireless communications, as illustrated in Table. II. In the first step, we try to provide a more realistic simulation of vehicular traffic in Simulation of Urban Mobility (SUMO) [116], therefore, we download downtown map data of Montreal city in Canada from "OpenStreetMap" [115] as an ".osm" file to import to SUMO. Then, we use SUMO (1.7.0) to generate urban vehicular traffic. Next, we employ a network simulator OMNet++ (5.6.2) [118] to work jointly with SUMO, Veins (5.2)/INET (4.4.0) [118], [117],and Simu5G (1.2.1) to simulate wireless communication. After that, DSRC technology is applied

for transmitting data in communication including V2V, V2R, and V2P. In addition, the 5G base station is used for long-range coverage of transmitting data in between infrastructures (I2I). In the end, we generate much load of data in VANET, to simulate the traffic situation. For this purpose, we defined an accident to be happened at a specific time of running the simulation scenario. We perform 1000 s duration for each run of the simulation. We calculate the Packet Delivery Ratio (PDR) values while the number of vehicles has increased and we assumed the values of PDR min as 0.4 and PDR max as 0.6 [122]. The transmitted data within 1000 s of the simulation scenario are considered to measure PDR. We recorded 90432 samples. Based on PDR value each data record with PDR less than 0.3 was labeled by 1 as a traffic condition in the network, otherwise, 0 as a non-traffic condition in the network.

Table 5.2 Configuration used to generate the simulated environment

Parameter	**Value**
Size of simulated Area	1000 m × 1000 m
Number of Lanes	4 (two in each direction)
Number of Vehicles	100
Number of Pedestrians	10
Number of RSU	4
Number of Base station (eNB)	1
Bandwidth (5G)	10 MHz
Uplink Frequency band (5G)	1 MHz
Downlink frequency band (5G)	10 MHz
Transmission Power(5G)	40 dBm
Bandwidth (IEEE 802.11P)	10 MHz
Minimum transmission power (IEEE 802.11p)	-20 dBm
Maximum transmission power (IEEE 802.11p)	32 dBm
Transmission rate (IEEE 802.11p)	6-27 Mbps
Spectrum band	5.895-5.925 GHz
Maximum Transmission range (IEEE 802.11p)	1000 m
Message Size	400 Bytes
Message generation rate	10 Hz
Vehicle Speed	0-40 km/h
Propagation model	Nakagami (m=3)
Simulation time	1000 s
Simulation runs	260s

Considering the real-world VANETs dataset, we used [**?**] which is based on DSRC communication of V2V and V2I and the experiment was performed in a realistic highway scenario. Moreover, it considered the northwest sector of Atlanta, GA along I-75 between Exit 250 and Exit 255 with five regular lanes and one High Occupancy Vehicle (HOV), which can be

representative of the most rods in the U.S cities [108]. The data is recorded by GPS during a day between 2 pm and 5 pm and consist of GPS-reported features every two seconds (e.g., location, longitude, latitude, speed, etc.) in 8022.11 ad-hoc networks. Moreover, IPerf cooperates with GPS aimed to read network parameters. The V2V communication was measured while both sender and receiver were placed in moving vehicles in the same lane. The V2I communication considered RSU that was placed on the bridge with different heights. The vehicles broadcasted the packets at a rate of 150 packets/s while moving in the rightmost lane and all the communication data (e.g., log time, location information, velocity, and packet sent/received) were collected [108]. When the number of packets increases through the VANET communication, traffic occurs in the network [3], [123], [12]. We considered 39,998 recorded data of the fusion of V2V and V2I communication with the aim of network traffic prediction and we labeled the dataset into classes including 1 when the packet is not received as a traffic condition and 0 when the packet is received as a non-traffic condition in the network.

5.4.2 Experimental Details

In this section, we explain the platform, programming language and the important AI-libraries that we use to implement our proposed model. Firstly, the platform we employ to execute the AI models is Google Colab, which is hosted on the Google Cloud Platform [53] and work based on Jupyter Notebook. Secondly, we apply Python (3.7.13) [111] for our programming language, which is open-source and can support a variety of AI libraries, such as TensorFlow and Keras that we used for the implementation of the AI models. Finally, we use some important libraries of Python which can simplify data analysis and visualization including Pandas for data analysis, NumPy for fundamental computation, Scikit-learn [111], [106] for scaling features, in addition to splitting the dataset into training and test sets, and Matplotlib for data visualization.

5.4.3 Comparative Analysis of the Proposed Model with Two Different Datasets and Baseline Models

In this section, we develop an effective neural network using a swarm intelligence method (PSO) for predicting traffic in VANET. Additionally, the evaluation is conducted on two different datasets, the first one is a real-world VANET dataset considering the merged of V2V and V2I communication with 39998 samples and 11 features and the second dataset is generated from V2X communication contains 90432 samples and 5 features. The purpose of this consideration is to gain insight into the effect of the size of training samples and the number

of input features on the proposed model and their performance compared to standalone ANN and DNN with different numbers of hidden layers. Furthermore, the evaluation is performed by applying important classification metrics (i.e., precision, recall, F1 score, AUC value, ROC curve and time) [58]. Table. III and Table. IV indicate the setting parameters of NN and PSO in the proposed model. In Table. III, BC stands for Basic Communication and AC stands for Advanced Communication.

In the experimental results, standalone ANN and DNN with 2 and 3 hidden layers are evaluated and compared to the proposed model for both the (V2V) (V2I) dataset and (V2X) dataset, which we named Basic Communication (BC) and Advanced Communication (AC), respectively, in the result analysis. As you can see in Table. V and Table. VI, the higher values of accuracy of our proposed model among other models are 93.37% for VANET with the BC dataset and 95.83% for VANET with the AC dataset. Moreover, considering the importance of correct prediction in our case study, a higher recall value means that most traffic situations in the network are predicted correctly. The proposed model also provided higher values in this metric and almost all of the metrics.

Table 5.3 Configuration Used to Generate Simulated Environment

Neural Network Parameter	**Value**
Input BC, AC and output	11,5,2
Optimizer	'adam'
Learning rate	0.26
Number of neurons	50,25
Activation function of hidden layer	'relu'
Activation function of output layer	'sigmoid'
Batch size	64
Number of epochs	100

Table 5.4 Parameter setting for the swarm intelligence (PSO) method

Parameter Name	**Variable**	**Value**
'Pbest 'and 'Gbest' are particle local and global best solution	c_1, c_2	0.5, 0.3
Inertia weight	ω	0.9
Number of particles	N	30
Number of iterations	t	100
Dimension (No. of parameters)	d	11,5

Finally, when it comes to AUC values that represent the stability of the classification model

[58]. Furthermore, the ROC curve employs to measure the performance of binary classifiers, which is another graphical tool, and it can be interpreted as an ideal model if the curve is closer to the upper left side of the plot and the distance from the middle-dotted line with a value closer to one [106], [58]. Based on Fig. 7, the proposed model shows good stability in distinguishing between negative and positive classes and obtain higher AUROC value obtains the highest values compared to other models which are 0.930 and 0.960 for the BC dataset and AC dataset, respectively.

Table 5.5 Comparison of the classification metrics for ANN, DNN and the proposed model in VANET with V2V and V2I (BC)

Prediction Models	**Precision**	**Recall**	**F1_Score**	**AUC**	**Accuracy**
ANN (MLP)	0.9168	0.9709	0.9431	0.8928	0.9206
DNN (2 hidden layer)	0.9026	0.9942	0.9462	0.8843	0.9234
DNN (3 hidden layer)	0.9122	0.9867	0.9480	0.8935	0.9266
Proposed Model(eSwaNN-NTP)	0.9190	0.9950	0.9570	0.9300	0.9337

Table 5.6 Comparison of the classification metrics for ANN, DNN and the proposed model in VANET with V2X (AC)

Prediction Models	**Precision**	**Recall**	**F1_Score**	**AUC**	**Accuracy**
ANN (MLP)	0.9546	0.9230	0.9300	0.9379	0.9434
DNN (2 hidden layer)	0.9470	0.9942	0.9530	0.9316	0.9540
DNN (3 hidden layer)	0.9483	0.9963	0.9570	0.9416	0.9550
Proposed Model(eSwaNN-NTP)	0.9563	0.9970	0.9590	0.960	0.9583

In Fig. 8, the train and validation loss curves are plotted in 100 iterations, which shows the model, has a good fit for the problem in both datasets. However, when it comes to the AC dataset, which provides more training samples resulted in lower loss values compared to BC datasets

The key point for NN based model is weight and the training cycle adjusts weights to improve the performance of the network. Therefore, the performance of the network critically relies on training performance.

Fig. 9 and Table. VII are shown for both datasets that in each iteration the cost of the model has been significantly minimized, which means the model is suitably fit for the training

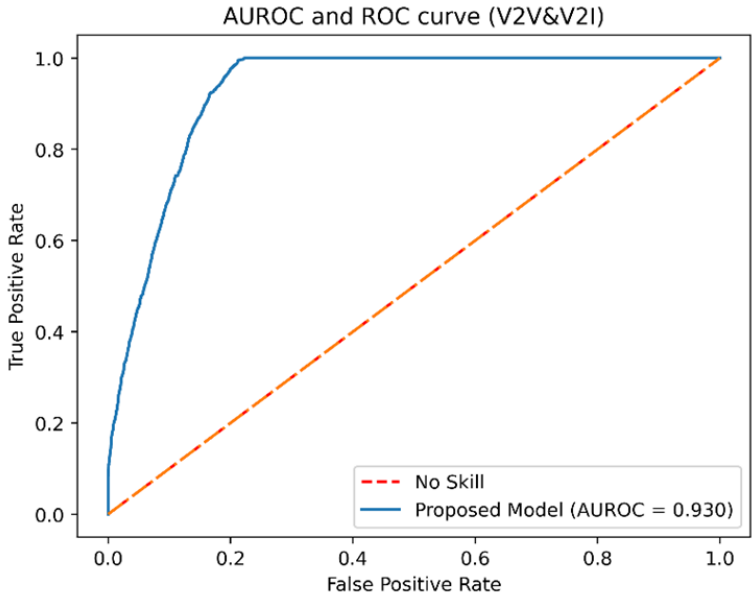

Figure 5.7 The ROC plot of the proposed model for VANET (a) Basic communication.

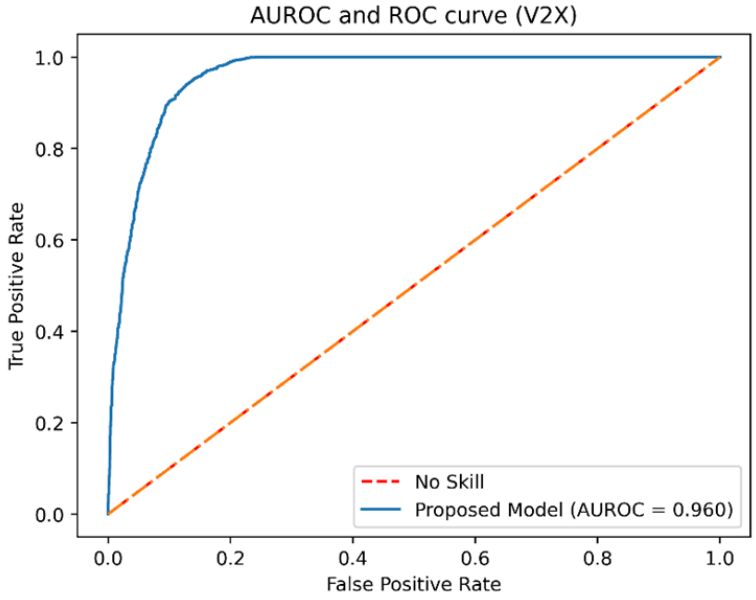

Figure 5.8 The ROC plot of the proposed model for VANET (b) advanced communication.

datasets. The local and global cost of the model at the of 100 iterations are pointed in the plots, which indicate better values for larger training samples in AC datasets.

Finally, we need to calculate time as another significant metric in ML-related studies. Since

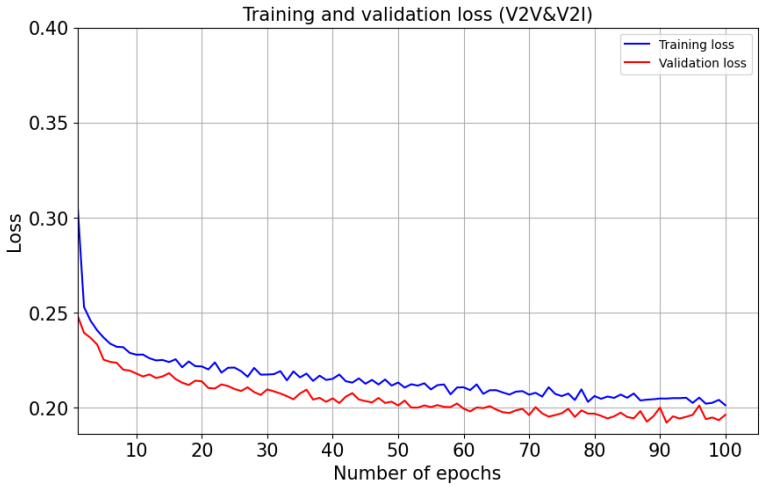

Figure 5.9 Comparison of the ROC curve of considered popular ML models and the proposed model (a) Basic communication.

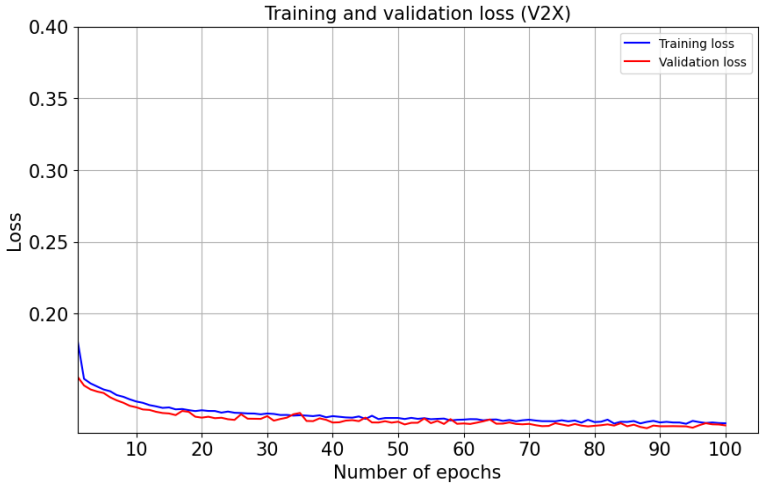

Figure 5.10 Comparison of the ROC curve of considered popular ML models and the proposed model b) advanced communication.

Table 5.7 The best values of the swarm intelligence (PSO) in the proposed model

ANN with PSO	**Value (V2V &V2I)**	**Value (V2X)**
Global best cost	0.259345	0.159076
time (min)	1.32	2.14
Accuracy	0.9337	0.9583

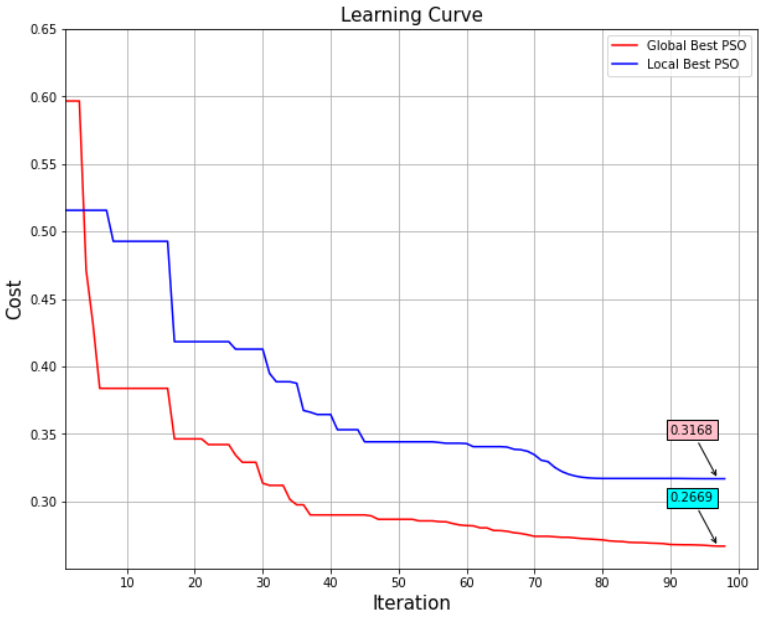

Figure 5.11 Local vs global Loss the Swarm Intelligence (PSO) (a) Basic communication.

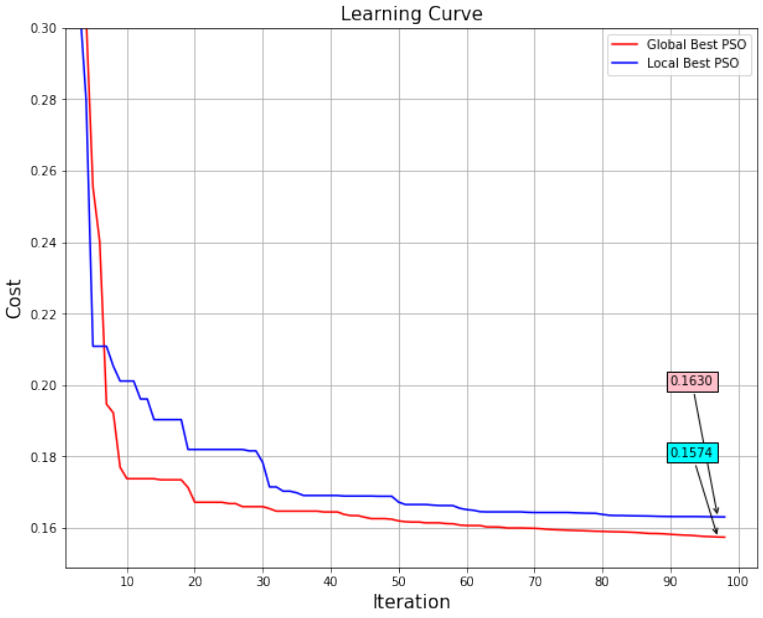

Figure 5.12 Local vs global Loss the Swarm Intelligence (PSO) (b) advanced communication.

higher accuracy at the cost of time might not be effective when it comes to time-sensitive applications. Therefore, in Fig. 9, we compared the training time that is required by different models in this paper including ANN, DNN with 2 and 3 hidden layers, and the proposed

model for both BC and AC datasets. According to the comparison, as shown in Fig. 10, the lowest training time belongs to the proposed model with the values of 1.32(min) and 2.14 (min) for BC and AC datasets, respectively. However, these values indicate that almost we require double the time for training the model when the size of the dataset becomes larger. Moreover, the DNN model could bring a significant weak point as a chosen model for traffic prediction, especially in time-sensitive problems.

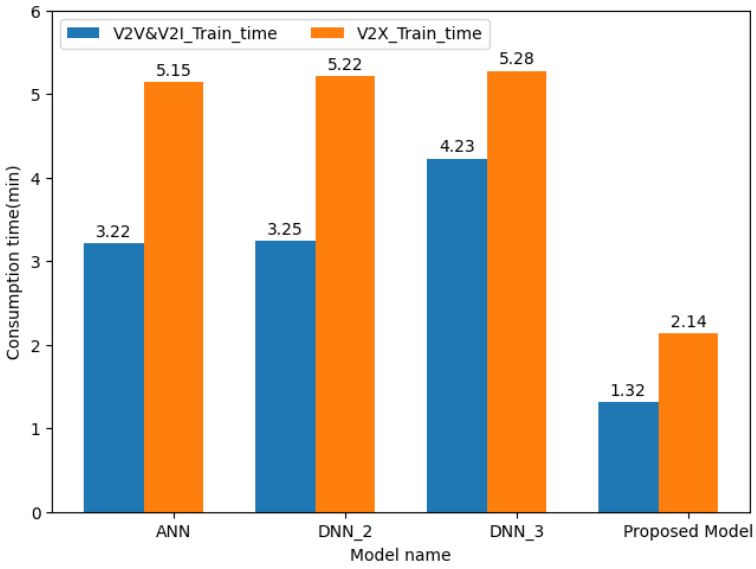

Figure 5.13 Comparison of the training time of the ANN, DNN for the proposed model in VANET with BC and AC datasets

Moreover, in this paper, we concentrate on ensemble of NN models. However, considering the previous work [2] based on ensemble of ML models, we can infer that our proposed method is simple in terms of strategy, and can achieve a better prediction accuracy (95.83%) than this previous work (94.10%) in the case of a bigger dataset. In addition, it provides more stability in prediction in both small and large datasets while reducing the complexity related to the integration of different ML models in advanced ensemble learning strategies such as stacking. Furthermore, the proposed model in this paper not only is simple and efficient but also more generalizable than the previous method.

In summary, in this study, we focus on enhancing the overall performance of the traffic prediction model by considering NN as the most popular model in this scop. Moreover, by considering two different datasets in VANET, we gain insight into the effect of the size of the dataset on the training time, accuracy, and optimization methods. Eventually, from the obtained results, we can infer that using SI (PSO) ensembled by ANN can give us better

accuracy, time and stability of network traffic prediction.

5.5 Conclusion and Future Work

In this paper, we proposed an efficient neural network model using an ensemble of swarm intelligence (PSO) and ANN for network traffic prediction in VANET. Moreover, the proposed model was applied to two different datasets including a real-world dataset with basic communication (V2V&V2I) and a realistic generated dataset with advanced communication (V2X). The purpose of this consideration is to validate the effectiveness of the proposed model. We consider our problem as a classification task. We used the packet delivery ratio as a target variable to categorize the non-traffic and traffic situation in the network. In addition, ANN and DNN with a different number of hidden layers, which are widely used in traffic prediction problems, have been applied to make a comparison with the proposed model in both datasets. The classification metrics including precision, recall, F1 score, accuracy, ROC curve and time are employed to evaluate the performance of the models. The simulation results emphasize that the proposed model performed well in all considered metrics and in both datasets. Furthermore, the proposed model provides the best accuracy with less time than the other models besides showing stable and reliable prediction results. This study aims to employ an efficient intelligent model that improves the prediction accuracy while not increasing the consumption time and provides overall better performance with stable prediction.

As future work, we will consider Flying Ad-hoc Networks (FANETs) as the future of ad-hoc networks for next-generation applications with big data that bring new challenges for intelligent and efficient AI-based traffic prediction models.

Acknowledgment

The authors would like to thank Dr. Franjieh El Khoury for the valuable comments and proofreading of this paper.

CHAPTER 6 GENERAL DISCUSSION

In this chapter, we discuss the proposed methods in this dissertation with the main objective of designing an efficient traffic prediction model with the integration of AI techniques and VANET. In VANETs, information about road conditions and other vehicles will be exchanged among communications, and when the number of sending and receiving packets through these communications increases (i.e., lots of vehicles on the road), traffic occurs in the network, network traffic causes a delay or decline in important services [4]. Furthermore, traffic will increase packet loss or decrease packet delivery ratio in the network [123]. Moreover, the implementation of ML techniques as a subset of AI could optimize the operation of the networks [10].

Accordingly, a more accurate and stable prediction of traffic in the network can help to identify the failure of the network and its dependent services. It can effectively help us to predict traffic in the network and mitigate that before declining the quality of services for the users. Specifically, vehicular networks that are related to important services like preventing road congestion and accident for road users will be more essential. Therefore, we need to infer network traffic from the performance of the network. Intelligent traffic prediction methods in vehicular networks have been considered by many research studies in this domain. However, still, there is lots of room for consideration in order to realize the full potential of AI methods that can be applied in VANET.

Considering the data which is a key component of the supervised algorithm. In this dissertation, we applied two datasets. The real-world VANET data including basic communication and the V2X data is generated using the cooperation of a traffic modelling simulator and wireless communication simulation tools. Therefore, vehicle movements and road traffic are simulated by SUMO which cooperates with OMNET++ and its extensions such as Veins Simu5G, and INET to create wireless communication in the network and then, generate traffic data. These simulators are selected based on our requirements in this dissertation including the heterogeneous types of vehicular networks, using DSRC and 5G simultaneously by all vehicles during their communications with each other and other entities and simulating vehicular traffic data.

In Chapter 3, the proposed method was designed with a stacking ensemble of heterogeneous ML algorithms in two layers structure consisting of a base layer with the integration of RF, KNN, and XGBoost and a meta layer with an optimized LR algorithm. The preliminary prediction is made through the basic learners (KNN, RF, Xgboost) of the first layer, and

then the prediction results of the first layer are aggregated using logical regression to get the final prediction results. Moreover, considering the importance of input data, Borota a LightGBM feature selection algorithm is applied to extract the most effective features of the dataset based on the target which is network traffic prediction. Then, the performance of the well-known machine learning models is evaluated and compared, which can fully demonstrate the advantages of the proposed model, that it is faster, more accurate and more stable than the other standalone ML models.

On the other hand, the development of VANET toward IoV brings some requirements and challenges. Therefore, considering V2X communication, integration of different access technology and simplify ensemble learning strategy while it provides higher accuracy and stability without imposing more time consumption and complexity, are the component of the design architecture and the proposed methods in chapters 4 and 5.

In chapter4, a simple soft-voting ensemble method is employed. It includes RF, KNN, and DT, for network traffic prediction in VNAET with V2X communication considering DSRC-based technologies for Communication between vehicles and other entities. EL strategies are capable of combining different AI models together to reinforce the flaws and limitations of each model and achieve a powerful traffic prediction model. The proposed method performs better than individual ML models considering classification metrics including accuracy, precision, F1-score and AUC-ROC value and time. In Chapter 5, the integrated DSRC and cellular-based network is designed in a V2X environment to deal with the issues related to coverage of short-range and long-range areas while we collect the traffic data through communication between vehicles, roadside units, infrastructure and pedestrians by applying a clustering method and intra-clustering communication. In this way, the nearest vehicles are grouped and communicated through selected vehicles as cluster heads with other entities. Moreover, NN is a well-selected method for traffic prediction integrated with the SI method to address the limitation of ANN and provide better performance of the model. Since, ML, DL and SI as a subset of AI in VANET traffic issues can reinforce each other to accomplish ideal solutions. Finally, the proposed model considers a trade-off between standalone models. It enhances the overall performance by increasing accuracy with minimum time and providing stability in the results.

The motivation of the combination of the selective algorithms can be justified with ML models in our approach require to be diverse due to the point that each single ML model has a different view about solving the prediction task and bring its advantage and disadvantage for traffic prediction. In summary, the efficiency of ML models relies on the size of datasets, the selected features, and the type of problems. Subsequently, we identified the best-fitted

model for network traffic prediction in VANET.

CHAPTER 7 CONCLUSION

7.1 Summary of Contributions

AI-based techniques are promising solutions for different domains of science because they can learn from data and solve problems. VANET challenges are also integrated with AI- methods to find an optimal solution. One of the VANET issues is related to maintaining the network performance for VANET applications, especially safety applications. Therefore, proposing an AI-based technique to predict traffic before it causes failure in the network performance is the main contribution of this dissertation. Considering, the limitations and challenges of single ML and DL models, proposing an efficient intelligent model that can address individual AI methods problems and achieve not only a more accurate but also stable network traffic prediction model. Another contribution is considering VANET with basic communication and V2X communication, which is an advanced type of communication toward IoV, and it can be more practical in real VANET application implementation. However, we need to deal with its requirement like the integration of different types of access technologies.

The network traffic prediction is considered a classification task. Therefore, we employ the most common classification evaluation metrics including precision, recall, F1 score, accuracy, the ROC curve and AUC value, the classification report, and the consumption time. We applied a supervised learning algorithm in this dissertation due to the best-match with our problem and datasets.

The Stacking Ensemble Learning with Booster Model (STK-EBM) is proposed as an efficient ensemble learning model for traffic prediction in the network. Firstly, because of the importance of the quality of input data, we apply an ensemble feature selection model with the aim of finding the most informative and effective features and parameters in the VANET dataset. We considered V2V and V2I as merged datasets as well. Secondly, the model employs a stacking strategy with a two-layer structure including the base layer and the meta layer. This model strategy needs to combine different ML models in the first layer, to realize which ML model can be a better choice, we require to theoretically and experimentally investigate on pros and cons of individual ML models in the literature that can perform well based on our datasets and problems. Therefore, we evaluate the performance of the baseline ML models considering different ML models and consider their limitations and advantages. Then, we integrate RF, KNN and XGBoost in the base layer to take advantage of each model while converting its drawbacks. In addition, XGBoost is capable of parallel learning ability and distributed computation that will not impose additional computational time on our model.

In the second layer, LR is selected as a combiner of prediction results from the first layer that can find the optimal combination of the prediction results while it is a simple model that can simplify the interpretation of the base learners' predictions. Moreover, we applied grid-search cross-validation that can optimize the hyperparameters of LR which lead to enhancing the accuracy of the model. Finally, the proposed model is compared with six commonly used ML models in terms of classification evaluation metrics to fully show the advantage of the proposed model that individual models.

It is interesting to note that identifying the best proper strategy to combine heterogeneous ML models and achieve enhanced performance is a challenging task. In the second contribution, the soft voting classification model is proposed for VANET considering V2X communication. This paper differentiates itself from the first objective, by using V2X communication which is more applicable to real-world traffic applications. Furthermore, the size of the data is doubled, therefore we propose a simple but effective ensemble model. The soft voting strategy model tried to take advantage of a simple ensemble strategy with the aim of reducing the complexity and enhancing the performance of the model than the standalone model. Therefore, it will not cause an increase in the computation resources. However, the STK-EBM model achieves much better stability in AUROC and recall metrics which are 0.955 and 0.98 respectively compared to the soft voting ensemble model in this paper with 0.907 for AUROC and 0.956 for recall. Eventually, vinous EL techniques can provide their advantages and disadvantages. We need to make the best decision regarding the selection of the ensemble learning methods based on our problem and dataset attributes, number of parameters, and size of recorded samples to achieve an efficient performance but not at the cost of time and complexity.

In the third objective, we take advantage of a well-chosen algorithm of traffic prediction based on the literature which is ANN, while addressing the limitation and challenges. On the other hand, we used simulated data that collected traffic data in an efficient way from V2X architecture. The V2X environment consists of V2V, V2R, R2I, and V2P communications that integrate DSRC and cellular-based technologies (5G) to provide short-range ad long-rang coverage for vehicular communications, unlike the second contribution which only applied DSRC communications. Moreover, the Cluster Head (CH) is elected among the nearest vehicles that are considered as a cluster to ease the communication among vehicles, roadside units and pedestrians and in turn data collection. In addition, the proposed model's effectiveness is validated by applying to a real-world VANET dataset with basic communication including V2V and V2I beside V2X simulated data with almost two times more recorded samples. In this way, we can analyze how NN and DNN can be effective considering different sizes of datasets and different numbers of features. Furthermore, conspiring DNN with two

and three hidden layers prove that the DL model can not be always a good solution for traffic prediction, especially in the case of small to medium size of datasets. Finally, comparative analysis of a single ANN and DNN with different numbers of layers, and two datasets with different sizes and attributes can provide us with the effect of input data on AI techniques which is highly important. However, the proposed model can achieve better performance in all metrics in both datasets which shows the generalizability of the model. Although, it works better for more data which is the proof point that the DL model needs more data to provide an efficient result.

The comparative analysis is based on the most common above-mentioned classification metrics that can comprehensively evaluate the performance of the proposed model. The experimental environment is based on the Google Colab platform and Python to execute the AI models. Different AI libraries including TensorFlow, Keras, PySwarms, Panda, NumPy, Matplotlib and Scikit-Learn are employed for different requirements of the implementation of AI models.

Finally, the ensemble learning (EL) model can be beneficial to provide a powerful prediction model and enhance the performance of the model, especially, if we cannot rely on using single ML models because of their limitations and challenges. However, realizing the best proposer strategy of combination among heterogeneous ML or DL models require lots of consideration that will not cause more problem in other aspects and achieve a trade-off in performance results.

7.2 Limitations

The main limitation of this dissertation for traffic prediction in VANET applications is the lack of real-world VANET-V2X publicly available dataset that consists of heterogeneous technologies (e.g., DSRC and 5G) and different communications between entities including vehicles, infrastructure, pedestrian, autonomous vehicles and so on. In this way, it can be more convenient for the real implementation of an application.

Regarding the real-world VANET dataset that we used, we are limited to the basic type of communication including V2V and V2I and DSRC-based technology for communication between vehicles and roadside units which is infrastructure in this dataset. However, the experiments were carried out in a realistic highway scenario in which the selected areas including the type of roads can be representative of the most roads in U.S cities.

Therefore, when we used publicly available datasets, we are limited to the consideration of the dataset including applied access technologies, communication types and provided features

and parameters. Accordingly, we should formulate our problem and methods based on the abovementioned considerations. In our case, we can consider our problem as a classification task to take full advantage of all provided features and parameters. In addition, we should employ a supervised learning method that applies labeled data to train the model and predict the results.

On the other hand, when it comes to the simulated environment and generated data, we are capable of designing the architecture based on what we require in terms of access technology, communication types, parameters and features for our problem. However, there are some limitations in simulation tools as well.

Although, the SUMO, OMNet++ and Veins are the most popular ad well-chosen simulators that can cooperate with each other and provide more realistic scenarios of traffic modelling and wireless communication. However, the data collected from running simulation scenarios, cannot produce big data in large-scale networks. Accordingly, for traffic prediction methods selection, we are limited to ML or simple DL models like ANN. In addition, the scalability of VNAET simulators that need to be adopted for more evolution of VNAET to ward IoV is a limitation.

Last but not least, In VANET communication, while we are employing AI techniques, require to ask about security authority to access the information about road users and their vehicle's location that might not desire to share with other entities on the road. This is why data privacy needs to be addressed in VANET communication data that are used by ML models.

7.3 Future Work

There are several considerations for future work based on this dissertation that are indicated in the following paragraphs.

- Flying ad Hoc Network (FANET), as the future of ad hoc networks can be employed for the next generation of VANET applications for efficient AI-based traffic prediction models. The ad hoc networking for flying vehicles is a subclass of VANET with significant beneficial characteristics including distributed nature, movement in three dimensions and high speed. Therefore, advanced applications are being designed for FANET which is the decentralized wireless communication between Unmanned Aerial Vehicles (UAVs) [131]. The emergence of UAVs that include GPS, ultrasonic sensors, and LiDAR [1], integrated with AI techniques can be effective for VANET applications to provide a greater perception of the surrounding environments for road users and adequate determination to control and manage network performance [132].

- When it comes to vehicular communication, there is a significant issue related to the energy efficiency communication system for VANET, that cause vehicular communication to be ineffective. Especially, in the case of large-scale networks. Although VANET brings many advantages, it also causes greenhouse gas emissions that have a serious negative impact on the environment and human life [133]. Moreover, energy efficiency is a significant challenge since the 6G/IoV system and AI combined in 5G/6G V2X application is one of the reasons for energy consumption because of big data processing [134]. Energy- efficiency in VANET can be accomplished by the traffic management system and AI-based techniques such as self-leaning and adaptive models that can be beneficial for providing less complex intelligent models while required energy is constantly monitored [135]. Green Vehicular Network (GVN) is the future of VANET. Green VANET infrastructure will sustain problems such as minimizing dangerous gas emissions, saving energy and decreasing environmental effects such as pollution [135].

- Regarding vehicular data, there are two aspects which are bring significant challenges including big vehicular data management and data security. These issues need to be considered to provide an efficient AI-powered model for VANET applications. VANET produces a huge amount of data from TB to PB that require to be managed for preserving bandwidth and channel capacity [1]. Furthermore, another significant challenge is linked with real-time data retrieval form such big data in VANET in terms of taking much time for analyzing, computing and data integration of various sensors and data storage [136]. Therefore, the promising solution to address the mentioned problems is the utilization of cloud and fog computing, that capable of using computing and storage resources in an efficient way. Big data management can be achieved by applying different ML techniques and data analysis approaches. However, real-time big vehicular data management is a considerable issue in future works. Modern architecture such as edge-cloud technology is a promising solution for managing latency and bandwidth especially in combined with a 5G/6G system that can offer new services and contribute to context-aware storage and computation in a distributed manner that led to not only keeping high bandwidth but also delay reduction for road users [132]. On the other hand, the next generation of V2X systems will bring new requirements in terms of managing a wide variety of applications, drastic standards and conditions for reliability, latency, power saving and so on [1]. In the case of a complex DL algorithm that needs more computational resources with respect to time especially for big data processing than traditional computing that performs a trade-off between efficiency and complexity [137], [138]. Quantum computing can be succeeded in achieving not only powerful computation and reduction the complexity issues, but also enhancing the

security of wireless communication, even in 6G-V2X that combined with secure data transmission, and security is highly critical that may cost human life [139].

- Finally, data security is considered a vital challenge in VANET [1]. Data need to be transmitted to the destination without any changes to its packet to provide reliable communication in VANET. This means the packet that is sent from the sender must be secure from a penetrator or attacker in a vehicular network that causes serious problems [1]. The solution is encryption and decryption of sensitive information by taking advantage of a cryptography algorithm. However, due to continuous changes in VANET topology resulted in significant issues in key management and key revocation which are solution techniques for covering keys from attackers [63].

REFERENCES

[1] N. H. Hussein, C. T. Yaw, S. P. Koh, S. K. Tiong, and K. H. Chong, "A comprehensive survey on vehicular networking: Communications, applications, challenges, and upcoming research directions," *IEEE Access*, vol. 10, pp. 86 127–86 180, Aug. 2022.

[2] P. A. D. Amiri and S. Pierre, "An ensemble-based machine learning model for forecasting network traffic in vanet," *IEEE Access*, vol. 11, pp. 22 855–22 870, Mar 2023.

[3] N. Taherkhani and S. Pierre, "Improving dynamic and distributed congestion control in vehicular ad hoc networks," *Ad Hoc Networks*, vol. 33, pp. 112–125, Oct. 2015.

[4] D. Alekseeva, N. Stepanov, A. Veprev, A. Sharapova, E. S. Lohan, and A. Ometov, "Comparison of machine learning techniques applied to traffic prediction of real wireless network," *IEEE Access*, vol. 9, pp. 159 495–159 514, Nov. 2021.

[5] S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, "A comprehensive survey on vehicular ad hoc network," *Journal of Network and Computer Applications*, vol. 37, pp. 380–392, Mar. 2014.

[6] F. Yang, J. Han, X. Ding, Z. Wei, and X. Bi, "Spectral efficiency optimization and interference management for multi-hop d2d communications in vanets," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6422–6436, June 2020.

[7] A. Mchergui, T. Moulahi, and S. Zeadally, "Survey on artificial intelligence (ai) techniques for vehicular ad-hoc networks (vanets)," *Vehicular Communications*, vol. 34, p. 100403, Apr. 2022.

[8] W. Tong, A. Hussain, W. X. Bo, and S. Maharjan, "Artificial intelligence for vehicle-to-everything: A survey," *IEEE Access*, vol. 7, pp. 10 823–10 843, Jan. 2019.

[9] Q. Pan, J. Wu, J. Nebhen, A. K. Bashir, Y. Su, and J. Li, "Artificial intelligence-based energy efficient communication system for intelligent reflecting surface-driven vanets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 714–19 726, Mar. 2022.

[10] A. Al-Dulaimi, S. Al-Rubaye, and Q. Ni, "Energy efficiency using cloud management of lte networks employing fronthaul and virtualized baseband processing pool," *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 403–414, Apr. 2019.

[11] J. E. Siegel, D. C. Erb, and S. E. Sarma, "A survey of the connected vehicle landscape—architectures, enabling technologies, applications, and development areas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 8, pp. 2391–2406, Aug. 2018.

[12] B.-M. Cho, M.-S. Jang, and K.-J. Park, "Channel-aware congestion control in vehicular cyber-physical systems," *IEEE Access*, vol. 8, pp. 73 193–73 203, Apr. 2020.

[13] K. Lazhar, N. Labraoui, A. Gueroui, and A. Ari, "Enhancing video dissemination over urban vanets using line of sight and qoe awareness mechanisms," *Annals of Telecommunications*, vol. 76, Jul. 2021.

[14] F. Tang, B. Mao, N. Kato, and G. Gui, "Comprehensive survey on machine learning in vehicular network: Technology, applications and challenges," *IEEE Communications Surveys Tutorials*, vol. 23, no. 3, pp. 2027–2057, June 2021.

[15] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," in *2018 15th Learning and Technology Conference (LT)*. Jeddah, Saudi Arabia: IEEE, 25-26 Feb 2018, pp. 40–45.

[16] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers - a tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul 2021.

[17] S. Cheng, F. Lu, P. Peng, and S. Wu, "Short-term traffic forecasting: An adaptive st-knn model that considers spatial heterogeneity," *Computers, Environment and Urban Systems*, vol. 71, pp. 186–198, Sept. 2018.

[18] L. Tong, S. Shi, and X. Gu, "Naive bayes classifier based driving habit prediction scheme for vanet stable clustering," *Mobile Networks and Applications*, vol. 25, 10 2020.

[19] S. S. Sepasgozar and S. Pierre, "A comparative study of artificial intelligence algorithms for network traffic prediction in vanet," in *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. Thessaloniki, Greece: IEEE, 10-12 Oct. 2022, pp. 431–436.

[20] J. Zhang, C. Chen, Y. Xiang, W. Zhou, and Y. Xiang, "Internet traffic classification by aggregating correlated naive bayes predictions," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 5–15, Jan. 2013.

[21] F. Falahatraftar, S. Pierre, and S. Chamberland, "A centralized and dynamic network congestion classification approach for heterogeneous vehicular networks," *IEEE Access*, vol. 9, pp. 122 284–122 298, Aug. 2021.

[22] H.-F. Yang, T. S. Dillon, and Y.-P. P. Chen, "Optimized structure of the traffic flow forecasting model with a deep learning approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2371–2381, Oct. 2017.

[23] Y.-J. Lee and O. Min, "Comparative analysis of machine learning algorithms to urban traffic prediction," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*. Jeju, Korea (South): IEEE, 18-20 Oct 2017, pp. 1034–1036.

[24] A. Dimara, D. Triantafyllidis, S. Krinidis, K. Kitsikoudis, D. Ioannidis, E. Valkouma, S. Skarvelakis, S. Antipas, and D. Tzovaras, "Mlp for spatio-temporal traffic volume forecasting," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. Toronto, ON, Canada: IEEE, 21-24 Apr. 2021, pp. 1–7.

[25] A. Boukerche, Y. Tao, and P. Sun, "Artificial intelligence-based vehicular traffic flow prediction methods for supporting intelligent transportation systems," *Computer Networks*, vol. 182, p. 107484, Dec. 2020.

[26] J. Tang, G. Liu, and Q. Pan, "A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 10, pp. 1627–1643, Oct. 2021.

[27] S. Bitam, A. Mellouk, and S. Zeadally, "Bio-inspired routing algorithms survey for vehicular ad hoc networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 843–867, Nov. 2015.

[28] B. Ranjan Senapati and P. Mohan Khilar, *Optimization of Performance Parameter for Vehicular Ad-hoc NETwork (VANET) Using Swarm Intelligence*. Cham: Springer International Publishing, Nov. 2020, p. 83—107.

[29] M. Bany Taha, C. Talhi, H. Ould-Slimane, and S. Alrabaee, "Td-pso: Task distribution approach based on particle swarm optimization for vehicular ad hoc network," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 3, p. e3860, 2022.

[30] R. R. Violanda and C. C. Bernido, "Modeling vehicular speed fluctuations as a stochastic process with exponentially decaying memory," *AIP Conference Proceedings*, vol. 2286, no. 1, Dec. 2020.

[31] F. Abbas and P. Fan, "Clustering-based reliable low-latency routing scheme using aco method for vehicular networks," *Vehicular Communications*, vol. 12, pp. 66–74, Apr. 2018.

[32] Y. Li, "Application of eos-elm with binary jaya-based feature selection to real-time transient stability assessment using pmu data," *IEEE Access*, vol. 5, pp. 23 092 – 23 101, Oct 2017.

[33] J. Li, N. Song, G. Yang, M. Li, and Q. Cai, "Improving positioning accuracy of vehicular navigation system during gps outages utilizing ensemble learning algorithm," *Information Fusion*, vol. 35, pp. 1–10, May. 2017.

[34] I. Nti, A. Adekoya, and B. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *Journal of Big Data*, vol. 7, Mar 2020.

[35] R. Das and P. M. Khilar, "Driver behaviour profiling in vanets: Comparison of ensemble machine learning techniques," in *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*. Chennai, India: IEEE, 04-06 July 2019, pp. 1–5.

[36] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2. Hangzhou, China: IEEE, 09-10 Dec 2017, pp. 361–364.

[37] M. Rasyidi, J. Kim, and K. Ryu, "Short-term prediction of vehicle speed on main city roads using the k-nearest neighbor algorithm," *Journal of Intelligence and Information Systems*, vol. 20, Mar. 2014.

[38] K. Jadaan, M. Al-Fayyad, and H. Gammoh, "Prediction of road traffic accidents in jordan using artificial neural network (ann)," *Journal of Traffic and Logistics Engineering*, vol. 2, pp. 92–94, Jan. 2014.

[39] W. Zhang, Y. Yu, Y. Qi, F. Shu, and Y. Wang, "Short-term traffic flow prediction based on spatio-temporal analysis and cnn deep learning," *Transportmetrica A: Transport Science*, vol. 15, pp. 1–45, July 2019.

[40] P. Sun, A. Boukerche, and Y. Tao, "Ssgru: A novel hybrid stacked gru-based traffic volume prediction approach in a road network," *Computer Communications*, vol. 160, pp. 502–511, July 2020.

[41] E. Khoza, C. Tu, and P. Owolawi, "Decreasing traffic congestion in vanets using an improved hybrid ant colony optimization algorithm," *Journal of Communications*, pp. 676–686, Jan 2020.

[42] H. Ding, H. Wu, L. Dong, and Z. Li, "Vehicle intersection collision monitoring algorithm based on vanets and uncertain trajectories," in *2018 16th International Conference on Intelligent Transportation Systems Telecommunications (ITST)*. Lisboa, Portugal: IEEE, 15-17 Oct 2018, pp. 1–7.

[43] J. Guo, B. Song, F. R. Yu, Y. Chi, and C. Yuen, "Fast video frame correlation analysis for vehicular networks by using cvs–cnn," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6286–6292, 2019.

[44] P. A. D. Amiri and S. Pierre, "Swarm-based ensemble model for network traffic prediction considering basic and v2x communication in vanet," *IEEE Access*, 2023.

[45] P. Ahmadi Doval Amiri and S. Pierre, "A soft voting classification model for network traffic prediction in vanet/v2x," in *2023 19th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. Montreal, QC, Canada: IEEE, 21-23 Jun. 2023, pp. 231–236.

[46] X. Feng, X. Ling, H. Zheng, Z. Chen, and Y. Xu, "Adaptive multi-kernel svm with spatial–temporal correlation for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2001–2013, Sep 2019.

[47] R. Sathya and A. Abraham, "Comparison of supervised and unsupervised learning algorithms for pattern classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, Feb. 2013.

[48] O. Aouedi, K. Piamrat, and B. Parrein, "Performance evaluation of feature selection and tree-based algorithms for traffic classification," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. Montreal, QC, Canada: IEEE, 14-23 Jun. 2021, pp. 1–6.

[49] J. Jenifer and R. Priyadarsini, "An ensemble based machine learning approach for traffic prediction in smart city," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*. Coimbatore, India: IEEE, 08-09 Oct. 2021, pp. 1–6.

[50] R.-C. Chen, C. Dewi, S. Huang, and R. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal Of Big Data*, vol. 7, p. 26, Jul 2020.

[51] J. Riihijarvi and P. Mahonen, "Machine learning for performance prediction in mobile cellular networks," *IEEE Computational Intelligence Magazine*, vol. 13, no. 1, pp. 51–60, Jan 2018.

[52] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. Tirunelveli, India: IEEE, 20-22 Aug. 2020, pp. 729–735.

[53] Google . (2022) Google colaboratory: Frequently asked questions. [Online]. Available: https://research.google.com/colaboratory/intl/en-GB/faq.html

[54] N. Stepanov, D. Alekseeva, A. Ometov, and E. S. Lohan, "Applying machine learning to lte traffic prediction: Comparison of bagging, random forest, and svm," in *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. Brno, Czech Republic: IEEE, 05-07 Oct. 2020, pp. 119–123.

[55] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-nearest neighbors and grid search cv based real time fault monitoring system for industries," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. Bombay, India: IEEE, 29-31 Mar. 2019, pp. 1–5.

[56] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and xgboost," *IEEE Access*, vol. 6, pp. 21 020–21 031, Apr 2018.

[57] R. Khanna and M. Awad, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Springer, Apr. 2015.

[58] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," vol. 06. ACM, Jun 2006.

[59] H. Khelifi, S. Luo, B. Nour, H. Moungla, Y. Faheem, R. Hussain, and A. Ksentini, "Named data networking in vehicular ad hoc networks: State-of-the-art and challenges," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 320–351, 2020.

[60] S. Sharma and B. Kaushik, "A survey on internet of vehicles: Applications, security issues solutions," *Vehicular Communications*, vol. 20, p. 100182, May. 2019.

[61] Y. Toor, P. Muhlethaler, A. Laouiti, and A. D. La Fortelle, "Vehicle ad hoc networks: applications and related technical issues," *IEEE Communications Surveys Tutorials*, vol. 10, no. 3, pp. 74–88, Sept. 2008.

[62] M. Lee and T. Atkison, "Vanet applications: Past, present, and future," *Vehicular Communications*, vol. 28, p. 100310, Oct. 2021.

[63] B. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, and D. Wang, "Survey on the internet of vehicles: Network architectures and applications," *IEEE Communications Standards Magazine*, vol. 4, no. 1, pp. 34–41, Mar. 2020.

[64] L. U. Khan, "Visible light communication: Applications, architecture, standardization and research challenges," *Digital Communications and Networks*, vol. 3, no. 2, pp. 78–88, 2017.

[65] A. Daniel, A. Paul, A. Ahmad, and S. Rho, "Cooperative intelligence of vehicles for intelligent transportation systems (its)," *Wirel. Pers. Commun.*, vol. 87, no. 2, pp. 461–484, Mar 2016.

[66] C. Lai, D. Zheng, Q. Zhao, and X. Jiang, "Segm: A secure group management framework in integrated vanet-cellular networks," *Vehicular Communications*, vol. 11, pp. 33–45, Feb. 2018.

[67] A. Bazzi, B. M. Masini, A. Zanella, C. De Castro, C. Raffaelli, and O. Andrisano, "Cellular aided vehicular named data networking," in *2014 International Conference on Connected Vehicles and Expo (ICCVE)*. Vienna, Austria: IEEE, 03-07 Nov. 2014, pp. 747–752.

[68] M. A. Al-Absi, A. A. Al-Absi, M. Sain, and H. Lee, "Moving ad hoc networks—a comparative study," *Sustainability*, vol. 13, no. 11, May. 2021.

[69] A. Srivastava and J. Prakash, "Future fanet with application and enabling techniques: Anatomization and sustainability issues," *Computer Science Review*, vol. 39, p. 100359, Feb. 2021.

[70] O. S. Oubbati, M. Atiquzzaman, P. Lorenz, M. H. Tareque, and M. S. Hossain, "Routing in flying ad hoc networks: Survey, constraints, and future challenge perspectives," *IEEE Access*, vol. 7, pp. 81 057–81 105, Jun. 2019.

[71] R. A. Nazib and S. Moh, "Routing protocols for unmanned aerial vehicle-aided vehicular ad hoc networks: A survey," *IEEE Access*, vol. 8, pp. 77 535–77 560, Apr. 2020.

[72] B. Hament and P. Oh, "Unmanned aerial and ground vehicle (uav-ugv) system prototype for civil infrastructure missions," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*. Las Vegas, NV, USA: IEEE, 12-14 Jan 2018, pp. 1–4.

[73] R. Frank, W. Bronzi, G. Castignani, and T. Engel, "Bluetooth low energy: An alternative technology for vanet applications," in *2014 11th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*. Obergurgl, Austria: IEEE, 02-04 Apr 2014, pp. 104–107.

[74] A. M. Said, M. Marot, A. W. Ibrahim, and H. Afifi, "Modeling interactive real-time applications in vanets with performance evaluation," *Comput. Netw.*, vol. 104, no. C, p. 66–78, Jul 2016.

[75] R. Khatoun, P. Gut, R. Doulami, L. Khoukhi, and A. Serhrouchni, "A reputation system for detection of black hole attack in vehicular networking," in *2015 International Conference on Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC)*. Shanghai, China: IEEE, 05-07 Aug 2015, pp. 1–5.

[76] A. Ullah, X. Yao, S. Shaheen, and H. Ning, "Advances in position based routing towards its enabled fog-oriented vanet–a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 2, pp. 828–840, Feb. 2020.

[77] A. Chattopadhyay, K.-Y. Lam, and Y. Tavva, "Autonomous vehicle: Security by design," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 7015–7029, 2021.

[78] 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)*. Jaipur, India: IEEE, 07-08 Feb 2020, pp. 145–148.

[79] J. Tong, X. Gu, M. Zhang, J. Wan, and J. Wang, "Traffic flow prediction based on improved svr for vanet," in *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. Changsha, China: IEEE, 26-28 Mar 2021, pp. 402–405.

[80] Z. Li, D. Cai, J. Wang, J. Fu, L. Qin, and D. Fu, "A stacking ensemble learning model for mobile traffic prediction," in *2020 IEEE/CIC International Conference on*

Communications in China (ICCC). Chongqing, China: IEEE, 09-11 Aug 2020, pp. 542–547.

[81] G. Zheng, W. K. Chai, and V. Katos, "An ensemble model for short-term traffic prediction in smart city transportation system," in *2019 IEEE Global Communications Conference (GLOBECOM)*. Waikoloa, HI, USA: IEEE, Dec. 2019, pp. 1–6.

[82] F. K. Oduro-Gyimah, K. O. Boateng, P. B. Adu, and K. Quist-Aphetsi, "Prediction of telecommunication network outage time using multilayer perceptron modelling approach," in *2021 International Conference on Computing, Computational Modelling and Applications (ICCMA)*. Brest, France: IEEE, 14-16 Jul 2021, pp. 104–108.

[83] F. Zhao, G.-Q. Zeng, and K.-D. Lu, "Enlstm-wpeo: Short-term traffic flow prediction by ensemble lstm, nnct weight integration, and population extremal optimization," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 101–113, 2020.

[84] C. Chen, Z. Liu, S. Wan, J. Luan, and Q. Pei, "Traffic flow prediction based on deep learning in internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3776–3789, 2021.

[85] S. S. Sepasgozar and S. Pierre, "Network traffic prediction model considering road traffic parameters using artificial intelligence methods in vanet," *IEEE Access*, vol. 10, pp. 8227–8242, Jan 2022.

[86] B. Dasarathy and B. Sheela, "A composite classifier system design: Concepts and methodology," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 708–713, May 1979.

[87] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, Sep 2022.

[88] M. Sabzevari, G. Martínez-Muñoz, and A. Suárez, "Building heterogeneous ensembles by pooling homogeneous ensembles," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 2, pp. 551–558, Feb 2022.

[89] K. A. Nguyen, W. Chen, B.-S. Lin, and U. Seeboonruang, "Comparison of ensemble machine learning methods for soil erosion pin measurements," *ISPRS International Journal of Geo-Information*, vol. 10, no. 1, Jan 2021.

[90] D. Ballabio, R. Todeschini, and V. Consonni, "Chapter 5 - recent advances in high-level fusion methods to classify multiple analytical chemical data," in *Data Fusion*

Methodology and Applications, ser. Data Handling in Science and Technology. Elsevier, 2019, vol. 31, pp. 129–155.

[91] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[92] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[93] L. Liang, H. Peng, G. Y. Li, and X. Shen, "Vehicular communications: A physical layer perspective," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 10 647–10 659, 2017.

[94] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeadally, "5g for vehicular communications," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 111–117, 2018.

[95] M. S. Sheikh, J. Liang, and M. A. Khan, "A comprehensive survey on vanet security services in traffic management system," *Wirel. Commun. Mob. Comput.*, vol. 2019, Jan 2019.

[96] X. Huang, D. Zhao, and H. Peng, "Empirical study of dsrc performance based on safety pilot model deployment data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2619–2628, 2017.

[97] M. Faezipour, M. Nourani, A. Saeed, and S. Addepalli, "Progress and challenges in intelligent vehicle area networks," *Commun. ACM*, vol. 55, no. 2, p. 90–100, Feb 2012.

[98] Y. Wu and H. Tan, "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework," *ArXiv*, vol. abs/1612.01022, 2016.

[99] Z. Zheng, Y. Yang, J. Liu, H.-N. Dai, and Y. Zhang, "Deep and embedded learning approach for traffic flow prediction in urban informatics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3927–3939, 2019.

[100] R. Goścień and A. Knapińska, "Efficient network traffic prediction after a node failure," in *2022 International Conference on Optical Network Design and Modeling (ONDM)*. Warsaw, Poland: IEEE, 16-19 May. 2022, pp. 1–6.

[101] Z. Liao, M. Su, G. Ning, Y. Liu, T. Wang, and J. Zhou, "A novel stacked generalization ensemble-based hybrid psvm-pmlp-mlr model for energy consumption prediction of copper foil electrolytic preparation," *IEEE Access*, vol. 9, pp. 5821–5831, Jan 2021.

[102] X. Zhang and F. Ren, "Improving svm learning accuracy with adaboost," in *2008 Fourth International Conference on Natural Computation*, vol. 3. Jinan, China: IEEE, 18-20 Oct 2008, pp. 221–225.

[103] Z. Jianjun, X. Yuanbiao, and F. Renhai, "Network traffic forecasting based on logistic iterative regression model," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*. Shanghai, China: IEEE, 12-15 Sep 2020, pp. 424–429.

[104] S. Xin, X. Yuanbiao, Z. Qijia, L. Zhimao, and F. Renhai, "Traffic forecasting of core network based on improved logistic regression," in *2021 IEEE 9th International Conference on Information, Communication and Networks (ICICN)*. Xi'an, China: IEEE, 25-28 Nov 2021, pp. 102–106.

[105] J. M. Hilbe, *Logistic Regression*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, p. 755—758.

[106] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Incorporated, 2019. [Online]. Available: https://books.google.ca/books?id=OCS1twEACAAJ

[107] M. P. H. H. W. M. P. J. L. R. M. Fujimoto, R. Guensler and J. Ko. (2006) Crawdad dataset gatech/vehicular retrieved from crawdad dataset. [Online]. Available: https://crawdad.org/gatech/vehicular/20060315

[108] H. Wu, M. Palekar, R. Fujimoto, R. Guensler, M. Hunter, J. Lee, and J. Ko, "An empirical study of short range communications for vehicles," ser. VANET '05. Association for Computing Machinery, 2005, pp. 83–84.

[109] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, "Using the jupyter notebook as a tool for open science: An empirical study," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Toronto, ON, Canada: IEEE, 19-23 Jun 2017, pp. 1–2.

[110] T. Carneiro, R. V. Medeiros Da NóBrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. R. Filho, "Performance analysis of google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61 677–61 685, 2018.

[111] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," p. 2825–2830, 2018.

[112] G. Figueroa, Y.-S. Chen, N. Avila, and C.-C. Chu, "Improved practices in machine learning algorithms for ntl detection with imbalanced data," in *2017 IEEE Power Energy Society General Meeting*. Chicago, IL, USA: IEEE, 16-20 Jul 2017, pp. 1–5.

[113] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system." San Francisco, California, USA: Association for Computing Machinery, 13 Aug 2016, p. 785–794.

[114] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2x access technologies: Regulation, research, and remaining challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1858–1877, 2018.

[115] O. contributors. (2017) Planet dump retrieved from https://planet.osm.org. [Online]. Available: https://www.openstreetmap.org.

[116] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. Maui, HI, USA: IEEE, 4-7 Nov 2018, pp. 2575–2582. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8569938

[117] C. Sommer, R. German, and F. Dressler, "Bidirectionally coupled network and road traffic simulation for improved ivc analysis," *IEEE Transactions on Mobile Computing*, vol. 10, no. 1, pp. 3–15, 2011.

[118] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A communications-oriented perspective on traffic management systems for smart cities: Challenges and innovative approaches," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 125–151, 2015.

[119] N. Ramakrishnan and T. Soni, "Network traffic prediction using recurrent neural networks," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Orlando, FL, USA: IEEE, 17-20 Dec 2018, pp. 187–193.

[120] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[121] L. Liu, X. Zhang, Y. Liu, W. Zhu, and B. Zhao, "An ensemble of multiple boosting methods based on classifier-specific soft voting for intelligent vehicle crash injury severity prediction," in *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*. Oxford, UK: IEEE, 03-06 Aug 2020, pp. 17–24.

[122] P. Ignaciuk and A. Bartoszewicz, *Congestion Control in Data Transmission Networks: Sliding Mode and Other Designs*. springer, 01 2013.

[123] A. Paranjothi, M. S. Khan, and S. Zeadally, "A survey on congestion detection and control in connected vehicles," *Ad Hoc Networks*, vol. 108, p. 102277, Nov. 2020.

[124] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2377–2396, 2015.

[125] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of dsrc and cellular network technologies for v2x communications: A survey," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9457–9470, 2016.

[126] R. Regin and T. Menakadevi, "A novel clustering technique to stop congestion occur vehicular ad-hoc networks using node density based on received signal strength," *Peer-to-Peer Networking and Applications*, vol. 14, pp. 2361 – 2371, 2020.

[127] R. R and T. Menakadevi, "Investigation relationship between network congestion and vehicle density in vanets," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2021, pp. 558–562.

[128] G. Beni and J. Wang, "Swarm intelligence in cellular robotic systems." Tuscany, Italy: Springer, Jun 1993.

[129] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4. Perth, WA, Australia: IEEE, 27 Nov-01 Dec. 1995, pp. 1942–1948 vol.4.

[130] D. B. Fogel and H.-G. Beyer, "A note on the empirical evaluation of intermediate recombination," *Evol. Comput.*, vol. 3, no. 4, p. 491–495, dec 1995.

[131] L. Zhao, M. B. Saif, A. Hawbani, G. Min, S. Peng, and N. Lin, "A novel improved artificial bee colony and blockchain-based secure clustering routing scheme for fanet," *China Communications*, vol. 18, no. 7, pp. 103–116, 2021.

[132] P. Arthurs, L. Gillam, P. Krause, N. Wang, K. Halder, and A. Mouzakitis, "A taxonomy and survey of edge cloud computing for intelligent transportation systems and connected vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6206–6221, 2022.

[133] Y. Yuan, Y. Zhao, Z. Baiqing, and S. Parolari, "Potential key technologies for 6g mobile communications," *Science China Information Sciences*, vol. 63, 03 2020.

[134] Y. Su, M. LiWang, L. Huang, X. Du, and N. Guizani, "Green communications for future vehicular networks: Data compression approaches, opportunities, and challenges," *IEEE Network*, vol. 34, no. 6, pp. 184–190, 2020.

[135] J. Wang, K. Zhu, and E. Hossain, "Green internet of vehicles (iov) in the 6g era: Toward sustainable vehicular communications and networking," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 1, pp. 391–423, 2022.

[136] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383–398, 2019.

[137] I. F. Akyildiz, A. Kak, and S. Nie, "6g and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133 995–134 030, 2020.

[138] P. Botsinis, D. Alanis, Z. Babar, H. V. Nguyen, D. Chandra, S. X. Ng, and L. Hanzo, "Quantum search algorithms for wireless communications," *IEEE Communications Surveys Tutorials*, vol. 21, no. 2, pp. 1209–1242, 2019.

[139] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6g," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020.

APPENDIX A CLASSIFICATION PERFORMANCE METRICS

$$Recall = \frac{TP}{TP + FN} \tag{A.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{A.2}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{A.3}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{A.4}$$