Exploring prediction strategies in vehicular networks through machine learning techniques and hybrid intelligence

Mamoudou Sangare

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE ABBÉ GRÉGOIRE
Laboratoire Cedric

THÈSE

présentée par : **Mamoudou SANGARÉ**

soutenue le : **11 juillet 2022**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au: **Conservatoire national des arts et métiers**

Discipline : **Section CNU 27**

Spécialité : **Informatique**

Exploration de Stratégies de Prédiction dans les Réseaux Véhiculaires en utilisant des Algorithmes d'Apprentissage et d'Intelligence Hybride

THÈSE dirigée par :
Mme BOUZEFRANE Samia

et co-encadrée par :
M. MUHLETHALER Paul

Jury

Mme Ndèye NIANG KEITA,	Maître de Conférences/HDR, Cedric, CNAM	Présidente
M. Cheikh Ahmadou Bamba GUEYE,	Professeur des universités, Université Cheikh Anta Diop,	Rapporteur
M. Saadi BOUDJIT,	Maître de Conférences/HDR, Université Sorbonne Paris Nord	Rapporteur
M. Thinh LE VINH,	Maître de Conférences, University of Technology and Education, Ho Chi Minh City,	Examinateur
Soumya BANERJEE,	Docteur, Trasna Solutions,	Invité

THÈSE

HESAM UNIVERSITÉ

le cnam

ABBE GREGOIRE DOCTORAL SCHOOL
Cedric Laboratory

THESIS

presented by : **Mamoudou SANGARÉ**

defended on the : **11 july 2022**

to obtain the rank of : **Doctor of HESAM University**

prepared at : **Conservatoire national des arts et métiers**

Discipline : **Section CNU 27**

Speciality : **Computer Science**

Exploring Prediction Strategies in Vehicular Networks Through Machine Learning Techniques and Hybrid Intelligence

THISIS supervised by :
Mme BOUZEFRANE Samia

and co-supervised by :
M. MUHLETHALER Paul

Jury

Mme Ndèye NIANG KEITA,	Associate Professor /ASR, Cedric, CNAM	President
M. Cheikh Ahmadou Bamba GUEYE,	University Professor, University of Cheikh Anta Diop,	Rapporteur
M. Saadi BOUDJIT,	Associate Professor /ASR, University of Sorbonne Paris Nord	Rapporteur
M. Thinh LE VINH,	Associate Professor, University of Technology and Education, Ho Chi Minh City,	Examiner
Soumya BANERJEE,	PhD, Trasna Solutions,	Guest

THESIS

To my very dear mother, Hawagbe CAMARA

Strong Woman ! Fighting Woman ! Humble Woman ! Courageous Woman ! Lovable Woman ! Whatever I express, it cannot describe you at your fair value. No matter what I do or say, I won't be able to thank you properly. Your affection covers me, your benevolence guides me and your presence by my side has always been my source of strength to face the various obstacles.

To my very dear father, Mohamed Lamine SANGARE

Disappeared at the dawn of my life. I hope that, from the world that is yours now, you appreciate this humble gesture as proof of gratitude from a son who has always prayed for the salvation of your soul. May Almighty God welcome you into His holy mercy !

To my late uncle Elhadj Kaab SANGARE.

Who supported and encouraged me during these years of study. May he find here the testimony of my deep gratitude.

To my dear brothers and sisters

May God give you health, happiness, courage and above all success

Remerciements

First and foremost I would like to thank You, God Almighty ; the most beneficent and merciful without you, nothing of this research work could have been achieved.

I would like to express my sincere gratitude and appreciation to my thesis advisors, Prof. Samia BOUZEFRANE, Dr. Paul MUHLETHALER and Dr. Soumya BANERJEE, for their invaluable guidance, constant encouragement, and endless patience during my Ph.D. study. Thanks to them, I had the opportunity to conduct and complete my PhD thesis between Cnam and Inria. I will be forever honored and grateful for working with them. Without their help and advice and constant guidance, this thesis would have gone for long and I would not have been able to write this thesis. They inspired me to do my best to accomplish my goals. They helped me to grow professionally even emotionally and I have learned many skills from them that will be of most significant assistance in my future career, especially hard work and patience.

My gratitude extends Prof. Cheikh Ahmadou Bamba GUEYE and Dr. Saadi BOUDJIT for accepting to review my manuscript. Also, my thanks go to the members of the PhD defense committee for the interest they showed in my research Dr. Ndeye NIANG and Dr. Thinh LE VINH. I would also like to thank Dr Soumya BANERJEE for accepting the invitation to participate in the defense.

I would moreover like to thank my parents especially my mom as my father passed away at early stage of my life. She was always supporting me and encouraging me. To all my brothers and sisters who always care and boost up my feelings. To my friends for your continuous support, help and encouragement.

To all the team of Eva at Inria and to the ROC team of the CEDRIC Lab/Cnam, staff

and colleagues, thank you for everything. Your optimism, kindness, and perseverance made my Ph.D. experience unique and unforgettable.

This thesis is dedicated to you, CAMARA Hawagbe, my dearest mother, who has always been present at my side to support and encourage me although we were distant from each other. Words cannot express how grateful I am for your support, encouragement, and kind wishes. Your love and your prayers have sustained me over the years.

Finally, I must take this opportunity to express my heartfelt gratitude to my little family, Mohamed-Daye SANGARÉ and his mother who suffered a lot due to my occupations in this thesis. But you knew it ; the more beautiful is coming for all of us ! Thank you my little family.

Résumé

Les accidents de la route sont l'un des problèmes majeurs auxquels sont confrontés les pays du monde entier. Les dommages qu'ils causent vont des blessures graves, des pertes économiques considérables à la mort, aux individus, à leurs familles et aux nations dans leur ensemble. Selon certaines statistiques, 1,3 million de personnes meurent chaque année des suites d'accidents de la route dans le monde, ce qui signifie que presque toutes les 25 secondes, une personne perd la vie. Bien qu'alarmants, ils devraient augmenter de 65 % et deviendront la cinquième cause de décès dans la décennie à venir. Il ressort de certaines statistiques que les erreurs humaines sont la principale cause des accidents de la route. Et si les conducteurs sont avertis en une demi-seconde (0,5 seconde) à temps et sont conscients des zones sujettes aux accidents, 60% de ces accidents de la route peuvent être évités. Dans ce contexte et pour pallier cette situation, un type particulier de réseaux appelés Vehicular Adhoc NETworks (VANETs) a vu le jour. L'objectif principal de ces réseaux est de diminuer le taux d'accidents sur la route et également d'assurer le confort des passagers. Dans les VANET, les véhicules communiquent entre eux et également avec l'infrastructure se trouvant le long de la route appelée RoadSide Unit (RSU). Dans ces scénarios de communication, différents types de données sont échangées entre les nœuds participants, allant des messages de sécurité routière à la gestion du trafic en passant par l'infodivertissement. Sur la base de la situation alarmante présentée ci-dessus, les applications de sécurité deviennent l'un des centres d'intérêt des acteurs majeurs dans ce domaine, des industries des constructeurs automobiles aux organisations gouvernementales en passant par les chercheurs, de nos jours.

Dans cette thèse, nous nous intéressons particulièrement au niveau des applications de sécurité qui sont conçues pour fournir une assistance aux conducteurs dans des situations

dangereuses et pour éviter les accidents. Nous présentons d'abord un aperçu des différentes fonctionnalités et caractéristiques des réseaux ad hoc véhiculaires (VANET). Nous présentons ensuite les projets les plus prometteurs et les plus importants entrepris dans le monde sur les VANET. Nous présentons également l'état des activités de normalisation dans le monde entier. En plus de cela, nous passons en revue l'état de l'art. Ensuite, nous décrivons et présentons la prédiction du positionnement du véhicule et la prédiction du succès de la transmission à l'aide des algorithmes d'apprentissage automatique (ML) dans les VANETs. De plus, nous étudions la prévision des accidents de la route à l'aide de ML dans les VANETs. Nous adoptons un modèle dans lequel nous présentons un modèle de prédiction qui combine deux approches. L'approche hybride proposée intègre les avantages des modèles génératifs (GMM) et discriminants (SVC). Par rapport au modèle statistique de base (GMM), notre modèle a obtenu jusqu'à 24% de précision. Nous proposons enfin un modèle de correspondance de contenu plus large de chaînes de confiance et de filtrage basé sur la blockchain pour les véhicules connectés, où un contenu et des en-têtes de sujet sont d'abord mis en correspondance, puis le résultat est consolidé par un mécanisme de vote par consensus de blockchain distribué pour toute décision prise concernant l'évaluation de la confiance.

Mots clés : Réseaux ad hoc véhiculaires(VANETs), CAMs (Cooperative Awareness Message/Message de sensibilisation coopérative),DENMs (Decentralized Environmental Notification Messages/Messages Décentralisés de Notification Environnementale), Positionnement, Prévision d'accidents, Apprentissage automatique avec VANETs, Protocole de transmission de messages avec Blockchain.

Abstract

Accidents on the road are one of the serious problems facing countries around the world. The damage that they cause ranges from severe injuries, considerable economic loss to death, and this is true for individuals, their families, and for nations as a whole. According to some statistics, 1.3 million people die each year as a result of road traffic accidents across the globe, which means that almost every 25 second a person loses his life. Already alarming, these statistics are expected to grow by 65% and to become the fifth greatest cause of fatalities in the coming decade. Some statistics highlight the fact that human error is the main cause of road accidents, and if drivers could be warned half a second (0.5 second) beforehand and were aware of the accident prone area, 60% of these road accidents could be prevented. In this context and to overcome this situation, a particular type of network, known as Vehicular Ad hoc NETworks(VANETs) has emerged. The primary goal of these networks is to diminish the number of road accidents and to provide comfort services to passengers. In VANETs, vehicles communicate with each other and also with the infrastructure along the road made up of RoadSide Units (RSU). In these communication scenarios, different kinds of data are exchanged among participating nodes, ranging from road safety messages, traffic management to infotainment. Due to the alarming situation presented above, safety applications are now becoming one of the focuses of major players in this field from automobile manufacturer industries to government organizations as well as researchers.

In this thesis, we particularly focus on the level of safety applications that are designed to provide assistance to drivers in dangerous situations and to avoid accidents. We fist present an overview of different features and characteristics of Vehicular Ad-hoc NETworks (VANETs). We then present the most promising and important projects undertaken worldwide on VA-

NETs. we also address and present the standardization activities status around the globe. In addition to that, we review the state of the art and then we describe and present prediction of vehicle positioning and predicting transmission success using machine-learning (ML) in VANETs. Moreover, we study road-accident forecasting using ML schemes in VANETs. We adopt a model in which we present a prediction model that combines two approaches. The proposed hybrid approach incorporates the advantages of both generative (GMM) and discriminant (SVC) models. Compared to the baseline statistical model (GMM), our model performed up to 24% better in terms of accuracy. Finally, we propose a broader content matching model of trusted strings and blockchain based filtering for connected vehicles, where a content and subject headings are first matched and then the outcome of that is consolidated by a distributed blockchain consensus voting mechanism for any decision taken with respect to trust evaluation.

Key words : Vehicular Ad hoc NETworks (VANETs), CAMs (Cooperative Awareness Message), DENMs (Decentralized Environmental Notification Messages), Positioning, Accident forecasting, Machine Learning with VANETs, Message Passing Protocol with Blockchain.

Résumé de la Thèse

Les accidents routiers constituent l'un des problèmes majeurs auxquels sont confrontés les pays du monde entier, engendrant des conséquences allant des blessures graves et des pertes économiques considérables à la perte de vies humaines, touchant ainsi à la fois les individus, leurs familles et les nations dans leur ensemble. Les statistiques indiquent qu'actuellement, environ 1,3 million de personnes perdent la vie chaque année en raison d'accidents de la route dans le monde, ce qui signifie qu'environ toutes les 25 secondes, une personne succombe à un tel accident. De manière préoccupante, ces chiffres devraient augmenter de 65 % au cours de la décennie à venir, plaçant les accidents de la route au cinquième rang des principales causes de décès.

Il est important de noter que les erreurs humaines sont identifiées comme la principale cause de ces accidents de la route. Une prise de conscience rapide des conducteurs, avec une marge de seulement 0,5 seconde, et une connaissance des zones à risque pourraient contribuer à éviter jusqu'à 60 % de ces accidents. C'est dans ce contexte qu'ont été développés les réseaux ad hoc véhiculaires, communément appelés VANETs. Ces réseaux visent principalement à réduire le nombre d'accidents routiers et à garantir le confort des passagers. Au sein des VANETs, les véhicules échangent des informations entre eux et avec les infrastructures le long des routes, connues sous le nom d'unités côté route (RSU).

Dans ces scénarios de communication, divers types de données sont échangés entre les nœuds participants, allant des messages de sécurité routière à la gestion du trafic en passant par les divertissements à bord. Compte tenu de l'urgence de la situation, les applications de sécurité deviennent un domaine d'intérêt majeur, impliquant des acteurs tels que les constructeurs automobiles, les organisations gouvernementales et les chercheurs.

Cette thèse se concentre particulièrement sur les applications de sécurité visant à assister les conducteurs dans des situations dangereuses et à prévenir les accidents. Elle commence par présenter une vue d'ensemble des caractéristiques des VANETs, des projets majeurs dans le monde entier, des activités de normalisation et des développements récents. La thèse explore ensuite la prédiction du positionnement du véhicule, la prédiction du succès de la transmission et la prévision des accidents de la route en utilisant des algorithmes d'apprentissage automatique (ML) au sein des VANETs. Un modèle de prédiction hybride est proposé, combinant les avantages des modèles génératifs (GMM) et discriminants (SVC), atteignant une précision allant jusqu'à 24 % par rapport au modèle GMM de base.

Contexte Le point central des futurs Systèmes de Transport Intelligent (ITS) devrait être la connexion avec les Véhicules Automatisés Connectés (VACs). L'avènement de ces VACs offrira l'opportunité d'améliorer la sécurité routière et la circulation. Les accidents de la route sont l'un des problèmes majeurs auxquels le monde est confronté. Par exemple, aux États-Unis, plus de 35 000 personnes meurent chaque année dans des accidents liés à des véhicules à moteur, selon l'Administration nationale de la sécurité routière (NHTSA) [5], tandis qu'en 2019, en France, plus de 3 000 décès ont été causés par des accidents de la route, ainsi que de nombreux blessés graves [6]. Ces statistiques devraient augmenter si rien n'est fait pour résoudre les problèmes actuels de sécurité. De plus, il a été souligné que 90% des accidents de la route sont dus à des erreurs humaines et que 60% de ces accidents pourraient être évités si le conducteur avait été averti au moins 0,5 seconde à l'avance [2]. Ainsi, les technologies d'automatisation ont un grand potentiel pour réduire ce nombre. Par conséquent, afin de faire face à l'augmentation continue des accidents de la route dans le monde, le développement de systèmes de transport intelligents et d'autres applications visant à améliorer la sécurité routière et le confort de conduite a été initié. Pour rendre ces applications possibles, un réseau de communication, appelé Réseau Adhoc Véhiculaire (VANET), dans lequel les véhicules sont équipés de dispositifs sans fil, a été développé. Récemment, les VANET ont attiré l'attention des chercheurs ainsi que des constructeurs automobiles en raison de leurs applications prometteuses. Bien que de nombreuses études et projets soient en cours, de nombreuses recherches et projets ont été entrepris dans ce domaine spécifique.

Dans un réseau VANET, la coopération entre les véhicules est la clé de la sécurité des

voitures connectées, des véhicules autonomes et d'autres voitures connectées sur les routes publiques. Les véhicules sont équipés d'une Unité Embarquée (OBU). Cette OBU supporte les Communications Dédiées de Courte Portée (Dedicated Short Range of Communication (DSRC)), la pile de protocoles de communication sans fils(Wireless Access in a Vehicular Environment (WAVE)) et/ou Cellulaire dans l'environnement de communication vehicules à tous objets(Vehicle to everything V2X (C-V2X))[7],[8]. L'OBU rend possible deux types de communication : (i) les communications entre vehicules et tous objets(Vehicle-to-Everything (V2X)) qui comprennent : la communication entre vehicules(V2V (Vehicle-to-Vehicle)), la communication entre vehicle et infrastructure(V2I (Vehicle-to-Infrastructure)), la communication entre vehicule et pieton(V2P (Vehicle-to-Pedestrian)) et la communication entre vehicule et le cloud (V2C (Vehicle-to-Cloud)). Dans ce scénario, des données sont échangées entre plusieurs véhicules. (ii) Les communications entre vehicule et infrastructure routieres de communication(Vehicle-to-Infrastructure (V2I)) où les données sont reçues de l'element appelé(Road Side Unit(RSU)) qui est unité intégrée à une intersection routière ou un feu de signalisation.

Les réseaux VANETs est un type de réseau qui possède ses propres spécificités, ce qui signifie une grande mobilité des nœuds avec des déplacements contraints et des nœuds mobiles disposant d'une quantité suffisante d'énergie et de puissance de calcul (c'est-à-dire de stockage et de traitement) [2]. Dans les VANETs, les véhicules et les unités côté route(RSU) utilisent le protocole IEEE 1609 WAVE (Wireless Access in Vehicular Environments) basé sur le protocole d'accès IEEE802.11p pour assurer la communication entre les véhicules (V2V) et entre les véhicules et l'infrastructure routière (V2I).

Les réseaux VANETs est principalement conçu pour transporter des communications liées aux applications de sécurité. Ces applications utilisent des transmissions périodiques de paquets qui contiennent la vitesse et la position des véhicules émetteurs. En Europe, nous avons deux types de messages de sécurité : les Messages d'alerte Automobile (Car Awarness Messages(CAMs)) [9] et les Messages de Notification Environnementale Décentralisée (Decentralized Environnment Notification MessagesDENMs) [10]. Les DENMs sont diffusés en multi-sauts lorsque survient un événement dangereux sur la route, tandis que les CAMs ne sont envoyés qu'en un seul saut et contiennent des informations sur les vitesses et les positions

des véhicules.

Les VANETs peuvent également être utilisés à d'autres fins. Par exemple, des informations sur l'état du trafic routier tel que le trafic fluide, les embouteillages, etc., peuvent être transmises aux véhicules. D'autres informations moins importantes peuvent être envoyées aux véhicules ou à leurs passagers, telles que de la publicité ou du divertissement, parfois appelé infodivertissement.

Une autre utilisation des réseaux VANET peut être comme système de positionnement. Pour des raisons de sécurité, les véhicules envoient périodiquement des messages CAMs qui contiennent leurs positions et leurs vitesses. Cette information est généralement obtenue à l'aide du GPS. Ainsi, si des Unités Côté Route(RSU) existent le long de la route où circulent les véhicules, une grande quantité de données de positionnement peut être mise à disposition pour l'apprentissage automatique. En supposant que les véhicules continuent d'envoyer leurs CAMs contenant peu ou pas d'informations de position, ces messages peuvent être utilisés pour déterminer la position du véhicule en fonction de la puissance à laquelle les RSU ou les véhicules reçoivent les balises. Des techniques d'apprentissage automatique peuvent être utilisées pour effectuer cette tâche.

De plus, avec l'émergence de dispositifs capteurs et de l'Internet des objets (IoT), il est devenu possible de configurer les véhicules futurs avec des capteurs de sécurité pour prévenir les accidents. Par conséquent, des études de recherche récentes ont été orchestrées pour examiner l'état de l'art de l'analyse de la sécurité des véhicules dans les réseaux ad hoc véhiculaires tout en abordant la question des accidents de la route. Il est devenu évident que la sécurité routière [11] implique différentes dimensions et paramètres. Par exemple, dans l'analyse de la sécurité routière, le style et le comportement d'un conducteur doivent être étudiés pour examiner un style de conduite non conventionnel et les conditions environnantes doivent être recueillies pour fournir un soutien intelligent aux conducteurs. Cependant, les principales contraintes de telles analyses sont l'absence de jeux de données substantiels et un moyen précis d'identifier les conditions environnantes sans l'incorporation de capteurs supplémentaires.

D'autre part, dans l'analyse de la sécurité routière, l'effet secondaire sur la sécurité routière de paramètres externes, tels que l'état de la route, la géométrie, la circulation, les conditions

météorologiques et le comportement des conducteurs et des piétons, doit être analysé. Malgré d'importants efforts de recherche, il n'a pas été possible de fournir le modèle le plus déterministe et le plus intelligent [12] pour prédire le contexte exact des accidents de la route en raison de données déséquilibrées à différents niveaux. La prédiction des accidents est l'un des aspects les plus cruciaux de la sécurité routière, où des mesures préventives sont prises pour éviter un accident avant qu'il ne se produise. Par conséquent, il est utile d'examiner les zones à risque d'accidents des villes et l'effet des facteurs externes afin de prévoir le niveau de sécurité des routes avec une granularité appropriée.

De plus, compte tenu de l'avancement des VANETs et de l'Internet des véhicules (IoV), la gestion de la confiance et l'évaluation de la confiance entre les nœuds de communication sont devenues deux techniques importantes dans le contexte de la lutte contre diverses attaques contemporaines. L'IoV est la fusion de trois réseaux. Il comprend un réseau inter-véhiculaire, un réseau intra-véhiculaire et un Internet mobile véhiculaire. En principe, il s'agit d'un système distribué à grande échelle pour la communication sans fil et l'échange d'informations entre les véhicules, la route, les humains et l'Internet, basé sur des protocoles de communication IP établis et des normes d'interaction des données (telles que le IEEE 802.11p WAVE standard, and cellular tech., e.g 4/5G). Une telle intégration de réseaux prend en charge la gestion intelligente du trafic, les services d'information dynamique intelligents et le contrôle intelligent des véhicules. Par conséquent, l'IoV est une application de l'IdO (Internet des objets) et un système de transport intelligent est l'une des principales façons de déployer pleinement le potentiel de l'IoV.

Enfin, l'IoV fait partie de l'informatique en périphérie, c'est-à-dire une approche visant à optimiser les systèmes d'informatique en nuage en éliminant le traitement des données du nuage et en le réalisant au bord du réseau, beaucoup plus près de la source des données. On pourrait dans ce cas dire que cela simplifie le flux de trafic local vers un serveur en nuage tout en permettant l'analyse des données en temps réel au sein des dispositifs en périphérie eux-mêmes [13]. Cette nécessité d'améliorer l'analyse en temps réel soulève à son tour la nécessité de définir l'évaluation de la confiance de manière plus efficace.

En examinant l'approche des politiques de gestion de la confiance pour les réseaux véhiculaires et les domaines connexes, il a été observé que d'importantes recherches ont été

menées sur les attaques et, dans une certaine mesure, sur les techniques d'atténuation de ces attaques. Par conséquent, il est crucial de concevoir un scénario d'évaluation de la confiance intelligent mais dynamique, capable d'améliorer la fiabilité du réseau et ainsi de réduire la possibilité d'attaque.

Motivée par les défis mentionnés ci-dessus, cette thèse couvre un large éventail de contextes de conduite dans les VANETS, notamment le positionnement des véhicules, la prédiction du succès transmission et la prévision du trafic.

Enoncé de la thèse

L'énoncé de thèse est la suivante : la sécurité routière dans des environnements dynamiques peut être améliorée de manière coopérative par des véhicules connectés et autonomes en :

— utilisant des mécanismes de coopération de véhicules décentralisés et des systèmes de positionnement des véhicules qui fournissent les emplacements des véhicules dans un VANET, et

— en utilisant des stratégies de conduite éco-autonome avec des sources de données statiques/dynamiques pour prévoir les accidents de la route.

Le reste de cette partie offre un aperçu de cette thèse. Nous commençons par décrire la portée du travail en présentant un aperçu des énoncés de problème. Deuxièmement, nous décrivons brièvement les principales contributions de notre travail. Troisièmement, nous présentons l'organisation du reste de cette thèse.

Portée de la thèse

Dans cette sous-section, nous commençons par une brève description des systèmes de véhicules et des environnements de trafic tels qu'ils sont pris en compte dans cette thèse.

Systèmes de véhicules connectés et autonomes

Les véhicules connectés et autonomes sont pris en compte dans ce travail, où nous supposons que les deux types de véhicules disposent d'une unité embarquée (OBU) qui prend en charge les communications dédiées à courte portée (DSRC), la pile de protocoles Wireless Access in a Vehicular Environment (WAVE) ou C-V2X[3][13]. Cette OBU est utilisée pour échanger des données entre plusieurs véhicules autonomes. La voiture autonome ou le véhicule autonome contient divers composants matériels et logiciels. Les principaux composants sont un système de navigation, un contrôleur de véhicule autonome, un système de perception, une capacité de

localisation, une base de données cartographique et une interface de communication sans fil. Pour le système à conduite autonome, nous considérons également dans cette thèse qu'il peut, à un moment donné, être sujet à une défaillance raisonnable. Par exemple, il pourrait s'agir d'une défaillance du système de contrôle, d'une défaillance de la synchronisation temporelle, d'une défaillance de la localisation et/ou d'une défaillance de la communication. Dans ce cas, nous supposons que, étant donné que la sécurité est le facteur le plus important dans les protocoles et les stratégies de conduite considérés ici, ces défaillances ne conduisent jamais à des accidents de la circulation.

Environnements de trafic

Le trafic homogène et le trafic hétérogène sont les deux types d'environnements de trafic différents pris en compte dans cette thèse. Cependant, nous gardons à l'esprit que pour remplacer les véhicules conduits par des humains par des véhicules entièrement autonomes, un point de transition est nécessaire.

Trafic Homogène

Dans le trafic homogène, les véhicules sont soit connectés, soit automatisés. Predemment nous avons présenté une brève description des véhicules connectés et autonomes (CAV) considérés dans cette thèse. Il existe des protocoles et des cadres proposés pour l'environnement de trafic qui peuvent utiliser à la fois les communications V2V et les systèmes de perception sur chacun des véhicules autonomes. Bien que les véhicules automatisés ne soient pas entièrement exempts d'opérateurs humains, les protocoles et les techniques de coopération entre véhicules peuvent également être appliqués aux entrepôts automatisés exploités par des robots mobiles, où il n'y a pas d'opérateurs humains.

Trafic Hétérogène

Le trafic hétérogène ou mixte est un flux qui contient différents types de véhicules. Ces véhicules peuvent être motorisés ou non motorisés. En d'autres termes, dans ce type de trafic, il y a des CAV ainsi que des véhicules conduits par des humains. Les véhicules conduits par des humains peuvent ne pas avoir de dispositifs de communication ni de capteurs embarqués, mais sont totalement contrôlés par des humains. Par conséquent, afin de garantir la sécurité routière, il y a deux principes fondamentaux pour concevoir un protocole de coopération entre véhicules :

— Les véhicules connectés et automatisés (CAV) ne peuvent pas se fier aux communications V2V et doivent plutôt effectuer les manœuvres requises uniquement par la perception ; et

— Les CAV ne doivent jamais mettre les véhicules conduits par des humains dans une situation inconfortable ou risquée. Ils ne doivent pas tromper les conducteurs humains et doivent donc avoir des comportements directs et confortables face aux véhicules conduits par des humains. Ici, deux choses très importantes doivent être prises en compte. Premièrement, pour coopérer en toute sécurité avec les véhicules conduits par des humains, les CAV doivent respecter les règles de circulation existantes. Deuxièmement, une interface de communication pour garantir la sécurité routière peut ne pas exister dans les véhicules conduits par des humains.

Résumé des contributions de la thèse

Dans cette sous-section, nous présentons et décrivons les principales contributions de cette thèse :

Contribution 1 : Positionnement des véhicules en utilisant la puissance du signal reçu des messages périodiques. Notre premier travail a porté sur les systèmes de positionnement des véhicules en utilisant la puissance du signal reçu (Received Signal Strength(RSS)) des messages périodiques en déterminant la position du véhicule grâce à des techniques d'apprentissage automatique. Dans cette partie, nous avons passé en revue différentes stratégies pour réaliser le positionnement des véhicules en utilisant le RSS des messages périodiques ainsi que les performances de transmission au sein du VANET. Nous avons procédé à revue suivant deux approches, comme suit :

— Le point suivant est la principale contribution dans la première partie : Tout d'abord, nous utilisons la puissance de réception pour prédire la position d'un véhicule. Deuxièmement, nous proposons et adaptons quatre techniques d'apprentissage automatique pour le positionnement des véhicules à savoir : K Nearest Neighbors (KNN), Neural Network (NN), Random Forest (RF) et Support Vector Machine (SVM). De plus, un outil de simulation simple est développé pour produire des données avec les positions des véhicules et les différentes puissances des messages envoyés par les véhicules et reçus à station de base. Enfin, nous analysons et comparons les performances de K

Nearest Neighbors (KNN), Neural Network (NN), Random Forest (RF) et Support Vector Machine (SVM) avec le jeu de données.

— Dans la seconde partie, nous étudions comment l'apprentissage automatique peut être utilisé pour prédire les performances de transmission au sein du VANET. Plus précisément, nous visons à calculer la probabilité de la réception réussie d'une transmission entre un véhicule et une unité de bordure située à une position donnée et connue. Cette enquête est consideré comme le point de départ important pour l'utilisation de techniques d'apprentissage automatique dans les VANET.

Contribution 2 : Prévision et analyse des accidents de la route. Dans cette partie, nous étudions et analysons les principales causes des accidents de la route, où nous indiquons un moyen de prévoir les accidents de la circulation. En d'autres termes, nous examinons l'état de l'art de l'analyse de la sécurité des véhicules dans les VANETs, en incluant la question des accidents de la route. Nous avons découvert que les techniques classiques de prévision de la circulation utilisent soit un modèle de mélange gaussien (Gaussian Mixture Model(GMM)) soit un classifieur à vecteurs de support (Support Vector Classifier (SVC)) pour modéliser les caractéristiques des accidents. D'une part, le GMM nécessite de grandes quantités de données et est peu coûteux en termes de calcul, d'autre part, un SVC fonctionne bien avec moins de données mais est coûteux en termes de calcul. Par conséquent, dans cette partie, nous présentons un modèle de prédiction qui combine les deux approches dans le but de prévoir les accidents de la circulation. Ainsi, une approche hybride est proposée, qui intègre les avantages du modèle génératif (GMM) et du modèle discriminant (SVC). Les échantillons de données brutes sont divisés en trois catégories : ceux représentant des accidents sans blessures, des accidents avec des blessures non incapacitantes et ceux avec des blessures incapacitantes. Le GMM est utilisé pour modéliser cette distribution trimodale, et les paramètres sont obtenus à l'aide de l'algorithme de maximisation de l'espérance(Expectation Maximisation (EM)). Les vecteurs moyens des gaussiennes composantes obtenus sont adaptés et utilisés comme entrée pour le modèle SVC, ce qui améliore encore la précision de la prédiction. Les résultats expérimentaux montrent que le modèle proposé peut amélio-

rer significativement les performances de la prédiction des accidents. Des améliorations de précision allant jusqu'à 24% sont rapportées, comparativement au modèle statistique de base (GMM).

Contribution 3 : Correspondance subjective basée sur un graphe de chaînes de confiance et filtrage basé sur la blockchain pour les véhicules connectés.
Dans cette partie, nous proposons un modèle de correspondance de contenu plus large de chaînes de confiance et un filtrage basé sur la blockchain pour les véhicules connectés, où un contenu et des en-têtes de sujet sont d'abord mis en correspondance, puis le résultat de cette mise en correspondance est consolidé par un mécanisme de vote par consensus de blockchain distribué pour toute décision prise concernant l'évaluation de la confiance. Pour établir la confiance pour un système distribué, le système doit être en mesure de mettre l'accent sur et d'assurer la participation des utilisateurs distribués. La procédure suit un mécanisme de consensus pour la mise en correspondance appropriée des entités de confiance. Les Communications à Courte Portée Dédiées (DSRC) sont obligatoires depuis 2016 pour les véhicules légers, et cette règle décrit un paquet de données défini avec un Message de Sécurité de Base (BSM) indiquant l'emplacement du véhicule, sa vitesse et d'autres paramètres sur la route. Cependant, le DSRC est incapable de spécifier les messages transmis et reçus en ce qui concerne une classification de confiance. Par conséquent, nous proposons une méthode unique pour étudier la correspondance de confiance optimale pour les messages entrants dans les véhicules connectés. De manière intéressante, l'étude ne considère pas la correspondance de mots clés (une approche mot par mot ou basée sur un dictionnaire). Au lieu de cela, le contenu thématique plus large et les en-têtes des messages communiqués sont pris en compte. Cela aidera à établir les catégories de contenu pour différents comportements non fiables tels que le comportement abusif, le branding forcé de produits, la diffusion d'informations trompeuses, le blocage des messages de sécurité routière, etc. Pour atteindre cette mise en correspondance proposée pour les dispositifs mobiles distribués, nous introduisons une procédure de transmission de messages suivie d'une décision de renforcement basée sur la blockchain. Les principales contributions de cette

étude sont les suivantes :

Nous proposons un schéma de transmission de messages pour les véhicules connectés. Dans ce schéma, nous ne considérons pas la correspondance de mots clés (une approche mot par mot ou basée sur un dictionnaire). Au lieu de cela, nous prenons en compte le contenu thématique plus large et les en-têtes des messages communiqués. Nous tentons d'améliorer l'évaluation de la confiance en utilisant un mécanisme de vote pour toute décision, qui est un concept qui utilise une décision de renforcement basée sur la blockchain. Nous visons à renforcer la sécurité et l'authentification des messages échangés entre les véhicules en introduisant le concept de correspondance de contenu et d'évaluation de la confiance dans une blockchain de voiture connectée en tant que perspective future.

Organisation du manuscrit

Dans le premier chapitre, nous avons introduit le contexte et la motivation de cette thèse, et nous avons également brièvement décrit nos contributions. Le reste de ce manuscrit est organisé comme suit : Dans le chapitre 2, nous donnons un aperçu des caractéristiques spéciales des VANETs. Nous présentons ensuite un aperçu de la normalisation de la communication inter-véhicules et des projets en cours de développement dans ce domaine. Dans le chapitre 3, nous discutons de l'état de l'art et de la prédiction de la position des véhicules en utilisant des techniques d'apprentissage automatique. Nous décrivons les quatre techniques d'apprentissage automatique utilisées : k-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM) et Neural Network (NN). Nous comparons ces quatre schémas d'apprentissage automatique et en tirons une conclusion basée sur leurs performances. De plus, nous étudions comment l'apprentissage automatique peut être utilisé pour prédire les performances de transmission dans les VANETs. Plus précisément, nous visons à calculer la probabilité de réception réussie d'une transmission entre un véhicule et une unité en bordure de route située à une position donnée et connue. Cette étude prend comme point de départ l'utilisation de techniques d'apprentissage automatique pour prédire les positions des véhicules en fonction de la force du signal reçu dans les VANETs.

Le chapitre 4 traite de la prévision des accidents de la route en utilisant des algorithmes d'apprentissage automatique dans les VANETs. Ici, nous étudions les zones à risque d'accidents dans les villes et l'effet de facteurs externes afin de prévoir le niveau de sécurité des routes avec une granularité appropriée. Nous adoptons un modèle dans lequel nous présentons un modèle de prédiction qui combine deux approches pour la prévision des accidents de la circulation. L'approche hybride proposée intègre les avantages des modèles génératifs (GMM) et discriminants (SVC). Les résultats expérimentaux montrent que le modèle proposé peut significativement améliorer la performance de la prédiction des accidents. Comparé au modèle statistique de base (GMM), notre modèle a surpassé le GMM jusqu'à 24% en termes de précision. Le chapitre 5 présente un protocole de passage de messages pour les VANETs. Dans ce chapitre, nous étudions et proposons un modèle de correspondance de contenu plus large de chaînes de confiance et de filtrage basé sur la blockchain pour les véhicules connectés, où un contenu et des en-têtes de sujet sont d'abord mis en correspondance, puis le résultat est consolidé par un mécanisme de vote par consensus de blockchain distribué pour toute décision prise concernant l'évaluation de la confiance. Enfin, nous concluons cette thèse, dans le chapitre 6, en résumant nos principales contributions et les résultats clés, puis présentons nos travaux futurs et discutons des problèmes de recherche ouverts concernant l'amélioration de la sécurité routière et des applications de sécurité dans les VANETs.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

General Introduction

1.1 Context

The core point of future Intelligent Transportation Systems (ITSs) is expected to be Connected with Automated Vehicles (CAVs). The advent of these CAVs will create an opportunity to improve road safety and traffic. Road accidents are one of the major issues that the world is facing. As an example, in the United States of America more than 35000 people die in motor vehicle-related crashes every year, according to the National Highway Traffic Safety Administration (NHTSA) [5] while in 2019, in France more than 3000 deaths occurred due to road accidents as well as many more serious injuries [6]. These statistics are expected to rise if nothing is done to solve current safety issues. Moreover, it has been pointed out that 90% of road accidents occur due to human error and 60% of these accidents could be avoided if the driver had been warned at least 0.5 seconds beforehand [2]. So, automation technologies have great potential to reduce that number. Therefore, to deal with the continuing increase in road traffic accidents worldwide, the development of Intelligent Transportation Systems and other applications to improve road safety and driving comfort have been initiated. In order to make these applications feasible, a communication network, called a Vehicular Adhoc NETwork (VANET), in which the vehicles are equipped with wireless devices, has been developed. Recently, VANETs have attracted the attention of researchers as well as automobile manufacturers due to their promising applications. Although many are on-going, a lot of studies and projects have been undertaken in that specific field.

In a VANET, vehicle cooperation is the key to the safety of connected cars and self-

driving vehicles and other connected cars on public roads. Vehicles are equipped with an On-Board Unit (OBU). This OBU supports Dedicated Short-Range Communications (DSRC), the Wireless Access in a Vehicular Environment (WAVE) protocol stack and/or Cellular V2X (C-V2X)[7],[8]. The OBU makes two types of communication possible : (i) Vehicle-to-Everything (V2X) communications that comprise : V2V (Vehicle-to-Vehicle), V2I (Vehicle-to-Infrastructure), V2P (Vehicle-to-Pedestrian), V2C (Vehicle-to-Cloud) and V2G (Vehicle-to-Grid) communications. In this scenario, data are exchanged among multiple vehicles. (ii) Vehicle-to-Infrastructure (V2I) communications where data received from a Road-Side Unit (RSU) embedded at a road intersection or a traffic signal.A VANET is a type of network that has its own specificity, which means high node mobility with constrained movements and mobile nodes that have ample energy and computing power (i.e., storage and processing) [2]. In VANETs, the vehicles and the Roadside Units use the IEEE 1609 WAVE (Wireless Access in Vehicular Environments) protocol built on the IEEE802.11p access protocol to provide communication between vehicles (V2V), and between vehicles and the roadside infrastructure (V2I).

A VANET is primarily designed to carry communications concerning safety applications. These applications use periodic packet transmissions which carry the speed and the position of the sending vehicles. In Europe, we have two types of safety messages : Car Awareness Messages (CAMs) [9] and Decentralized Environmental Notification Messages (DENMs) [10]. DENMs are multi-hop broadcasted when a hazardous event occurs on the road, whereas the CAMs are sent only at one-hop and carry information about the vehicles' velocities and positions.
VANETs can also be used for other purposes. For instance, information about the status of the vehicular traffic such as fluid traffic, traffic jams, etc. can be sent to vehicles. Other less important information can be sent to vehicles or their passengers such as advertising or entertainment, which is sometimes called infotainment.

Another use of a VANET can be as a positioning system. For safety reasons, the vehicles send periodic CAM messages that carry their positions and speeds. This information is usually obtained with the GPS. Thus if Roadside Units exist along the road where the vehicles are moving, a huge quantity of positioning data can be made available to machine-learning

algorithms. Assuming that the vehicles continue to send their CAMs carrying no (or very little) position information, these messages can be used to establish the vehicle's position by using the power at which the RSUs or the vehicles receive the beacons. Machine-learning techniques can be used to perform this task.

Moreover, with emerging sensor devices and the Internet of Things (IoT), it has become feasible to configure future vehicles with safety sensors to prevent accidents. Therefore, recent research studies have been orchestrated to investigate the state-of-the-art of vehicle safety analysis in vehicular ad-hoc networks while including the issue of road accidents. It has become clear that driving safety [11]and road safety involve various dimensions and parameters. For example, in driving safety analysis, the style and behavior of a driver must be studied to investigate an unorthodox driving style and neighboring conditions are fetched to arrange intelligent support for the drivers. However, the major constraints of such analyses are the absence of substantial data sets and a precise means of identifying the neighboring conditions without incorporating additional sensors.

On the other hand, in road safety analysis, the subsidiary effect on road safety from external parameters, including the road surface, geometry, traffic flow, weather conditions and both drivers' and pedestrians' behavior should be analyzed. In spite of considerable research efforts, it has not been possible to provide the most deterministic and computationally intelligent model [12] to predict the exact context of road accidents due to unbalanced data instances at various levels. Accident prediction is one of the most crucial aspects of road safety wherein precautionary measures are taken to avoid an accident before it occurs. Therefore, it is worth investigating the accident–prone areas of cities and the effect of external factors in order to forecast the safety level of roads with appropriate granularity.

In addition, considering the advancement of VANETs and of the Internet of Vehicles (IoV), trust management and trust evaluation across communicating nodes have become two important techniques in the context of dealing with various contemporary attacks. The IoV is the fusion of three networks. It comprises an inter-vehicle network, an intra-vehicle network and vehicular mobile Internet. In principle, it is a large-scale distributed system for wireless communication and information exchange between vehicles, the road, humans and the Internet based on established IP communication protocols and data interaction standards

(such as the IEEE 802.11p WAVE standard, and cellular tech., e.g 4/5G). Such network integration supports intelligent traffic management, intelligent dynamic information services, and intelligent vehicle control. Hence, the IoV is an application of the IoT and an intelligent transportation system is one of the prominent ways to deploy the full potential of the IoV. Finally, IoV is a part of edge computing, i.e., an approach to optimizing cloud computing systems by eradicating data processing from the cloud and performing it at the network edge, much closer to the source of the data. It might be argued that this streamlines the flow of local traffic to a cloud server as well as enabling data to be analyzed in real-time within the edge devices themselves [13]. This necessity for improved real-time analytics in turn raises the need to define trust evaluation in a more effective way.

While investigating the approach of trust management policies for vehicular networks and relevant areas, it has been observed that significant research has investigated attacks and, to some extent, techniques for mitigating these attacks. Hence, it is crucial to design an intelligent yet dynamic trust evaluation scenario, which can enhance the reliability of the network and thus be able to reduce the possibility of attack.

Motivated by the above-mentioned challenges, this dissertation covers a wide range of driving contexts in VANETS, including vehicle positioning, prediction of transmission success, and traffic forecasting. It also deals with Graph-Based Subjective Matching of Trusted Strings and Blockchain-Based Filtering for Connected Vehicles.

1.2 Thesis Statement

The thesis statement is as follows : road safety in dynamic environments can be cooperatively enhanced by connected and autonomous vehicles by :

— utilizing decentralized vehicle cooperation mechanisms and vehicle positioning systems that provide vehicles' locations in a VANET, and

— utilizing eco-autonomous driving strategies with static/dynamic data sources to forecast road accidents.

The remainder of this chapter provides an overview of this dissertation. We first describe the scope of the work by presenting an overview of the problem statements. Secondly, we briefly describe the key contributions of our work. Thirdly, we present the organization of the rest of this dissertation.

1.3 Scope of the Thesis

In this subsection, we start with a brief description of vehicle systems and traffic environments as they are considered in this dissertation.

1.3.1 Connected and Automated Vehicle Systems

Connected and automated vehicles are considered in this work, where we assume that both types of vehicle have an On-board Unit (OBU) that supports Dedicated Short-Range Communications (DSRC), the Wireless Access in a Vehicular Environment (WAVE) protocol stack or C-V2X[3][13]. This OBU is used to exchange data among multiple automated vehicles. The automated car or self-driving vehicle contains various hardware and software components. The main components are a navigation system, an autonomous vehicle controller, a perception system, a localization capability, a map database, and a wireless communication interface. For the autonomous driven system, we also consider in this dissertation that it may be, at a given time, subject to a reasonable failure. For instance, it could be a control system failure, a time-synchronization failure, a localization failure, and/or a communication failure. In that case, we assume that, as safety is the most important factor in the protocols and driving strategies considered here, those failures never lead to traffic accidents.

1.3.2 Traffic Environments

Homogeneous traffic and heterogeneous traffic are the two different kinds of traffic environment considered in this thesis. However, we keep in mind that in order to replace human-driven vehicles with fully self-driving vehicles, a transition point is required.

Homogeneous Traffic

In homogeneous traffic, the vehicles are either connected or automated. Section 1.3.1 presents a brief description of the Connected and Automated Vehicles (CAVs) considered in this dissertation. There are protocols and frameworks proposed for the traffic environment that can use both V2V communications and the perception systems on each of the autonomous vehicles. Although automated vehicles are not fully free of human operators, the vehicle cooperation protocols and techniques can also be applied to automated warehouses operated by mobile robots, in which there are no human operators.

Heterogeneous Traffic

Heterogeneous or mixed traffic is a stream that contains various types of vehicles. Those vehicles can either be motorised or non-motorised. In other words, in this type of traffic there are CAVs as well as human-driven vehicles. The human-driven vehicles may not have any communication devices or on-board sensors, but are totally controlled by humans. Therefore, in order to guarantee road safety, there are two main principles to design a vehicle cooperation protocol :

— Connected and Automated Vehicles (CAVs) cannot rely on V2V communications and rather they should conduct required maneuvers only by perception ; and

— CAVs should never put human-driven vehicles in an uncomfortable or risky situation. They should not mislead human drivers and therefore they should have straightforward and comfortable behaviors in front of human-driven vehicles. Here, two very important things need to be considered. Firstly, to safely cooperate with human-driven vehicles, CAVs are required to obey the existing traffic rules. Secondly, a communication interface to guarantee road safety may not exist in human-driven vehicles.

1.4 Summary of the thesis contributions

In this sub-section, we present and describe the main contributions of this thesis :

1.4.1 Contribution 1 : Vehicles' positioning using the Received Signal Strength of periodic messages Our first work dealt with vehicles' positioning systems using the Received Signal Strength (RSS) of periodic messages by determining the vehicle's location through machine-learning techniques. In this part, we survey different strategies to carry out vehicle positioning using RSS of periodic messages and also the transmission performances within the VANET. We proceed with these two parts as follows :

— The following point is the main contribution in the former part : First, we use the reception power to predict a vehicle's position. Second we propose and adapt four machine-learning techniques to the positioning of vehicles : K Nearest Neighbors (KNN), Neural Network (NN), Random Forest (RF) and Support Vector Machine (SVM). In addition, a simple simulation tool is developed to produce data with the positions of vehicles

and the different powers of the messages sent by the vehicles and received at the base stations. Finally, we analyze and compare the performance of K Nearest Neighbors (KNN), Neural Network (NN), Random Forest (RF) and Support Vector Machine (SVM) with the given data-set.

— In the latter part, we study how machine-learning can be used to predict the transmission performances within the VANET. More specifically, we aim at computing the probability of the successful reception of a transmission between a vehicle and a Roadside Unit located at a given and known position.

This survey considers the important starting point for the use of machine-learning techniques in VANETs.

1.4.2. Contribution 2 : Forecasting and Analyzing Road Accidents

In this part, we study and analyze the main causes of road accidents where we point out a way to forecast traffic accidents. In another words, we investigate the state-of-the-art of vehicle safety analysis in VANETs, while including the issue of road accidents. We discovered that conventional traffic forecasting techniques use either a Gaussian Mixture Model (GMM) or a Support Vector Classifier (SVC) to model accident features. A GMM on the one hand requires large amounts of data and is computationally inexpensive, SVC on the other hand performs well with less data but is computationally expensive. Therefore in this part, we present a prediction model that combines the two approaches for the purpose of forecasting traffic accidents. A hybrid approach is proposed, which incorporates the advantages of both the generative model (GMM) and the discriminant model (SVC). Raw feature samples are divided into three categories : those representing accidents with no injuries, accidents with non-incapacitating injuries and those with incapacitating injuries. GMM is used to model this trimodal distribution, and the parameters are obtained using the expectation- maximization (EM) algorithm. Mean vectors of the component Gaussians obtained are adapted and used as input to the SVC model which further improves the prediction accuracy. Experimental results show that the proposed model can significantly improve the performance of accident prediction. Improvements in accuracy of up to 24% are reported, compared to the baseline statistical model (GMM).

1.4.3 Contribution 3 : Graph-Based Subjective Matching of Trusted Strings and Blockchain-Based Filtering for Connected Vehicles

In this part, we propose a broader content matching model of trusted strings and blockchain-based filtering for connected vehicles, where a content and subject headings are first matched and then the outcome of that is consolidated by a distributed blockchain consensus voting mechanism for any decision taken with respect to trust evaluation. To establish trust for a distributed system, the system should be able to emphasise and assure from distributed users. The procedure follows a consensus mechanism for the appropriate matching of trusted entities. Dedicated Short Range Communications (DSRC) have been mandatory since 2016 for light vehicles and this rule describes a defined data packet with a Basic Safety Message (BSM) indicating the location of the vehicle, its speed and other on-road parameters. However, DSRC is unable to specify transmitted and received messages with respect to a trusted classification. Therefore, we propose a unique method to investigate the optimal trusted matching for incoming messages in connected vehicles. Interestingly, the study does not consider key word matching (a word-by-word or dictionary-based approach). Rather, the broader thematic content and headings for communicated messages are taken into account. This will help to establish the content categories for different untrustworthy behaviors like abusive behavior, forced branding of products, misleading information, blocking of road safety messages, etc. In order to achieve this matching proposed for distributed mobile devices, we introduce a message passing procedure followed by a blockchain-based reinforcement decision. The key contributions of this study are as follows :

— We propose a message passing scheme for connected vehicles. In this scheme, we do not consider key word matching (a word by word or dictionary-based approach). Rather, we take into account the broader thematic content and headings for the messages communicated.

— We attempt to improve trust evaluation by using a voting mechanism for any decision, which is a concept that makes use of a blockchain-based reinforcement decision.

— We aim to enhance securing and authenticating messages exchanged between vehicles by introducing the concept of content matching and trust evaluation in a connected car blockchain as a future perspective.

1.2. Manuscript organization

In the present chapter, we have introduced the context and the motivation of this thesis and we have also briefly described our contributions. The rest of this manuscript is organized as follows :

In Chapter 2, we provide an overview of the special features of VANETs. We then give an insight into inter-vehicle communication standardization and projects that are being developed in the field. In Chapter 3, we discuss the state of the art and prediction of vehicle positioning using machine-learning techniques. We describe the four machine learning techniques used : k-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN). We compare those four machine-learning schemes and draw a conclusion based on their performances. In addition, we study how machine-learning can be used to predict the transmission performances in VANETs. More specifically, we aim to compute the probability of the successful reception of a transmission between a vehicle and a Roadside Unit located at a given and known position. This survey will take as its starting point the use of machine-learning techniques to predict vehicle positions based on the received signal strength in VANETs.

Chapter 4 deals with road-accident forecasting by using machine-learning schemes in VANETs. Here, we investigate the accident–prone areas of cities and the effect of external factors in order to forecast the safety level of roads with appropriate granularity. We adopt a model in which we present a prediction model that combines two approaches to forecasting traffic accidents. The proposed hybrid approach incorporates the advantages of both generative (GMM) and discriminant (SVC) models. The experimental results show that the proposed model can significantly improve the performance of accident prediction. Compared to the baseline statistical model (GMM) : our model outperformed GMM by up to 24% in terms of the accuracy.

Chapter 5 presents a message passing protocol for VANETs. In this chapter, we study and propose a broader content matching model of trusted strings and blockchain-based filtering for connected vehicles, where a content and subject headings are first matched and then the outcome of that is consolidated by a distributed blockchain consensus voting mechanism for any decision taken with respect to trust evaluation.

Finally, we conclude this thesis, in Chapter 6, by summarizing our main contributions and the key results, and then present our future work and discuss open research issues regarding the improvement of road safety and safety applications in VANETs.

1.3 List of publications

Throughout the research work proposed in this thesis, a number of research articles have been published or submitted for publication to journals and conferences as follows :

— Mamoudou sangare, Soumya Banerjee, Paul Muhlethaler, Samia Bouzefrane "Graph Based Subjective Matching of Trusted Strings and Blockchain Based Filtering for Connected Vehicles" MSPN 2020. 20 January 2021. Lecture Notes in Computer Science, vol 12605. Springer, Cham. https ://doi.org/10.1007/978-3-030-67550-91.

— Mamoudou Sangare, Sharut Gupta, Soumya Banerjee, Paul Muhlethaler,Samia Bouzefrane "Exploring the Forecasting Approach for Road Accidents ; An Analytical measures with Hybrid Machine Learning" Expert Systems with Applications Volume 167 , 1 April 2021, 113855.

— Mamoudou Sangare, Dinh-Van Nguyen, Soumya Banerjee, Paul Muhlethaler,Samia Bouzefrane "Comparing different Machine-Learning techniques to predict Vehicles' Positions using the received Signal Strength of periodic messages". Conference, Sep 2019, Paris, France. Hal-2178360.

— Mamoudou Sangare, Soumya Banerjee, Paul Muhlethaler,Samia Bouzefrane " Predicting transmission success with Machine-Learning and Support Vector Machine in VANETs

An Approach using an analytical model of CSMA" IFIP/IEEE 7th International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (IFIP/IEEE PEMWN 2018).

— Mamoudou Sangare, Soumya Banerjee, Paul Muhlethaler , Samia Bouzefrane "Predicting Vehicles' Positions using Roadside Units : a Machine- Learning Approach" 2018 IEEE Conference on Standards for Communications and Networking (CSCN).

Chapitre 2

General Architecture and Characteristics, Related Research and Development projects and Standardization Activities In Vehicular Adhoc NETworks

Contenu

2.1 Introduction

Vehicular ad hoc networks (VANETs) are classified as an application of mobile ad hoc networks (MANETs). In other terms, VANETs are designed to apply the principles of MANETs by using the wireless network of mobile devices in the domain of vehicles. VANETs have been developed with the objective of improving road safety by allowing drivers to be more aware of

the surrounding vehicles and providing travellers with comfort services as well as navigation and other roadside services. They are seen as a key part of the Intelligence Transportation System (ITS) framework. They are also referred to as Intelligent Transportation Networks where they are understood as having evolved into a broader "Internet of Vehicles" which itself is expected to ultimately evolve into an "Internet of autonomous vehicles"[14, 15]. Due to the variety of applications in ITS, VANETs have become an emerging technology.

In this chapter, we present the state of art and set out more clearly the context of this thesis by giving an overview of VANETs. In addition to the communication architectures and general characteristics of VANETs, we then, according to their requirements and functions, classify VANET applications and highlight inter-vehicle communication standardization and projects that are currently under development in the field. Finally, we give a brief summary of different standardization activities and we also focus on the vehicle positioning and traffic forecasting challenges that need to be overcome to make such networks' predictions usable in practice.

2.2 Vehicular networks

2.2.1 Definition and architecture

A VANET is a type of a network that is derived from MANET technology [14, 15]. In VANETs, wireless communication devices installed on vehicles called On Board Units (OBUs) are the key element to establish communication between vehicles without being dependent on a centralized system. A communication scenario where vehicles interact with each other is termed vehicle-to-vehicle communication (V2V). In this communication, vehicles exchange information in order to help drivers to be aware of their surroundings or, specifically, to be aware of the presence of other vehicles in their environment. Nowadays, the progress in the technology has made possible other types of communication scenarios that can be grouped into two categories : communication between infrastructures (I2I), and vehicle to everything (V2X) communication. In I2I communication scenarios, information is exchanged between road infrastructures such as Roadside Units (RSUs), intelligent traffic lights, etc.

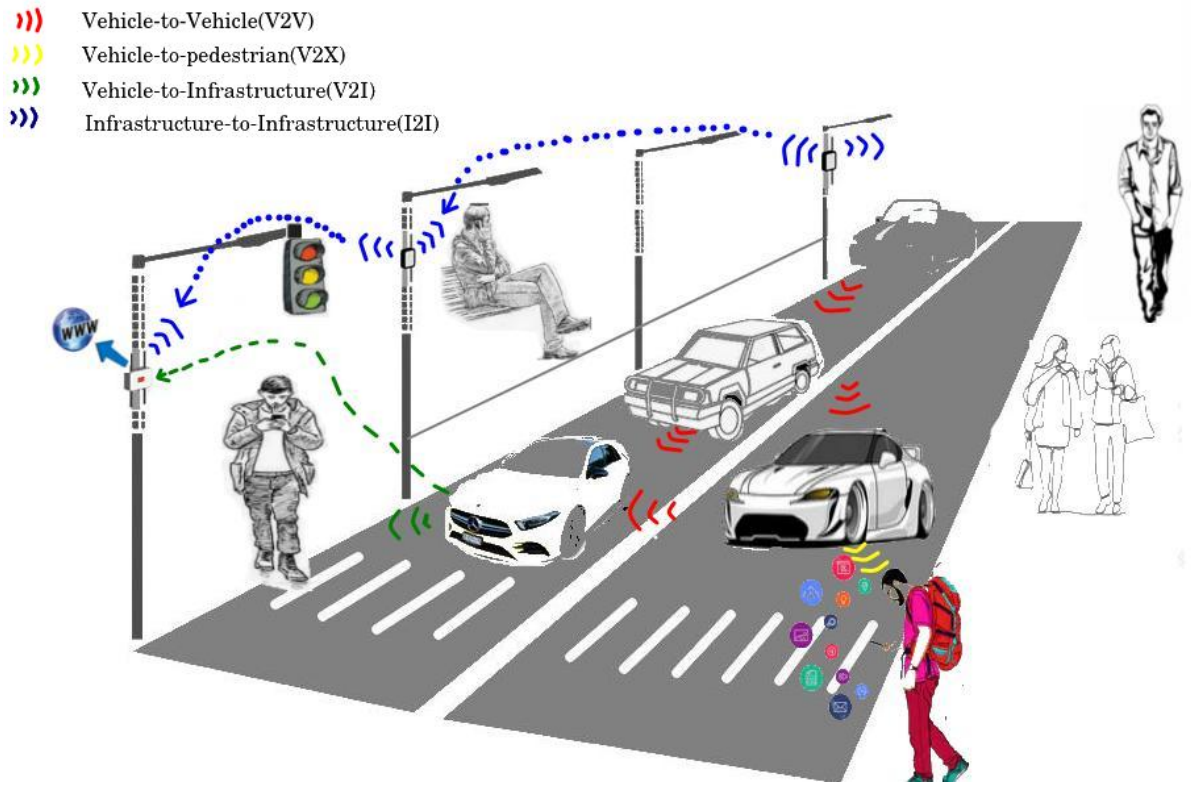

FIGURE 2.1 – An Overview of a VANET network

However, a V2X communication scenario that includes the vehicle to infrastructure (V2I) communication scenario, the vehicle to pedestrian (V2P) communication scenario, vehicle to cyclist (V2C), etc, aims at extending the communication to involve all types of road users. In this communication, data is exchanged between vehicles, RSUs and all other types of road users in order to provide traffic efficiency to all road users [16, 17, 18, 19]. In addition to the OBU, vehicles are equipped with another device called an Application Unit (AU). The OBU is used to exchange information with other OBUs (and/or with RSUs) within the same ad-hoc domain, the AU however executes ITS applications that use the communication capabilities of the OBU. Moreover, in VANETs there are Hot-Spots installed along the road which can allow vehicles to get access to the Internet.

Research in the field of VANETs is currently very active and varied as it touches on several axes at the same time, namely : wireless communications, protocols for physical and MAC layers, routing protocols, positioning and road accident forecasting. The following section will

detail some characteristics related to vehicular networks that should be taken into consideration when proposing a new solution that meets the requirements of VANET standards.

2.2.2 General characteristics

VANETs are derived from MANETs, however, they differ in certain of their characteristics. For instance, the existing MANET solutions such as MAC protocols, routing strategies, congestion control algorithms are not systematically compatible with VANETs. In addition, the mobility model, high node mobility, quality of service (QoS) requirements, etc. make VANETs unique. In the following, we highlight some specific VANET characteristics that should be taken into consideration when designing a new solution.

— **High node mobility :** Vehicles, termed nodes in VANETs, are known to change their position and speed frequently. As a result, the network topology also changes so as to make the network dynamic. The change of vehicles' speeds affects the wireless signal as the nodes move rapidly. This may lead to packet loss and result in high communication delay. In urban areas, the density of the vehicles may be great, but they may become quickly sparse when they move towards highways or rural areas. Based on these motion characteristics of VANETs, it is obvious that the high mobility of the nodes impacts the routing strategy as well as the communication performance. Therefore, it is necessary to take into consideration adaptive and efficient MAC protocols when developing VANETs.

— **Availability of Geographical position :** With technological advancements, recent VANET applications such as safe driving and emergency rescue often require high position accuracy. For several geographic protocols in VANETs, vehicles in motion knowing their position as well as the position of other vehicles in their surrounding area is very important. For instance, geographic protocols in VANETs such as Geo-Networking protocols are capable of routing information between vehicles based on their geographical position when accurate real-time three dimension positions (latitude, longitude and altitude), direction, velocity and precise time are provided. Therefore, making geographical positioning available needs to be taken into consideration and it is crucial for the deployment of VANETs.

— **Mobility model :** In VANETs, when it comes to evaluating protocol behaviours, the mobility model is considered to be one of the most important factors. When designing a mobility model, it should reflect reality (speed variations, traffic lights, crossroads, and traffic-jams) as accurately as possible. In order to define or to get a suitable mobility model, it is important to start by identifying the environment of the test scenario. Basically, there are three main environments that are conventionally considered :

1. **Highway :** This is a type of road that is specifically designed to connect major towns or cities. Depending sometimes on the time of the day and the day of the week, it is characterised by high speeds and a variable density of vehicles.
2. **City :** A city is a large human settlement. It can be defined as a permanent and densely settled place with administratively defined boundaries and it has several main and secondary roads, characterized by lower speeds with a high density of cars during rush hours.
3. **Countryside :** In general, this is considered as a geographic area that is located outside towns and cities, characterized by average speeds with a low density of cars.

— **No energy constraint :** Energy constraint means a limitation on the ability of a generating unit or group of generating units to produce active power due to the restrictions in the availability of fuel or other necessary expendable resources. However, VANET nodes have ample energy and computing power for both storage and processing compared to many MANET setups. There is therefore no constraint on energy consumption. In fact, both On Board Units (OBUs) and Road Side Units (RSUs) are directly powered by vehicles and road platforms, respectively.

— **Different QoS requirements :** Known as technical specifications that specify the system quality of features such as performance, availability, scalability, and serviceability, in VANETs, they vary significantly depending on the service. For instance, real-time applications involving services related to road safety and traffic management require guaranteed access to the channel and have strict requirements regarding end-to-end delay and packet loss ratio ; in contrast, however, infotainment applications have more flexible requirements both in terms of transmission rates and delay [20, 7, 21, 18, 22].

To sum up, as mentioned above, VANETs differ from MANETs due to their specific characteristics and requirements. These differences represent a challenge for the design of low-access delay, high throughput, scalable and robust MAC protocols. Nevertheless, there are other VANET characteristics, such as the ample electrical power and the limited degrees of freedom in the nodes' movement patterns, that can help to design and develop efficient MAC protocols. However, that is beyond the scope of this thesis.

2.2.3 VANET applications

The primary goal of VANETs was, and remains, to increase traffic safety and efficiency by reducing the risk of road accidents. However, like other technologies developed through scientific progress, the VANET has rapidly embraced other sub domains. Even though a VANET was initially considered as a subset of a MANET, it is now seen as an entirely separate field, and the development of VANET-specific applications has grown greatly. Nowadays, VANETs support a wide range of applications from simple one-hop information dissemination of, e.g., cooperative awareness messages (CAMs) to multi-hop dissemination of messages over vast distances. These networks are now used for a wide range of applications which can be grouped into the following three categories : safety, traffic management and user-oriented services [19, 17, 6, 23]. From the COMe-Safety point of view, applications in VANETs can be classified under three main fields as follow :

— **Traffic Safety :** As the primary goal of VANETs, traffic safety applications aim at reducing the number of accidents and enhancing driver and passenger safety. To do so, they enable each vehicle to provide a warning in real time when a critical event is detected in inter-vehicular communications (V2V). Drivers are then warned through a received warning which could be displayed on the dashboard of vehicle or take the form of a seat vibration, etc. or a combination of different indicators. As an illustration, Figure 2.2 and Figure 2.3[1] below show examples of safety applications that are based on V2V communications. Figure 2.2 shows when a potentially hazardous situation has occurred, a V2V-equipped vehicle ahead broadcasts that information or puts its hazard lights on, and other drivers will receive a "Hazard Lights Ahead" alert with the vehicle's estimated distance. This can allow drivers to safely maneuver away from the hazard.

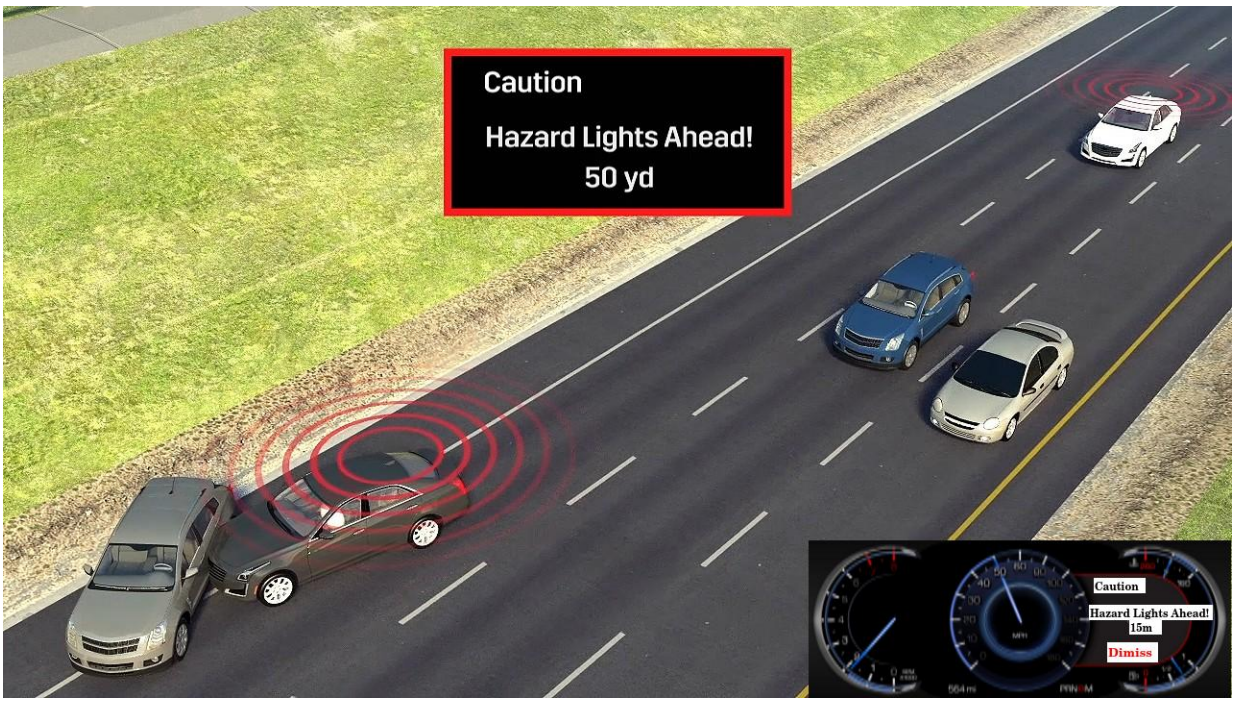

FIGURE 2.2 – Crash detection through V2V communication[1]

Another safety application is presented in Figure 2.3[1] where a V2V-equipped vehicle ahead is detected to have a StabiliTrak, traction control or anti-lock brake event, then the vehicle broadcasts that information about its current status (i.e., position, speed, deceleration, etc.) and drivers will get a "Slippery Road Ahead" alert allowing them to slow down and proceed carefully.

— **Traffic efficiency :** Another application of VANETs is traffic efficiency. These applications deal with traffic flow improvement as well as optimizing route management. This can reduce travellers' fatigue by decreasing the time spent on the road. It also reduces fuel consumption as well as air pollutants. By improving traffic flow, these applications can help to decrease the number of road accidents. Figure 2.4[1] shows a Road Side Unit (RSU) notifying drivers of the recommended speed according to the current traffic conditions.

— **Value-added services :** In addition to safety applications, non-safety applications, which are called value-added services, aim to provide on-board infotainment, comfort, and convenience to both drivers and passengers. They are intended to offer va-

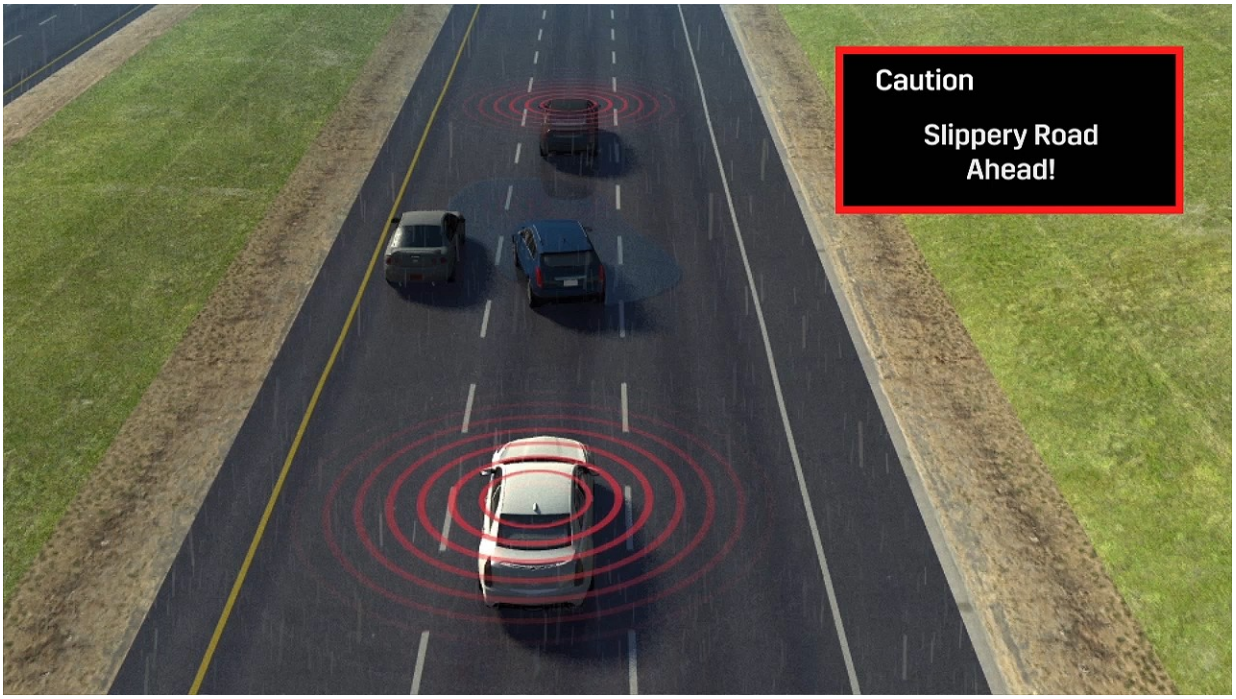

FIGURE 2.3 – Slippery Road Ahead warning[1]

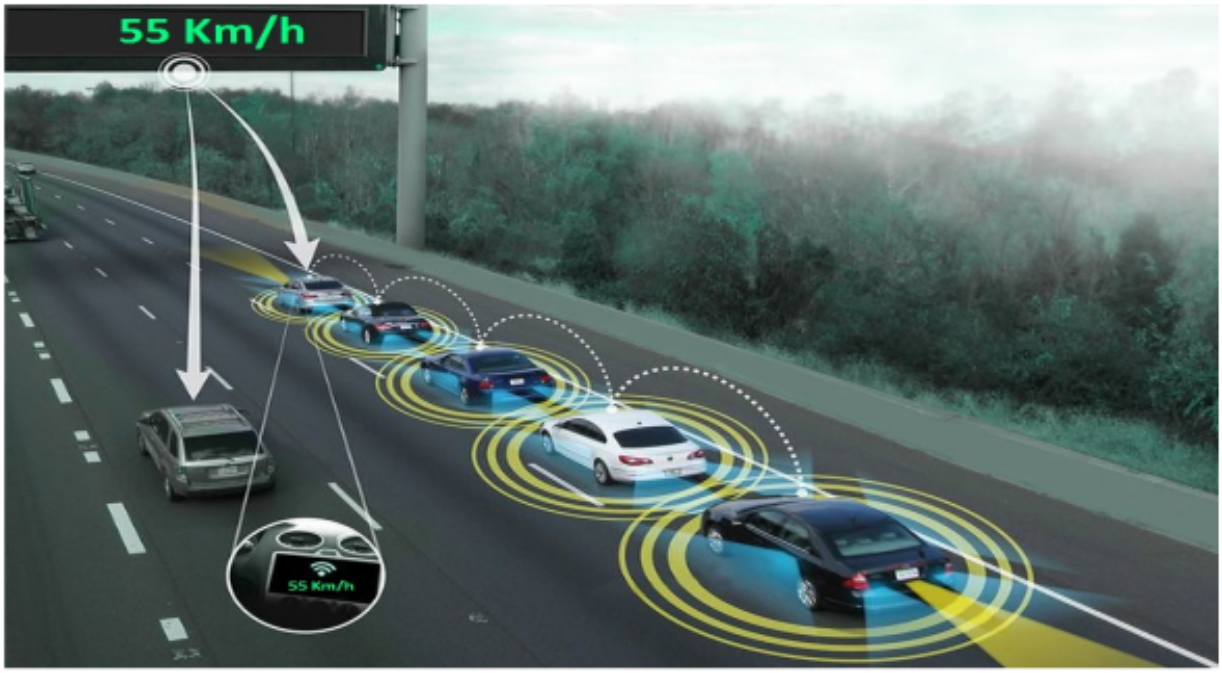

FIGURE 2.4 – Traffic notification in traffic management[2]

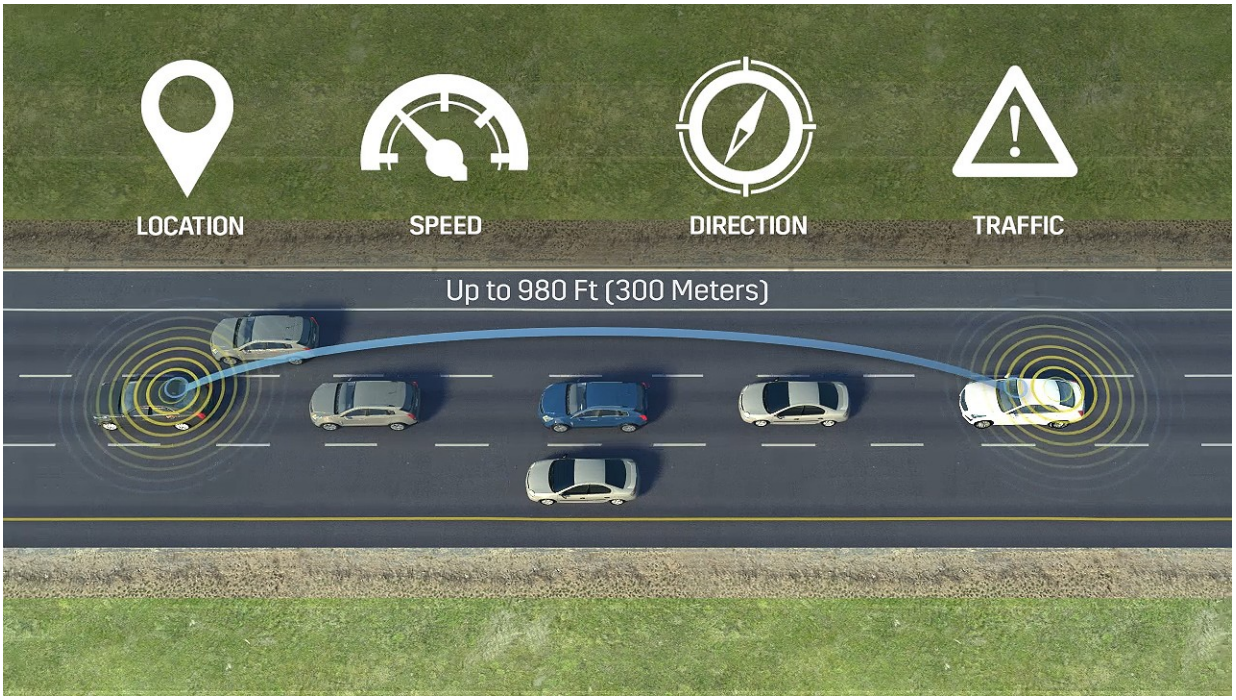

FIGURE 2.5 – File and video streaming sharing in V2V communication[1]

rious entertainment services to drivers and passengers that include maps, Internet access, navigation, instant messenger, toll payment service, electronic advertisements and entertainment information, free parking places, video streaming sharing, etc. In vehicle-to-infrastructure communication, Internet access can be provided. By doing so, business services will be available. In addition, in V2V communication, file sharing and video streaming can also be provided, as shown in Figure 2.5[1]. These file sharing and video streaming services can make the journey more enjoyable by making it less tiring and stressful. However, guaranteeing real-time and reliable communications for delay-sensitive applications without impacting throughput-sensitive applications can be an extremely challenging task because this category of applications has different QoS requirements in terms of bandwidth and delay.

2.2.4 Common VANET Units and Entities

In a VANET, vehicles that are not necessarily within the same radio transmission range are able to communicate with each other as VANETs also allow vehicles to connect to RSUs.

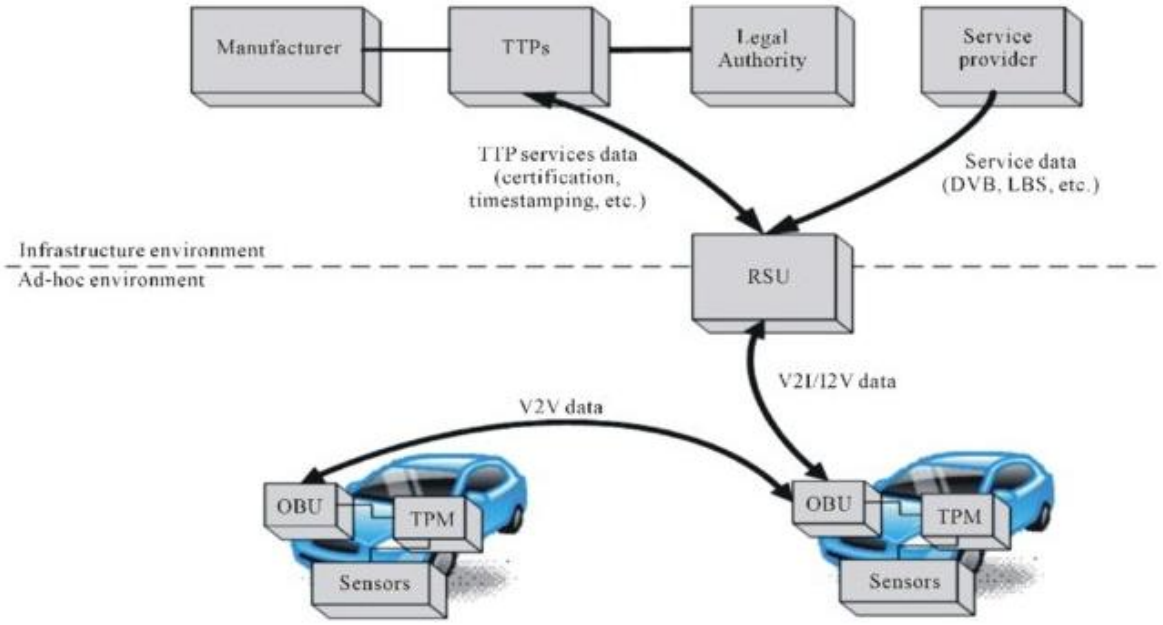

FIGURE 2.6 – VANET units and entities[3]

Besides vehicles and RSUs, there are other different units involved in the deployment. Although the majority are nodes (vehicles), there are other units or entities that keep the basic operations functioning in the network. Figure 2.6 illustrates the VANET units and entities that make up the VANET model. It also describes the infrastructure and Ad hoc environments that form a simplified VANET network, explained in detail in the section below. Specifically, there are generally two different environments in a VANET, namely : the Infrastructure environment, and the Ad-Hoc environment, as shown in Figure 2.6. The Ad hoc part is mainly composed of vehicles equipped with sensors, the OBU and Trusted Platform Module (TPM), whereas the infrastructure part involves the manufacturer, the third unit : Trusted Third Party (TTP), service providers and legal authorities or trusted authority [3].

— **Infrastructure Environment**

VANET units or entities can be permanently interconnected in this environment. This environment mainly contains the entities that manage traffic and also gives access to external services. Manufacturers are inside this environment of the VANET model because during manufacturing they identify each vehicle uniquely.

The legal authority or trusted authority is responsible for managing the entire VANET

system such as registering the RSUs, OBUs, and the vehicle users. Moreover, it is responsible for ensuring the security management of VANETs by verifying vehicle authentication, user ID, and OBU ID in order to avoid danger to any vehicle.

The Trusted Third Party (TTP) is also part of this environment. It provides various services such as time stamping and credential management. Manufacturers and the Authority are related to (TTP) because their services are needed, for example : issuing electronic credentials. Service providers are also in this environment because they offer services that can be accessed via the VANET, such as Location Based Services (LBS) or Digital Video Broadcasting (DVB) etc [24].

— **Ad-Hoc Environment**

In this environment, ad-hoc communications are created between vehicles. Vehicles are mainly equipped with 3 different devices namely : On-Board Unit (OBU), Trusted Platform Module (TPM) and sensors.

The OBU is a device that enables V2V and V2I communication. It is a GPS-based tracking device that allows vehicles to share information with RSUs and other OBUs. Its main function is to connect with RSUs or other OBUs through the wireless link of IEEE 802.11p and is responsible for communication with other OBUs or RSUs in the form of messages. The vehicle's battery is the main source of energy for the OBU. A vehicle also contains sensors such as a global positioning system (GPS), event data recorder (EDR), and forward and backward sensors that provide input to the OBU.

Trusted Platform Module (TPM, also known as ISO/IEC 11889) is an international standard for a secure crypto-processor, a dedicated micro-controller designed to secure hardware through integrated cryptographic keys. A TPM is always installed on the vehicles, and such devices are dedicated to security and also for computation and reliable storage [25, 3].

Sensors are also part of the ad-hoc environment. Nowadays, when a smart or an intelligent vehicle is designed it incorporates a set of sensors (front radar, reversing radar, etc.) that receive useful environmental information that generally the driver alone is unable to perceive. Figure 2.7 shows a smart vehicle equipped with various sensors. The main sensors are the global positioning system (GPS), event data recorder

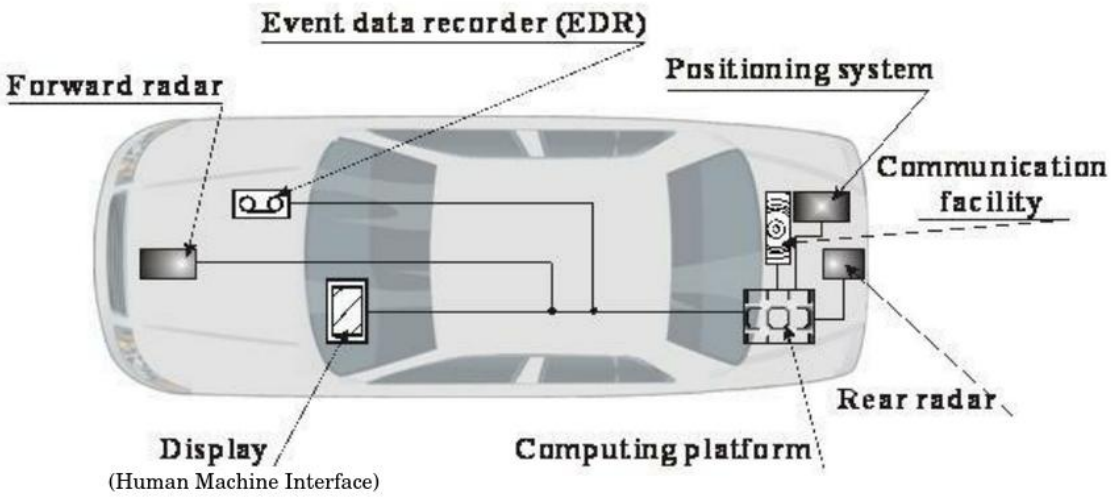

FIGURE 2.7 – A smart vehicle[4]

(EDR), and forward and backward sensors. They capture and determine the status regarding the vehicle and its environment e.g. (Fuel Consumption, Slippery Road, Safety Distance, etc.). Data from sensors are used to input other devices in order to improve road safety.

In the following section, we review recent VANET standardization efforts as well as some research projects in the field of VANETs in different parts of the world.

2.3 Relevant Research and Development projects

Vehicular Ad hoc Networks, known as VANETs, have gained significant interest around the world in recent years as they represent a promising technology to improve road safety. Policies have been developed for sustainable mobility, one of which, for instance, in Europe defined a political framework to ensure a high level of mobility, allied with the protection of humans and the environment, technological innovation, international cooperation and most importantly to reduce fatalities. Many projects have been funded and several standardization activities have been initiated in Europe, the US, South Korea and Japan, etc. Some of the projects have been jointly initiated and carried out by industry and research organizations. Nowadays, research, development and innovation are being intensified in the field of VANETs.

FleetNet-Internet on the Road [26] is a European project that is regarded as an initial feasibility study for inter-vehicular communication. It aims at developing a wireless multi-hop ad hoc network among vehicles to improve the safety and comfort of the driver and the passengers. Network on Wheels (NoW), a successor of FleetNet, has the particular characteristic of integrating non-safety applications and infotainment in a single system. PReVENT [27] was a research and development integrated project initiated by the European automotive industry and co-funded by the European Commission to increase road safety by developing and demonstrating preventive safety applications and technologies. Another integrated research project co-funded by the European Commission Information Society Technologies is SAFESPOT [28] . It aims to create dynamic cooperative networks between vehicles and between vehicles to the roadside in order to develop a system that helps to detect potentially dangerous situations in advance but also to increase drivers' awareness about their surrounding environment in time and space. CVIS-Cooperative Vehicle Infrastructure System, whose main goal is to enable continuous communication and cooperative services between vehicles and the infrastructure to improve road safety and traffic efficiency, was also funded by the European Commission. Apart from the above-mentioned projects, several others have been jointly initiated and carried out on the basis of cooperation. A non-profit organization, the Car2Car Communication Consortium (C2C-CC), was launched by European vehicle manufacturers and supported by the automobile industry. Moreover, the specifications of the European ITS standard foundation were laid down and developed through the cooperation of ETSI, ISO\TC, and the C2C-CC. In addition to the aforementioned R&D activities, mainly in Europe, there are other several research projects that have been initiated and carried out at national and international scales to develop and prepare the deployment of efficient vehicular communication protocols. Figure 2.8 gives an overview of the projects related to inter vehicular communications in Europe as well as in the US and Japan. These include AKTIV, AIDE, COM2React, SEVECOM, CarTALK CyberCar-2, DAIDALOS-II, i-IMPACT, FRAME, GST, MORYNE, REPOSIT, i-Way, Watchover [29], COMeSafety [30], GeoNET [31], coopers [32], euroFOT [33], PRE-RIVEC2X [34] and evita [35] which are sponsored by the European Union, Advanced Highway Technologies in the USA and the Advanced Safety Vehicle Program (ASV) sponsored by the government of Japan and more details of them can

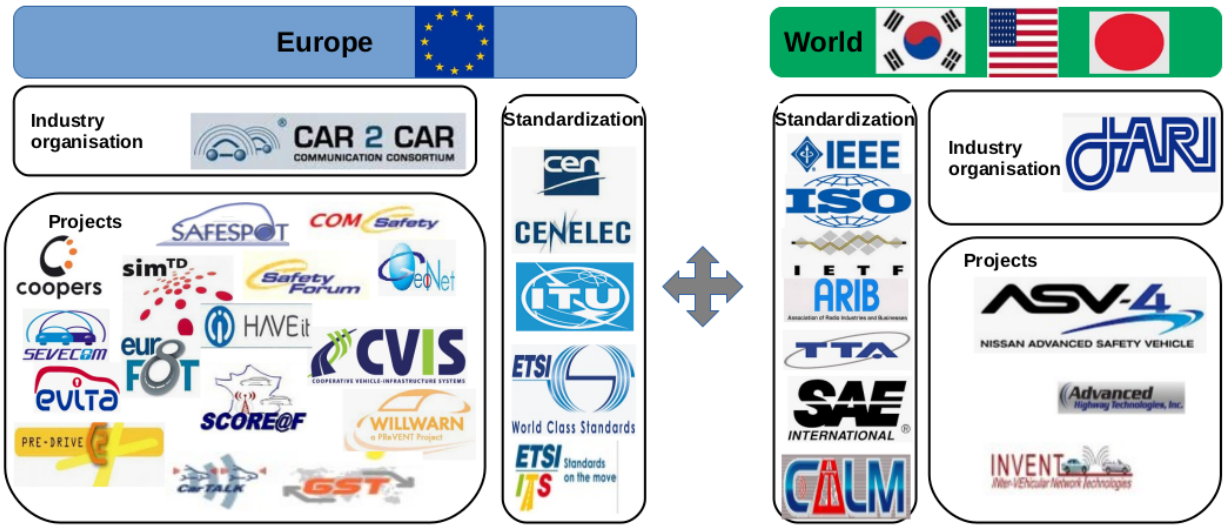

FIGURE 2.8 – Overview of ITS activities in Europe and other parts of the world

be found in [36].

2.4 Standardization activities

2.4.1 Standardization

Vehicular ad hoc networks are created by applying the principles of mobile ad hoc networks, which simply means the spontaneous creation of a wireless network of mobile devices in the domain of vehicles. VANETs were first mentioned and introduced [37] in 2001 under "car-to-car ad-hoc mobile communication and networking" applications, where networks can be formed and information can be relayed among cars. Whereas in the early 2000s, VANETs were seen as a mere one-to-one application of MANET principles, they have since developed into a field of research in their own right. In the past decade, VANETs have increasingly attracted researchers' attention and many research studies have been undertaken as well as standardization and development projects. In this section, the recent standardization efforts and the related activities in the field of VANETs are presented in details.

— **Dedicated Short Range Communication :**

Defined in the frequency band of 5.850 GHz to 5.925 GHz with a total bandwidth of 75 MHz, DRSC is one-way or two-way short-range to medium-range wireless communication channels specifically designed for automotive use and a corresponding set of

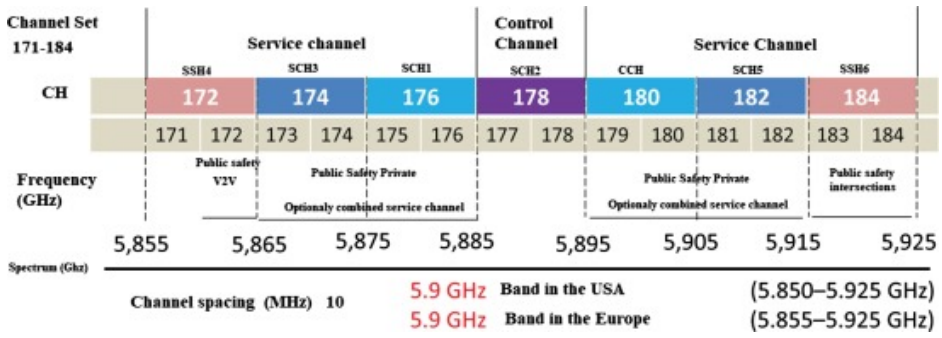

FIGURE 2.9 – DSRC frequency spectrum[[2]

protocols and standards. Created in the USA, it was proposed by the United States Federal Communication Commission (FCC) to support Intelligent Transportation Systems (ITS). The ITS Joint Program Office of the US Department of Transportation conducts research on DSRC and other wireless communication technologies and their uses in vehicle safety. DSRC technology operates on the communication range of [300 -1000] meters. It also supports a data rate of [6 - 27] Mbps, and a vehicular speed up to 190 Km/h. Operating on the 5.9 GHz band of the radio frequency, DRSC has a total bandwidth of 75 MHz that is divided into 7 channels of 10 MHz each. These channels are divided functionally into one control channel and six service channels, as presented in Figure 2.8. The control channel, CCH, is reserved for the transmission of network management messages (resource reservation, topology management) and it is also used to transmit high priority messages (critical messages relating to road safety). The six other channels, SCHs, are dedicated to data transmission for different services.

DRSC is also used outside of the USA. It is used for safety and other purposes. For instance, the European Telecommunications Standards Institute designated a 30 MHz band of radio frequency for DSRC in 2008, although the technology was used earlier than that for electronic tolling. Other countries, such as Japan, are also using the technology for safety and billing purposes [38]. A comparison between different regional standards for DSRC is presented in Table 2.1. For more details about regional standards for DSRC, we can refer to [39][40].

— **IEEE 802.11p :**

Features	Europe	Japan	North America
Duplex	Half	Half-duplex(OBU) Full-duplex(RSU)	Half
Radio Frequency	5.8GHz	5.8GHz	5.8-5.9GHz
Bandwidth	20MHz	80MHz	75MHz
Channels	4	7	7
Channel separation	5MHz	5MHz	10MHz
Data rate	Up link :250Kps Down link : 500Kps	Up/Down link : 1 or 4 MPS	Up/Down link : 6-27 Mps
Coverage(m)	15-20	30	1000
Modulation	RSU : 2-ASK OBU : 2-PSK	RSU : 2-ASK OBU : 4-PSK	OFDM

TABLE 2.1 – Comparison of DSRC of different regions

IEEE 802.11p : Designed for the purpose of enhancement, the IEEE 802.11p is an approved amendment to the IEEE802.11 standard to add wireless access in vehicular environments (WAVE), a vehicular communication system. It defines improvements to 802.11 (the basis of products marketed as Wi-Fi) required to support ITS applications. This includes data exchange between high-speed vehicles and between vehicles and the roadside infrastructure, so called V2X communication, in the licensed ITS band of 5.9 GHz (5.85–5.925 GHz). IEEE1609 is a high-layer standard based on the IEEE 802.11p. It is also the basis of a European standard for vehicular communication known as ETSI ITS-G5 that supports the Geo-Networking protocol for vehicle-to-vehicle and vehicle-to-infrastructure communication in Europe. The functionalities of Enhanced Distributed Channel Access (EDCA) derived from the IEEE 802.11e standard to improve the QoS are used by the IEEE 802.11p. Among those functionalities, there is prioritization. EDCA, allows safety messages with high priority level to be sent in priority. The Arbitration inter-frame spacing (AIFS) which is an optional technique used to prevent collisions in IEEE 802.11e based WLAN standard (Wi-Fi), in the medium access control (MAC) layer, and the Contains Windows (CWs) are varied to achieve prioritization. The probability of successful medium access is increased by using this scheme of adaptation which is important for real-time communication. As presented in Figure 2.9, the channel access time follows the time division multiple access (TDMA) channel access scheme which is based on the time division multiplexing (TDM) scheme.

TDMA provides different time slots to different transmitters in a cyclically repetitive frame structure. Similarly, the channel access time is equally divided into repeating synchronization intervals of 100 ms [41]. Each synchronization interval is divided into control channel intervals (CCHI) of 50ms each, and service channel intervals (SCHI) of 50ms each. Within the selected CCHI interval, the control channels of the vehicles are tuned in sender or receiver mode. Vehicles then may send or receive high-priority safety messages or to transmit other non-safety messages to their specific channels. For radio switching at the start of each channel, there is a 4μs Guard Interval used. This specification is defined by the standard.

— **Wireless Access in Vehicular Environment(WAVE) :**

A wireless access in vehicular environments (WAVE) system provides inter-operable, efficient, and reliable radio communications to support applications offering safety and convenience in an intelligent transportation system. In many of the ITS application scenarios, the WAVE system is designed to provide vehicles with direct connectivity to other vehicles (V2V), or to infrastructures (V2I) such as to RSUs through dedicated short-range communications (DSRC). The IEEE 802.11p DSRC\WAVE protocol stack is shown in Figure 2.10. It incorporates a number of protocols in conjunction with the family of the IEEE 1609 standards [42]. These protocols include the IEEE 1609.1 WAVE resource manager, the IEEE 1609.2 WAVE security services for applications and management messages, the IEEE 1609.3 WAVE networking services, and the IEEE 1609.4 WAVE multi-channel operation.

— **ISO :TC204\WG16-CALM :** Communications Access for Land Mobiles (CALM) has been proposed by the International Organization for Standardization (ISO) TC 204/Working Group 16 to define a set of wireless communication protocols and air interfaces for a variety of communication scenarios spanning multiple modes of communications and multiple methods of transmissions in ITS [43]. CALM was designed to support different methods of transmission such as 2G/3G/LTE cellular telephone communication technology, DSRC 5.8-5.9 GHz (legacy systems), various evaluations of the IEEE 802.11 standard including WAVE (IEEE P1609.3/D23), M5(ISO 21215), WIMAX- IEEE 802.16e etc. It uses these wireless access technologies to provide broad-

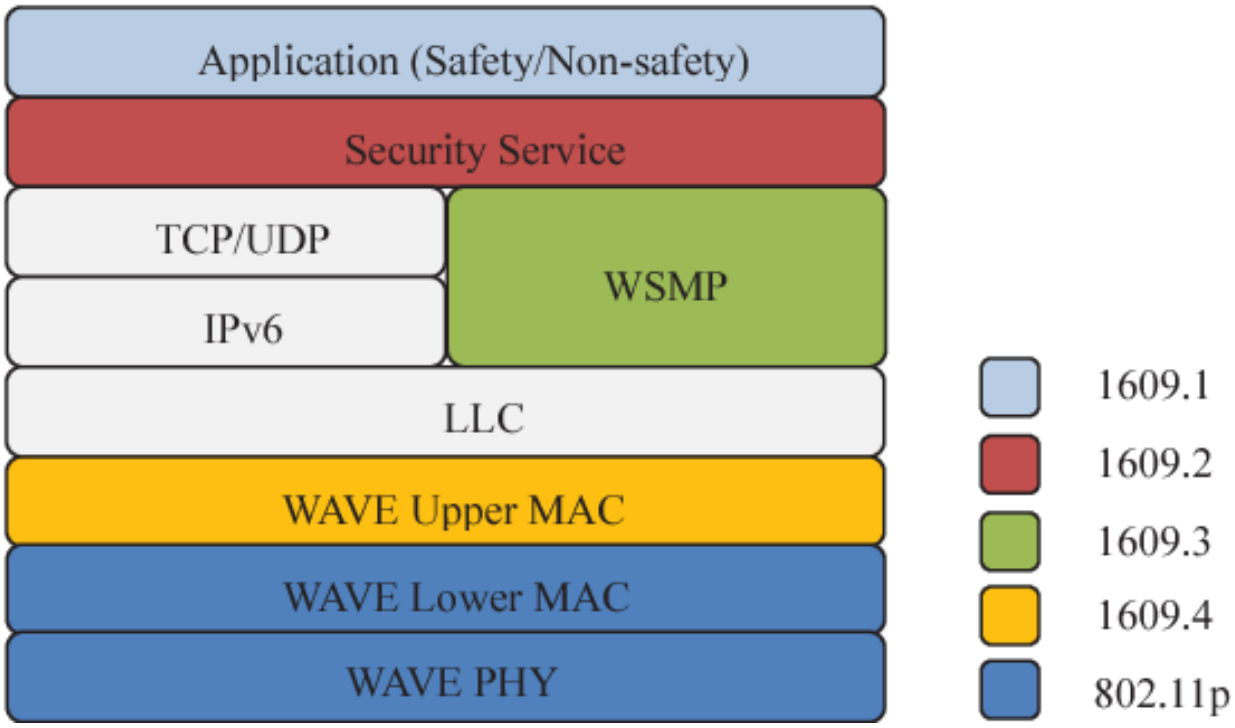

FIGURE 2.10 – IEEE 802.11p DSRC-WAVE Protocol stack[[2]]

cast, uni-cast, and multi-cast communications between mobile nodes, between mobile nodes and the infrastructure, and between fixed infrastructures. The CALM architecture provides an abstraction layer for vehicle applications, managing communication for multiple concurrent sessions spanning all communication modes, and all methods of transmission. CALM M5 was developed based on IEEE 802.11p to support vehicular ad hoc networking. It enables V2I communication that is initiated by either roadside or vehicle (e.g. toll booth), V2V communication that is peer to peer ad hoc networking amongst fast moving objects following the idea of MANET's/VANET's and I2I communication which is point-to-point connection where conventional cabling is undesirable (e.g. using lamp posts or street signs to relay signals). The main objective of CALM M5 is to provide real-time road safety applications requiring bounded access channel delays and low communication overhead. For these applications, a dedicated frequency band is allocated. However, for non-safety applications, there is another frequency band allocated that requires less latency constraints [44, 43, 45].

— **ETSI TC ITS :**

European Telecommunication Standardization Institute (ETSI) is a European Standards Organization of the telecommunication industry in Europe. It is well known for its standards for GSM, TISPAN and others. It has established a technical committee (TC) ITS (Intelligent Transportation System), in order to develop standards and specifications for ITS service provision in Europe [46, 10]. TC ITS is organized into the following five working groups :

1. WG 1 : User and Application requirements
2. WG 2 : Architecture and cross layer issues
3. WG 3 : Transport and Networks
4. WG 4 : Media and related issues
5. WG 5 : Security

ETSI TC ITS has also converged in harmonization with ISO TC204 WG16 towards the ITS communication architecture. The scope of the committee is not limited by V2X communication. It goes beyond that by encompassing various other modes of transportation (motorbike, pedestrians, aeronautics, railway, etc.) and also wireless communication technologies (5GHz, 60GHz, Infrared, and GSM)[47].

2.4.2 Summary and Outlook

Many research and development efforts as well as standardization activities have been initiated and funded in Europe, in the USA, South Korea and Japan to establish and develop efficient vehicular communication systems. Some of the projects have been jointly initiated and carried out by industry and research organizations. Nowadays research, development and innovation are becoming intensified in the field of VANETs. Moreover, at the international level, many standardization organizations, such as TTA, ETSI ITS, IEEE , ISO and IETF [47, 48, 49, 50, 51, 52], are cooperating together. These organisations are aiming to establish a collaborative convergence towards the deployment of the expected intelligent transportation systems (see Figure 2.10). However, challenges and open issues in this field remain to be studied, addressed and overcome.

On the one hand, the IEEE 802.11p under its current status still fails to guarantee the strict QoS requirements necessary for safety applications, in terms of delay and loss rate, especially

in heavy traffic conditions [53]. In addition, QoS for safety applications is still not satisfied when using CSMA/CA. This is because, to organize the channel access, the main IEEE 802.11p MAC layer is based on the CSMA/CA [1] protocol which is well-known to be able to handle packet collisions in broadcast communications. So in dense scenarios, the CSMA/CA system does not guarantee bounded channel access delays under its current state, which means that the QoS requirements for safety applications are still not satisfied.
Moreover, on the control channel, no acknowledgment messages are sent-back to confirm the reception when safety messages are transmitted in broadcast mode. Furthermore, there is an increase in the collision probability in the presence of hidden terminals as no RTS [2]/CTS [3] exchange is used. Indeed, potential collisions between messages that have the same Access Category (AC) can happen when employing the EDCA technique [54].

On the other hand, it is necessary to address the issue of analysing traffic safety based on the external parameters. Despite tremendous research efforts, there is no deterministic and computationally intelligent model [12]to predict the exact context of road accidents due to unbalanced data instances at various levels. Therefore, there remains a need to forecast road traffic in order to determine accident–prone areas.

2.5 Conclusion

Research in the field of intelligent transportation system in general and in VANETs in particular has attracted significant interest around the world over past decade as a promising technology to improve road safety. However, due the specific characteristics of VANETs, which include a high density of vehicles, there are challenging problems when designing communication protocols as well traffic management solutions in VANETs. This chapter provided an overview of the fast changing field of vehicular communications around the world. The chapter first gave an introduction to ITS in general and to VANETs in particular. It described the features of VANETs and the different types of vehicular communications. Then, VANET applications were presented and classified into three main groups according to their

1. Carrier Sense Multiple Access with Collision Avoidance
2. Request To Send
3. Clear To Send

requirements in terms of delay and throughput and their improvement in safety on the road. Moreover, the chapter also provided a summary of recent and current research and standardization activities, and the development of technologies in the field. Finally, it identified the shortcomings of these R&D projects and standardization activities. The chapter concluded with an outlook of vehicular adhoc network communication around the world. The next chapter presents the advancements of VANETs with artificial intelligence in general and with machine learning in particular.

Chapitre 3

Advancements of VANETs with Machine Learning

Contenu

3.1 General Introduction

Recent progress in the field of Artificial Intelligence (AI) or Machine Learning (ML) techniques has opened up new opportunities for Intelligent Transportation Systems. Manufacturers are also making vehicles' communicating units smarter, resulting in the ability of the vehicles to better assess the environment. This development has led to the possibility of realizing autonomous driving that is based on the idea of imitating human driving behavior while mitigating human faults. Nowadays, a number of applications have been developed starting from active and passive road safety to the optimizing of traffic, ranging from autonomous vehicles to the Internet of vehicles.

In ITS, the V2X concept is the most pronounced and refers to a paradigm that is essentially based on sharing information in the form of Vehicle-to-Infrastructure (V2I), Vehicle-to-Vehicle (V2V), Vehicle-to-Pedestrian (V2P), Vehicle-to-Self (V2S) and Vehicle-to-Road side units (V2R). In V2X communication systems, a large number of vehicles may require efficient and reliable resource allocation to reserve the resource pool in the frequency-time domain. In addition, the environmental conditions also affect the resource pool, such as the density of vehicles on the road, the changing from one area to another (from urban area to freeway or vice versa), the speed of the vehicles, or the multi-use case of a vehicle. The high speed of vehicles can cause the loss of signals and change the road condition. The learning algorithm can take the vehicle use case, the channel condition, and the environmental conditions as the input of the intelligent algorithm which satisfies the QoS in terms of reliability, latency, and throughput of the future communications. AI can be applied to V2X applications to improve the safety, security, or navigation of vehicles, and machine learning (ML) or AI can be applied to predict the available resources or duration that the vehicles are in or out of coverage. It can also be used to predict the channel condition to maximize the number of transmissions while ensuring the QoS requirements. Moreover, it can also be used to forecast incidents on the road.

In this chapter, we review the root cause of the vehicle positioning issue in VANETs, as well as its effects on the communication performances. We then survey the fundamentals of the research studies that have utilized AI to address various research challenges in V2X systems.

We then present a survey work on predicting vehicle position and also predicting transmission success using machine learning schemes. Finally, we conclude with a global summary of the advancement of VANETs with AI.

3.2 Positioning in VANETs with Machine Learning

Positioning is necessary to provide stable and precise location information for VANETs. In VANETs, the vehicles and the Road Side Units (RSU) use the IEEE 1609 WAVE [55] protocol built on the IEEE802.11p access protocol to provide communication between vehicles (V2V), and between vehicles and roadside infrastructure (V2I).
VANETs are primarily designed to carry communications concerning safety applications. These applications use periodic packet transmissions which carry the speed and the position of the sending vehicles. In Europe, there are two types of messages for safety : Car Awareness Messages (CAMs) [56] and Decentralized Environmental Notification Messages (DENMs) [57]. DENMs are multihop broadcasted when a hazardous event occurs on the road, whereas the CAMs sent only at one-hop, carry information about the vehicles' velocities and positions. Moreover, VANETs can be used for other purposes, such as to provide information about the status of the vehicular traffic or information about infotainment, which could be advertising or entertainment. VANETs can also be used as a positioning system. As mentioned in the previous section, for safety reasons the vehicles send periodic CAM messages that carry their positions and speeds. This information is usually obtained with the GPS ; thus if RoadSide Units exist along the road where the vehicles are moving, a huge quantity of positioning data can be made available to machine-learning algorithms. The Vehicles continue to send their CAMs which carry no (or very little) position information. These messages can be used to establish the vehicle's position by using the received signal strength at which the RSUs or the vehicles receive the beacons. ML techniques can be used to perform this task.

3.2.1 Main techniques used for outdoor positioning

3.2.1.1 Time-Of-Arrival (TOA)-based techniques.

Time of arrival (TOA or ToA) is the absolute time instant when a radio signal emanating from a transmitter reaches a remote receiver. ToA techniques rely on the measurements of the distance between the receiver of a signal and the base station. The position is computed using a triangulation technique. Precise time measurements are not possible for VANETs, mostly because these techniques require a perfect synchronization between the clocks of the base stations and the receivers. To obtain this result, the use of atomic clocks, which are very expensive, is required. ToA is however the basis of GPS and other similar positioning systems [58].

3.2.1.2 Techniques based on the round trip delay

A small packet is sent by the transmitter to the base station and the sender waits for a reply. The distance between the base station and the sender is proportional to the time elapsed. The transmitter can compute its position by triangulation if it can compute three round trip delays from three different base stations. A precise estimation of the location requires a large distance between the sender and the base station. However, like ToA schemes, techniques based on round trip delay require very accurate delay evaluation, which is very difficult unless dedicated transmission modems are used. VANETs use off-the-shelf IEEE 802.11p communication units ; thus, techniques based on the round trip delay are generally not suitable for these networks.

3.2.1.3 Received Signal-Strength (RSS) based techniques

Received signal strength (RSS) or received power has been used in many wireless sensing applications such as in WiFi, ZigBee, and narrow-band wireless networks. Based on the received signal strength from several wireless access points, the vehicles can compute an estimation of their location. In this case, the road area must be completely covered by the access points. Reported results based on this technique show poor accuracy [59][60].

3.2.1.4 Global Positioning System (GPS)

The Global Positioning System (GPS), originally Navstar GPS, is a satellite-based radio-navigation system that provides geolocation and time information to a GPS receiver anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites. Obstacles such as mountains and buildings can block the relatively weak GPS signals. In vehicular networks, the GPS or other similar positioning systems are the most widely used positioning techniques even though they have three main drawbacks : limited accuracy, incomplete coverage and security problems. The accuracy of civilian GPS is around 20 meters, which is not suitable for many VANET applications e.g. lane tracking, collision avoidance, autonomous driving, etc [61]. The best accuracy claimed by GPS vendors is plus/minus 5 meters but this accuracy is achieved for only 95% of the time, leaving the remaining 5% with much larger margins. Thus, GPS alone is not suitable for critical applications. GPS coverage is incomplete ; GPS has a high accuracy only when four signals can be detected from four different satellites and this situation is quite unlikely even if we do not consider the obvious case of obstruction by, for example, tunnels. Moreover attackers can use strong fake GPS signals that the vehicles are forced to lock on to, which can lead to large errors in the vehicles' positions. Several contributions have promoted an enhancement of GPS called Differential GPS (DGPS) where transmitters whose locations are precisely known complement the signals sent by the satellites. However, DGPS and other similar techniques do not work when the signals are too weak, for instance underground, in tunnels, or in densely built-up areas.

3.2.2 Machine-Learning Schemes

Defined as the study of computer algorithms that can improve automatically through experience and by the use of data, Machine learning (ML) is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. They can be used to perform positioning in VANETs [62][63]. There are four widely machine-learning schemes, namely K-Nearest Neighbors (KNN), Neural Networks (NN), Random Forest (RF) and Support Vector Machine (SVM), that are used for predicting vehicle positions in VANETs. These four techniques and their suitability to perform positioning in

VANETs are described in the section below.

In ML schemes, we often have a vector $X_j = \{x_j^1, \ldots, x_j^n\}$ of observations and these observations of X_j are linked to the variables Y_j. The problem is to infer Y_j knowing the vector X_j. In general (but not always), the algorithm has to be trained. In this case, the algorithm must work on a given number of observations $\{Y_j, X_j\}_{1 \leq j \leq K}$ to build a model which will be used to perform the predictions. Building this model is equivalent to computing a function $\hat{Y} = f(X)$. Then, given an observation X_i the model can compute $\hat{Y}_i = f(X_i)$. When Y_i is known, the prediction error can then be computed $\epsilon_i = Y_i - \hat{Y}_i = Y_i - f(X_i)$. With the same situation, ML algorithms can perform classification ; in this case X belongs to a class Y_l for $l \in \{1, \ldots, p\}$ [1]. As there is an observation X_i, the prediction algorithm will have to predict the most probable class given the observation X_i. As can be observed here that the issue is a positioning problem, it thus comes down to a regression problem.

3.2.2.1 k Nearest Neighbors (KNN)

K Nearest Neighbors (KNN) is one of the simplest ML algorithms which was first described, to the best of our knowledge in [64]. As previously stated, we have a given number of observations $\{Y_j, X_j\}_{1 \leq j \leq K}$ where X_j is usually a vector and Y_j is a real number.

Assuming that we have an observation X_i, we want to predict Y. The KNN algorithm must select the k nearest observations of X_i in $\{Y_j, X_j\}_{1 \leq j \leq K}$.

Let $i_1, \ldots, i_k$ be the k values that provide the k minimum values of the function

$$g(j) = d(X_j - X_i).$$

In other words, $i_1, \ldots, i_k$ are the indexes of the k minimum values of $g(j) = d(X_j - X_i)$. These minimum values can be equal if there are multiple values of X_j at the same distance from X_i.

We have at least the three possibilities for the distance, the most often used being the Euclidean distance.

1. We can observe that regression and classification problems are very close.

$$
\begin{aligned}
d(X_j - X_i) &= \sqrt{\sum_{l=1}^{n}(x_l^j - x_l^i)^2} \quad \text{Euclidean} \\
d(X_j - X_i) &= \sum_{l=1}^{n}|x_l^j - x_l^i| \quad \text{Manhattan} \\
d(X_j - X_i) &= \Big(\sum_{l=1}^{n}|x_l^j - x_l^i|^q\Big)^{1/q} \quad \text{Minkowski}
\end{aligned}
$$

The value predicted for Y_i will be the mean value of the k values Y_j for the k nearest neighbors of x_i.

$$\hat{Y_i} = \frac{1}{k}\sum_{1}^{k} Y_{i_k}$$

3.2.2.2 Neural Networks

Neural networks take a given number of observations $\{Y_j, X_j\}_{1\leq j\leq K}$ where X_j is usually a vector and Y_y is a real number. The idea is to build a network consisting of successive layers that combine the coordinates of X_j in successive layers to obtain the output, which is the real number Y_y. The successive layers generally perform linear combination of the previous layer and the result in the neurons is obtained with an activation function which is usually a Sigmoid function, see Figure 3.1.

To boost the performance of neural networks, we can use ensemble neural networks. The idea is to randomly pick observations in the test set and create a neural network based on these observations. We can thus create many different sets of observations and for each set derive the associated neural network. Our final prediction network will be the average of these neural networks' predictions. The theory shows that the individual neural networks have a small bias and a high variance but the final prediction network will have a small bias and a small variance.

3.2.2.3 Random Forest

We still have a given number of observations $\{Y_j, X_j\}_{1\leq j\leq K}$ where X is usually a vector and Y is a real number. The first step, in a random forest scheme, is to create a selection

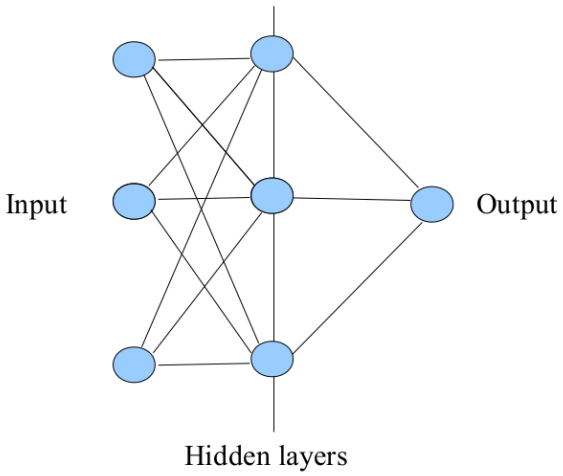

FIGURE 3.1 – A neural network with, in this case, one hidden layer and three neurons.

tree. Using the observations $\{Y_j, X_j\}_{1\leq j\leq K}$, we build different sets using different splitting criteria which operate on the vectors $\{X_j\}_{1\leq j\leq K}$. Each criterion allows the initial subset to be divided into two subsets. For instance, in Figure 3.2 the criterion $X_j < A$ provides the first splitting of the observations. The following two criteria complete the selection tree which ends with four final leaf nodes.

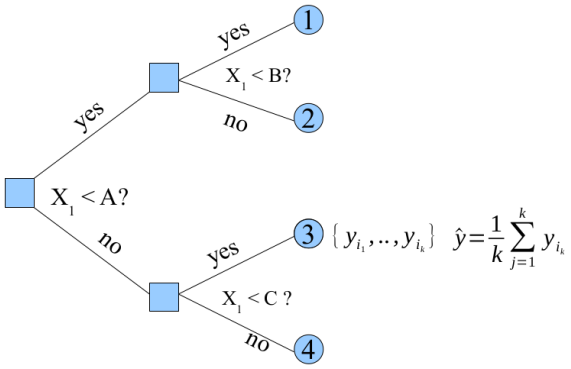

FIGURE 3.2 – Regression tree

Suppose now that we have a vector X_i and that we want to predict $\hat{Y}_i$. We will use the previous selection tree and determine in which final node the vector X_i is classified. Let us assume that X_i is classified in node 3 as are $X_{i_1}, \ldots, X_{i_k}$. In this case, the prediction of $\hat{Y}_i$

will simply be :

$$\hat{Y_i} = \frac{1}{k}\sum_{j=1}^{k} Y_{i_k}$$

The idea of Random Forest is to correct the error obtained in one selection tree by using the predictions of many independent trees and by using the average value predicted by all these trees. This Random Forest technique was first introduced in [65].

3.2.2.4 The Support Vector Machine regression technique

Generally, the positions y_i and the related values x_i are known (thus we know $(y_i, x_i)_{1\leq i\leq N}$) and we have to predict the positions using other values : $(x'_i)_{1\leq i\leq N}$. We assume that

$$y_i = w^T\phi(x_i) + b \tag{3.1}$$

where w and b are two unknown vectors and $\phi(x)$ an unknown function of a vector x.

To solve these equations, we introduce the following convex optimization problem :

$$\text{minimize} \quad \frac{1}{2}||w||^2$$

$$\text{subject to } -\epsilon \leq w^T\phi(x_i) + b \leq \epsilon. \tag{3.2}$$

This problem assumes that the function given in 3.1 can approximate the set of points that is given *i.e.* $(y_i, x_i)_{1\leq i\leq N}$ with an accuracy of ϵ. Sometimes, this is not possible and some errors must be accepted. In this case, slack variables which allow us to cope with impossible constraints, are introduced. This relaxation procedure uses a cost function. The convex problem then becomes :

$$\text{minimize} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$\text{subject to } -\epsilon - \xi_i^* \leq w^T\phi(x_i) + b \leq \epsilon + \xi_i \ \textit{with} \ \xi_i^*, \xi_i > 0 \tag{3.3}$$

This problem can be solved by using Lagrange multipliers. The problem becomes :

$$\begin{aligned} L: &= \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\nu_i\xi_i + \nu_i^*\xi_i^*) \\ &- \sum_{i=1}^{N}\alpha_i(\epsilon + \xi_i - y_1 + w^T\phi(x_i) + b) \\ &- \sum_{i=1}^{N}\alpha_i^*(\epsilon + \xi_i^* + y_1 - w^T\phi(x_i) - b) \end{aligned}$$

where L is the Lagrangian and $\nu_i, \nu_i^*, \xi_i^*, \xi_i^*$ are the Lagrangian multipliers which are thus positive *i.e.* $\nu_i, \nu_i^*, \xi_i^*, \xi_i^* > 0$

We know that the minimum of L is attained when the partial derivatives are zero, thus :

$$\partial L/\partial b = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$$

$$\partial L/\partial w = w - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i) = 0$$

$$\partial L/\partial \xi_i^{(*)} = C - \alpha_i^{(*)} - \nu_i^{(*)} = 0$$

The substitution of these equations in the Lagrangian leads to the following problem :

$$\begin{aligned} maximize &- \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi(x_i)^T\phi(x_j) \\ &- \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N}y_i(\alpha_i - \alpha_i^*) \end{aligned}$$

subject to

$$\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C].$$

Thus we have :

$$w = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\phi(x_i)$$

and

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\phi(x_i)^T \phi(x) + b$$

This formula is called the Support Vector expansion. The complexity of the function representation only depends on the dimensionality of the input space.

The rest of the analysis uses the Karush-Kuhn-Tucker (KKT) conditions. These conditions imply that at the solution, the product between the constraints and the dual variable must vanish. In other words, we have

$$\begin{aligned} \alpha_i(\epsilon + \xi_i - y_i + w^T\phi(x_i) + b) = 0 \\ \alpha_i^*(\epsilon + \xi_i^* + y_i - w^T\phi(x_i) - b) = 0 \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} (C - \alpha_i)\xi_i = 0 \\ (C - \alpha_i^*)\xi_i^* = 0 \end{aligned}$$

We can deduce that only the samples which do not satisfy the constraint of 3.7 have $\alpha_i^{(*)} = C$. Moreover, since the two values of the right part of 3.8 cannot be simultaneously 0 then we have $\alpha_i \alpha_i^* = 0$. Thus, after some observations we have :

$$\begin{aligned} \max(-\epsilon + y_1 - w^T\phi(x_i) | \alpha_i < C \text{ or } \alpha_i^* > 0) \leq b \leq \\ \min(-\epsilon + y_1 - w^T\phi(x_i) | \alpha_i < C \text{ or } \alpha_i^* > 0) \end{aligned}$$

This part is adapted from [66] a tutorial by Smola.

3.2.3 Numerical results

The numerical results are obtained on a straight road of length 600m. The position on the road is given by $x \in [0, 600]$. We assume that we have three RSUs located at $x = 0m$, $x = 305m$ and $x = 600m$.

Since we consider that there are no obstacles on the road to hinder free propagation, the signal strength received by the vehicles (or by the RSUs) depends solely on the distance between the vehicles and the RSUs. The power received is given by the following equation :

$$P = \frac{P_0}{r^\beta} \quad with \quad \beta \in [2, 4]$$

We measure the power received in dB as in the following :

$$P^{dB} = 10\frac{log(P)}{log(10)}.$$

Moreover, errors in the measurements are taken into account. We assume a Gaussian noise of zero mean and with a variance 0.05. This can also be interpreted by a log-normal fading which would affect the reception. This assumption is realistic when the transmissions between the vehicles and the RSUs are in line of sight. In the second analysis, we assume a Raleigh fading of rate $\mu = 1$. This corresponds to the case where there are multi-path links between sources and destinations as, for example, in built up areas in a town.

The database is obtained by 20 different measurements at each location of the vehicle, the locations being 15 meters apart. Thus, the data consist of 780 sets with three different powers, each of them corresponding to the power received by the three roadside units respectively.

Even if it were possible to do otherwise, for the sake of simplicity, we assume that the vehicles send beacons which are received by the three RSUs. The RSUs fuse these data and perform the machine-learning process. The location of the vehicles having been established, it can then be sent to them by one of the RSUs.

For the KNN algorithm, we use the data set directly derived from the power measurements in dB ; we do not perform any data processing before using the KNN algorithm. The code of KNN is found in the R software [67] and in its KNN library. We use the Euclidean distance.

For the Neural Network we use a library coded in Visual C++. The neural network has one hidden layer with three neurones. The activation function used is a sigmoid. We create an ensemble of 50 neural networks using a bagging technique and the final prediction is that of the average of the 50 neural networks previously obtained.

For the Random Forest algorithm, we use the data set directly derived from the power measurements in dB ; we do not process the data before using the algorithm. We use the Random Forest library found in the R software [67].

For the Support Vector Machine, the libsvm library [68] is used. The data set (powers in dB) is not directly processed. A linear transformation of these powers is performed so that the minimum power becomes 0 and the maximum power 1. The following values for the parameters are used : $C = 10$, $\epsilon = 10^{-6}$ and an exponential kernel. This means that we have to significantly increase the penalization of not respecting the bounds for the estimation since the default value for C is 1.

3.2.4 Direct link and log-normal fading

The comparison between KNN, NN, Random Forest and Support Vector Machine is presented in Figure 3.3. We observe that except for KNN, the position error remains in the interval $[-20m, 20m]$, which shows that when having a direct link, the machine-learning techniques offer reasonably good predictions. We also observe that the NN and RF techniques seem to perform the best and apparently there is no clear segment on the road where the prediction is better. We nonetheless observe a notable degradation of the prediction close to $x = 300m$ for the KNN and the RF techniques.

Table 3.1 proposes a quantitative evaluation of the four methods with the mean absolute error and the root mean square deviation. From these results, we note the NN approach provides the best performances with a mean absolute error of $3.3m$ and a root mean square deviation of $4.7m$ followed by the RF technique with an absolute error of $5.2m$ and a root mean square deviation of $7.1m$. We then have the KNN scheme with an absolute error of $7.3m$ and a root mean square deviation of $10.8m$ and the least effective scheme is SVM with an absolute error of $9.4m$ and a root mean square deviation of $11.1m$

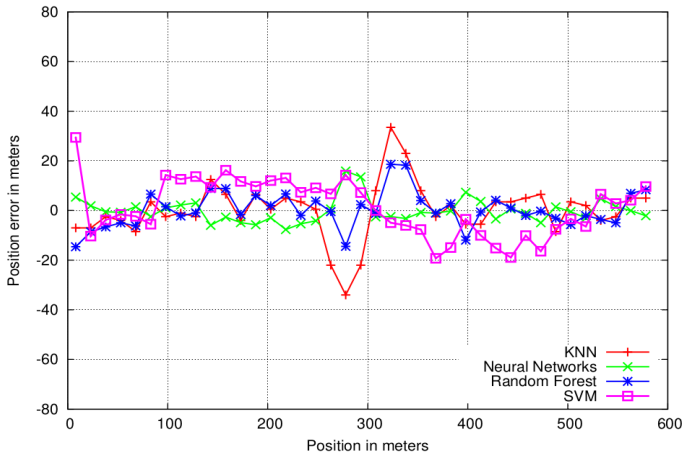

FIGURE 3.3 – Position errors versus position $x \in [0, 600m]$ on the roads with the different machine-learning techniques $\sigma = 0.05$.

TABLE 3.1 – Mean error and root mean square deviation versus prediction techniques (direct path propagation)

$KNN\ (m, \sigma)$	$NN\ (m, \sigma)$	RF (m, σ)	SVM (m, σ)
(7.31 , 10.8)	(3.32 , 4.67)	(5.24 , 7.06)	(9.44 , 11.14)

3.2.5 Direct link and log-normal fading but measurements only every 30m

Here, we perform the same comparison as in the previous section but we only have training data every $30m$ instead of every $15m$ as in the previous subsection. The predictions of the four algorithms remain acceptable and the ranking is still : first Neural Network, second Random Forest, third KNN and fourth SVM. Table 3.4 provides the quantitative results ; NN offers the best estimation with an absolute mean error of $4.7m$ with an RMS of $7.8m$, then comes RF with an absolute mean error of $5.85m$ and an RMS of $7.3m$. The two last places are for KNN : absolute mean error of $8.8m$ and an RMS of $13.8m$ followed by SVM : absolute mean error of $10.1m$ and an RMS of $11.3m$.

The degradation of precision of the localization is roughly 50% in terms of absolute error for NN and RF and there is no significant degradation for KNN and SVM.

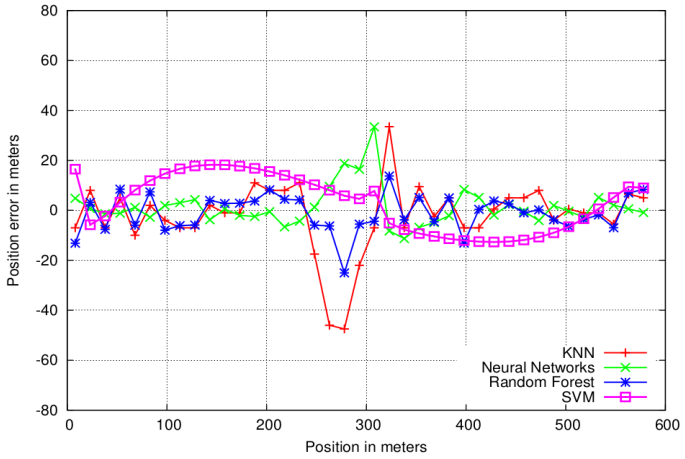

FIGURE 3.4 – Position errors versus position $x \in [0, 600m]$ on the roads with the different machine-learning techniques $\sigma = 0.05$. Training data measurements available only every 30m

TABLE 3.2 – Mean error and root mean square deviation versus prediction techniques (direct path propagation but measurements only every 30m)

$KNN\ (m, \sigma)$	$NN\ (m, \sigma)$	RF (m, σ)	SVM (m, σ)
(8.81 , 13.79)	(4.75 , 7.78)	(5.85 , 7.33)	(10.13 , 11.27)

3.2.6 Direct link and log-normal fading but no measurement in the segment [30m,105m]

Here, we study the performance of our algorithms when we have no training data for a given section of the road for $x \in [30m, 105m]$. We observe (except for NN) that the algorithms exhibit larger estimation errors in the segment where there is no training data available and the mean absolute error also increases. We still have the following ranking of the four algorithms : NN, RD, SVM and KNN. We observe that the least effective scheme in this scenario is KNN but we did not change the number of nearest neighbors (which might be considered unfair).

Table 3.3 provides a qualitative analysis of the scenario with no data for $x \in [30m, 105m]$. For the NN algorithm, the performance degradation is around 30% whereas the degradation is around 60% for the RF algorithm. There is almost no degradation for SVM and the degradation is around 130% for KNN.

A significant loss in the training data leads to a significant degradation in the estimation

of the position, except for NN.

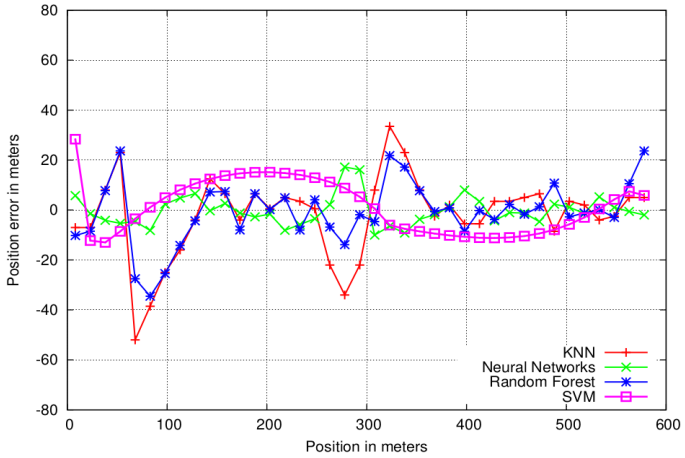

FIGURE 3.5 – Position errors versus position $x \in [0, 600m]$ on the roads with the different machine-learning techniques $\sigma = 0.05$. No training data available in [30m,105m]

TABLE 3.3 – Mean error and root mean square deviation versus prediction techniques (direct path propagation but no training data available in [30m,105m])

KNN (m, σ)	NN (m, σ)	RF (m, σ)	SVM (m, σ)
(18.35 , 23.73)	(4.33 , 5.78)	(8.69 , 12.17)	(9,23, 10.60)

3.2.7 No prominent path and Rayleigh fading (no direct path Rayleigh fading)

In the following, we present the results of our four algorithms when the power received is more affected by the fading. A Rayleigh fading (of rate 1) is assumed, which means that there is no prominent direct path between the vehicle and the roadside units. The power received consists in a random combination of many independent paths. In these conditions, the predictions, without filtering the measurement, lead to really poor results. Thus, they are not included in our manuscript.

The comparison between KNN, NN, Random Forest and Support Vector Machine is presented in Figure 3.6. We observe that, except for KNN, the position error remains in the interval $[-60m, 60m]$, which shows that when having no direct link, the machine-learning techniques offer only average predictions.

Table 3.4 proposes a quantitative evaluation of the four methods with the mean absolute

error and the root mean square deviation. From these results, we note that the NN and SVM approaches provide the best performances with mean absolute errors of respectively $17.63m$ and of $17.45m$ and with root mean square deviations of respectively $23.08m$ and $23.51m$. These techniques are followed by the RF technique with an absolute error of $26.9m$ and a root mean square deviation of $37.1m$. The least effective scheme is KNN with an absolute error of $30.30m$ and a root mean square deviation of $35.63m$

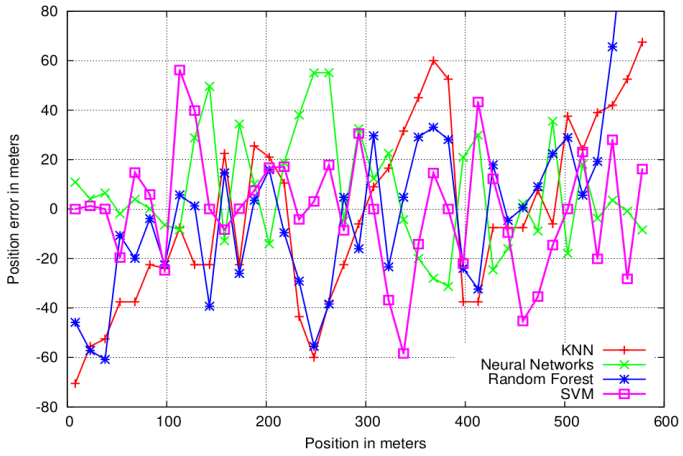

FIGURE 3.6 – Position errors versus position $x \in [0, 600m]$ on the roads with the different machine-learning techniques with Rayleigh fading .

TABLE 3.4 – Mean error and root mean square deviation versus prediction techniques (no direct path propagation)

KNN (m, σ)	NN (m, σ)	RF (m, σ)	SVM (m, σ)
(30.30 , 35.63)	(17.63 , 23,08)	(26,94 , 37,15)	(17,45 , 23.51)

3.2.8 Summary

Various ML algorithms, namely K-Nearest Neighbors, Neural Networks, Random Forest and Support Vector Machine have been used to predict the position of a vehicle using the reception power of packets sent to fixed nodes whose positions are precisely known. Among the four techniques, KNN happens to be the simplest method. In the data set, the scheme selects the k closest samples of the actual measurement. The neural scheme presented in this chapter consists of one hidden layer with three neurons. To boost this technique, an ensemble

neural network with 50 elements built with a bagging algorithm was used. In the Random Forest scheme, we use a classification tree to generate different classes according to a random classification tree. The location, in each class, is assumed to be the average location of the points in this class. The tree is then used for the prediction ; the location predicted being that of the training samples at the same leaf of the random trees. The Support Vector Machine is an approximation technique that usually uses kernels as base functions. The main goal is to maintain, as far as possible, the samples and their approximations with a bounded error. In general, the base functions are exponential functions. The numerical experiments presented in this study demonstrate that a precise prediction can only be obtained when there is a main direct path of propagation. The prediction is altered when the training is incomplete or less precise but the precision remains acceptable. In contrast, with Rayleigh fading, the accuracy obtained is much less striking. It is observed that the Neural Network is nearly always the best approach. With a direct path, the ranking is : Neural Network, Random Forest, KNN and SVM except in the case when we have no measurement in [30m, 105m] where the ranking is Neural Network, Random Forest, SVM and KNN. When there is no direct path, the ranking is SVM, NN, RF and KNN but the difference in performance between SVM and NN is small.

The next section deals with the use of Support Vector Machine in predicting transmission success in VANETs.

3.3 Predicting transmission success with Support Vector Machine in VANETs

3.3.1 Introduction

Recently, great progress in wireless transmission technology has paved the way for large networks with massive transmission patterns. Such networks can be Wireless Sensor Networks (WSNs) or VANETs. VANETs are considered to be one of the main tools to help reduce traffic accidents and fatalities. In these networks, vehicles exchange packets between each other but can also send packets to RSUs or receive packets from them.

VANETs can produce a huge volume of data which can be exchanged and aggregated. VANETs are thus conducive to the use of ML techniques, which aim to achieve various goals. Machine learning is used for vehicular networks in the following areas :

— Positioning : The information received or sent by the vehicles in VANETs can be used to establish their positions. Such a situation is ideal for the application of ML approaches [69, 70]. In vehicles, there are many sources of information such as GPS, map analysis, radar, lidar, etc. A huge volume of data can be collected, analyzed and shared by the vehicles and the RSUs ;

— Automatic Incident Detection (AID) can be performed by ML algorithms. The Support Vector Machine technique has been used in recent studies, such as [71, 72]. Older studies such as [73, 74] use Neural Networks and Fuzzy logic. A Wavelet scheme is used for AID in [75] ;

— an ML algorithm can analyze vehicle trajectories in order to predict dangerous situations and warn drivers when such situations are detected ;

— Routing in VANETs can benefit from machine learning ; examples of such a use are given in [76, 77, 78] ;

— Security in VANETs : Machine learning has been proposed to improve security in VANETs. The collection of packets sent in VANETs can be used to find compromised vehicles which do not perform suitable actions such as relaying when it is appropriate or sending forged messages over the network. We find such a use of machine learning in [66, 79, 80].

In the first areas, the ML schemes offer other services in addition to those that help the functioning of the VANETs, whereas in the second areas the ML techniques are used solely to help the functioning of the VANETs.

In this thesis, we study how machine learning can be used to predict the transmission performances within VANETs. More specifically, we aim to compute the probability of the successful reception of a transmission between a vehicle and a Roadside Unit located at a given and known position and with a given transmission rate [2]. This is very important since the performance of the transmission is a key parameter for the safety applications use transmissions of packets ; tuning the transmission of safety messages is crucial in order to optimize their efficiency. In Europe, ETSI standards define two safety messages for VANETs : Car Awareness Messages (CAMs) and Decentralized Emergency Notification Messages (DENMs). DENMs are sent in broadcast and generally relayed (in multihop mode) to warn vehicles of hazards. Each vehicle sends CAMs to neighboring vehicles ; for a given vehicle, these messages contain its position (obtained by GPS) and speed. RoadSide Units can easily compute the statistics of the transmission, i.e., for a given location and a given transmission rate, the probability of success. This probability can be computed by averaging the number of successful receptions. These statistics can be used by an ML scheme which can produce a model used by the RSU and also by the the vehicles themselves around the RSU.

Since it is very difficult to obtain real data in VANETs, for a first step we use an analytical model to compute the reception probability in a CSMA network and to build our database. The main mechanism for access in a VANET is the carrier Sense Multiple Access scheme (CSMA). The main idea is that before transmitting a packet, a node (in our case a vehicle) senses the channel to determine whether there is already a transmission on the channel.

We model the location of the vehicles as an homogeneous Poisson Point Process, which eases the computation and has been proved a good assumption. To build the model of the simultaneously transmitting vehicles, we use the Matern selection process [81] which was first used in [82] to evaluate the pattern of simultaneous transmissions in CSMA. The model of [82] was improved by [83, 84], which is the model that we used and extended in our solution.

2. This can also be used to compute the probability of the successful reception of a transmission between two vehicles.

3.4 System model for the VANET performance

3.4.1 Network nodes

The nodes are randomly deployed according to a Poisson Point Process Φ. We denote by λ the intensity of the process. In this work, we consider a 2D infinite plan, $\mathcal{S} = \mathbb{R}^2$ or a 1D infinite line, $\mathcal{S} = \mathbb{R}$. The 2D model is for MANETs or WSNs. The 1D model is more relevant to VANETs.

3.4.2 Propagation law, fading and capture model

We suppose that the signal received in a transmission is the result of a random fading F and a power-law in the distance decay $1/r^\beta$ where β is the decay factor and is generally between 3 and 6. In our study, the fading will be Rayleigh i.e exponentially distributed with parameter μ and thus is of mean $1/\mu$. Hence, the signal received when the transmitter and the receiver are at distance r from each other is $F/l(r)$ [3] with $l(r) = r^\beta$.

We use the well-accepted SIR [4] (Signal over Interference Ratio) with a capture threshold T.

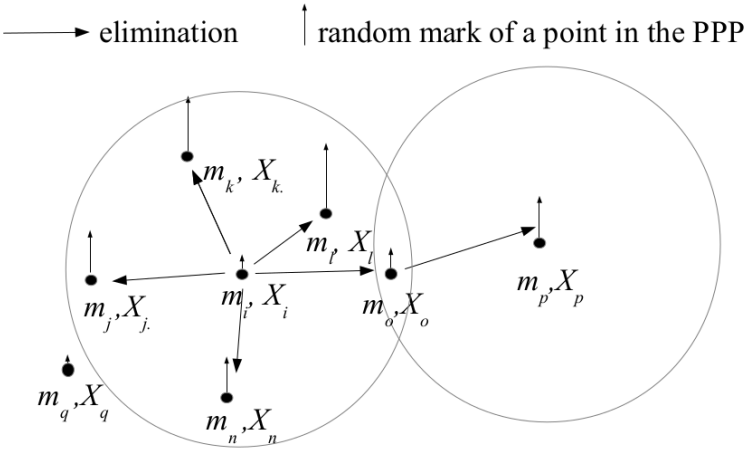

FIGURE 3.7 – Matern CSMA selection process and an example of over-elimination.

3. The power received $P = \frac{P_0 F}{l(r)}$ and we set $P_0 = 1$
4. We omit the thermal noise but it could be easily added, as is explained below.

3.4.3 Model for CSMA

Using the model developed in [83], we adopt a Matern selection process to mimic the CSMA selection process. The points X_i in Φ receive a random mark m_i. We also call $F_{i,j}$ the fading for the transmission between X_i and X_j. The idea of the Matern selection is to select the points X_i with the smallest random marks m_i in their neighborhood. To define the neighborhood of a point X_i, we need to introduce the carrier sense threshold P_{cs} which is the power threshold under which the channel is considered as busy. We define $\mathcal{V}(X_i) = \{X_j \in X_i \ F_{i,j}/l(|X_i - X_j|) > P_{cs}\}$ the neighborhood of X_i. X_i will be selected in the Matern selection process if and only if $\forall X_j \in \mathcal{V}(X_i) \ \ m_i < m_j$. In other words, this means that X_i has the smallest mark m_i in its neighborhood. The Matern selection is illustrated in Figure 3.7. Node i has the smallest mark m_i within its neighborhood. Although node q does in fact have a smaller mark, it is not within node i's neighborhood. We should point out that, for the sake of simplicity, here we have not taken into account any Rayleigh fading ($F \equiv 1$) and thus the neighborhood of node i is a disc.

The technique based on marks used by the Matern selection process results in an over-elimination of nodes. When a node is eliminated by a node with a smaller mark, the node that has the smallest back-off in its neighborhood can start transmitting. The nodes that have been eliminated should not eliminate other nodes. But this over-elimination can occur, as shown in Figure 3.7. Node o is eliminated by node i, but node o eliminates node p in the Matern selection process, whereas in a CSMA system, node o is correctly eliminated by node i, but, being eliminated, node o cannot eliminate another node. We do not take this case into account in our model.

We note the medium access indicator of node X_i $e_i = \in (\forall X_j \in \mathcal{V}(X_i) m_i < m_j)$

The mean number of neighbors of a node is :

$$\mathcal{N} = \lambda \int_{\mathcal{S}} P\{F \geq P_{cs} l(|x|)\} dx.$$

In a 2D network, we have :

$$\mathcal{N} = \frac{2\pi\lambda\Gamma(2/\beta)}{\beta(P_{cs}\mu)^{2/\beta}}.$$

In a 1D network, we have :

$$\mathcal{N} = \frac{2\lambda\Gamma(1/\beta)}{\beta(P_{cs}\mu)^{1/\beta}}.$$

This result is very simple. Let F_j^0 be the fading between the node at the origin X_i and node X_j

This is just the application of Slivnyak's theorem and Campbell's formula, see [85, 83]

$$\begin{aligned}\mathcal{N} = & \quad E^0\Big[\sum_{X_j\in\phi} \in (F_j^0 l(|X_j - X_i|) \geq P_{cs}\Big] \\ = & \quad \lambda\int_{\mathbb{S}} P\{F \geq P_{cs} l(|x|)\}dx\end{aligned}$$

A straightforward computation provides the explicit value of $\mathcal{N}$ in the 1D and 2D cases.

The probability p that a given node X_0 transmits i.e. $e_0 = 1$ is :

$$p =^0 [e_0] = \frac{1 - e^{-\mathcal{N}}}{\mathcal{N}}.$$

Démonstration. The proof is obtained by computing the probability that a given node X_0 at the origin with a mark $m = t$ is allowed to transmit. The result is then obtained by deconditioning on t. The details of the proof can be found in [83].

□

Thus p measures the probability of transmission in a CSMA network. If p is close to 1 it means that the carrier sense does not restrain transmissions. In contrast, if p is small, this means that the carrier sense imposes a severe restriction on transmissions.

The probability that X_0 transmits given that there is another node $X_j \in \Phi$ at distance r is p_r with

$$p_r = p - e^{-P_{cs}\mu l(r)}\Big(\frac{1 - e^{-\mathcal{N}}}{\mathcal{N}^2} - \frac{e^{-\mathcal{N}}}{\mathcal{N}}\Big)$$

Démonstration. The proof is the same as that of Proposition 3.4.3. □

Let us suppose that X_1 and X_2 are two points in Φ such that $|X_1 - X_2| = r$. We suppose that node X_2 is retained by the selection process. The probability that X_1 is also retained is :

$$h(r) = \frac{\frac{2}{b(r)-\mathcal{N}}\left(\frac{1-e^{-\mathcal{N}}}{\mathcal{N}} - \frac{1-e^{-b(r)}}{b(r)}\right)\left(1 - e^{-P_{cs}\mu l(r)}\right)}{\frac{1-e^{-\mathcal{N}}}{\mathcal{N}} - e^{-P_{cs}\mu l(r)}\left(\frac{1-e^{-\mathcal{N}}}{\mathcal{N}^2} - \frac{e^{-\mathcal{N}}}{\mathcal{N}}\right)}$$

with

$$b(r) = 2\mathcal{N} - \lambda \int_{\mathcal{S}} e^{-P_{cs}\mu(l(|x|)+l(|r-x|)} dx.$$

In a 2D network, we have :

$$b(r) = 2\mathcal{N} - \lambda \int_0^{\infty} \int_0^{2\pi} e^{-P_{cs}\mu(l(\tau)+l(\sqrt{\tau^2+r^2-2r\tau cos(\theta)}))} d\tau d\theta.$$

In a 1D network, we have :

$$b(r) = 2\mathcal{N} - \lambda \int_{-\infty}^{\infty} e^{-P_{cs}\mu(l(\tau)+l(|r-\tau|))} d\tau$$

Démonstration. The proof can be found in [83] □

3.5 The support vector machine technique

The idea is to use the Support Vector Machine technique to compute the reception probability of p_i versus $x_i = (r_i, T_i)$. In the first step, we assume that we know the p_i and the related value x_i (thus we know $(p_i)_{1\leq i\leq N}$ and $(r_i, T_i)_{1\leq i\leq N}$) and we have to predict the reception probability using other values : $(p'_i)_{1\leq i\leq N}$ knowing $(r'_i, T'_i)_{1\leq i\leq N}$. We assume that

$$p_i = w^T \phi(x_i) + b \tag{3.5}$$

where w and b are two unknown vectors and $\phi(x)$ an unknown function of a vector x.

To solve these equations, we introduce the following convex optimization problem :

$$\text{minimize} \quad \frac{1}{2}||w||^2$$

$$\text{subject to } -\epsilon \leq p_i - w^T\phi(x_i) - b \leq \epsilon. \tag{3.6}$$

This problem assumes that the function given in 3.5 can approximate the set of points that is given *i.e* $(p_i, T_i)_{1\leq i\leq N}$ with an accuracy of ϵ. Sometimes this is not possible and some errors must be accepted. In this case, slack variables, which allow us to cope with impossible constraints, are introduced. This relaxation procedure uses a cost function. The convex problem then becomes :

$$\text{minimize} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$\text{subject to } -\epsilon - \xi_i^* \leq p_i - w^T\phi(x_i) - b \leq \epsilon + \xi_i \text{ with } \xi_i^*, \xi_i > 0 \tag{3.7}$$

This problem can be solved by using Lagrange multipliers. The problem becomes :

$$\begin{aligned} L: &= \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\nu_i\xi_i + \nu_i^*\xi_i^*) \\ &- \sum_{i=1}^{N}\alpha_i(\epsilon + \xi_i - p_i + w^T\phi(x_i) + b) \\ &- \sum_{i=1}^{N}\alpha_i^*(\epsilon + \xi_i^* + p_i - w^T\phi(x_i) - b) \end{aligned}$$

where L is the Lagrangian and $\nu_i, \nu_i^*, \xi_i^*, \xi_i^*$ are the Lagrangian multipliers which are thus positive *i.e* $\nu_i, \nu_i^*, \xi_i^*, \xi_i^* > 0$

We know that the minimum of L is attained when the partial derivatives are zero, thus :

$$\partial L/\partial b = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$$

$$\partial L/\partial w = w - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i) = 0$$

$$\partial L/\partial \xi_i^{(*)} = C - \alpha_i^{(*)} - \nu_i^{(*)} = 0$$

The substitution of these equations in the Lagrangian leads to the following problem :

$$\begin{aligned} maximize \quad & - \quad \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi(x_i)^T\phi(x_j) \\ & - \quad \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \end{aligned}$$

subject to

$$\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C].$$

Thus, it follows that :

$$w = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i)$$

and

$$p_i(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\phi(x_i)^T\phi(x) + b$$

This formula is called the Support Vector expansion. The complexity of the function representation only depends on the dimensionality of the input space.

The next step of the analysis uses the Karush-Kuhn-Tucker (KKT) conditions. These conditions imply that at the solution the product between the constraints and the dual variable must vanish. In other words, we have :

$$\begin{aligned} \alpha_i(\epsilon + \xi_i - p_i + w^T\phi(x_i) + b) = 0 \\ \alpha_i^*(\epsilon + \xi_i^* + p_i - w^T\phi(x_i) - b) = 0 \end{aligned} \tag{3.8}$$

and

$$(C - \alpha_i)\xi_i = 0$$
$$(C - \alpha_i^*)\xi_i^* = 0$$

We can deduce that only the samples that do not satisfy the constraint of 3.7 have $\alpha_i^{(*)} = C$. Moreover, since the two values of the right part of 3.8 cannot be simultaneously 0, then we have $\alpha_i \alpha_i^* = 0$. Thus, after some observations we have :

$$\max(-\epsilon + p_i - w^T\phi(x_i)|\alpha_i < C \text{ or } \alpha_i^* > 0) \leq b \leq$$
$$\min(-\epsilon + p_i - w^T\phi(x_i)|\alpha_i < C \text{ or } \alpha_i^* > 0)$$

This part is adapted from [66] a tutorial by Smola.

3.6 Numerical Results

In this part, we use the results of Section 3.4 to build the database with which we train the SVM model as defined in Section 3.5.

For the radio model we use a Rayleigh fading with $\mu = 10$ and a density of vehicles $\lambda = 0.02$ vehicles per meter. We assume that $\beta = 4$ and we use the 1D model, which is more suitable for VANETs than the 2D model.

The database is built as follows : r varies from $10m$ to $250m$ by steps of $10m$ and we vary T from 1 to 10 by steps of 0.2. For each of these couples (r, T), we use the analytical model to compute the probability p of successful reception. We also assume that $P_{cs} = 4.510^{-10}$ which is approximately the value that optimizes the density of successful transmissions for $r = 50m$ and for $T = 10$.

3.6.1 Results with no errors in the database

The database is built as described above. Initially, we do not include any errors in the measurement of the probability of successful transmission given by the analytical model. For

the Support Vector Machine algorithm, we use the libsvm program in regression mode with the option :

'-s 3 -c 10 -p 0.00000001 -e 0.000000001'

to build the SVM model. Thus, the kernel used is an exponential kernel. In Figures 3.8 and 3.9, we present the probability of successful reception versus T for respectively $r = 75m$ and $r = 125m$. The results compare the prediction of the SVM algorithm with the results of the analytical model. We observe a perfect matching between the two approaches. Thus, we can foresee that SVM (with an exponential kernel) is very suitable to predict the probability of successful transmission in a VANET using RSUs and a CSMA-based access scheme.

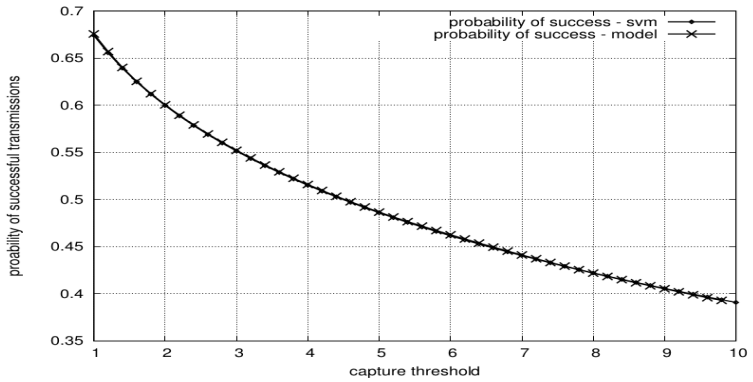

FIGURE 3.8 – Probability of successful reception versus T (x= 75m , $\mu = 10$, $\beta = 4$).

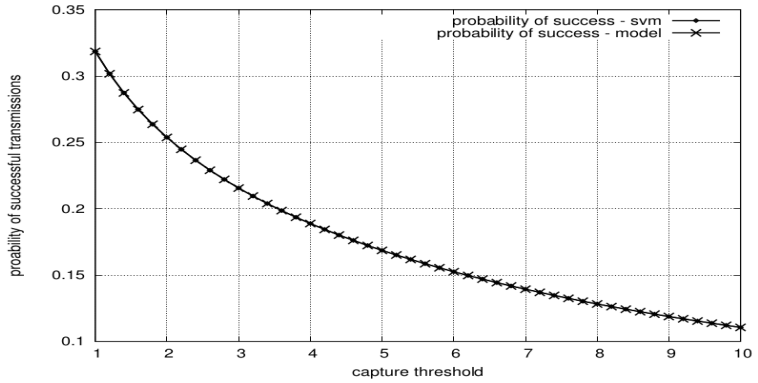

FIGURE 3.9 – Probability of successful reception versus T (x= 125m , $\mu = 10$, $\beta = 4$).

3.6.2 Results with errors in the database

We assume that the database includes errors in the measurements. We assume that the values in the database are the real values multiplied by $(1 + N(0, 0.05))$ where $N(0, 0.05)$ is the normal random variable of mean 0 and variance 0.05.

Here again, the prediction by the SVM model is very good. Even with noise measurements, SVM remains very good at predicting the transmission probability in VANETs using an access protocol based on CSMA.

In Figures 3.10 and 3.11, we present the probability of successful reception versus T for respectively $r = 75m$ and $r = 125m$ with these assumptions and we can compare the prediction of the SVM model with the direct analytical model.

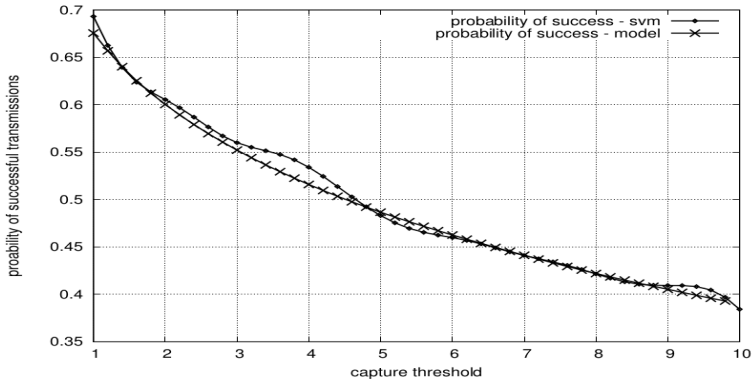

FIGURE 3.10 – Probability of successful reception versus T with error in the database (x= 75m , $\mu = 10$, $\beta = 4$).

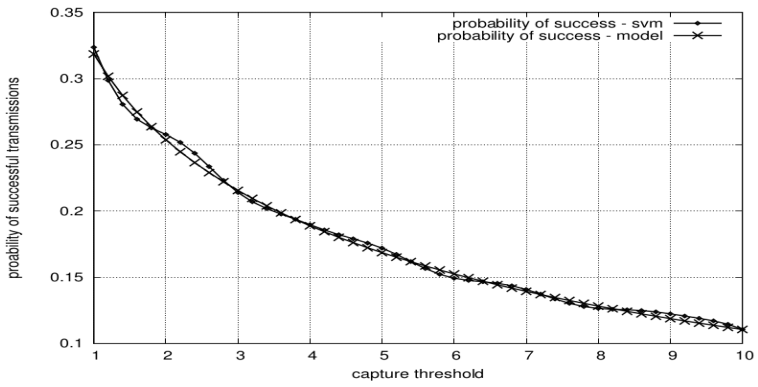

FIGURE 3.11 – Probability of successful reception versus T with error in the database (x= 125m , $\mu = 10$, $\beta = 4$).

3.6.3 Optimal data rate with SVM

Here, the idea is to use the prediction of SVM to predict the transmission probability. Then it is possible for a vehicle to compute the expected throughput : $Wp_c(r, T, P_{cs})$ versus the capture threshold used T and to choose the best transmission rate W corresponding to a given value of T. We assume a throughput of $W = 6Mbits/s$ with $T = 1$, a throughput of $W = 9Mbits/s$ with $T = 4.65$ and a throughput of $W = 12Mbits/s$ with $T = 7$. These values of T are compatible with Shannon's law $W = k \log_2(1 + T)$.

In Figure 3.12, we show the optimal data-rate algorithm which optimizes the expected throughput. We observe that up to $154m$ the best data-rate is $12Mbits/s$ and then we have to use the lowest data rate of $6Mbits/s$.

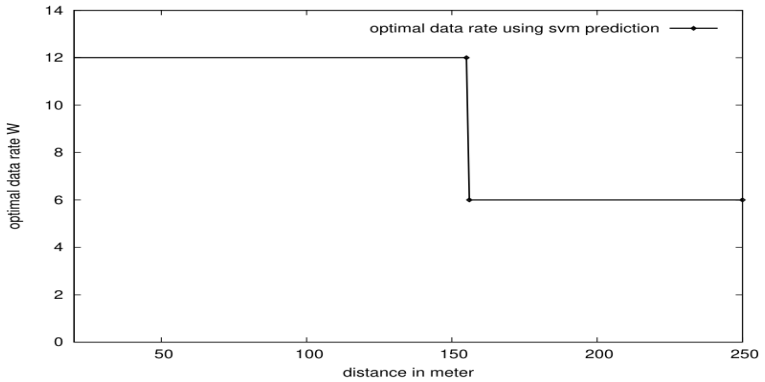

FIGURE 3.12 – Optimal data rate versus distance vehicle-RSU, $\mu = 10$, $\beta = 4$).

3.7 Conclusion

In this study, we use SVM to predict the probability of successful transmissions in a VANET using a CSMA access scheme such as IEEE 802.11p. The prediction of the SVM technique is excellent when we use a database built with an analytical model. We observe that even with errors, the prediction of the SVM technique is very good. The SVM technique can be used to build a dynamic rate control algorithm that optimizes the VANET's average throughput. Although this initial study is very encouraging, further investigations such as introducing more errors, changing the fading law, using real figures for the database, etc. should be carried out to confirm our first results. Moreover, other ML techniques such as

Random Forest, K Nearest Neighbor (KNN), etc. could be used to optimize the transmission in VANETs. We think that similar good results would be obtained. This study and a careful comparison of these techniques could be used to explore further work.

3.8 General Summary

Intelligent traffic safety, comfort and efficiency solutions have greatly benefited from advancements in communications, intelligent transportation systems. Recently, the advancements in artificial intelligence have opened up opportunities to address various types of issues in many research fields. Artificial intelligence, in different areas of the scientific research, is widely used to optimize traditional data-driven approaches. Artificial intelligence, used in Vehicle-to-everything (V2X) systems together can acquire information from diverse sources in order to enhance the driver's perception, predicting vehicle positions, as well as estimating transmission success, thereby enhancing the comfort, safety, and efficiency of driving. In this chapter, we presented a comprehensive survey of the research work that has used ML to address vehicle positioning in vehicular networks and predicting transmission success. We applied a number of techniques, namely Artificial Neural Network, K-Nearest Neighbour, Support Vector Machine, and Random Forest. We have summarized the contribution of these research studies by comparing their performances. Finally, we presented open issues and research challenges that need to be addressed in order to realize the full potential of AI to advance V2X systems.

Chapitre 4

Accident forecasting using Machine Learning and hybrid intelligence in VANETs

Contenu

4.1 Introduction

Unfortunately, the number of road traffic accidents continues to rise due to rapid urban growth and the ever-increasing density of vehicles in cities and the surrounding areas. According to statistics from the World Health Organization, each year approximately 1.25 million people lose their lives in road traffic accidents worldwide, which means that one person is killed in every 25 seconds. The statistics predict that road accidents will grow by 65% and

become the fifth greatest cause of fatalities by 2030. However, with emerging sensor devices and the IoT, it has become feasible to configure future vehicles with safety sensors to prevent accidents.

Therefore, recent research efforts have been orchestrated to investigate the state-of-the-art of vehicle safety analysis in VANETs, including the issue of road accidents. It has become clear that different aspects of driving safety [11] and road safety involve various dimensions and parameters. For example, in driving safety analysis, the style and behavior of a driver must be studied to identify and investigate unorthodox driving styles and neighboring conditions are analyzed to enable the provision of intelligent support for drivers. However, the major constraints of such analyses are the absence of substantial data sets and a precise means of identifying the neighboring conditions without incorporating additional sensors.

On the other hand, in road safety analysis, the subsidiary effect on road safety from external parameters, including the road surface, geometry, traffic flow, weather conditions and both drivers' and pedestrians' behavior is analyzed. In spite of considerable research efforts, it has not been possible to provide the most deterministic and computationally intelligent model [86] to predict the exact context of road accidents due to unbalanced data instances at various levels. Accident prediction is one of the most crucial aspects of road safety where precautionary measures are taken to avoid an accident before it occurs.

Therefore, it is worth investigating the accident–prone areas of cities and the effect of external factors in order to forecast the safety level of roads with appropriate granularity.

There are huge variations in road traffic accidents in terms of the seriousness of the accident and the damage to people and their property, which is also referred to as accident severity [87]. It is necessary to investigate the relationship between traffic accident severity and the related risk factors such as the traffic volume, the driver's age, the maximum velocity possible, geometrical factors like the type of vehicle, the distance from the nearest intersection, details of the intersections, etc. and the environmental features of traffic sites like weather conditions, lighting conditions, type of roadways, etc. Previous studies, in this area, are broadly classified into two categories : statistical modeling and machine-learning modeling. Initially, accident severity analysis for traffic accident forecasting was primarily done using statistical techniques [88, 89, 90, 91, 92, 88]. Statistical models have been prefer-

red over machine-learning models due to their solid theoretical base and strong mathematical formulation [93]. Gaussian mixture models (GMMs) are the most common, and they have been successfully used in a wide variety of fields, such as speaker recognition systems, video image processing, and pattern classification. In studies related to traffic flow and accidents, GMMs have been repeatedly used to model raw time-to-collision (TTC) samples for traffic safety prediction [94, 95], severity detection of traffic accidents along with Hidden Markov models [96] and in traffic flow forecasting [97]. However, these models assume some inherent properties about the data patterns like the assumption of risk factors influencing accident severity linearly which might not always be the case [98, 99] and hence they inevitably induce inaccurate results.

Machine-learning models, on the other hand [100], are highly adaptable with no or little presumption about the input features and offer higher flexibility to outliers, inaccurate and missing data. Some of the popular ML models applied to traffic accident related studies are Decision Trees [101, 102], Support vector machines (SVMs) and support vector classifiers (SVCs)[103, 104], K-means clustering [105, 106], Artificial Neural Networks(ANNs) [107, 108, 109], etc. Amongst all the above models, support vector machines have been increasingly used in traffic related studies to address traffic flow prediction [110, 111, 112, 113], crash frequency analysis [114, 115, 116, 117], and to analyze accident severity in a crash [118, 119, 120]. The major drawback of ML models is their performance as a 'black-box' which leads to unclear inference of the function that actually correlates the input variables with the target class.[121]. The purpose of this study is to combine the Gaussian Mixture Model from statistical modeling and the Support Vector Classifier from machine learning modeling to overcome the disadvantages of one model by the advantages of the other and hence improve the overall accuracy. Favoured by its innate discriminative power, even in the case of non-linearly separable classes using kernels, the SVM presents an attractive way of enhancing the baseline generative model (GMM) [122]. Hence the GMM serves as a parametric basis for the support vector classifier. Since SVCs perform poorly on unbalanced data and cannot select relevant attributes with respect to the target variable, data preprocessing using re-sampling techniques and feature importance ranking methods are applied.

4.2 Literature Review

We review road safety prediction that incorporates various levels of machine learning and intelligent algorithms. Then, we analyze the advantages and drawbacks of various methods, and formulate the missing value problem, which is a substantial challenge in this area of research.

For traffic flow forecasting, various ML methods have been employed. The most prominent among these are : Autoregressive Integrated Moving Average (ARIMA) which belongs to time series categories [123] ; probabilistic graphical models, such as Bayesian Network [124], Markov Chain [125], and Markov Random Fields (MRFs) [126] ; and nonparametric approaches, such as Artificial Neural Networks (ANNs) [125], Support Vector Regression (SVR) [127], and Locally Weighted Learning (LWL) [126]. However, as seen in the literature, there are multiple reasons for fluctuations in the traffic flow. In addition to that, the patterns in the data are multimodal. These multimodal properties make it difficult to learn. Moreover, in order for these shallow network approaches to be able to model complex mapping, they require a high dimensional space, which requires a huge amount of annotated data. Therefore, in the high-dimensional space, the overfitting problem becomes acute. In order to overcome this issue, we use a multilayer nonlinear structure since deep-learning approaches have a strong ability to express multimodel patterns in data using a reduced number of dimensions.

An ANN (Artificial Neural Network) [128] is a type of network in machine learning that has been widely used for road incident prediction in different environments (freeway, highway, urban and non urban roads, etc.) in order to minimize injury and loss of life on the roads.

An ANN aims to reproduce and simulate human behaviour and cognitive functions. It uses a network of nodes, often called neurons, that contain configurable weights and these weights can be trained to produce a desired output [129]. Many kinds of pattern recognition problems can be solved by configuring the layers and the weights of the network. Today, ML techniques have found different applications in the fields. These include, for instance, road safety where they have been used for collision detection. As illustration, in the study done in [128] by Chang, an ANN is implemented to predict collisions on a National Freeway in Taiwan. Road features were used as input in their model and it is claimed that the model accepted

those features and provided the number of collisions as output. However, the ANN model has many local minimums, which makes it difficult to find the global optimum solution. This highlights that even though ANN models are easy to understand, their weights solution space is non-convex. This is considered to be one of its drawbacks. Another shortcoming of an ANN is that it is supervised based learning, which means that the model requires training data, which limits its applicability in real-world situations. In order to overcome these issues, back propagation (BP) such as Bayesian regularization has been proposed, although it has been found that Bayesian regularization has led to great improvements, it still requires training data, which makes its applicability in the real world limited.

Banyesian Networks (BNs) have also become very popular in traffic prediction. When forecasting, BNs enable prediction of traffic to consider multiple inputs of data. It is known to have applications that can take many forms. It has been pointed out that the inputs of BNs sometimes show less relativity than that of neural networks [130][97]. This specific characteristic of BNs offers more possibilities in the usage of combining different prediction factors.

Research has revealed that, for traffic prediction, there is no single method that can be considered to be the best for every situation. Thus, in traffic forecasting, researchers are constantly trying to combine different models. It has been observed that almost all the research studies undertaken in using hybrid models (HMs) in traffic prediction yield higher prediction accuracy compared to those that use a single model [131][132][133]. To get better accuracy of prediction, Jiaming Xie and Yi-King Choi conducted research on designing and implementing a hybrid model that can forecast the traffic flow of the city of Hong Kong by using historical and real-time data. The question that arises is how one can balance the importance of historical data and real data. This is because it is obvious that the traffic situation changes over time and that continuous changes make the traffic status dynamic [134].

As no single model can be suitable for prediction in all kinds of situation ; the main objective of this research is to build a prediction model that combines two approaches (Gaussian mixture model and support vector classifiers) in order to predict traffic accidents. The improvement in terms of accuracy is very notable compared to other models.

As the aim of this work is to derive analytics towards the prediction of road accidents, it is important to include a data acquisition mechanism (see Figure4.1) and inter-process communication between vehicles on the road. For a standard use case, we consider a segment of a fairly densely populated city road where this type of acquisition model can be placed. To formulate such a model, the physical components of the VANETs can be one of the parameters. However, inter-process communication protocol of 5G and beyond can establish a more reliable process exchange mechanism.

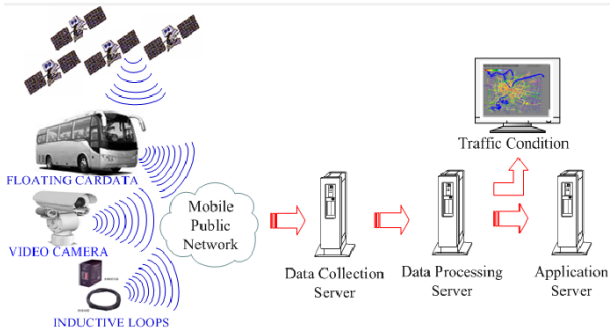

FIGURE 4.1 – Traffic-data-acquisition -using-different-detectors [135]

In conventional usage, a Message Application Programming Interface (MAPI) can be suitable for use cases to collect and to propagate the data for that particular segment of road. In this work, certain realistic scenarios are considered to formulate the data acquisition and the message passing mechanism.

In the overall system flow for message passing and broadcasting, we recognize four layers :

— Application layer (message wrapping mechanism) ;

— Transport layer (with handshaking between the sender and the receiver) ;

— Network layer (message distribution mechanism) ;

— Physical layer (connection and devices).

Thanks to these layers, the system can collect on-road data to pass them towards a particular nearest cloud center or adjacent vehicle, so that the neighbouring vehicle can receive an alert containing several features.

The format of the message is the following :

1. transaction ID,

2. previous transactions ID if any
3. sender ID
4. reputation ID (this number will present the reputation or likelihood for a given vehicle to have an accident)
5. receiver ID
6. message content (message text, location, direction of sender vehicle)
7. message type (this information will deliver the warning or the prediction of the message after using a hybrid analytical algorithm such as intersection movement, tendency for forward collision, deviation from the lane, extreme conditions of the road surface and other relevant features).

Including all these message fields, each and every message will be followed by the action requested. For example, the requested action indicates the message from the sender vehicle to all receivers for a particular alert if the adjacent vehicle could change lane abruptly. Similarly, if a vehicle accelerates more than normal, in spite of the traffic congestion on the road, then there will be an alert message for other neighbouring vehicles.

All these features of message access and distribution can be implemented in a traffic environment simulator (SIMO) which can be deployed by another event-based simulator like OMNET++. However, since the objective of our work is to analyze the action requested by the given message distribution format, we did not use these kinds of simulation tools.

4.2.1 Mathematical symbols at glance

4.3 Dataset description

The dataset was obtained from data.govt.uk [136] which is a United Kingdom Government project providing open source data published by central government, local authorities and public bodies. The road traffic accident database for the year of 2017 was utilized in this novel research. These dataset files provide detailed road safety data about the environmental, physical, geometrical, geographical and personal information related to accidents as shown in 4.2. The data points correspond only to the accidents whose information was reported to the police or the authorities. The dataset was compiled from the information recorded

Symbols used in the Gaussian mixture model	
Symbol	Semantics
$p(x/\lambda)$	Conditional probability of x given λ
$N_i(x)$	Gaussian probability density function
w_i	Mixing weights
K	Number of component gaussians
x	Observed data
μ_i	Mean vector of component gaussians
Σ_i	Co-variance matrix of the component gaussians
λ	Set of tuples of model parameters w_i, μ_i and Σ_i
M	Number of training examples
$L(x/\lambda)$	Log likelihood
Symbols used in the Support vector classifier	
Symbol	Semantics
(x_i, y_i)	Data examples
$f(x)$	Classification function
w	Normal direction cosines to the line
$\|\|a\|\|$	Vector norm
ξ_i	Slack variable
C	Penalty index for outliers
α_i	Linear combination weights
K	Kernel
$\langle x, y \rangle$	Inner product between x and y
d	degree of polynomial kernel
σ	Variance in the RBF kernel

TABLE 4.1 – Mathematical symbols

in the STATS19 accident reporting form. The entire dataset is mainly composed of three main categories : accident, vehicle and casualty data. The accident variables have 31 features, including the weather conditions, lighting, time, day of the week, number of vehicles involved, etc. The Vehicle-Driver database consists of 22 features like the age of the vehicle, sex and age of the driver, etc., and the casualty dataset contains 15 feature variables such as the type of victim, sex of the casualty, age of the casualty, etc.

The output class or accident severity is divided into 3 categories, namely ”no injury in the accident” encoded as 1, ”non-incapacitating injury in the accident” encoded as 2 and ”incapacitating injury in the accident” encoded as 3. The detailed encoding of every variable has been listed in Table 4.3.

Dataset variables		
Accident variables	Vehicle and driver variables	Casualty variables
Index of the crash		
Police Force		
Accident Severity		
Vehicles involved		
number of victims involved	Index of the crash	
Date	Vehicle Id code	
Day of the Week	Kind of vehicle	
Time	Towing and Articulation	Index of the crash
Longitude	Vehicle Manoeuvre	Vehicle Id code
Latitude	Position of vehicle	Casualty Id code
District of Local Authority	Location of intersection	Type of victim
Local Highway Authority	Skidding and Overturning	Gender of victim
1st Road Class	Hit Object in Carriageway	Age of victim
1st Road Number	Vehicle Leaving Carriageway	Age Band of victim
Kind of roadway	Hit Object off Carriageway	Intensity of the fatality
maximum velocity possible	1st Point of Impact	Position of the pedestrian
Details of intersection	Was Vehicle Left Hand Drive	Motion of the pedestrian
2nd Road Class	Journey Purpose of Driver	Position of victim in car
2nd Road Number	Gender of the Driver	Position of victim in Bus or coach
Pedestrian Crossing-Human Control	Age of the Driver	ype of fatality
Pedestrian Crossing-Physical Facilities	Age band of the Driver	Casualty IMD Decile
Lighting	Motor power	Location of victims house
Road conditions	Vehicle fuel type	
raod surface conditions	age of the vehicle	
Special characteristics of accident location	Rider IMD Decile	
Carriageway Hazards	Rider Home Area Type	
Type of area		
Police attention		

Table 4.2 – Dataset Variables

Variable name	Variable categories	Code	Frequency	Percentage	Class 1	Class 2	Class 3
Day of the week	Sunday	1	21836	12.20%	20.4%	11.4%	14.7%
	Monday	2	24653	13.77%	12.4%	14.1%	12.2%
	Tuesday	3	25911	14.48%	10.4%	14.7%	13.6%
	Wednesday	4	25837	14.44%	14.4%	14.5%	14.1%
	Thursday	5	27014	15.09 %	11.0 %	15.4 %	14.2%
	Friday	6	29227	16.33%	13.1%	16.5%	15.8%
	Saturday	7	24440	13.66%	18.2%	13.1%	15.3%
Kind of roadway	Traffic circle	1	10328	5.77%	1.05%	3.67%	6.35%
	Single direction traffic	2	2740	1.53%	0.68%	1.28%	1.60%
	Divided highway	3	33739	18.85%	22.9%	16.62%	19.27%
	Undivided highway	6	128608	71.88%	74.8%	77.12%	70.61%
	Slip road	7	1904	1.06%	0.37 %	0.82%	1.13%
	Unknown	9	1599	0.89%	0.06%	0.46%	1.01%
	Single direction traffic/Slip road	12	0	0%	0.00%	0.00%	0.00%
	Erroneous data	-1	0	0%	0.00%	0.00%	0.00%
Human pedestrian crossing	None within 50 metres	0	175819	98.26%	0.12%	0.42%	0.60%
	School crossing patrol	1	497	0.27%	0.06%	0.31%	1.03%
	By another official	2	1013	0.56%	99.5%	99.0%	99.0%
	Erroneous data	-1	1589	0.88%	0.27%	0.26%	0.28%
Details of intersection	No intersection within 20m	0	78468	43.85%	72.01%	49.80%	41.86%
	Traffic circle	1	13524	7.55%	1.33%	4.79%	8.32%
	Mini-Traffic circle	2	1770	0.98%	0.12%	0.68%	1.07%
	Staggered intersection	3	50618	28.29%	15.54%	27.82%	28.67%
	Slip road	5	3502	1.95%	2.32 %	1.52%	2.04%
	Crossroads	6	18403	10.28%	5.05%	8.67%	10.77%
	More than 4 arms (not Traffic circle)	7	1663	0.92%	0.34%	0.60%	1.01%
	Private road	8	5102	2.85%	1.73%	3.28%	2.77%
	Other intersections	9	5567	3.11%	1.52%	2.73%	3.23%
	Erroneous data	-1	301	0.16 %	?%	0.04%	0.20%
Lighting	Daylight	1	126049	70.45%	60.28%	67.95%	71.24%
	Dark with lights	4	36489	20.39%	17.18%	20.02%	20.55%
	Dark with dimmed lights	5	1123	0.62%	0.86%	0.59%	0.63%
	No illumination at all	6	10314	5.76%	20.81%	8.94%	4.70%
	Unknown illumination	7	4941	2.76%	0.83 %	2.48 %	2.86 %
	Erroneous data	-1	2	0.001 %	?%	?%	0.001%
Weather conditions	Breeze	1	144368	80.68%	84.02%	81.45%	80.44%
	Rain with breeze	2	20948	11.70%	8.99%	11.93%	11.71%
	Snow with breeze	3	915	0.51%	0.49%	0.29%	0.56%
	Fine with gale	4	2032	1.13%	1.55%	1.31%	1.08%
	Rain with gale	5	1705	0.95%	1.27 %	1.16%	0.89%
	Snow with gale	6	163	0.09%	?%	0.04%	0.10%
	Fog or mist	7	940	0.52%	0.77%	0.64%	0.49%
	Others	8	3410	1.90%	1.42%	1.55%	1.99%
	Unknown	9	4435	2.47%	1.45%	1.59%	2.70%
	Erroneous data	-1	2	0.001 %	?%	?%	0.001%
Road surface	Dry	1	124151	69.39%	68.38%	69.23%	69.44%
	Wet	2	49632	27.74%	28.88%	28.55%	27.52%
	Snow	3	707	0.39%	0.31%	0.23%	0.43%
	Icy or snow	4	3172	1.77%	2.23%	1.61%	1.79%
	Flood over 3cm. deep	5	172	0.09%	0.18 %	0.17%	0.07%
	Oily	6	0	0%	0.00 %	0.00 %	0.00 %
	Silt or mud	7	0	0%	0.00%	0.00%	0.00%
	Erroneous data	-1	1084	0.60%	?%	0.16%	0.72%
Sex of the driver	Male	1	114282	63.87%	79.42%	69.89%	62.15%
	Female	2	51848	28.97%	18.33%	24.15%	30.31%
	Not known	3	12784	7.14%	2.23%	5.95%	7.52%
	Erroneous data	-1	4	0.002 %	?%	0.003%	0.002%
Type of victim	Driver	1	121870	68.11 %	63.382%	66.20%	68.65%
	Passenger	2	43204	24.14%	25.00%	22.74%	24.441%
	Pedestrian	3	13844	7.73%	11.60%	11.05%	6.89%
Police officer attend?	Yes	1	143847	80.39 %	95.90%	89.27%	78.02%
	No	2	35055	19.59 %	4.09%	10.71%	21.96%
	No - accident was reported using a self completion form	3	16	0.009 %	?%	0.006%	0.10%

FIGURE 4.2 – Variable descriptions

4.4 Research Methodology

4.4.1 Algorithm description

The models used for traffic accident forecasting are discussed in this section. The accident data including vehicle, casualty and drivers' features are collected from data.govt.uk[136]. These higher dimensional features are then preprocessed to remove any kind of erroneous entries and balance the dataset. Moreover, if the dataset has a highly unequal distribution of the number of data points corresponding to each class, the SVC model tends to predict every data sample as the majority class. In order to achieve an unbiased performance, it is

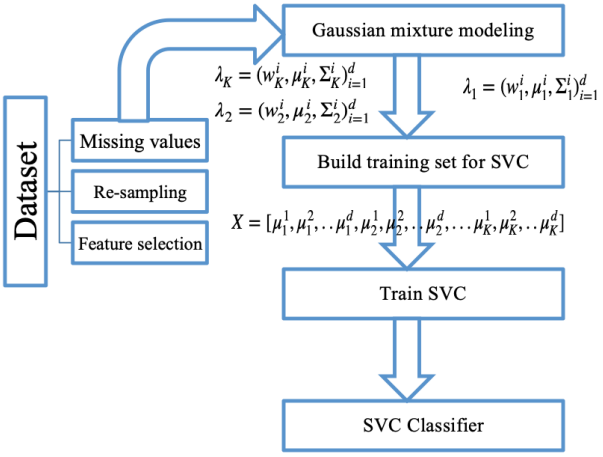

FIGURE 4.3 – Algorithm description

necessary to balance the dataset with respect to the output class.

Since the dataset is high dimensional, dimensionality reduction techniques must also be used. Furthermore, similar to Bayesian network (BN) models [88], SVMs lack the capability to automatically select the relevant features. Feature or attribute selection helps to target both of the above-mentioned disadvantages. Variable importance ranking methods are deployed and the data are further cleaned. This processed dataset is then used as input to the Gaussian mixture model [137], [138] and [139] which estimates the parameters of the various mixture gaussians using expectation maximization. Out of all the parameters i.e. the mean, variance and the mixing probability, the vector of means is adapted and used as input to the SVC model [140]. The SVC treats the accident severity modelling as a classification problem i.e. the accident data is classified into various categories based on the severity classes. This trained hybrid model is then evaluated with respect to the performance metrics and sensitivity analysis is performed. The model is also compared to the baseline GMM model and the results are reported.

A brief description of the hybrid algorithm is shown in Figure 4.3.

4.4.1.1 Data pre-processing

0. The data has been obtained from data.govt.uk. [136] The variable names have been changed, keeping the semantics the same as before.

As there was a considerable amount of missing and erroneous data, data pre-processing was performed prior to the application of the hybrid model. One can either remove the examples with erroneous data or remove the attributes with corrupted data. For the former, data processing was carried out using the *Filter Examples operator* of the *RapidMiner Studio* [1][141] software. This operator filters out the data entries as per the conditions specified by the user. This is achieved with the help of *The Select Attributes operator* in *RapidMiner*. *The Filter Examples operator* reduces the number of data entries in a dataset but it has no effect on the number of attributes. On the other hand, the *Select Attributes operator* chooses the attributes with no missing or corrupted values and has no effect on the number of examples in the example set.

4.4.1.2 Data Re-sampling

The dataset used consists of 2044 data points for accident severity Class 1, 21098 data points for Class 2, and 93321 data points for Class 3, as shown in the form of a distribution curve in Figure 4.4. This accounts for the severe imbalance in the data, causing the prediction results to be skewed significantly in favour of the majority class. This causes poor classification rates on minor classes and extreme biasing towards the majority class. In addition, it is also possible that the classifier predicts everything as a major class and ignores the minor class. To tackle this issue, one has to use re-sampling techniques to balance the data. We used the Synthetic Minority Oversampling (SMOTE) [122] up-sampling technique, which works by creating synthetic observations based upon the existing minority observations.

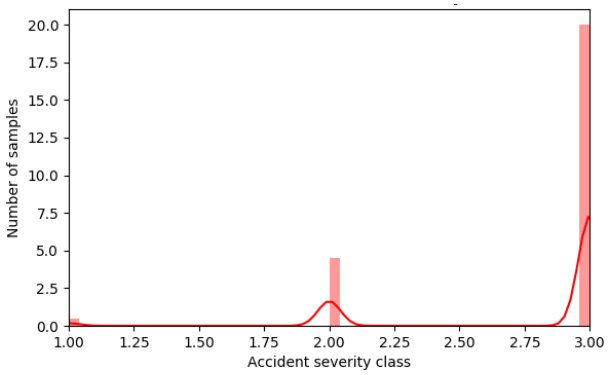

FIGURE 4.4 – Accident severity class distribution

1. Rapidminer is an open source statistical and data mining tool.

4.4.1.3 Feature/Attribute selection

In machine learning and statistics, feature selection or attribute selection is the process of selecting a subset of features that are used to build the model. It is performed to get rid of any unnecessary, irrelevant, or redundant features from the dataset, consequently resulting in improving the accuracy of the model. This also leads to better interpretability of the underlying relationship between input variables and target class. In this study, feature relevance analysis was performed using the *RapidMiner studio. The Weight by Tree Importance operator* was used to find the relevant features. The weights of the attributes are calculated by analyzing the split points of a Random Forest model. Each node of each tree is visited and the benefit created by the respective split is retrieved, which is further summed per attribute. The importance ranking is done by calculating the mean benefit over all trees. This approach was implemented following the idea from the seminal work by Menze, Bjoen H et al (2009) [142]. The higher the weight of the attributes is, the greater is their relevance. The Information Gain method was used to find the weights by the tree importance operator. Information Gain (IG) measures how much "information" a feature gives us about the class which is the entropy of the distribution before the split minus the entropy of the distribution after it. Mathematically, the information gain is given by the equation below :

$$IG = E(p) - w * E(c) \tag{4.1}$$

where IG is information gain, E is entropy, p is parent node, c stands for children and w corresponds to the average of the weights.

4.4.1.4 Gaussian Mixture Model

Accident severity data can be formulated as a weighted sum of K component Gaussian distributions :

$$p(x/\lambda) = \sum_{i=1}^{K} w_i N_i(x) \tag{4.2}$$

where x is a d dimensional vector, $N_i(x)$ are the component multivariate Gaussian densi-

ties and w_i is the mixing proportion or the mixture weights with $\sum_{i=1}^{K} w_i = 1$.

Each component multivariate Gaussian density function is given by

$$N_i(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{-1}{2}}} e^{\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)} \tag{4.3}$$

with μ_i, Σ_i as the mean vector and the co-variance matrix respectively. The above-mentioned parameters, namely w_i, μ_i and Σ_i, are represented by

$$\lambda = (w_i, \mu_i, \Sigma_i) \quad \forall i = 1, 2, 3...K \tag{4.4}$$

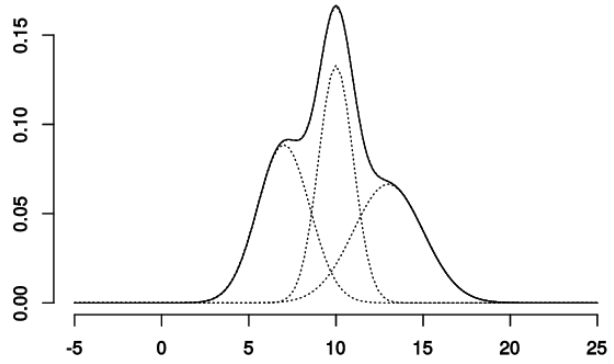

FIGURE 4.5 – Gaussian mixture model with K=3

Given the M training vectors x = $(x_1, x_2, x_3...x_M)$, the GMMs are trained with parameter evaluation using Maximum Likelihood estimation. Assuming all the training vectors are independent, the likelihood function and the log likelihood function turn out to be

$$p(x/\lambda) = \prod_{j=1}^{M} p(x_j/\lambda) \tag{4.5}$$

with the log likelihood using Equation 4.2 as

$$L(x/\lambda) = \sum_{j=1}^{M} \log\left(\sum_{i=1}^{K} w_i N_i(x)\right) \tag{4.6}$$

The maximization of the likelihood function in Equation 4.5 is achieved by using the expectation-maximization (EM) algorithm. In the expectation (E) step, a function for the expectation of the log-likelihood is constructed while in the maximization (M) step, model

parameters like the mean, variance and the mixing probability are estimated by the maximizing function found in the E step. The formulas obtained after simplification performed at each E-M step are :

E step : Posterior probability estimation

$$p(i/x_j, \lambda) = \frac{w_i N_i(x_j)}{\sum_{l=1}^{K} w_l N_l(x_j)} \tag{4.7}$$

M Step : Updating the parameters

$$w_i = \frac{1}{M} \sum_{j=1}^{M} p(i/x_j, \lambda) \tag{4.8}$$

$$\mu_i = \frac{\sum_{j=1}^{M} p(i/x_j, \lambda) x_j}{\sum_{j=1}^{M} p(i/x_j, \lambda)} \tag{4.9}$$

$$\Sigma_i = \frac{\sum_{j=1}^{M} p(i/x_j, \lambda) x_j^2}{\sum_{j=1}^{M} p(i/x_j, \lambda)} - \mu_i^2 \tag{4.10}$$

4.4.1.5 GMM and Traffic prediction

The observations (x), including features like weather conditions, lighting conditions, age of the driver, distance from the junction, etc., are assumed to be a mixture of three Gaussians (λ) which correspond to the three accident classes. Hence, our objective is to find a model that maximizes the posterior probability

$$\max_{1 \leq k \leq K} p(\lambda_k / x) \tag{4.11}$$

which by Bayes's rule is

$$\max_{1 \leq k \leq K} \frac{p(x/\lambda_k) p(\lambda_k)}{p(x)} \tag{4.12}$$

Assuming all the Gaussians to be equally likely and taking the log, we have our likelihood

function as :

$$\max_{1\leq k\leq K} \sum_{j=1}^{M} \log p(x_j/\lambda_k) \tag{4.13}$$

which is further reduced to (4.6) and solved using expectation maximization.

4.4.1.6 Support vector classification

A support vector classifier or SVC is a discriminative model that makes decisions by constructing an optimal hyper-plane or a line among linearly or non-linearly separable classes [143].

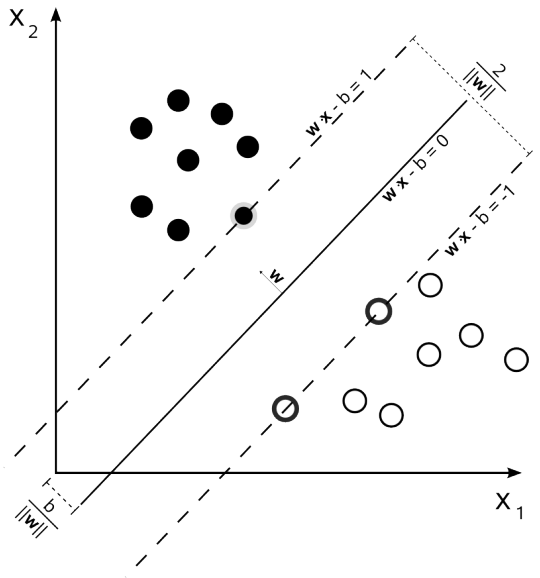

FIGURE 4.6 – Concept of optimal hyper-plane

For linear support vector classifiers on the data (x_i, y_i) with i=1,2...n, the classification function is represented as :

$$f(x) = w^T(x) + b \tag{4.14}$$

The margin according to Figure 4.6 is given by

$$\frac{|w^T(x)+b|}{|\ w}\bigg|_{w^T(x)+b=1} + \frac{|w^T(x)+b|}{|\ w}\bigg|_{w^T(x)+b=-1} = \frac{2}{|\ w} \tag{4.15}$$

since $w^T(x) + b = \pm 1$ for the support vectors.

Maximizing the margin (the minimum distance of the hyper-plane from these points), the problem can be formulated as :

$$\min \frac{1}{2} \mid w^2 y_i(x_i w + b) \geq 1 \tag{4.16}$$

The solution for the optimal w turns out to be a linear combination of support vectors i.e. which satisfies $y_i(x_i w + b) = 1$.

In the case of a non-linearly separable dataset, no hyper-plane exists that satisfies the above mentioned constraints. In that case, a new model is introduced [144] :

$$\begin{aligned} \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\ s.t. y_i(x_i w + b) \geq 1 - \xi_i \quad i = 1, 2...n \\ \xi_i \geq 0, \quad i = 1, 2..n \end{aligned} \tag{4.17}$$

where ξ_i is a non-negative factor called the slack variable responsible for allowing the functional value of certain samples to be negative. The factor 'C' is used to penalize the outliers and expresses the degree to which they are not acceptable.

The solution for the optimal w is a linear combination of all points ($\sum_i \alpha_i y_i x_i$) in the feature space that have $\xi_i > 0$ and lie on the margin ($\alpha_i \neq 0$) and hence the classification function becomes :

$$\begin{aligned} f(x) &= sign\left[\left(\sum_{i=1}^{n} \alpha_i y_i x_i \right)^T x + b \right] \\ &= sign\left[\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \right] \end{aligned} \tag{4.18}$$

The non-linear classifier can be extended using the kernel function (K) satisfying Mercer's condition to map the input features to a higher dimensional space where it is linearly separable [145, 146], as represented in Figure 4.7 . Then all the inner products are replaced with the

kernel function and hence the classification function becomes :

$$f(x) = sign\left[\sum_{i=1}^{n} \alpha_i y_i K\Big(x_i, x\Big) + b \right] \tag{4.19}$$

The most commonly used kernel functions are :

1. Polynomial kernel of degree d

$$K(x, y) = (\langle x, y \rangle + 1)^d \tag{4.20}$$

2. Radial basis function (RBF)

$$K(x, y) = exp(-\frac{\|x - y\|^2}{2\sigma^2}) \tag{4.21}$$

3. Hyperbolic tangent (Sigmoid) kernel

$$K(x, y) = tanh(\alpha.\langle x, y \rangle + c) \tag{4.22}$$

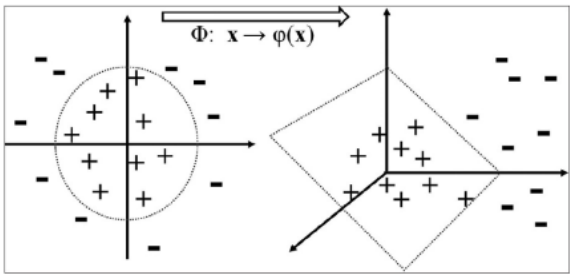

FIGURE 4.7 – Kernel trick

4.4.1.7 Multiclass SVC

1. One-against-all method :

   This method [147] considers N classifiers where N is the number of classes and trains the i^{th} classifier with all other examples considering the instances of the i^{th} class as positive and all other instances as negative labels.

2. One-against-one method :

   This method [148] constructs $N(N-1)/2$ classifiers and trains the i^{th} classifier with every j^{th} classifier considering the instances of the i^{th} classifier as positive and those of the j^{th} classifier as negative.

4.4.2 The Need for a Hybrid Model

A hybrid model refers to an approach that combines different methods that could produce better results than any of those methods applied individually. A hybrid research model breaks down the boundaries between models in order to tackle issues that a single method of research would be unlikely to solve and give better results. As different cases may be tackled in different ways, research has shown that, for traffic prediction, there is no single method that can be considered to be the best for every situation, and so this has led researchers to attempt to combine different models. It has been observed that almost all the research undertaken using hybrid models (HMs) in traffic prediction yields higher prediction accuracy than those using a single model [131][132][149][150]. For instance, to get better accuracy of prediction, Jiaming Xie and Yi-King Choi conducted a research on designing and implementing a hybrid model that can forecast the traffic flow of the city of Hong Kong by using historical and real-time data.

As mentioned above, it has been shown that no single model can be suitable for prediction in all kinds of situation.

In this work, the need for HMs is justified by the use of statistical modeling method i.e., a Gaussian Mixture Model (GMM) and a machine-learning modeling scheme i.e., the Support Vector Classifier (SVC). As the the baseline generative model (GMM) could only classify with maximum likelihood, the SVM presents an attractive way of enhancing it. This is because the SVM is favoured by its innate discriminative power, even in the case of non-linearly separable classes using kernels. However, SVCs also perform poorly on unbalanced data. Hence the GMM serves as a parametric basis for the support vector classifier. Therefore, in this work, the use of HM was needed in order to overcome the disadvantages of one model by the advantages of the other and hence to improve the accuracy.

4.5 DISCUSSION OF RESULTS

4.5.1 Data pre-processing results

Erroneous and missing data entries were removed using the *RapidMiner Studio*. By removing all the missing values, the data was reduced from 178918 examples and 69 attributes to 116463 examples and 69 attributes. By removing the attributes with missing values, the number of features was reduced from 69 to 62. The variables and the number of missing values are listed in Table 4.3 below.

Variable name	Number of missing values
Index of the crash	53701
LSOA of crash location	17736
Longitude	59
Latitude	59
Location Easting OSGR	35
Location Northing OSGR	35
Time	3

TABLE 4.3 – Attributes with missing and erroneous values

4.5.2 Data Re-sampling results

The imbalanced dataset is balanced using SMOTE from the 'imblearn' module of python. This is an upsampling technique which balances the data by increasing the number of data points for the minority class. After applying SMOTE on our dataset, we received a total of 279963 samples with 93170 samples from Class 1, 93756 samples from Class 2 and 93037 from Class 3, as listed in Table 4.4.

Accident severity class	Training samples before SMOTE	Training samples after SMOTE
Class 1	2044	93170
Class 2	21098	93756
Class 3	93321	93037

TABLE 4.4 – Data re-sampling results

4.5.3 Feature selection results

The variable importance ranking based on the three accident severity levels was conducted using the *RapidMiner Studio*. Weight by the tree importance ranking operator was used after applying the Random forest model on the processed data. In addition to this, the information gain method was used to find weights using tree importance. In this, the variable with the largest score is normalized to 1 and the scores of all the others are calculated with respect to the best performing variable. The results obtained are shown below in Table 4.5 and Figure 4.8.

Variable	Score	Variable	Score	Variable
Intensity of the fatality	1	Age of victim	0.092	Did Police Officer Atte
Location Northing OSGR	0.699	Lighting	0.091	Type of area
Latitude	0.693	Location of intersection	0.086	Hit Object off Carriage
Longitude	0.661	Location of victims house	0.078	Type of fatality
Location Easting OSGR	0.654	Kind of road	0.075	Vehicle Id code x
Date	0.439	Age Band of victim	0.073	Gender of the driver
1st Road Number	0.385	Kind of vehicle	0.069	Rider Home Area Type
District of Local Authority	0.317	Skidding and Overturning	0.067	Vehicle Id code y
Vehicles involved	0.288	Carriageway Hazards	0.067	Casualty Id code
Day of Week	0.262	road surface conditions	0.063	special characteristics c
Area of police responsible	0.222	2nd Road Class	0.063	Gender of victim
maximum velocity possible	0.21	Vehicle IMD Decile	0.062	Position of the pedestri
number of victims involved	0.21	1st Point of Impact	0.06	Hit Object in Carriage
Details of intersection	0.175	Casualty IMD Decile	0.055	Propulsion Code
1st Road Class	0.129	Age band of the driver	0.052	Towing and Articulatio
How old is the vehicle ?	0.12	Journey Purpose of Driver	0.052	Position of victim in B
Age of the driver	0.117	Type of victim	0.051	Position of vehicle
Traffic control at intersection	0.113	Pedestrian Crossing-Physical Facilities	0.05	Motion of the pedestria
Rider IMD Decile	0.11	Vehicle Manoeuvre	0.049	Was Vehicle Left Hand
Vehicle Leaving Carriageway	0.109	Pedestrian Crossing-Human Control	0.046	Pedestrian Road Maint
2nd Road Number	0.108	Position of victim in car	0.045	
Motor power (CC)	0.106	weather conditions	0.042	

TABLE 4.5 – Variable relevance scores

It can be seen that among all the variables, Intensity of the fatality is the most related to accident severity with a score of 1. The location attributes like Location Northing OSGR, Latitude, Location Easting OSGR, Longitude follow in the list. The Pedestrian Road Maintenance Worker variable was of the least importance with a score of 0.0. Surprisingly, weather

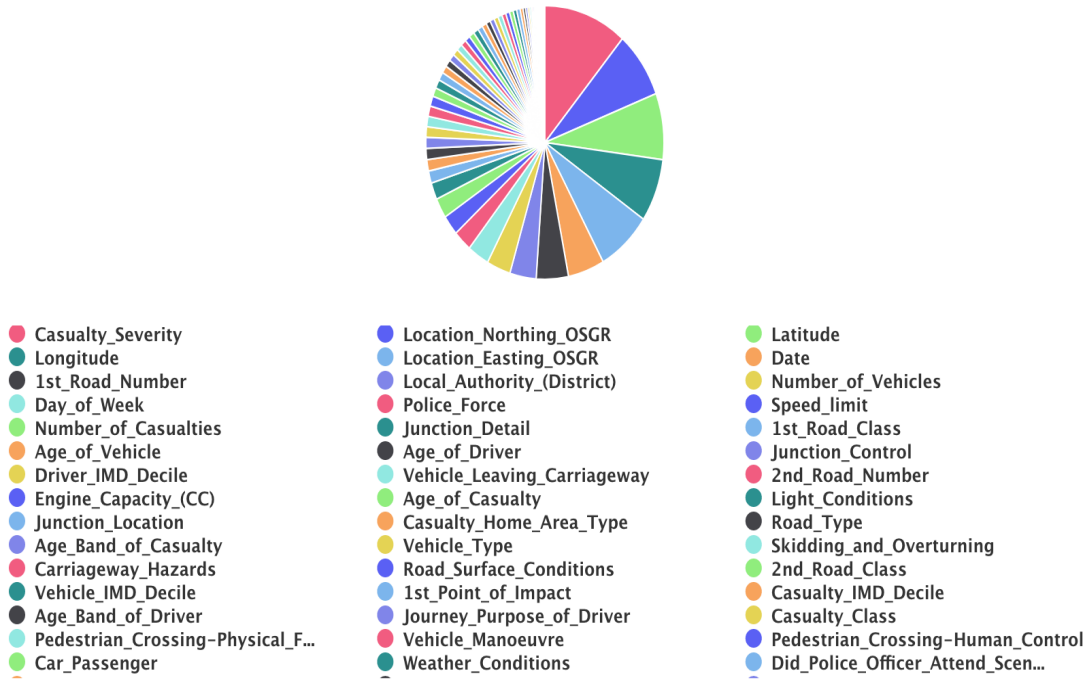

FIGURE 4.8 – Variable importance score distribution

conditions had a relevance of 4.2%, which is quite insignificant. Factors like maximum velocity possible, day of the week, Vehicles involved in the accident, date, etc. contributed significantly with a score above 20%. Features like Details of intersection, Location of intersection, Kind of road, lighting conditions, age of the casualty, etc., also turn out to be quite important to the hybrid model. All features varying from environmental, physical, geometrical, geographical and historical were included in the top features ranked using this technique. The results obtained align with one's personal experience and knowledge about the risk factors related to accidents.

4.5.4 Hybrid Gaussian mixture model and support vector classifier results

After data pre-processing and re-sampling, 120000 data samples with 39.996 samples from Class 1, 39.998 samples from Class 2, and 40,006 samples from Class 3 were used as input to the Gaussian mixture model. The top 25 features according to the variable importance ranking results were chosen as features of the input data entries. The data were fitted with a mixture of three Gaussians which correspond to the three accident severity classes. Moreover, principal component analysis (PCA) was applied to visualize the results in 2 dimensions. The

results with clustering based on the predictions by the Gaussian mixture model are shown in Figure 4.9. The mean matrices obtained for all three classes are also listed in Table4.6.

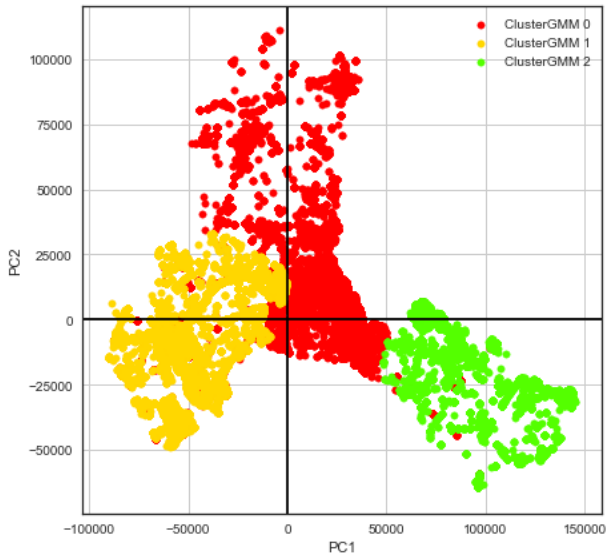

FIGURE 4.9 – Gaussian mixture modelling results

Accident class	Mean matrix
Class 1	[2.50 , 16247.158 , 51.35 , -0.53 , 50264.661 , 381.98 , 460.20 , 3.60 , 4.03 , 44.59 , 48.21 , 3.11 ,
Class 2	[2.43 , 11990.362 , 50.98 , -1.17 , 45831.188 , 400.35 , 497.09 , 3.61 , 3.12 , 44.00 , 55.88 , 3.40 ,
Class 3	[2.59 , 15912.665 , 51.30 , 0.74 , 59138.743 , 234.60 , 537.71 , 2.85 , 7.29 , 46.00 , 90.77 , 2.94 ,

TABLE 4.6 – Mean vectors using Gaussian mixture model

These results suggest that for a particular class of accident severity, the mean vector is the average value of the observed features. To interpret these results, let us consider an example : if a person is driving at a longitude of -0.527 and has a speed of nearly 48.2 km/hr, then he is very likely to have an accident of class 1, i.e. a no-injury accident. On the other hand, a person travelling at a longitude of 0.743 with a speed of 90.7 km/hr has very high chance of having an accident of class 3, i.e. an incapacitating injury. Similarly with respect to the day of the week, the mean value observed rounded to the nearest integer for classes 1, 2 and 3 are 4, 3 and 7 respectively. This implies that accidents of the highest severity or fatal accidents are likely to happen on Saturdays. One possible explanation for this could be that on Saturdays, there are more cars on the road and more drivers are impaired by alcohol due to weekend celebrations and parties than on any other day. These results are fairly consistent with one's

personal experience and logical reasoning.

It is also observed that some values in the mean vectors of classes 2 and 3 are very close to each other. The reason for this is the uneven distribution of the values inside every feature. For example, the feature, 'Lighting' has 126.049 data points corresponding to daylight while only 1123 and 10314 for darkness-light unlit and darkness -no lighting respectively. Thus, daylight itself accounts for 70.45% of the example set which is a very high number. This sub-skewing of data leads to biasing in favour of the majority class.

The overall accuracy of the Gaussian mixture model was 64.68%.

These 3 mean vectors were used as input to the support vector classifier. For such a large dataset with 120.000 examples, 3 data points for training would be insufficient and would lead to overfitting. As a result, we used some extra data points alongside the mean vectors for the purpose of training the support vector classifier and further decreasing the testing data. Had there been more classes of accident severity in the dataset, one could have directly used the mean vectors as input for the SVC and hence improve the model performance like the technique applied in text independent speaker identification using both SVM and GMM. [140].

The SVC with radial basis function produced a total accuracy of 84.35%. Precision, recall and the F1-score were calculated to quantify the performance of our classifier. In order to compute these parameters, 4 performance metrics given below are evaluated from the confusion metric :

— True Positives (TP) - These are the examples with 'yes' as their actual class as well as the class predicted by the model.
— True Negatives (TN) - These are the examples with 'no' as their actual class as well as the class predicted by the model.
— False Positives (FP) – These are the examples with 'no' as their actual class, but are predicted as 'yes' by the model.
— False Negatives (FN) – These are the examples with 'yes' as their actual class, but are predicted as 'no' by the model.

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

TABLE 4.7 – Confusion metric

Subsequently, the performance estimation parameters are defined as :

— Accuracy (A) : This is defined as the ratio of the number of correctly predicted examples over the total number of examples. Mathematically it is, $\frac{TP+TN}{TP+FP+FN+TN}$

— Precision (P) : This is defined as the the ratio of correctly predicted positive observations to the total predicted positive observations. Mathematically it is, $\frac{TP}{TP+FP}$
  — Macro precision : Precision found by calculating metrics for each label, and then finding their un-weighted mean.
  — Micro precision : Precision found by calculating metrics globally by counting the total true positives, false negatives and false positives

— Recall (Sensitivity) (R) : Recall is the ratio of correctly predicted positive observations to the all positive observations in actual class. Mathematically it is, $\frac{TP}{TP+FN}$
  — Macro Recall : Recall found by calculating metrics for each label, and then finding their un-weighted mean.
  — Micro Recall : Recall found by calculating metrics globally by counting the total true positives, false negatives and false positives.

— F1 score : F1 Score is the weighted average of Precision and Recall and is used to combine precision and recall in a single metric. Mathematically it is, $2 * \frac{P*R}{P+R}$

The performance scores including precision, recall and F1-score are listed below in 4.8. The radial basis function or the RBF achieved an accuracy of 88.52%, outperforming the linear kernel, which was 59.89% accurate.

Figure 4.10 shows the confusion matrix obtained with the hybrid model. As can be observed, there is a clear separation between accidents without any injury (Class 1) and accidents

Class	Precision	Recall	f1-score
Class 1	1.00	0.9342	0.9659
Class 2	1.00	0.7214	0.8381
Class 3	0.7437	1.00	0.8530
Macro average	91.4595%	88.5220%	88.5748%
Micro average	88.5166%	88.5166%	87.2895%
Weighted average	91.4588%	88.5166%	88.5711%
AUC	0.99	0.97	0.97
Accuracy		88.5167%	

TABLE 4.8 – RBF kernel performance metrics

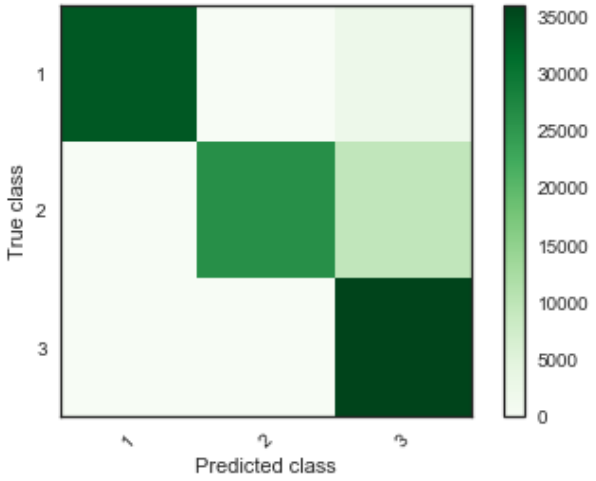

FIGURE 4.10 – Confusion matrix for accident dataset

with injury (Class 2 and 3). Most of the confusion occurs between non-incapacitating injury accident and incapacitating injury accidents.

Furthermore, the ROC (Receiver Operating Characteristics) curve and AUC/AUROC (Area Under the Receiver Operating Characteristics) were determined using the above parameters. ROC is a probability curve with TPR (y) plotted against the FPR (x) which is FP/TN+FP. The area under the ROC curve quantifies the model's ability to identify the classes correctly and distinguish between them [122].

The AUC-ROC curve for this model is shown in Figure 4.11. The AUC values for Classes 1, 2 and 3 are 0.99, 0.97 and 0.97 respectively. These values are very close to 1 and reflect the good discriminative power of the classifier.

Class	Precision	Recall	f1-score
Class 0	0.71	0.89	0.78
Class 1	0.69	0.56	0.62
Class 2	0.73	0.57	0.64

TABLE 4.9 – Linear kernel performance metrics

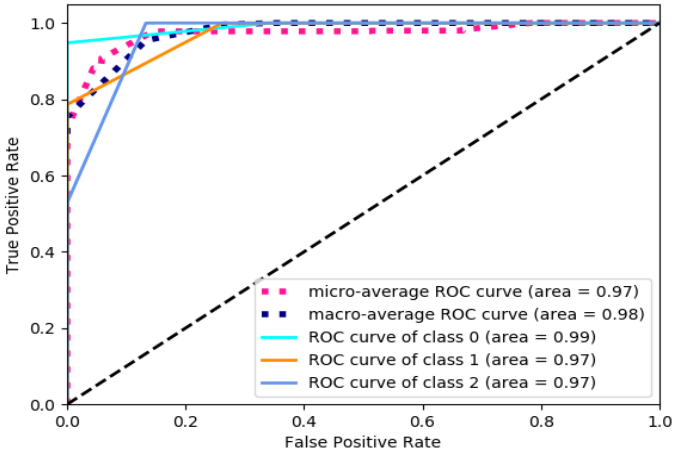

FIGURE 4.11 – ROC curve for the accident dataset

Similarly to AUC-ROC, an area under precision/recall curve (AUC-PR) can also be calculated to show the tradeoff between precision and recall as a function of varying a decision threshold. The higher the area under the curve is, the higher are the values of precision and recall, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate [151]. For the hybrid model, the AUC-PR curve micro averaged over all classes is shown in Figure 4.12. The AUC-PR curves for each class represented over the iso-F1 curves are plotted in Figure 5.2 where an iso-F1 curve is a curve containing all the points in the precision-recall space whose F1 scores are the same.

4.6 Conclusion

Road traffic accidents have become a major cause of injury and death. With increasing urbanization and populations, the volume of vehicles has increased exponentially. As a result, traffic accident forecasting and the identification of accident prone areas can help reduce the risks of traffic accidents and improve overall life expectancy.

The data about circumstances of personal injury in road accidents, the types of vehicles

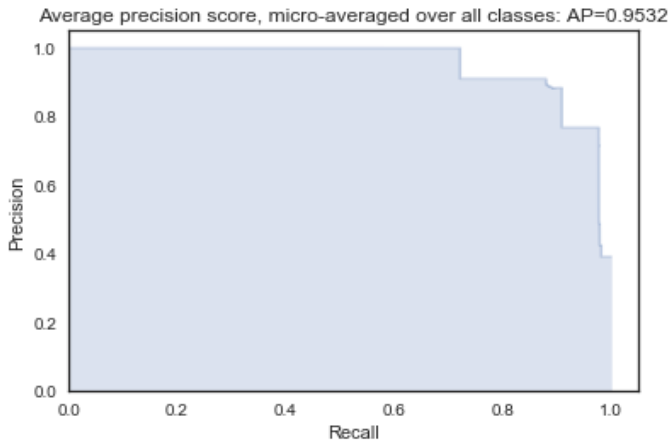

FIGURE 4.12 – AUC-PR curve micro averaged over all classes for the accident dataset

involved and the consequential casualties were obtained from data.govt.uk [136]. The output or the accident severity class was divided into three major categories namely : no injury in the accident, non-incapacitating injury in the accident and an incapacitating injury in the accident. In this chapter, a hybrid classifier was proposed which combined the descriptive strength of the baseline Gaussian mixture model (GMM) with the high performance classification capabilities of the support vector classifier (SVC). A new approach was introduced using the mean vectors obtained from the GMM model as input to the SVC. The model was supported with data pre-processing and re-sampling to convert the data points into a suitable form and avoid any kind of biasing in the results. Feature importance ranking was also performed to choose relevant attributes with respect to accident severity. This hybrid model successfully took advantage of both the models and obtained a better accuracy than the baseline GMM model. The radial basis kernel outperformed the linear kernel by achieving an accuracy of 85.53%. Data analytics performed including the area under the receiver operating characteristics curve (AUC-ROC) and the area under the precision/recall curve (AUC-PR) indicates the successful application of this model in traffic accident forecasting.

Although a significant improvement in accuracy has been observed, this study has several limitations. The first concerns the dataset used. This research is based on a road traffic accident dataset from the year 2017 which contains very few data samples for the no injury and non-incapacitating injury types of accident. The data was unbalanced not just with respect to the output class but also with respect to the sub features of various attributes. Moreover,

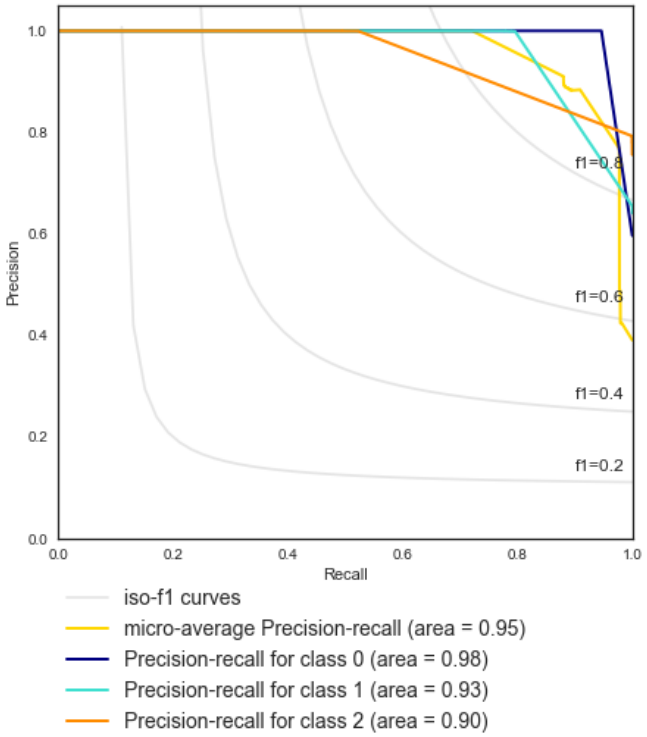

FIGURE 4.13 – AUC-PR curves for each class on various ISO-F1 curves for the accident dataset

aggregating the accident severity into just three categories limits the scope of the study and the results obtained. The greater the number of severity classes, the less is the amount of extra training data required to feed in the SVC to avoid overfitting. Thus, datasets with sufficient records corresponding to each class are desirable and should be used for further study.

The second limitation concerns the dependence of the SVC model on parameters and attribute selection. In this study, the performance of SVC relies heavily on the feature selection results and the mean vectors obtained from the GMM. In order to improve the accuracy of the support vector classifier, other approaches like particle swarm optimization (PSO), ant colony optimization, genetic algorithms, etc. could be used for effective parameter selection. In addition to this, more kernels like the polynomial kernel and the sigmoid kernel could be tested in order to improve future model performances.

Chapitre 5

Graph Based Subjective Matching of Trusted Strings and Blockchain Based Filtering for Connected Vehicles

Contenu

5.1 Introduction

Increased safety and a higher level of comfort have led to efforts to adapt conventional vehicular access network to the world of connected vehicles. Considering the wide spectrum of connected vehicles, which can communicate in five different methodologies : Vehicle to Vehicle (V2V), Vehicle to Infrastructure (V2I), Vehicle to Cyclist (V2C), Vehicle to Pedestrian (V2P) and Vehicle to Everything (V2X), it is worth consolidating the cooperative safety and collected mobility management from different distributed devices. However, the prime

objective of connected vehicles is not only to impose security and trust measures for individual vehicles, instead the strategy of connected vehicles should concentrate the cooperative and collective environment on a fleet of vehicles. Therefore, keeping simple authentication and access control may not be efficient to evaluate trust and assurance for all the distributed stake holders. As trust is an important entity for this entire system, hence a strategy for trust evaluation also becomes crucial. There are many instances in distributed systems, where trust for multiple parties may not follow the same benchmark for transmission and reception of messages. This phenomenon could be more prominent, when distributed users carry different mobile edge oriented devices and media. For each of those devices, the transmission and reception strategies with protocols may be different. For example, a text message sent to the mobile device through social media might not be the same as sending the same message through mailing or through other types of online media communication. These observations raise some challenges to synchronize distributed mobile edge devices and media from eavesdropping and intended spam injection procedures. To establish trust for a distributed system, the system should be able to establish trust from distributed users. The procedure follows a consensus mechanism approach for the appropriate matching of trusted entities. DSRC has become mandatory since 2016 for light vehicles and this rule describes a defined data packet with a Basic Safety Message (BSM) indicating the location of the vehicle, its speed and other on-road parameters. However, DSRC is unable to specify transmitted and received messages with respect to a trusted classification. Therefore, this chapter proposes a unique method to investigate the optimal trusted matching for incoming messages between connected vehicles. Interestingly, this work does not consider the key word matching (word by word or dictionary based approach). Instead, the broader thematic content and headings for communicated messages are taken into account. This will help to fetch the content categories for different untrustworthy behaviors like abusive behavior, forced branding of products, misleading information, blocking of safety message on road, etc. In order to achieve this proposed matching under distributed mobile devices, this chapter introduces a message passing procedure followed by a blockchain-based reinforcement decision. Thus, this study comprises two parts, where the first part describes content based message passing. The second part, after matching the content and subject headings, consolidates the distributed consensus or

voting mechanism for any decision with respect to the trust evaluation.

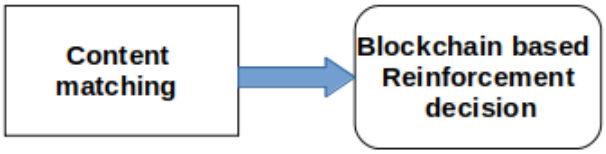

FIGURE 5.1 – System diagram

The key contributions of this research work is summarized below :

— We propose a message-passing scheme under connected vehicles. In this scheme, we do not consider the key word matching (word by word or dictionary based approach). Instead, we take into account the broader thematic content and headings for communicated messages.

— We attempt to improve the trust evaluation by using the voting mechanism for any decision, which is a concept called blockchain-based reinforcement decision.

— We aim to enhance the securing and authentication of messages exchanged among vehicles by introducing the concept of content matching and trust evaluation in connected cars blockchain as a future perspective.

5.2 Related work

Connected-vehicle based applications use both unicast and broadcast communications. However, as for all mobile and wireless networks, these communication scenarios suffer from various security issues that hinder the functionality of such communication protocols. Existing trust-based security solutions are usually classified into entity-based, data-based, and hybrid trust models depending on the revocation target, which can be dishonest entities, malicious messages, or both [152]. In addition, for a message passing protocol in a VANET, especially between connected cars, blockchain is seen as the most promising technique to provide a secured distributed network among different frameworks [153]. In the following, we survey message content matching procedures for connected cars as well as providing background details on blockchain technology for secure message dissemination using a voting mechanism.

5.2.1 Message content matching

In general, string matching has been explored by researchers using different techniques. A technique for detecting phishing attacks was proposed by the authors in [154]. As an objective of that study, this technique was meant to specify the similarity grade between a URL with blacklisted URLs. Consequently, it can be classified as phishing or non-phishing based on the textual properties of a URL. In this work, a well-known string matching algorithm, the so-called Longest Common Subsequence (LCS), was implemented by the authors in the hostname for comparison. With an accuracy found to be 99.1%, it is regarded as very efficient in detecting phishing attacks. It also achieved very low false positive and false negative rates. Similarly, the same algorithm was used in [155]. The authors used it on biological files to discover sequence resemblance between genetic codes. In this test carried on a sequence of DNA that was generated randomly, the accurate DNA sequence similarity was found by the algorithm. This comparison is a path to implement codes of genetics from one DNA sequence to another. When the algorithm was tested on 50 samples with two input DNA genetic code sequences, it performed well and produced good results.

The authors in [156] carried out an investigation on the use of string matching algorithms for spam email detection. Particularly, their work examined and compared the efficiency of six well-known string matching algorithms, namely Longest Common Subsequence (LCS), Levenshtein Distance (LD), Jaro, Jaro-Winkler, Bi-gram, and term frequency inverse document frequency (TFIDF) on two datasets, which are the Enron corpus and the CSDMC2010 spam dataset. After observations based on the performance of each algorithm, they found that the Bi-gram algorithm performed best in spam detection on both datasets. While the authors claim that all six methods have shown good results in terms of efficiency, they suffer from time performance.

A Levenshtein distance algorithm has been used by K. Beijering et al. in [157]. They used it to calculate phonetic distances between 17 Scandinavian language varieties and standard Danish. When comparing phonetic transcriptions of two pronunciations, the Levenshtein distance is defined as the number of procedures necessary to convert one transcription to another. The strength of the Levenshtein distance consists in minimising the overall number

of string operations when converting one pronunciation to another.

5.2.2 Blockchain technology

A blockchain can be defined as a growing list of records, called blocks, that are linked using cryptography. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data. In other words, it is a distributed and decentralized public database of all transactions or digital events that have been accomplished or shared among participating nodes. Each event in the public database is validated based on the agreement of a large number of nodes in the blockchain network. The popularity of the blockchain is due to its advantages, which include decentralization, anonymity, chronological order of data, distributed security, transparency, immutability and suitability for trust-less environments [158].

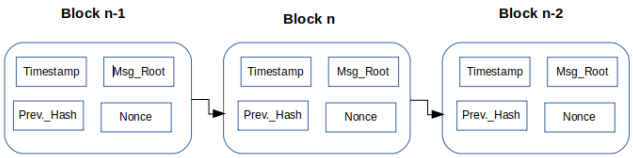

FIGURE 5.2 – The structure of blocks in a blockchain

The blockchain consists of two types of nodes. A full node is a node that stores and maintains the complete history of blockchain transactions. It begins a transaction directly and independently, and it authoritatively verifies all the network transactions. Every node in the blockchain network knows the genesis block's hash. Every node in the network builds a trusted blockchain based on the genesis block that acts as a secure root. The genesis block does not have the hash of a previous block. If a node is new, then it only knows the genesis block, and it will have to download all blocks starting from the genesis block to synchronize with the blockchain network and is constantly updated when new blocks are found. The chaining of blocks is performed by appending hashes of the previous blocks to the current block so that the hash of the current block is in a sequential manner to the following block. Then, it is shared with other nodes in a distributed P2P network in a secure way without the need for a central authority. The sequential hashes of blocks ensure a sequential order of transactions. Then, previous transactions cannot be modified without modifying their blocks and all subsequent blocks. The blockchain is verified by the consensus of anonymous nodes

in the generation of blocks. It is considered secure if the aggregated computational power of malicious nodes is not larger than the computational power of honest nodes. In the case of Bitcoin, the concept of Proof of Work (PoW) makes sure that a miner is not manipulating the network to make fake blocks. A PoW is a mathematical puzzle that is very hard to solve and easy to verify so that it protects the blockchain from double-spending attacks. In the research on VANETs, some previous studies related to secure event message dissemination are based on voting. Most voting approaches attempt to solve the issues of node security by asking for the opinions of other nodes to determine the trustworthiness of a particular node.

However, this type of approach has the problem of whether the nodes providing the feedback can be trusted. Generally speaking, limited work has been done to study connected vehicles using the blockchain. The authors in [159] used a basic blockchain concept to simplify the distributed key management in heterogeneous vehicular networks. The authors in [160] combined the VANET and Ethereum's blockchain-based application concepts and enabled a transparent, self-managed and decentralized system. They used Ethereum's smart contract system to run all the types of applications on an Ethereum blockchain.

In contrast, our proposed work applies a different type of blockchain for secure message dissemination for connected cars. In [152], the authors proposed a blockchain technology for automotive security by using an overlay network in the blockchain and additional nodes called overlay block managers. The overlay network nodes are clustered by cluster heads, and these cluster heads are accountable for handling the blockchain and operating its main functions. However, the introduction of additional overlay nodes might cause high latency and could be the center point of failure if the cluster head is compromised.

5.3 Architecture of mobile edge search process

In the architecture of the mobile edge entity search process, the mobile edge entity initiates the handshaking by specifying the sensor observation sequence to be queried by the terminal, and sends the search request to the mobile edge computing (MEC) server. In return to that request, the cloud server is responsible for responding to the user's search request, and publishing the search request to the MEC server according to the searched content. The

MEC server is responsible for fitting the raw data uploaded by the sensor and calculating its similarity with the search conditions published by the cloud server. The sensor layer is responsible for collecting environmental data and uploading it to the MEC server. Figure5.3 shows the mobile edge entity search process. The steps are as follows :

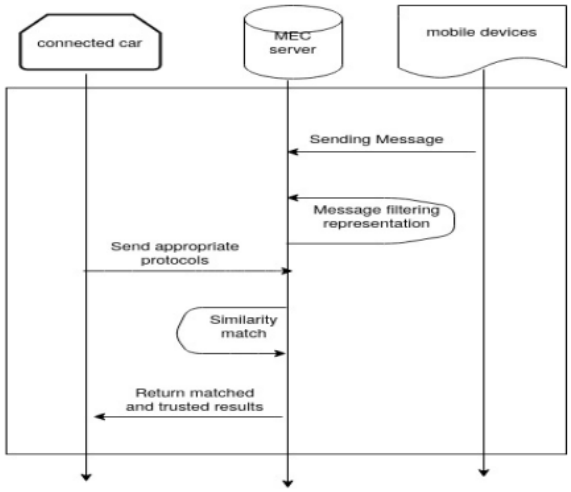

FIGURE 5.3 – Mobile Edge Entity Search Process

1. The mobile device reports the environmental message observed to the MEC server.
2. The MEC server fits the reported message of the mobile device, and stores the processed message.
3. The connected car sends a request for an appropriate protocol to the MEC server.
4. After receiving the request for an appropriate protocol, the MEC server computes the similarity between the search condition and the mobile device message stored internally.
5. Finally, the MEC server returns matched and trusted results that match with the connected car's request to the connected cars.

5.3.1 Graph Representation of Connected Vehicles

A graph is a structure amounting to a set of objects in which some pairs of the objects are in some sense "related". The objects correspond to mathematical abstractions called vertices (also called nodes or points) and each of the related pairs of vertices is called an edge (also called link or line). Figure5.4 illustrates a graph representation of vehicles. The nodes (cars) of the graph are in topological order. For instance in Figure (4b) we have 1, 4, 6, 5, 2, 3, 7

(visual top-to-bottom, left-to-right) or 3, 1, 5, 2, 4 (arbitrary) in Figure(4a). Each car has an identification number (ID).

5.4 The proposed Solution

The proposed solution of this research will perform trust enhancement among communicating nodes of connected vehicles. The nodes in the graph representation are in topological order. The components are content matching under thematic matching operations reinforced by a graph-based blockchain mechanism, as in Figure5.4.

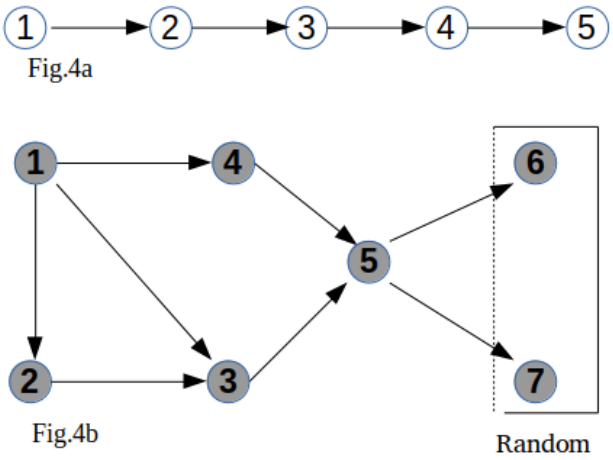

FIGURE 5.4 – Graph representation of connected cars

The following themes and content are included in the model proposed :
a) Exhaustive themes (dangerous products, adult content, gambling and games, inappropriate messaging, personalized promotions, forced promotion)
b) Non-exhaustive (affiliating the message against the program rules, promoting the same content from multiple accounts, trying repeatedly to push brand promotion, brand disinvestment, intentional and manipulation to switch the messages towards inappropriate content).
Under these two heads or leads, the service provider of the connected car can clearly differentiate the two types of content and their thematic message strings. The dictionary is not subjected to one-to-one mapping but it defines lexical matching either in the message head (a) or in the message head (b). This is respective of any theme or content which maybe outside these message heads. This constraint may be a limitation for this model.

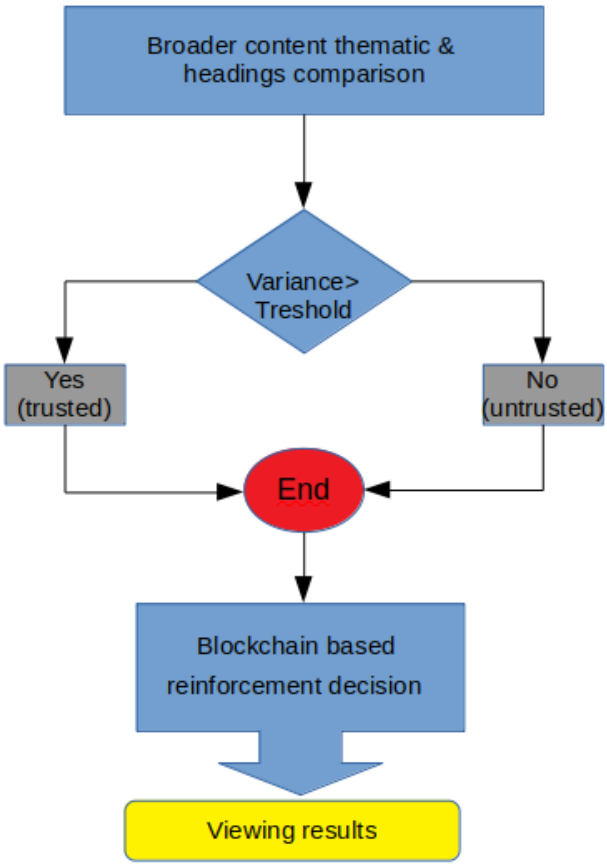

FIGURE 5.5 – Flowchart of the proposed Solution

Data elements	0.6, 1.2, 1.8, 2.4, 3.0, 3.6, 4.2, 4.8, 6.4, 6.0
Mean	3.3
Max	6.0
Variance	2.97

TABLE 5.1 – Summary of statistical parameters and values.

5.4.1 Function Matching Trust

The function of matching-trust is described below : (I, S_i , d_I , d_{Si} , d_{min}, β)

Input :

- I is the identifier of the priority string (”xxxx”) on the trust graph edge.

- S_i is the string identifier of the moving car transmitting D_s

- d_I represents the distance I to the terminating node in case $I \neq$None (availability steady but trusted).

- d_{Si} represents the distance I to the terminating node in case $S_i \neq$ None (not trusted)

- $d_min > 0$ minimum distance of connected cars to perform D_s

- $\beta > 1$ co efficient to transmit the target string

Output :

if ($I \neq None$ & $S_i \neq None$ & $d_I > \beta.d_{min}$ & $d_Si < d_{min}$) or ($I == None$ & $S_i \neq None$ & $d_Si < d_min$)

then match string I

else

terminate

return Match string I untrusted.

5.5 Experimentation and Results Analysis

5.5.1 Assumptions

— All connected car members are under the same network service provider on their edge devices.

— Out of the total number of members registered in the network, only the agreement of older members (>1 year) could be considered.

— To avoid the physical consensus, the proposed blockchain prototype will deploy a graph-based referencing. It implies that based on the subjective terms of untrusted message leads, the service providers will predefine a maximum high-positive mutual agreement of trusted messages. This typical graph-driven direction will help to prevent latency and delay for the reply of the message block through participants and it also avoids self-biasing to manipulate consensus, if some groups of participants are known to the victim of untrusted acts.

Intersections in Figure5.6 conceptually define that it is the association between certain immediate past values of a car with ID transmitted in the message to the neighbors. This includes the values of the car ID at present participating in the message transmission. In Figure5.6, the red dotted lines indicate the length of the graph formed either by the transmitting car or its affected neighbors. Therefore, they are not trusted. However, distinctly in

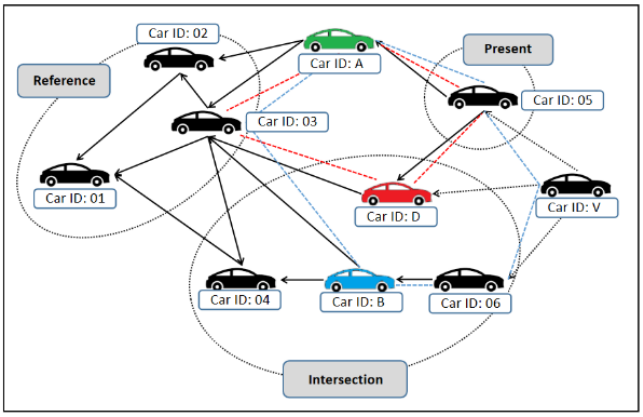

FIGURE 5.6 – Graph based referencing

this cluster, blue dots represent safe messages, where one of the affected cars is placed in the same cluster. In this context, graph referencing is used to investigate the variance and the degree of trust distorted by the odd entry to that cluster.

5.5.1.1 Graph-based referencing towards trusted consensus

The concept here is designed to estimate the temporal inconsistency (ambiguity) between two messages. If two message-clusters are contradictory to each other, their temporal order cannot be determined. This means that the message clusters might be from isolated connected cars. However, the time discrepancy of two message-clusters is bounded by their nearest common ancestor and nearest common descendant, since the real creation time of a message-block is bounded by its ancestors and descendants. The untrusted message-clusters always intend to hide or counterfeit their real creation time in order to carry out spam message generation such as repeat occupancy (conventionally known as double spending). Therefore, the consensus agreement between the untrusted message block and the most trusted message-clusters should be very large, otherwise the real creation time of the distrusted block would be bounded by some trusted message-clusters into a small interval. On the other hand, the agreement of two trusted message-clusters is normally much smaller. If the links between message-clusters are not artificially manipulated, the agreement of two message-clusters should only depend on the network propagation speed and the block creation rate. When the network propagation speed or the block creation rate increases, the time discrepancy between the nearest

common ancestor and the nearest common descendant will decline. However, the length of the shortest path between two message-clusters will increase and cancel out the decline in time discrepancy to some extent. Therefore, the agreements are not very sensitive to the network propagation speed and the block creation rate. The analysis demonstrates that the agreements between two trusted message-clusters are mostly smaller than 10 while the agreements between the trusted block and the distrusted block might be higher by two or more orders of magnitude. Figure5.7 shows the relationship between the block creation rate and the maximum agreement between trusted blocks. In each case of the block creation rate, 16 simulations were conducted. The statistical analysis is shown in Table5.1 Even when the message block creation rate reaches 6 message-clusters per second, the agreement between trusted message-clusters still does not increase too greatly. Therefore, the agreements can be utilized to filter the suspect distrusted blocks. In this section, we give a proposed framework named MsgBlock Filter for identifying the trusted message-clusters based on the agreement. Given a block DAG, we first calculate the agreements for every pair of message-clusters and get the agreement of reference matrix. Then, the agreement matrix is converted into a binary matrix where each element is 1 if the corresponding element in the agreement matrix is larger than a preset threshold d, and 0 otherwise. By using the binary matrix obtained as the adjacency matrix, we can construct an undirected graph, in which each vertex represents a block. This graph is called the d-agreement graph of the given block DAG. Intuitively, if a block DAG only contains trusted blocks, the degrees in the d-agreement graph will be very small since the agreements between most trusted message-clusters are zero and the remaining non-zero agreements are also very small. Considering the trusted message-clusters to be the majority, the trusted block identification problem can be addressed by identifying the maximum subset of vertexes with small degrees. Considering a graph $G = (V, E)$, the k-independent set of G refers to the vertex subset V' in which the maximum degree in the induced sub-graph does not exceed k. The maximum k-independent set problem is to find the k-independent set with maximum size which is a generalization of classical maximum independent set problem. The maximum k-independent set can be formulated as the following integer programming, in which **xs** represents whether a certain vertex s is selected and **aij** denotes the element of adjacency matrix of the graph G.

5.5.1.2 Direct acyclic graph

Considering a Direct Acyclic Graph (DAG), it is worth formulating some statistical analysis with respect to message spreading strength (including trusted and untrusted messages) precision and recall. However, due to the legacy of the consensus protocol it becomes more stringent to model the same for different participants in a connected car environment. The concept for finding the trusted messages and the untrusted or distrusted messages is to find the interval graph from the first cycle of the message repeat, although the graph here is referred to as an acyclic graph as no cycle exists for the repetition of the message. Therefore, the only measure to identify the interval of the message is to find out the variance of the message repeat from one node to another in terms of time. Here, we calculate primarily three values for a given message creation rate (Msg_block/second) that is 0.6, 1.2, 1.8, 2.4, 3.0, 3.6, 4.2, 4.8, 5.4, 6.0 respectively. Under these message creation rates, we find the mean to be 3.3, the max to be 6 and the variance to be 2.97. The different steps to calculate the mean, the max and the variance are as follows.

Specific points : the variance of any dynamic quantity is the sum of the square difference between each data point and the mean divided by the data value. Hence sigma squared should be the sum of the squared difference divided by the total number of items in the given problem. This variance will help to trace the closeness of trusted and untrusted blocks, assuming that the untrusted message must be repeated more than once.

- Step 1 : we find the mean of the dataset
- Step 2 : we add all the data values divided by the sample size : $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Step 3 : we find the sum of the squared difference : $SS = \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Step 4 : we calculate variance of sigma squared accordingly : $\sigma^2 = \overline{x} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$

In order to identify and grab the described concept of this process, we refer to Figure5.7 which shows the maximum number of honest messages versus the message generation rate. Here, the variance gives the idea that the density of trusted messages in ideal conditions is always higher. Therefore, even when the message block creation rate reaches 6 message clusters/second, the variance between trusted messages and clusters become 2.97.

Figure5.7 also indicates that the trust level agreement cannot differ too much with respect to untrusted messages. Hence, the intersection could be used as a filter for reference to create

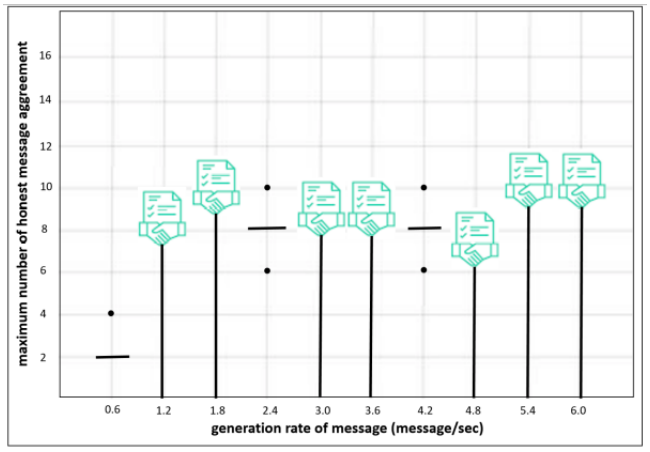

FIGURE 5.7 – Relationship between the block creation rate and the maximum agreement between trusted blocks

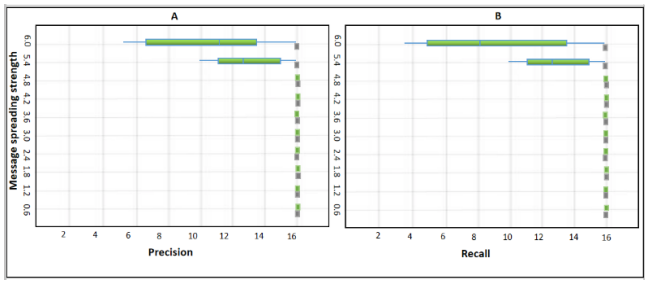

FIGURE 5.8 – Precision and recall

the predefined trusted and untrusted message blocks. Two major technical specifications are considered :

1. The predefined referencing of the service provider can prevent a delay in the legitimate reply to the consensus or group messages.

2. Self-biasing or personal manipulation can also avoided.

Figure5.8 provides an interesting observation with respect to the precision and the recall by which the strength of the damaging messaging can be highlighted. The left-hand side of Figure5.8 is divided almost with the same intervals. Here, we also calculate the quantiles of the given data-set from 0.6 to 6.0 to find the exact interval of the precision and recall of trusted messages (there is no memory or learning in the recall, only topological ordering has been investigated). Statistically, quantiles are cut points, dividing the range of the data sample of the probability distribution into continuous intervals with equal probabilities. Here in Figure5.8,

we started calculating the message repeat strength from the median, first quarterly, third quarterly, first decile, last decile, one percentile as maximum level of 6 Msg_Block/second. The flow is to identify the median towards one percentile which is actually the maximum value of the data sample. The analysis helps to correlate the importance of the variance so that repeat messages and the variance can support it as a consensus filter.

5.6 Conclusion

In this work, a message matching model and the conceptual level of graph referencing blockchain have been proposed. The model can filter the trusted and untrusted messages in connected car scenarios, analogous to a conventional blockchain mechanism. However, as participants proceed with a voting mechanism, unwanted delay and self-biasing can occur in the process. In order to avoid that, a distributed blockchain consensus voting mechanism for any decision taken with respect to trust evaluation is used, making this method more feasible for collective decisions. This work raises more open research issues challenging the blockchain mechanism. This is because the security is questionable due to group and collective decision-making and repeat occupancy of the message. This is equivalent to a double spending attack in a normal blockchain. As a future extension, therefore, a DAG and the descendants can be integrated in the blockchain, consolidating its security and spoofing mechanism.

Chapitre 6

Conclusion and perspectives

To conclude this thesis, we briefly summarize our contributions and outline the perspectives for future work.

6.1 Introduction

The main objective in designing vehicular ad-hoc networks is to improve safety on the roads. VANETs can support a wide range of safety applications through V2V and V2R communications. These applications range from cooperative collision warning, emergency braking, cooperative driving, and accident detection. An efficient and reliable broadcast mechanism is the requirement of most of these applications, if not all, in order to inform neighboring drivers about a dangerous situation in a timely manner, that is, exploring vehicles' positioning, traffic forecasting, and evaluating trust in vehicular contexts.

6.1.1 Evaluation

This thesis explores prediction strategies problem in the context of VANETs where different numbers of vehicles need to broadcast their position and communicate securely together and with the infrastructure to disseminate safety/other messages in the network. However, there are some challenges : first, the use of decentralized vehicle cooperation mechanisms and vehicle positioning systems that provide the vehicles locations in the vehicular ad-hoc network. Second, the question of how to utilize eco-autonomous driving strategies with static/dynamic data sources to forecast road accidents, still needs to addressed. And the hybrid

trustworthiness evaluation based on centralized and distributed cooperation needs to be explored. Many researchers are investigating how to deal with these issues in the context of the enhancing safety within VANETs. It is well-known that these problems are difficult. Thus, this thesis aims at proposing and applying a novel framework and techniques to handle the positioning issue in VANETs, forecasting road accidents and evaluating trust.

In the introductory part of this thesis (Chapter 2), we presented and summarized the general aspects of VANETs and pointed out the regulatory aspects worldwide and different projects targeting this area. In Chapter 3, We explored the literature on existing of vehicle positioning systems and solutions within vehicular networks where we first worked on vehicle positioning systems using the Received Signal Strength (RSS) of periodic messages by determining the vehicle location through machine-learning techniques. In this work, we surveyed different strategies to cope with vehicle positioning using RSS of periodic messages and also the transmissions performances within VANETs. We used the reception power to predict a vehicle's position. Moreover, we proposed and adapted four machine learning techniques to the positioning of vehicles namely K-Nearest Neighbors (KNN), Neural Network (NN), Random Forest (RF) and Support Vector Machine (SVM). In addition to that, a simple simulation tool was developed to produce data with the positions of vehicles and the different powers of the messages sent by the vehicles and received at the base stations. And finally, we analyzed and compared the performance of the above mentioned ML with the given data-set. Finally, we studied how machine-learning can be used to predict the transmission performances within the VANETs. More specifically, we explored and presented an approach to computing the probability of the successful reception of a transmission between a vehicle and a Roadside Unit located at a given and known position. Some issues that remain to be investigated and beyond the scope of this thesis were highlighted and they may be the subject of research efforts.

In Chapter 4, we investigated the state-of-the-art of vehicle safety analysis in VANETs while including the issue of road accidents. We found that conventional traffic forecasting techniques use either a Gaussian Mixture Model (GMM) or a Support Vector Classifier (SVC) to model accident features. A GMM on the one hand requires a large amount of data and is computationally inexpensive, whereas SVC, on the other hand, performs well with less data

but is computationally expensive. Therefore in this part, we presented a prediction model that combines the two approaches for the purpose of forecasting traffic accidents. We propose a hybrid approach,that incorporates the advantages of both the generative model (GMM) and the discriminant model (SVC). Raw feature samples are divided into three categories : those representing accidents with no injuries, accidents with non-incapacitating injuries and those with incapacitating injuries. GMM is used to model this trimodal distribution, and parameters are obtained using the expectation-maximization (EM) algorithm. Mean vectors of the component Gaussians obtained are adapted and used as input to the SVC model which further improves the prediction accuracy. Experimental results show that the proposed model can significantly improve the performance of accident prediction. Improvements of up to 24% are reported in the accuracy compared to the baseline statistical model (GMM).

Finally, in Chapter 5 we proposed the message passing scheme for VANETs : an adaptive strategy for a group-based system. We proposed an efficient framework for message content matching that is entitled Graph Based Subjective Matching of Trusted Strings and Blockchain Based Filtering for Connected Vehicles. It is based on the assumption of a hierarchical grouping structure within the network based on vehicles, RSUs and the infrastructure, such as assuming that all connected car members are under the same network service provider on their edge devices, and that only those members whose membership has lasted more than one year could be considered. We proposed improving secure message passing dissemination by using blockchain prototype that will deploy a graph-based referencing in order to avoid the physical consensus. This implies that based on the subjective terms of untrusted message leads, service providers will predefine a maximum high-positive mutual agreement of trusted messages which will help to prevent latency and delay on the reply of the message block through the participants, and it also avoids self-biasing to manipulate consensus, if some groups of participants are known to the victim of untrusted acts. Simulation scenarios were carried out showing the advantages of the proposed message string matching framework.

6.1.2 Perspectives

The exploration of prediction strategies in using ML techniques proposed in this manuscript have presented an advanced method for vehicle positioning and road traffic forecasting in VANETs. Traffic forecasting is based on an advanced method of expectation-maximization algorithm. The proposed scheme can still be enhanced to cover untackled issues in mid-term and long-term future research.

6.1.3 Mid-term perspectives

Research being a continuous process leading to further improvements and innovation, there is still room for enhancement of our proposed methods. So in the near future, several future research directions could be undertaken in order to enhance these schemes, enabling them to perform better in various realistic scenarios.

Regarding the positioning issue discussed in Chapter 3, we have used some machine-learning techniques for vehicle positioning and the probability of successful transmissions in a VANET using a CSMA access scheme such as IEEE 802.11p. It might be worthwhile considering other future work scenarios including other machine-learning schemes over the proposed techniques and investigating it from different angles.

The prediction of the SVM technique has given very good results when using a database built with an analytical model. Moreover, after observation, we noticed that even with errors, the prediction of the SVM technique is still very good. The SVM technique can be used to build a dynamic rate control algorithm that optimizes the VANET's average throughput. Although this initial study is very encouraging, further investigations such as introducing more errors, changing the fading law, using real figures for the database, etc. should be carried out to confirm our first results. Moreover, other ML techniques such as Random Forest, K Nearest Neighbor (KNN), etc. could be used to optimize the transmission in VANETs. This study and a careful comparison of these techniques could be explored.

6.1.4 Long-term perspectives

In the long run, and to enhance security for VANETs, it would be interesting to investigate the following topics, as many open issues still need to be investigated, such as :

Use of different algorithms

— Regarding accident forecasting, while there is already a significant improvement in terms of accuracy, in order to extend this study, different algorithms could be used for effective parameter selection along with different kinds of kernels to enhance the model's performance.

Data context trust management and evaluation

— The main objective of VANETs is to provide safety and cooperative driving information as well as comfort to passengers. To do so, a VANET needs to provide information to drivers and vehicles, and the exchanged information needs to be secured, checked and verified in order to avoid any kind of security issue. Therefore, it is important as a continuation of this work to investigate data-centric trust management evaluation.

Consolidating security with blockchain

— In Chapter 5, with the proposed conceptual model, the security issue is still questionable due to group and collective decision-making and repeat occupancy of the message. Exploring security challenges by using Direct Acyclic Graphs along with Blockchain could be a means to consolidate security.

Bibliographie

[1] "V2v safety technology now standard on cadillac cts sedans," 2017. [En ligne]. Disponible : https://media.cadillac.com/media/us/en/cadillac/photos.detail.html/content/Pages/news/us/en/2017/mar/0309-v2v/_jcr_content/rightpar/galleryphotogrid.htm

[2] Y. BOUCHAALA, "Handling safety messages in vehicular ad-hoc networks (vanets)," *PhD thesis of the University Paris-Saclay prepared at the University of Versailles IT and Communications Sciences and Technologies PhD speciality*, 2017.

[3] A. Sari *et al.*, "Review of the security issues in vehicular ad hoc networks (vanet)," *International Journal of Communications, Network and System Sciences*, vol. 8, n°. 13, p. 552, 2015.

[4] J.-P. Hubaux, S. Capkun et J. Luo, "The security and privacy of smart vehicles," *IEEE Security & Privacy*, vol. 2, n°. 3, p. 49–55, 2004.

[5] O. N. I. de la Sécurité Routière– Bilan des chiffres définitifs 2020, "Accidentalité routière 2020 données définitives„" *https ://www.onisr.securite-routiere.gouv.fr*, 2021.

[6] N. H. T. S. Administration *et al.*, "Overview of motor vehicle crashes in 2019," *US Department of Transportation : Washington, DC, USA*, 2020.

[7] Z. Shafiq, M. H. Zafar et A. B. Qazi, "Qos in vehicular ad hoc networks–a survey," *Journal of Information Communication Technologies and Robotic Applications*, p. 48–58, 2018.

[8] R. Kumar et M. Dave, "Mobile agent as an approach to improve qos in vehicular ad hoc network," *arXiv preprint arXiv :1108.2095*, 2011.

[9] ETSI, "Etsi en 302 637-2 intelligent transport systems (its) - vehicular communications - basic set of applications - part 2 : Specification of cooperative awareness basic service,"

ETSI EN 302 637-2, vol. 1, 2014.

[10] E. E. . 637-2, "Etsi en 302 637-2, "intelligent transport systems (its) - vehicular communications - basic set of applications - part 3 :specifications of decentralized environmental notification basic service," vol. 2, 2014.

[11] M. Jeong, B. C. Ko et J.-Y. Nam, "Early detection of sudden pedestrian crossing for safe driving during summer nights," *IEEE transactions on circuits and systems for video technology*, vol. 27, n°. 6, p. 1368–1380, 2016.

[12] M. Aghagholizadeh et N. Catbas, "Comparative analysis and evaluation of two prestressed girder bridges," *arXiv preprint arXiv :1907.13014*, 2019.

[13] S. Eichler, "Performance evaluation of the ieee 802.11 p wave communication standard," dans *2007 IEEE 66th Vehicular Technology Conference*. IEEE, 2007, p. 2199–2203.

[14] J. Guerrero-Ibáñez, C. Flores-Cortés et S. Zeadally, "Vehicular ad-hoc networks (vanets) : architecture, protocols and applications," dans *Next-generation wireless technologies*. Springer, 2013, p. 49–70.

[15] S. Al-Sultan *et al.*, "A comprehensive survey on vehicular ad hoc network," *Journal of network and computer applications*, vol. 37, p. 380–392, 2014.

[16] A. M. Vegni *et al.*, "Smart vehicles, technologies and main applications in vehicular ad hoc networks," *Vehicular Technologies—Deployment and Applications*, p. 3–20, 2013.

[17] W. Liang *et al.*, "Vehicular ad hoc networks : architectures, research issues, methodologies, challenges, and trends," *International Journal of Distributed Sensor Networks*, vol. 11, n°. 8, p. 745303, 2015.

[18] M. S. Sheikh, J. Liang et W. Wang, "A survey of security services, attacks, and applications for vehicular ad hoc networks (vanets)," *Sensors*, vol. 19, n°. 16, p. 3589, 2019.

[19] I. Ali, A. Hassan et F. Li, "Authentication and privacy schemes for vehicular ad hoc networks (vanets) : A survey," *Vehicular Communications*, vol. 16, p. 45–61, 2019.

[20] A. Indra et R. Murali, "Routing protocols for vehicular adhoc networks (vanets) : A review," 2014.

[21] P. K. Pandey, A. Swaroop et V. Kansal, "A concise survey on recent routing protocols for vehicular ad hoc networks (vanets)," dans *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE, 2019, p. 188–193.

[22] E. Talavera, A. D. Álvarez et J. E. Naranjo, "A review of security aspects in vehicular ad-hoc networks," *IEEE Access*, vol. 7, p. 41 981–41 988, 2019.

[23] M. S. Sheikh et J. Liang, "A comprehensive survey on vanet security services in traffic management system," *Wireless Communications and Mobile Computing*, vol. 2019, 2019.

[24] M. N. Mejri, "Securing vehicular networks against denial of service attacks," Thèse de doctorat, Université Sorbonne Paris Cité ; École nationale d'ingénieurs de Tunis (Tunisie), 2016.

[25] G. C. Obasuyi, A. Sari *et al.*, "Security challenges of virtualization hypervisors in virtualized hardware environment," *International Journal of Communications, Network and System Sciences*, vol. 8, n°. 07, p. 260, 2015.

[26] "Fleetnet-internet on the road." [En ligne]. Disponible : http://www.et2.tu-harburg.de/fleetnet.

[27] "Preventive and active safety applications." [En ligne]. Disponible : http://www.prevent-ip.org.

[28] "Safespot (cooperative vehicles and road infrastructure for road safety)." [En ligne]. Disponible : http://www.safespot-eu.org/.

[29] S. Olariu et M. C. Weigle, *Vehicular networks : from theory to practice*. Chapman and Hall/CRC, 2009.

[30] "Comesafety project (communication for esafety)." [En ligne]. Disponible : http://www.comesafety.org.

[31] "Geonet project (geo-addressing and geo-routing for vehicular communications)." [En ligne]. Disponible : http://www.geonet-project.eu.

[32] "Coopers project." [En ligne]. Disponible : http://www.coopers-ip.eu.

[33] "Eurofot eu project." [En ligne]. Disponible : http://www.eurofot-ip.eu.

[34] "Pre-drivec2x eu project." [En ligne]. Disponible : http://www.pre-drive-c2x.eu.

[35] "E-safety vehicle intrusion protected applications (evita)." [En ligne]. Disponible : http://www.evita-project.org.

[36] S. Zeadally *et al.*, "Vehicular ad hoc networks (vanets) : status, results, and challenges," *Telecommunication Systems*, vol. 50, n°. 4, p. 217–241, 2012.

[37] V. Kumar et R. Kumar, "Smart re-route routing protocol for vehicular ad-hoc networks," dans *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, p. 758–763.

[38] "Fcc dsrc (dedicated short range communications)." [En ligne]. Disponible : http://wireless.fcc.gov/services/its/dsrc.

[39] Y. J. Li, "An overview of the dsrc/wave technology," dans *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*. Springer, 2010, p. 544–558.

[40] K. Bilstrup, "Vehicular communication standards-dsrc, calm m5, wave and 802.11 p," dans *SAFER Seminar*, 2009.

[41] Q. Chen, D. Jiang et L. Delgrossi, "Ieee 1609.4 dsrc multi-channel operations and its implications on vehicle safety communications," dans *2009 IEEE vehicular networking conference (VNC)*. IEEE, 2009, p. 1–8.

[42] D. Jiang et L. Delgrossi, "Ieee 802.11 p : Towards an international standard for wireless access in vehicular environments," dans *VTC Spring 2008-IEEE Vehicular Technology Conference*. IEEE, 2008, p. 2036–2040.

[43] I. T. WG16, "Iso 21217 : 2010 : Intelligent transport systems-communications access for land mobiles (calm)-architecture," 2010.

[44] ——, "Iso 21215 : 2010 : Intelligent transport systems-communications access for land mobiles (calm)-m5," 2010.

[45] ——, "Iso 29281 :2011 : Intelligent transport systems-communications access for land mobiles (calm)-non-ip networking, 2011," 2011.

[46] S. Hess *et al.*, "Towards standards for sustainable its in europe," dans *16th ITS World Congress and Exhibition, Stockholm, Sweden*, 2009.

[47] "Etsi tc its." [En ligne]. Disponible : http://www.etsi.org.

[48] "Ieee (institute of electrical and electronics engineers)." [En ligne]. Disponible : http://www.ieee.org/index.html.

[49] "Ietf (internet engineering task force)." [En ligne]. Disponible : http://www.ietf.org.

[50] "Iso (international standard organization)." [En ligne]. Disponible : http://www.iso.org/iso.

[51] "Internet its consortium." [En ligne]. Disponible : http://www.internetits.org.

[52] "Tta." [En ligne]. Disponible : http://www.tta.or.kr/eng/contents.do?key=161

[53] M. Abu-Rgheff, G. Abdalla et S. Senouci, "Softmac : Space-orthogonal frequency-time medium access control for vanet," dans *IEEE Global Information Infrastructure Symposium (GIIS)*, 2009, p. 1–8.

[54] I. C. S. L. S. Committee *et al.*, "Ieee standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements part 11 : Wireless lan medium access control (mac) and physical layer (phy) specifications," *IEEE Std 802.11ˆ*, 2007.

[55] *Task Group p. IEEE 802.11p*, Wireless Access in Vehicular Environments (WAVE) Draft Standard, 2007.

[56] ETSI EN 302 637-2, "Intelligent Transport Systems (ITS) - Vehicular Communications - Basic Set of Applications - Part 2 : Specification of Cooperative Awareness Basic Service," *History*, vol. 1, p. 1–44, 2014.

[57] ETSI EN 302 637-3, "Intelligent Transport Systems (ITS) ; Vehicular Communications ; Basic Set of Applications ; Part 3 : Specifications of Decentralized Environmental Notification Basic Service," vol. 2, p. 1–73, 2014.

[58] M. Khalaf-Allah, "Time of arrival (toa)-based direct location method," dans *2015 16th International Radar Symposium (IRS)*. IEEE, 2015, p. 812–815.

[59] B.-C. Liu et K.-H. Lin, "Sssd-based mobile positioning : On the accuracy improvement issues in distance and location estimations," *IEEE Transactions on Vehicular Technology*, vol. 58, nº. 3, p. 1245–1254, 2008.

[60] S. E. Mohamed, "Why the accuracy of the received signal strengths as a positioning technique was not accurate ?" *International Journal of Wireless & Mobile Networks (IJWMN)*, vol. 3, n°. 3, p. 69–82, 2011.

[61] M. Porretta *et al.*, "Location, location, location," *IEEE Vehicular Technology Magazine*, vol. 3, n°. 2, p. 20–29, 2008.

[62] W. Tong *et al.*, "Artificial intelligence for vehicle-to-everything : A survey," *IEEE Access*, vol. 7, p. 10 823–10 843, 2019.

[63] H. Ye *et al.*, "Machine learning for vehicular networks : Recent advances and application examples," *ieee vehicular technology magazine*, vol. 13, n°. 2, p. 94–101, 2018.

[64] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, n°. 3, p. 175–185, 1992. [En ligne]. Disponible : https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879

[65] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, n°. 1, p. 5–32, oct. 2001. [En ligne]. Disponible : https://doi.org/10.1023/A:1010933404324

[66] J. Grover *et al.*, "Machine Learning Approach for Multiple Misbehavior Detection in VANET," dans *Advances in Computing and Communications*, A. Abraham *et al.*, édit. Berlin, Heidelberg : Springer Berlin Heidelberg, 2011, p. 644–653.

[67] R. C. Team *et al.*, "R : A language and environment for statistical computing," 2013.

[68] C.-C. Chang et C.-J. Lin, "Libsvm : a library for support vector machines. http," *www, csie. ntu. edu. tw/-cjlin/libsvm*, 2011.

[69] D.-V. Nguyen *et al.*, "Improving poor gps area localization for intelligent vehicles," dans *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2017, p. 417–421.

[70] K. Shi *et al.*, "Support vector regression based indoor location in ieee 802.11 environments," *Mobile Information Systems*, vol. 2015, 2015.

[71] L. Chen *et al.*, "A svm-based approach for vanet-based automatic incident detection," *International Journal of Simulation–Systems, Science and Technology*, vol. 17, n°. 30, p. 1–5, 2016.

[72] L. Chen, Y. Cao et R. Ji, "Automatic incident detection algorithm based on support vector machine," dans *2010 Sixth International Conference on Natural Computation*, vol. 2. IEEE, 2010, p. 864–866.

[73] S. I. Khan et S. G. Ritchie, "Statistical and neural classifiers to detect traffic operational problems on urban arterials," *Transportation Research Part C : Emerging Technologies*, vol. 6, n°. 5, p. 291 – 314, 1998. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/S0968090X99000054

[74] S. Lee, R. A. Krammes et J. Yen, "Fuzzy-logic-based incident detection for signalized diamond interchanges," *Transportation Research Part C : Emerging Technologies*, vol. 6, n°. 5, p. 359 – 377, 1998. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/S0968090X99000042

[75] H. Adeli et A. Karim, "Fuzzy Wavelet RBFNN model for freeway incindent detection," dans *J. Transp. Eng. , Volume 126, pp 464–471*, 2000.

[76] W. K. Lai, M.-T. Lin et Y.-H. Yang, "A Machine Learning System for Routing Decision-Making in Urban Vehicular Ad Hoc Networks," *International Journal of Distributed Sensor Networks*, vol. 11, n°. 3, p. 374391, 2015. [En ligne]. Disponible : https://doi.org/10.1155/2015/374391

[77] L. Zhao *et al.*, "A SVM based routing scheme in VANETs," dans *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, Sept 2016, p. 380–383.

[78] M. Slavik et I. Mahgoub, "Applying machine learning to the design of multi-hop broadcast protocols for VANETwei," dans *2011 7th International Wireless Communications and Mobile Computing Conference*, July 2011, p. 1742–1747.

[79] F. A. Ghaleb *et al.*, "An effective misbehavior detection model using artificial neural network for vehicular ad hoc network applications," dans *2017 IEEE Conference on Application, Information and Network Security (AINS)*, Nov 2017, p. 13–18.

[80] T. Zhang et Q. Zhu, "Distributed privacy-preserving collaborative intrusion detection systems for vanets," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, n°. 1, p. 148–161, March 2018.

[81] D. Stoyan, W. S. Kendall et J. Mecke, *Stochastic geometry and its applications. 2nd edition.* Wiley, 1995.

[82] P. Muhlethaler et A. Najid, "Throughput optimization in multihop csma mobile ad hoc networks," dans *EW 2004. The 5th European Wireless Conference*, February 24 - 27. Barcelona 2004.

[83] F. Baccelli et B. Błaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume II — Applications*, ser. Foundations and Trends in Networking. NoW Publishers, 2009, vol. 4, No 1–2.

[84] P. M. Nadjib Achir, Younes Bouchaala et O. Shagdar, "Optimisation of spatial csma using a simple stochastic geometry model for 1d and 2d networks," dans *IWCMC 2016*, September 5-9th, 2016, Paphos, Cyprus 2016.

[85] F. Baccelli et B. Błaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume I — Theory*, ser. Foundations and Trends in Networking. NoW Publishers, 2009, vol. 3, No 3–4.

[86] C. Dong *et al.*, "An improved deep learning model for traffic crash prediction," *Journal of Advanced Transportation*, vol. 2018, 2018.

[87] H. S. Manual, "American association of state highway and transportation officials," *Washington, DC*, vol. 19192, 2010.

[88] C. Chen *et al.*, "A multinomial logit model-bayesian network hybrid approach for driver injury severity analyses in rear-end crashes," *Accident Analysis & Prevention*, vol. 80, p. 76–88, 2015.

[89] ——, "Hierarchical bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations," *Accident Analysis & Prevention*, vol. 85, p. 186–198, 2015.

[90] J. Liu *et al.*, "How big data serves for freight safety management at highway-rail grade crossings ? a spatial approach fused with path analysis," *Neurocomputing*, vol. 181, p. 38–52, 2016.

[91] D. Lord et F. Mannering, "The statistical analysis of crash-frequency data : A review and assessment of methodological alternatives," *Transportation research part A : policy and practice*, vol. 44, n°. 5, p. 291–305, 2010.

[92] Q. Wu *et al.*, "Exploratory multinomial logit model–based driver injury severity analyses for teenage and adult drivers in intersection-related crashes," *Traffic injury prevention*, vol. 17, n°. 4, p. 413–422, 2016.

[93] J. Tang *et al.*, "Crash injury severity analysis using a two-layer stacking framework," *Accident Analysis & Prevention*, vol. 122, p. 226–238, 2019.

[94] S. Jin, X. Qu et D. Wang, "Assessment of expressway traffic safety using gaussian mixture model based on time to collision," *International Journal of Computational Intelligence Systems*, vol. 4, n°. 6, p. 1122–1130, 2011.

[95] S. Jin *et al.*, "Short-term traffic safety forecasting using gaussian mixture model and kalman filter," *Journal of Zhejiang University SCIENCE A*, vol. 14, n°. 4, p. 231–243, 2013.

[96] Ö. Aköz et M. E. Karsligil, "Severity detection of traffic accidents at intersections based on vehicle motion analysis and multiphase linear regression," dans *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2010, p. 474–479.

[97] S. Sun, C. Zhang et G. Yu, "A bayesian network approach to traffic flow forecasting," *IEEE Transactions on intelligent transportation systems*, vol. 7, n°. 1, p. 124–132, 2006.

[98] E. Hauer, J. C. Ng et J. Lovell, "Estimation of safety at signalized intersections," *Transportation Research Record*, vol. 1185, p. 48–61, 1988.

[99] D. Mahalel, "A note on accident risk," *Transportation Research Record*, vol. 1068, p. 85–89, 1986.

[100] J. Tang *et al.*, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Systems with Applications*, vol. 91, p. 452–463, 2018.

[101] J. de Oña, G. López et J. Abellán, "Extracting decision rules from police accident reports through decision trees," *Accident Analysis & Prevention*, vol. 50, p. 1151–1160, 2013.

[102] J. Abellán, G. López et J. De OñA, "Analysis of traffic accident severity using decision rules via decision trees," *Expert Systems with Applications*, vol. 40, n°. 15, p. 6047–6054, 2013.

[103] N. Dong, H. Huang et L. Zheng, "Support vector machine in crash prediction at the level of traffic analysis zones : assessing the spatial proximity effects," *Accident Analysis & Prevention*, vol. 82, p. 192–198, 2015.

[104] A. Iranitalab et A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, vol. 108, p. 27–36, 2017.

[105] T. K. Anderson, "Kernel density estimation and k-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, n°. 3, p. 359–364, 2009.

[106] R. Mauro, M. De Luca et G. Dell'Acqua, "Using a k-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis," *Modern Applied Science*, vol. 7, n°. 10, p. 11, 2013.

[107] H. T. Abdelwahab et M. A. Abdel-Aty, "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections," *Transportation Research Record*, vol. 1746, n°. 1, p. 6–13, 2001.

[108] D. Delen, R. Sharda et M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident Analysis & Prevention*, vol. 38, n°. 3, p. 434–444, 2006.

[109] Q. Zeng et H. Huang, "A stable and optimized neural network model for crash injury severity prediction," *Accident Analysis & Prevention*, vol. 73, p. 351–358, 2014.

[110] R. L. Cheu *et al.*, "Forecasting shared-use vehicle trips with neural networks and support vector machines," *Transportation research record*, vol. 1968, n°. 1, p. 40–46, 2006.

[111] M.-L. Huang, "Intersection traffic flow forecasting based on ν-gsvr with a new hybrid evolutionary algorithm," *Neurocomputing*, vol. 147, p. 343–349, 2015.

[112] D. Wei et H. Liu, "An adaptive-margin support vector regression for short-term traffic flow forecast," *Journal of Intelligent Transportation Systems*, vol. 17, n°. 4, p. 317–327, 2013.

[113] Y. Rong *et al.*, "Urban road traffic condition pattern recognition based on support vector machine," *Journal of Transportation Systems Engineering and Information Technology*, vol. 13, n°. 1, p. 130–136, 2013.

[114] X. Li *et al.*, "Predicting motor vehicle crashes using support vector machine models," *Accident Analysis & Prevention*, vol. 40, n°. 4, p. 1611–1618, 2008.

[115] R. Gang et Z. Zhuping, "Traffic safety forecasting method by particle swarm optimization and support vector machine," *Expert Systems with Applications*, vol. 38, n°. 8, p. 10 420–10 424, 2011.

[116] A. S. Sánchez *et al.*, "Prediction of work-related accidents according to working conditions using support vector machines," *Applied Mathematics and Computation*, vol. 218, n°. 7, p. 3539–3552, 2011.

[117] R. Yu et M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation," *Accident Analysis & Prevention*, vol. 51, p. 252–259, 2013.

[118] L. Guo *et al.*, "Pedestrian detection for intelligent transportation systems combining adaboost algorithm and support vector machine," *Expert Systems with Applications*, vol. 39, n°. 4, p. 4274–4286, 2012.

[119] R. Yu et M. Abdel-Aty, "Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data," *Safety science*, vol. 63, p. 50–56, 2014.

[120] Z. Li *et al.*, "Using support vector machine models for crash injury severity analysis," *Accident Analysis & Prevention*, vol. 45, p. 478–486, 2012.

[121] C. Chen *et al.*, "Investigating driver injury severity patterns in rollover crashes using support vector machine models," *Accident Analysis & Prevention*, vol. 90, p. 128–139, 2016.

[122] N. V. Chawla *et al.*, "Smote : synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, p. 321–357, 2002.

[123] M. Van Der Voort, M. Dougherty et S. Watson, "Combining kohonen maps with arima time series models to forecast traffic flow," *Transportation Research Part C : Emerging Technologies*, vol. 4, n°. 5, p. 307–318, 1996.

[124] Z. Li *et al.*, "Building sparse models for traffic flow prediction : An empirical comparison between statistical heuristics and geometric heuristics for bayesian network approaches," *Transportmetrica B : Transport Dynamics*, 2017.

[125] D. Huang *et al.*, "A short-term traffic flow forecasting method based on markov chain and grey verhulst model," dans *2017 6th Data Driven Control and Learning Systems (DDCLS)*. IEEE, 2017, p. 606–610.

[126] M. Shuai *et al.*, "An online approach based on locally weighted learning for short-term traffic flow prediction," dans *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, 2008, p. 1–4.

[127] M. Castro-Neto *et al.*, "Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert systems with applications*, vol. 36, n°. 3, p. 6164–6173, 2009.

[128] L.-Y. Chang, "Analysis of freeway accident frequencies : negative binomial regression versus artificial neural network," *Safety science*, vol. 43, n°. 8, p. 541–557, 2005.

[129] G. Pan, L. Fu et L. Thakali, "Development of a global road safety performance function using deep neural networks," *International journal of transportation science and technology*, vol. 6, n°. 3, p. 159–173, 2017.

[130] J. Wang, W. Deng et Y. Guo, "New bayesian combination method for short-term traffic flow forecasting," *Transportation Research Part C : Emerging Technologies*, vol. 43, p. 79–94, 2014.

[131] M. J. Lawrence, R. H. Edmundson et M. J. O'Connor, "The accuracy of combining judgemental and statistical forecasts," *Management Science*, vol. 32, n°. 12, p. 1521–1532, 1986.

[132] S. Makridakis, "Why combining works ?" *International Journal of Forecasting*, vol. 5, n°. 4, p. 601–603, 1989.

[133] F. Guo *et al.*, "Predictor fusion for short-term traffic forecasting," *Transportation research part C : emerging technologies*, vol. 92, p. 90–100, 2018.

[134] J. Xie et Y.-K. Choi, "Hybrid traffic prediction scheme for intelligent transportation systems based on historical and real-time data," *International Journal of Distributed Sensor Networks*, vol. 13, n°. 11, p. 1550147717745009, 2017.

[135] Q. MA *et al.*, "Traffic condition on-line estimation using multi-source data," *Journal of Computational Information Systems*, vol. 8, n°. 6, p. 2627–2635, 2012.

[136] D. for Transport, "Road safety data," 2017. [En ligne]. Disponible : https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

[137] J. A. Bilmes *et al.*, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, n°. 510, p. 126, 1998.

[138] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, n°. 6, p. 47–60, 1996.

[139] A. K. Jain, R. P. W. Duin et J. Mao, "Statistical pattern recognition : A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, n°. 1, p. 4–37, 2000.

[140] B. Houcine, K. A. Cherif et D. Rafik, "Novel approach in speaker identification using svm and gmm," *Journal of Control Engineering and Applied Informatics*, vol. 15, n°. 3, p. 87–95, 2013.

[141] I. Mierswa et R. Klinkenberg, "Rapidminer studio (9.2)[data science, machine learning, predictive analytics]," *Retrieved from rapidminer. com*, 2018.

[142] B. H. Menze *et al.*, "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC bioinformatics*, vol. 10, n°. 1, p. 1–16, 2009.

[143] J. A. Suykens et J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, n°. 3, p. 293–300, 1999.

[144] C. Cortes et V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, n°. 3, p. 273–297, 1995.

[145] T. Hastie *et al.*, *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2009, vol. 2.

[146] J. Friedman, T. Hastie et R. Tibshirani, "The elements of statistical learning, volume 1 springer series in statistics springer," 2001.

[147] C. M. Bishop et N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, n°. 4.

[148] S. Knerr, L. Personnaz et G. Dreyfus, "Single-layer learning revisited : a stepwise procedure for building and training a neural network," dans *Neurocomputing*. Springer, 1990, p. 41–50.

[149] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, p. 159–175, 2003.

[150] H. Dhahri et A. M. Alimi, "The modified differential evolution and the rbf (mde-rbf) neural network for time series prediction," dans *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, 2006, p. 2938–2943.

[151] J. Davis et M. Goadrich, "The relationship between precision-recall and roc curves," dans *Proceedings of the 23rd international conference on Machine learning*, 2006, p. 233–240.

[152] A. Dorri *et al.*, "Blockchain : A distributed solution to automotive security and privacy," *IEEE Communications Magazine*, vol. 55, n°. 12, p. 119–125, 2017.

[153] C. A. Kerrache *et al.*, "Rita : Risk-aware trust-based architecture for collaborative multi-hop vehicular communications," *Security and Communication Networks*, vol. 9, n°. 17, p. 4428–4442, 2016.

[154] D. Abraham et N. S. Raj, "Approximate string matching algorithm for phishing detection," dans *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2014, p. 2285–2290.

[155] A. Murugan et U. Udayakumar, "Sequence similarity between genetic codes using improved longest common subsequence algorithm," *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, vol. 5, n°. 7, p. 57–60, 2017.

[156] C. Varol et H. M. T. Abdulhadi, "Comparision of string matching algorithms on spam email detection," dans *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*. IEEE, 2018, p. 6–11.

[157] K. Beijering, C. Gooskens et W. Heeringa, "Predicting intelligibility and perceived linguistic distance by means of the levenshtein algorithm," *Linguistics in the Netherlands*, vol. 25, n°. 1, p. 13–24, 2008.

[158] R. Shrestha *et al.*, "A new type of blockchain for secure message exchange in vanet," *Digital communications and networks*, vol. 6, n°. 2, p. 177–186, 2020.

[159] A. Lei *et al.*, "A secure key management scheme for heterogeneous secure vehicular communication systems," *ZTE Communications*, vol. 14, n°. S0, p. 21–31, 2019.

[160] M. Awais Hassan *et al.*, "A secure message-passing framework for inter-vehicular communication using blockchain," *International Journal of Distributed Sensor Networks*, vol. 15, n°. 2, p. 1550147719829677, 2019.