CHAPTER 1

INTRODUCTION

1.1 Background

The global number of motor vehicles is rapidly increasing, with projections suggesting that it will surpass two billion by 2050 [1]. This surge has significantly contributed to the rise in collisions and accidents, resulting in approximately 1.25 million fatalities and 40 times as many injuries occurring annually due to road accidents. The World Health Organization (WHO) predicts that by 2030, vehicle collisions will rank as the fifth leading cause of death. These accidents not only affect those directly involved but also lead to traffic congestion and diversions, causing substantial economic losses in the billions of dollars each year due to treatment costs, property damage, lost work hours, and increased fuel consumption. Human error is estimated to be responsible for 95% of all road accidents [2].

In response to these growing concerns, Cooperative Intelligent Transport Systems (C-ITS) have been developed and evolved with the aim of enhancing driver safety and reducing traffic accidents [3], [4]. C-ITS includes a variety of subsystems such as personal, vehicle, roadside, and central systems, which work together cooperatively. These systems enable communication between vehicles and other subsystems through Vehicle-to-Everything (V2X) message broadcasts over the vehicular network [5]. There are four types of V2X communications: Vehicle-to-Vehicle (V2V), Vehicle-to-Network (V2N), Vehicle-to-Infrastructure (V2I), and Vehicle-to-Pedestrian (V2P) [6]. However, despite these promising results, C-ITS remains vulnerable to misbehaviours, which defined as any abnormal occurrence in the C-ITS network, caused by one or more participating nodes, that disrupts the normal functioning of C-ITS services [4], [7]. Various mechanisms for detecting mis behavior in C-ITS have been proposed, including machine learning-based approaches [8].

Machine learning (ML), particularly supervised learning, has been extensively applied to analyze communication message patterns and identify anomalies. These techniques, often implemented by individual C-ITS elements or centralized servers, are crucial for detecting complex misbehaviors. Various machine learning models have been proposed for data-centric Misbehaviour detection in C-ITS, including Deep Belief Network (DBN), which utilize stacked layers of Restricted Boltzmann Machines for both unsupervised and supervised tasks [9], and Multi-Layer Perceptron (MLP), a type of Feedforward Neural Network used for classification and prediction tasks [5]. Support Vector Machines (SVM) are also employed for classification and outlier detection in Misbehaviour analysis [10]. Autoencoders (AEs), another model, are particularly effective in learning efficient data coding for tasks such as Sybil attack detection [8]. Long Short-Term Memory (LSTM) networks, known for their ability to remember extended patterns, have been applied to global Misbehaviour detection using raw beacon data [11]. Other techniques include K-Nearest Neighbors (KNN), a non-parametric method for classifying new data based on proximity to known data [10], and logistic regression, which models the probability of discrete outcomes to predict Misbehaviour [12]. Additionally, bagging, an ensemble technique that aggregates multiple predictors to improve model accuracy, has been used for detecting Misbehaviours like false alerts [13].

Nonetheless, these advancements are accompanied by a shared drawback—substantial reliance on computational resources, resulting in heightened time complexity. In VANET scenarios, where computational capabilities are limited [14], primarily due to vehicles' constrained computing power and memory, the efficient operation of ML-Based Misbehaviour Detection System (MLMDS) algorithms becomes imperative [15]. Traditional machine learning models, which typically rely on a single monolithic structure, also can be insufficient when dealing with complex data and intricate relationships, further exacerbating these challenges [16]. The swift movement of vehicles necessitates MLMDS algorithms to function within strict time constraints to ensure timely detection and response to security threats [17]. Additionally, to mitigate computation overhead, the complexity of VANET applications must be minimized [18].

To address these limitations, a more sophisticated approach is required. Cascaded ML offers a hierarchical architecture that breaks down complex problems into smaller,

more manageable subtasks. Each subtask is tackled by specialized models within the cascade, which enhances processing efficiency and focus [19], [20]. This hierarchical structure allows for capturing intricate data relationships and patterns, thereby improving system resilience, performance, and adaptability. For instance, previous studies have integrated multiple ML techniques, such as combining MLP with Artificial Neural Networks (ANN) [21], or using CNN for feature extraction followed by threshold-based detection [22]. The core principles of Cascaded ML—hierarchical decomposition, sequential processing, and error propagation control—contribute to a more robust and effective detection system [23]. Moreover, the ANN method is recognized as offering distinct advantages over other techniques. In comparison to other deep learning architectures, ANNs are characterized by simplicity and efficiency. With fewer layers and parameters, ANN is less complex, easier to train, and requires less computational power [24]. Its straightforward structure minimizes overfitting, ensuring stable performance across cascade layers, and has been shown to achieve over 97% accuracy while maintaining consistent computational demands [25]. This makes ANN an ideal choice for the Cascaded ML framework in VANET environments, where balancing low complexity with high detection accuracy is crucial.

The objective of this thesis is to address the challenge of achieving high detection accuracy in ML methods while minimizing computational resource consumption. To enhance precision and practicality, feature engineering, data cleansing, and data transformation will be employed during preprocessing. The proposed solution involves utilizing a Cascaded ANN methodology on the BurST-ADMA dataset. This approach classifies incoming data traffic into two main categories—normal and Misbehaviour—in the initial layer, with further classification of Misbehaviour traffic by trajectory type in the second layer. To ensure robust evaluation and reduce overfitting risk, a 5-fold cross-validation approach will be used. The goal is to develop a low-complexity MLMDS that maintains strong performance across key detection and complexity metrics by applying the Cascaded ANN approach, and ensure reliable obtained performances by through the evalution by the aforementioned validation technique.

1.2 Problem Identification

There are various challenges identified in existing research on MLMDS for VANETs, with the main issues being as follows:

- High computational resource consumption: Existing MLMDS approaches often require extensive computational resources, resulting in a trade-off between accuracy and efficiency. This trade-off is especially problematic in real-time applications, where delays can severely impact system performance [14]. For instance, Anyanwu et al. [26] and Idris et al. [27] achieved detection performances of over 98%, but their models took 240.87 seconds and 443.42 seconds, respectively, to simulate the entire BurST-ADMA dataset. Such processing times can be impractical for resource-constraint environments like VANET.
- Less reliable simulation results: Some studies use basic preprocessing techniques, such as splitting the dataset once for training and testing, without fully accounting for the data's variability. Additionally, classification is often performed only a single time, which may not adequately represent different possible outcomes. This can lead to unreliable results, especially when working with datasets that contain diverse traffic types [28]. Despite achieving high accuracy, the reliability of these models' outcomes is questionable due to potential data imbalances and the lack of robustness in their evaluation methods.
- Noisy and inconsistent datasets: The dynamic nature of VANET environments results in datasets that are not only large but also highly variable, with significant spatial and temporal dependencies. These datasets are often noisy and inconsistent, containing outliers and class imbalances due to fluctuating traffic conditions and diverse driver behaviors. These inconsistencies complicate the data analysis and model training processes, making it challenging to develop models that can generalize well across different scenarios [29].

To address these challenges, the proposed framework in this thesis, as illustrated in Figure 1.1, aims to optimize the trade-off between computational efficiency and accuracy, enhance the reliability of simulation results through more robust preprocessing and classification strategies, and develop techniques to handle the noisy and inconsistent nature of VANET datasets.

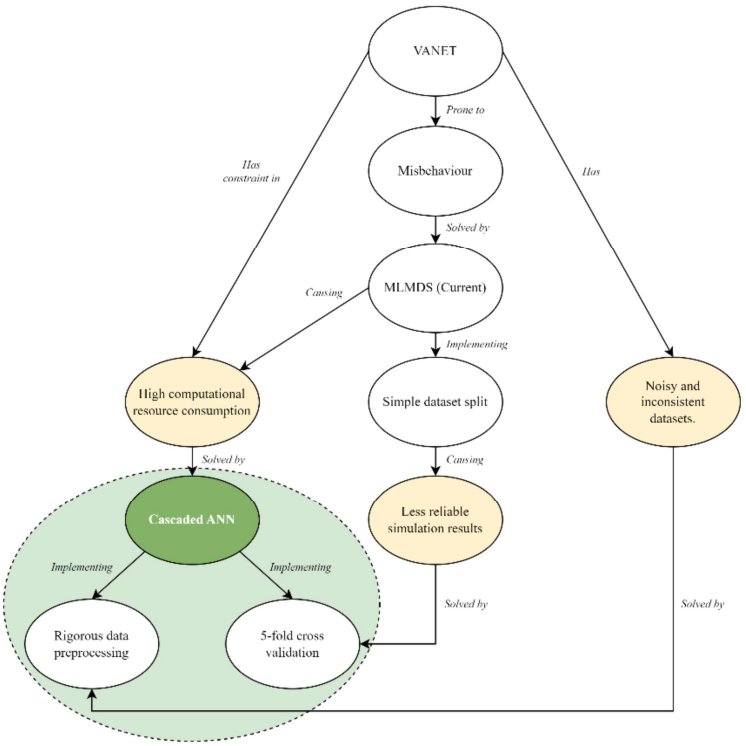

Figure 1.1 Problem Identification Mind Map

1.3 Objectives

This thesis considers the following assumptions:

1. This thesis designs MLMDS with a simple architecture that can achieve high detection metrics without compromising computational resource consumption.
2. This thesis optimizes the dataset for the simulation process by applying rigorous preprocessing.
3. This thesis optimizes the reliability of training and testing results by providing more generalizable assessments and minimal random fluctuations.

1.4 Scope of Work

The research in this thesis will be conducted with the following scope:

1. Cascaded ANN will be used as the MLMDS classification method.
2. Misbehaviour detection will be simulated on the BurST-ADMA dataset.
3. The traffic dataset will be preprocessed using feature engineering, data cleansing, and data transformation.
4. Simulation results will be validated using a 5 fold cross-validation approach.

5. The performance of simulation results will be measured using detection metrics (accuracy, precision, recall, and F1-score) and complexity metrics (simulation time and consumed memory).
6. The simulation will be conducted through Google Colabs on a notebook device with AMD Ryzen A5 6600H CPU @ 3.3 GHz 16 GB configuration utilizing 12.7 GB of RAM and 107.7 GB of disk space from Google's cloud infrastructure.

1.5 Expected Results

In previous works [30], [31], [32], misbehaviour detection in VANET environments was performed using various modifications and combinations of ML methods. The detection results demonstrated high performance in several aspects of detection metrics. However, these results were obtained using high computational resource consumption, leading to high complexity. This thesis proposes an MLMDS method with a simple architecture using Cascaded ANN. Furthermore, this thesis proposes applying rigorous preprocessing methods encompassing feature engineering, data cleansing, and data transformation and a 5 fold cross-validation approach to improve simulation results' reliability. The proposed method is anticipated to yield high accuracy, precision, recall, and F1-score performance metrics while exhibiting low complexity.

1.6 Research Methodology

To carry out this thesis, fundamental studies and experiments were conducted, divided into the following Work Packages (WP):

WP1. Build an MLMDS model based on a Cascaded ANN architecture.

WP2. Simulate the model on the BurST-ADMA dataset and validate the results with a 5 fold cross-validation approach.

WP3. Evaluate the detection performance using detection metrics (accuracy, precision, recall, and F1-score) and complexity metrics (simulation time and consumed memory).

WP4. Compare and analyze the simulation results of the proposed method with related works from previous studies.