Explainable Deep Learning Models for Intrusion Detection in Autonomous Vehicle Networks

By Dr. Sebastian Panisello

Professor of Industrial Engineering, University of Chile

1. Introduction

The anomaly classification component applies decision-making tree (DT) classifier and generates interpretable rules related to adversarial instances. The subsequent verification (VM) component uses sequential feature from data instances that are classified as adversarial to check whether a packet is benign or false. The HEDL model allows the use of suitable algorithms wherein the ID part can use hidden Markov Model (HMM), conditional restricted boltzmann machine (CRBM) and long short-term memory (LSTM) algorithms. The open-source NSL–KDD dataset and NASA datasets are leveraged for training and testing of the proposed model. The experimental results demonstrate that the proposed hybrid model offers improved results including precision, recall and accuracy as compared with various existing hybrid and sequential models in the literature [1] . The model also allows for interpretation by means of easy to understand rules and unidentified attack services can be detected through ID agent which is not available in majority of the existing models. Future scientific work covers verify the performance of the proposed model on real data by means of software-in-the-loop (SiL) in vehicle network.

ref: 6770478a-3412-41b9-a7d4-46345d6ef5e0 ref: 345e7ee7-b4f3-4fa3-9f0d-7c9138e1404c The extensive computer networks of an autonomous vehicle (AV) communicate each other and require a sleuth of cybersecurity measures to avoid adversarial impacts. Intrusion detection (ID) and machine learning (ML) are installed to meet this goal; however, a drawback of the classification-based ID is that it is unable to explain, comprehend, and respond to detected anomalies caused by adversarial instances. We propose an explainable deep learning (XDL) based system to better capture, interpret, and manage adversarial instances in the context of a vehicle network. A novel Hybrid Explainable Deep Learning (HEDL) model is proposed to classify the data samples as per the following predefined categories: normal, normal but

having adversarial instance, and adversarial instance introduced to cause attack [2]. The proposed model consists of an intrusion detection (ID) component that uses sequential data for detection of adversarial instances present in normal data and then applies deep learning (DL) algorithms to obtain probability-based feature vector.

1.1. Background and Significance of Intrusion Detection in Autonomous Vehicle Networks

The importance of ultra-secure communication and the risks entailed by communication failure cannot be overemphasized as the AVs provide a service where errors are often unacceptable. The majority of AV applications offer services mainly in urban environments where complex radio propagation, interference and mobility are highly variable. In 15% of intrusions, the intrusion attempt targets the network communication infrastructure. Due to the travel aspect of traffic, ground and cross-country connectivity are very dear. With the advent of advanced hexadecimal connectivity and current thruster technology, the kinds of vehicles are expected to be connectivity increasing from the tens available today. As AVs are operated by artificial intelligence systems, communications exhibit asynchronous message exchange patterns established over end-to-end connections. To model the semantic-centric behavior that is communication patterned, natural satisfaction response methods are executed in AV IThunder. As always, ornate changes may be performed and as a traffic behavior obtains the undemanding' passenger. Requirements directed at vehicle data streams include the severe timeliness demands of AVs including the high time resolution of remote sensors. [3] This L3–L4 safety apparatus should provide fault detection. The mainstay of rendering up-to-date services to passengers and frequent downsizing (section units) for communications is the wireless radius of the rear of the vehicle. The wireless radius in which they seamlessly connect is a category of irritability suppression element bringing up global layer. For any two connected vehicles, the network defines a cyclic graph with the vehicles as storeys. The networks should offer connected vehicles support for advanced diagnostic trouble code detection (the last ten data points from databases/vehicles). In this congregate encumbrance detection approach, the vehicle network needs to provide consistent deterministic paths with known data rate under contention to the application layer of cloud computations. The quantum devices or entities cannot be measured with any certainty at a given instant of time. On-board computing is prohibited for extended transmission power and interference. [4]

One of the main reasons why IDS assumes greater importance in AV network operation than in other networks is that a functional compromise of an AV system can lead to catastrophic consequences [5]. In contrast to regular vehicles, connected AVs need real-time communication to access traffic facilities and make driving decisions based on a broad set of environmental and road information collected from servers. The availability of AVs itself is guaranteed by middleware communication technologies allowing the provision of uninterrupted exchange of information between vehicles. Avionics and remote sensing (AVs) are real-time systems that work at microscale. Persistence occurs when a time process is not able to provide timely responses to the input. Already, a variety of middleware technologies are available to allow cars to access both roadside equipment (V2I) and services via communication with roadside infrastructure. Accessed vehicles (V2V) exchange information between each other, in addition to accessing these roadside equipment (V2I). This kind of connected AVs' architecture requires radial parameters for flawless operation.

1.2. Motivation for Using Deep Learning Models

[6] [5]Cybersecurity attacks have become a very big threat to the functioning of the vehicles in a connected environment. In an automobile scenario, cybersecurity risks can have life-threatening consequences. Therefore, reliable security countermeasures need to be implemented to ensure safe operation of the vehicles. The amount of data generated in train-based maneuvers is large and multi-modal. Detection of novel attacks that may streamline the normal car network traffic is important in mentioned multi-modal traffic. An intrusion detection system (IDS) for in-vehicle communication needs to be intelligent and proactive. In-vehicle communication networks are engineered to support communication among the various controllers in an automobile, each with its own customized message format and unique encoding. Even though communication is partial, a cyber-attack can occur not only over-the-air but also within the vehicle itself, making it difficult to protect the network security.[3]In-vehicle networks (IVNs) pose a great security threat to vehicles, as a large number of vehicles are susceptible to cyber security threats. Intrusion detection is the main mechanism used to ensure network security. Existing intrusion detection systems (IDSs) in IVNs mainly adopt supervised learning algorithms, which require manual labelling and have difficulty extracting traffic features. Furthermore, IVN traffic is not only spatially correlated but also temporarily correlated. Current research mainly focuses on the attack detection of the last snapshot. In this paper, a new autoencoder-based unsupervised deep learning model is

applied to detect intrusions without human assistance. We empirically evaluate our multi-scale convolutional recurrent network on the UNSW-NB15 dataset, which is generated using the known cyber-attack scenarios in IVNs and compare it with the vanilla deep learning model. The experimental results demonstrate that our proposed model accurately classifies the cyber-attack categories/fluctuates and has better feature extraction.

1.3. Importance of Explainability in Deep Learning Models

They are difficult to be used in the field of medicine, finance, or autonomous vehicles. Moreover, a reliable performance prediction by any model is required to be able to remain confident on the model's predictions even on unseen data. This paper gives a comprehensive review of the existing research in the field of XDL to ensure the usage of these models in autonomous vehicles. Deep neural network-based models are a highly popular choice as a DL model in use for various applications as well as autonomous vehicles as inference engine models. Neural networks have achieved state-of-the-art performance over other traditional learning schemes, such as support vector machines (SVM) and decision trees. LIME is basically a post-hoc method, though it can be used not just for DL models, but also for other non-linear models. LIME overcomes this limitation by approximating the complex decision boundaries learned by the black box model around the area where it is supposed to be explaining its predictions.

[7], [8] The role of machine learning in autonomous vehicles is continuously growing for tasks such as path planning, computer vision, and object detection. However, in order to be seamlessly integrated into autonomous vehicles, these learner pro-posed by leveraging deep learning models, especially using convolutional neural network (CNN) architectures for computer vision tasks. While such schemes tend to yield state-of-the-art performance in comparison to traditional learning schemes, they hold the downside of being opaque in the workings, i.e., the decision-making process is not transparent and is difficult for humans to understand. Transparency is considered to be an essential factor to ensure trust in the decision-making process being made by these models. The increasing usage of these models in sensitive applications necessitates the need for a system that helps in interpreting the decisions made by these models. Not being able to explain deep learning (DL) model decisions affects their deployment in real-world settings.

2. Fundamentals of Intrusion Detection

In the automobile industry, safety is ranked as the most important factor in a vehicle; a zero magnitude of accidents is expected to be achieved because of autonomous vehicle (AV) technology. The security of in-vehicle networks (IVNs) has become a serious issue because of the vulnerabilities of the traditional CAN protocol, which is based on the controller area network bus. External and internal attackers exploit these vulnerabilities and launch several security attacks, for example, invasion of privacy, damage, and theft of the vehicle. Hence, in-vehicle network intrusion detection systems (IVN-IDSs) have been evolved and considered as an essential security solution against the intrusion attacks entering in a vehicle's network [5].

Cyber Security in a Connected and Autonomous Vehicle (CAV) domain has attracted a lot of attention in recent years. According to a McKinsey study, by 2026, the global market for connectivity hardware in vehicles is expected to reach $70 billion, with nearly all car models launched expected to include connected hardware [9]. An automotive network, or automotive in-vehicle network, connects automobile components in safety, infotainment and vehicle performance. A vehicle network consists of communication data that is transmitted between electronic control units (ECUs). The controller area network (CAN) is the most common communication network used in vehicles. It was developed by Bosch and released in 1984. The traditional CAN network does not possess any inherent countermeasures to protect the integrity of the messages [1].

2.1. Types of Intrusions in Autonomous Vehicle Networks

Data-Level Intrusions: Attacker behaves as 'node' on the CAN network and it tries to alter data values or forging new CAN messages to mislead the entire network or to initiate extra tasks [10]. A specific example for this kind of target is done by Ferrari and Priemer by introducing spoofing attacks in autonomous vehicles in their paper [6]. Control-Level Intrusions (CLI): Intruders perform various attacks after compromising the ECU (Electronic Control Unit) or ID of the target node. CLI attacks are also more dangerous cause single node-level intrusion might tend to affect all the ECU functions via the consequent associations [11]. Nevertheless there are some different types of target nodes instead of ECU, main goal for the intruder is to skew the normal driving scenarios where affects vehicles functions.

What do they actually try to detect in autonomous vehicle networks? Currently, intrusion detection studies mainly focus on the CAN bus network. Here we give some information about CAN network topology and the targeted attacks related to each area.

2.2. Traditional Intrusion Detection Methods

An alternate approach to intrusion detection is the use of statistical means to capture the normality based on the overview of the data. A simple approach is the use of pre-segmented data, even though in the vehicular environment this could create an overhead which is difficult to accept, specifically in an edge-cloud continuum with resource scarcity. Still, it is an approach that is often used, just because the profile of the devices often resides in the binned space of the raw data, and the search for this profile characteristics is greatly simplified. It is important to emphasize a technique related to traditional as well as deep learning models. An autoencoder is an unsupervised machine learning model that learns compact representations of data without supervision to minimize a reconstruction error. Slowly, the model is taught to ignore slight deviations in the input such that these remains apparent in the error. The distribution of these errors can thus be used to recognize interfering behavior [12]. Furthermore, adversarially trained models, such as generative adversarial networks, were used to capture the normal behavior better and adversarial activity can thus be recognized more easily. A combination of this approach with the domain transfer or invariant representations are not often used in the automobile environment, even though the wide variety of vehicles could benefit from such extraction of non-vehicle related behavior.

[13] Traditional intrusion detection methods are commonly used because of their advantages such as easy implementability. The traditional signature-based intrusion detection systems are used to monitor databases and match observed activities against the known database of attack patterns. All the techniques first learn model about the profiles of normal activities and then uses these models to identify abnormal behavior. The use of expert rules provides the transparency and /or explanations. The care for anomalies in industrial processes is important because of the potential damages that may result from an unknown perception and classification of the origin of equipment or process disturbances [6]. In vehicle networks, such as roadside units and autonomous vehicles, legitimate transmission removed in an unknown way can severely disturb the overall message exchange. Intrusion detection is a potential strategy to capture and classify such thefts of data exchange. In the vehicle environment, classical methods make use of several stable features to allow profiling of the intended device, like timing or length differences, to allow a classification into attack or no attack.

2.3. Challenges in Intrusion Detection for Autonomous Vehicles

Recent studies have also applied deep learning techniques such as deep learning with explainable embeddings and FreezeOut network to enhance detection performance. Moreover, various industrial and real-life problems have been discussed for supervised and unsupervised explainability and the appropriate explanation for not handled cases. Besides above discussion, IVN, and ADs have already been presented for IV systems and AVs. However, existing studies have not been able to address the limitations of AV network security, especially in the context of cyber-physical system AV design [14].

An autonomous vehicle (AV) network has a complex and dynamic environment along with stringent security requirements to ensure safety, comfort, and efficiency. Attacks against AV networks can prevent them from achieving these goals [6]. In this section, we discuss various challenges and limitations involved in enhancing the security of AV networks using deep learning-based intrusion detection systems (DL-IDSs). Anomaly-based DL-IDSs are currently a popular choice due to their high detection rate and flexibility in identifying both known and unknown attacks. However, such systems are inherently vulnerable to adversarial attacks that are specifically designed to fool the deployed ML-models [15]. These vulnerabilities can be triggered by ensuring that the adversarial examples are misclassified by attacking one or more independently trained ML-models. Similarly, the availability of the worst-case and white-box scenarios makes the deep-learning-based intrusion detection systems more susceptible to such threat.

3. Deep Learning for Intrusion Detection

This chapter gives an overview of different works for detection of anomaly in autonomous vehicle networks. It also discusses the works related to developing IDS's based on machine learning. It gives good comparison of the techniques being used for machine learning. It also provides ideas and concepts which can be applied for evaluating the performance of the existing machine learning models. As the new scale of terrorism in the form of cyber attacks has created devastating scenarios for the Governments and businesses, there is a strong need of such systems that should be able to handle the traffic and can ensure that all the vehicles including autonomous and non-autonomous are protected at the best levels. Again, the goal should be handled the security-related issues at the level that they should never impact the safety of the rides. So, our model will satisfy these needs by doing the anomaly detection and evaluation both.

Since the Internet and the world become more connected, the security and privacy issues become more prominent. Intrusion detection systems (IDS), as the major tools to protect computer systems from various cyber threats, can be more widely applied in various systems. Most of the research for Intrusion Detection Systems (IDS's) for autonomous vehicle networks [10] lack a mechanism that explains the basis of the output and conclusions they make. This makes it difficult to understand, trust, and utilize machine learning models by cybersecurity professionals for these important tasks. To the best of our knowledge, still no comprehensive research methodology has been designed yet to analyze and compare Deep Learning (DL) models for IDS in autonomous vehicle networks with a focus on their Explainability.

3.1. Overview of Deep Learning

Regardless of the reasons, DL eventually completes the learning process and outputs the result, and does not allow to examine its own response. In the real-life applications, specifically in computer networks and AI-based autonomous vehicle networks, each system cannot be standalone [ref: bdae7f1f-b100-4020-ad5a-096b26307065]. Besides, real computer network data are complex, dynamic, and noisy in real-time, so the models cannot provide very good results. If the DL models are not easily understandable, and it is not using with other methods in computer intrusion detection systems, then false negatives and false positives occur so the detection rates of log files does not efficiently provide security.

It is well known that deep learning (DL) models have developed rapidly and become an important role in AI technology [ref: 6bdd7e76-ad6b-4c6b-b74f-cfb7a88230ba]. Besides, deep learning technology has been implemented in several applications of computer network security. As one of deep learning techniques, deep learning has played an important role in the success of detecting network intrusion and is mainly aimed at creating a benchmark [ref: c8bf97f5-7073-4652-88f2-5d816e4ace1f]. The key advantage of this technique is that it automatically recognizes the feature and provides good results even if raw data is given without any human intervention. However, not all the applications in DL for intrusion detection are successful, and they may increase the false positive rates and have limitations for the number of layers less than 35. In other words, extremely deep layers may have failures.

3.2. Applications of Deep Learning in Intrusion Detection

Inference adaptive strength of the vehicle threat model distinguishes between racV4R_12D and DNN. from the word adaptive, one can recognize the point that models employ- ing this scheme will gradually adjust themselves from real-time ongoing CAN data from vehicles and the same will lead to realistic, scalable, and prac- tical systems with low false positive data as seen in the context of CVT-based detection. In comparison, it is shown that from the perspective of the benchmark, the features and the detection algorithm of the validation set, which have mostly been kept slightly hidden, are adaptable models founded upon RacV4R_12D yielding nearly equivalent or better answer accuracy . Notice that, in real-time, it can gather vehicle traffic data only from all the source devices coupled with a diagnostic tool, a protocol sniffer, a hardware device, or a remote vehicle interface etc. from the software/IT perspective.

There are various applications of deep learning models in the context of intrusion detection in autonomous vehicle networks. A most special mention can be given to the common execution in the majority of the retrieved documents i.e. "InceptionNet"dpt, "LSTM"fc, -DNNfc, -HTMfc, and -GRUfc in one form or the other – namely, this work is highly inspired, motivated, and is designed accordingly with these reference works [16] [17]. In the same way, RacV4R_12Dfc is adapted from VGG16 pre- trained model, however, it is significantly modified and is very different from the pre-trained version. There is no predefined vocabulary on attack patterns in the domain of in-vehicle networks i.e. CAN messages other than conventional and few evasive attacks have been studied in context of secure TrustZone-based attack detection yongzhi2021 detecting. From the viewpoint of resisting generic attacks on scaled experiment and field data, DNN and RacV4R_12D have distinct advantage due to usage of STSTM-based scheme and signature- based approach where no predefined attack pattern in prior is used while training model [12] .

3.3. Advantages and Limitations of Deep Learning Models

CNNs on the other hand are strong in learning the spatial dependencies in network traffic, that it, they are strong in analyzing raw data independent of time. In a few works CNN has been adapted for capturing temporal dependencies as well, making it also useful for sequence based classification. In our study we take advantages of both the RNN, capturing time dependencies, and the CNN, capturing the spatial dependencies, to provide a more complete feature set for our dreams of achieving the best possible performance for an IDS in

autonomous vehicle networks [13]. While Regularized CNN has shown to be the best adaptation from general image recognition to network intrusion detection among the state-of-the-art approaches, its development during the years has not reached the stage where it can satisfy the tight requirements from autonomous vehicle networks. And in order to meet the strong real-time requirement of network intrusion detection, it is crucial to massively decrease the model complexity.

Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are among the most widely used deep learning models for detection systems owing to their ability to capture the time dependencies and spatial dependencies between network features [ref: 5f5bc4c3-97fc-4c2e-8439-1e5dc20e1b8d,8dc6d9c4-75d1-401c-b68a-370f08c1fb37]. RNNs are strong in learning and extracting temporal dependencies in network traffic, i.e., they are powerful in classifying sequences of network packets and capturing flows of data. However, in recent years, CNN Backbones of RNN, Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM), have been proposed by researchers to enhance its functionality in cybersecurity. Compared to RNN, GRU has fewer parameters and runs faster in practice, but with a slightly lower performance, whereas LSTM has better performance but with a very large amount of parameters [3]. They are powerful in data analysis because as a whole, they capture the whole data information which is very useful for anomaly detection [5]. In our study, both GRUs and LSTMs are leveraged to create our proposed RNN backbones. While GRUs have slightly better accuracy than LSTMs, LSTMs has more robust performance, which we believe is a primary factor to consider during intrusion detection in an operation environment where we cannot rely on always detecting intrusion with the highest accuracy.

4. Explainable AI and Interpretability

Existing research on interpretation for the IDS aims at improving the decision making process by making it more understandable and manageable. Some initial research has proposed AI methods to make AI models more interpretable. Moreover, some work is looking at using interpretable classification models for intrusion detection instead of black-box models. However, most work on interpretability focuses on the use of this property to improve decision-making and/or reveal vulnerabilities. Finally, few studies have explored interpretability using deep-learning-based Ai at all. Specifically, in the domain of network security, the common applications of AI for interpretability were the study or litiga-tion of

standard AI models. So, the AI did not contribute to the effectiveness of the AIs for problem-solving. Indeed, AI does not stop at interpretation methods as there is also the concept of explainable AI (XAI) which means being able to explain to a user, just why an AI decision was made. Toward XAI, thereby, we aim to help the user validate the robustness of the explanations provided [18].

As in other domains of Artificial Intelligence (AI), the interpretability aspect of Deep Learning models is crucial for guaranteeing transparency and effectiveness. This interpretability issue must be addressed in autonomous vehicle networks to certify their security. Intrusion Detection Systems (IDS) teach models based on different features extracted from the traffic, among which explaining features and attacks remains complex. In order to face this complexity and increase the explainability of the IDS systems used in autonomous vehicle networks, we propose to investigate the applicability of artificial intelligence (AI) interpretation methods. The underlying objective is to benefit from the boosting power of DL in terms of detection accuracy, while offering capabilities for accounting for its decisions [19]. Indeed, these AI interpretation methods have the potential to highlight specific aspects of the data, which are the basis for learning a model's decision process.

4.1. Importance of Explainability in AI Systems

Different explanation models have their unique advantages and disadvantages. The selection of a framework will depend on the domain characteristics, application constraints, required hidden layer involvement, and restrictions including reliability. There are different technologies to enhance py justice and reliability beyond a reduction of coherence and transparency in AI processes. Research is now moving towards enhancing explainability through human-interpretable explanations provided in the form of generative models that give a mental model of explainable representations, symbolic approaches to integrate interpreter-based controllers for explaining learned representational models, and few-shot contrastive explanations for enhancing explainability by positioning objects to prove the decision of the AI system. The presentation of concept-driven explanations for human-interpretable relevance spaces provided multifactorial explanations as used by concepts for visual explanation. In contrast, a reliable system can be understood by the decision logics and the structures of multimedia events. In human science, the automatizing and automating of explainable AI are multi-interdisciplinary areas. Explaining the decision making of an AI

system facilitates trust for clients, provides apparent system capabilities, prevents overfitting, allows system error identification, and gives the opportunity to enhance the stability and generalization of an AI system. The idea of the interaction and the cooperation, being a question of explainability, presented aspects of two-hand psychology on the design of theories as for a computational model of the world, which could come into existence in cognitive science. [20]

[8] The explainability of an AI system plays a crucial role in interacting with human users which includes trustworthiness, credibility, acceptance, and safety in areas of autonomous vehicle networks. In addition, regulations like the General Data Protection Regulation (EU) and the California Consumer Privacy Act (USA) require the consideration of the black-box problem in AI systems. As shown in Figure 2, the explainability of AI systems can be broadly categorized into the following AI explanation model types: post hoc explanations, inherently interpretable explanations, transferring representation insights, and human-interpretable explanations. Post-hoc explanation models explain the output decision of the black-box model, also known as glass-box models or faithful explanations. Different framework families such as probing, perturbation, behavioral testing, gradient-based methods, local model approximation-based methods, or simpler and more interpretable models are used to generate post-hoc explanations. Examples of post-hoc explanation techniques are Layer-wise Relevance Propagation, LIME, SHAP, Gradient-based Saliency maps, and Integrated Gradients. Inherently interpretable explanation models ensure the transparency and interpretability of each individual component in the decision-making process. Examples of inherently interpretable explainability models are decision trees, rules, decision sets, rules sets, linear models, generalized linear models, and generalized additive models. Transferring representation-insight explanations quantify the difference between local patches to determine their relevance on the output. Human-interpretable explanations are extracted directly from the feature space of the input and thus are employment designed to annotate features as a specification for a given AI model used [21].

4.2. Interpretability Techniques for Deep Learning Models

All gradient based methods have at least two things in common. First, the gradient of the output class with respect to the input sample is used as an indicator of the importance of the input elements and second they are relational and former is less reliable and misleading, and

because of that weaker assertiveness and trustworthiness [22]. In this regard, for a practical consideration eSolo is the most suitable data-driven method for off-line monitoring process. It is completely based on the input data and the model's predictions without any need to use information about the training data and the weights of the considered neural network. Also, eSolo shows its worth with weakest assertiveness.

Another form of explainability is post-hoc explainability, where an explanation is provided after the network has been trained using saliency methods, or input perturbation methods such as LIME, DeepLIFT, and SHAP, or by producing a visual map over the input image by projecting the gradient of the output class with respect to the input sample [23]. Grad-CAM is an example of the first kind, LIME, DeepLIFT, and SHAP represent the second kind, and XAI methods like eSolo and DeepLIFT are the third type of post-hoc explainability methods. In Grad-CAM, for monitoring, the network's final convolutional layer is modified to compute the behaviour of the final convolutional layer (MLC) for a given feature and the target class.

The goal of interpretability techniques is to make black box models such as deep learning networks easier to understand. One form of explainability is intrinsic explainability, where the decision boundaries in the dataset are human-relevant [24]. When a model's decision is concordant with a humans, the model is intrinsically explainable.

4.3. Trade-offs between Accuracy and Interpretability

Often, the depth of a CNN is considered to influence the sensitivity of the model to adversarial features. It is natural to reckon that a deep-learning model with higher layers would extract more intrinsic adversarial features. Conversely, we can surmise that the sensitivity of a L = 2 model towards noise is less than L = 3 models. However, Figure 3 shows that when a network contains a lot of adversarial features, the first few layers can be forced to identify adversarialities at high sensitivity [25]. It should be noted that linear training and robustness testing in a complex network are effective in understanding these shrewd attacks when the malicious instructions have a uniform distribution. Otherwise, it will be more challenging to train a deep CNN with robust inputs, as the training set would be dominated by attacks that have a high similarity and common characteristics which only pertained successfully de-novo on generation.

In cybersecurity applications, understanding how an attack might propagate by leveraging on the intricacies of the network architecture is complex. On the one hand, a system should be able to detect zero-day attacks with high accuracy, which requires models to be robust and to have deep-layers for high-level feature extraction [7]. At the same time, these models should be simple and interpretable, mainly for the ease of problem-solving and easy root-cause diagnosis. It has been observed that the hierarchical nature of the adversarial instructions in phishing networks can be exploited to improve the accuracy of a given NN. Conversely , a) greater layers could increase the sensitivity of the algorithm to 0-day adversarial instructions and b) the trade-off, considering the former point, could be compensated with ensemble learning, but ensemble learning significantly involves complexity, as every time new labels are added.

5. Explainable Deep Learning Models

One possibility is to use the complex deep networks and transform upper-layer abstraction to the structured under-level and combined it with the existing machine learning (ML)/Deep Learning (DL) technique to achieve a better performance vehicle single-step or end-to-end autonomous systems with explainability. As other models of expression using Deep learning and other de-facto machine learning models for prediction and classification in cortical networks were very idiomatic and complacent to concerns like data bias, outliers, adversarial attacks etc. Hence, a feasible solution that rectifies the issue is explained. This review paper highlights the various methods of Vitality Network which looked solely into the explain ability with deep learning. A study investigating explainable deep learning models towards Midwifes Support System was proposed. A study utilizing the ResNet and the SqueezeNet to predict which images contain diseases towards Fundoscopy Image Systems was proposed. To Improve the Communication layer Security for Vehicular Ad Hoc Network an IDS was proposed for using VIPNet and UNet. A study dealing with Intrusion Handling in Time Multiplexed CAN Network connected to Autonomous Vehicles via LSTM and PCA Models were introduced. Not only this, but the same model has been extensively used to model the CAN network layer.

Vehicular networks have become vital in current times and are widely used in smart vehicles to make the driving experience safer and easier. These smart vehicles consist of various components such as sensors, controllers, and actuators connected through wires. The

transmission of data through these networks results in various security threats. Thus, the intrusion detection system (IDS) was proposed to mitigate these security concerns [26]. Vehicular networks are prone to different types of attacks such as the denial of service (DoS), the man in the middle (MitM), and the replay attack, which are deteriorating the system performance and leads to catastrophic events on the road. The Cyber-intrusion anomalies are categorized into misuse detection systems and anomaly detection systems commonly used within autonomous vehicle networks. With the advent of Deep Learning and Data Analytics, numerous networks (DT, LR, SVM, XYZ etc.) were proposed used to learn the signature of attacks and were able to result in high missed detection rate during the detection of new, unknown attacks. The use of hybrid IDS systems where the sensors to anticipate the anomalies or how the anomalies interact with the protocols to mitigate the attack [17]. Smart Systems designer riotously use black-box models, such as Support Vector Machines, Neural Networks, and Deep Learning techniques to classify certain outcomes without understanding the output model. Deep Neural Networks (DNN) are proficient at learning hierarchical data explanation extract abstraction handcrafted at a higher level [27].

5.1. Interpretable Neural Networks

Recently, a novel approach that combines the attention mechanism, CNN and Long Short-Term Memory units (CNN-BiLSTM) on network intrusion detection (NID) method in a cloud computing environment has been proposed to overcome these existing shortcomings. In this approach, new attack changes, such as DoS or using previously unseen services, altered or unknown tactics of the attack, may not use the full state transition operations of the modeled CNN, so the attention mechanism is required to determine the importance of each region in the inputs. Two sets of separate CNNs are trained from the original and constructed data [25]. Specifically, CNN-BiLSTM models are able to learn and integrate the varied uncovered features in spatial manner. The attention mechanisms capture local and global features on feature maps and feature sequences produced by the CNNs and BiLSTMs. The attention mechanisms enable CNN-BiLSTM models to identify informative regions and hidden states for predicting the network intrusions, respectively. A cloud computing system is created as an extension of the well-known KDD99 to evaluate the CNN-BiLSTM intrusion detection method. The proposed CNN-BiLSTM is validated by experimental results wherein it outperforms other state-of-the-art network intrusion detection systems such as DLSTM and MIML. The source code and all the datasets are available for reproducibility purposes. The

experiments show that the proposed system is promising in anomaly detection and does a good job in classifying the types of attacks [27].

Commonly used deep learning approaches for intrusion detection recordings include distributed neural networks, BP neural networks, CNN-based approaches, RNN-based approaches, and deep belief network-based approaches [28]. However, these approaches have some shortcomings such as feature extraction, learning, and adaptability. As an important branch of the state-of-the-art deep learning approaches, the attention mechanism effect in network intrusion detection is still weak. Attention mechanism has emerged as a powerful and general concept for modelling variable-length variants by modelling dependencies in the input. Inheriting the generalization power of CNN, the global CNN captures essential text representations that are useful for predicting local context, local CNN captures local context features that should be preserved as individual patterns, whereas RNN captures an interaction between the features and mixing the information from all positions of the sequence. Transformer-based methods have been employed extensively to solve network intrusion detection before.

5.2. Decision Trees and Rule-Based Models

Another research model investigated K-Nearest Neighbour (KNN) which enhanced the VisaNetIDS accuracy for fraud detection but these traditional ML-based IDS models were not sufficient to detect new unseen attacks. This paper extends the ideas from, which was ranged the predictive performance of RF and SVM as a learning model to detect cyber-attacks on the network security. To address these challenges, many researchers developed deep learning networks (DLN) for network intrusion detection to improve the learning process and detection rate capabilities [28]. For instance, Authors proposed that deep belief network outperforms all the traditional and advanced machine learning algorithms to detect cyber-attacks. Moreover, developed an auto-encoder based approach to detect unseen attacks to overcome dependency on the training dataset. The work in this paper is closest to the work of with improved intrusion detection model (IDM) predictive capabilities and accuracy based on two intrusion detection deep learning models such as bi-directional long short-term memory (BiLSTM) and deep convolutional neural network (CNN).

Several research studies applied traditional ML-based intrusion detection models but with different learning paradigms to detect cyber-attacks [10]. These include optimised Random

Forest-based multi-stage ML model to detect different types of attacks. Besides, the SVM-based model to detect unsupervised network traffic was examined applying PCA during the feature extraction phase. The work of is closest to this study, but the proposed IDS approach was developed for detecting attacks occurring in the intra-vehicle networks.

5.3. Attention Mechanisms in Deep Learning

In general, by aggregating multiple sources of information with different importance in multimodal tasks, attention mechanism-based methods produce models with higher performance. This remarkable trait is the reason why it has been exercised in several domains such as medical diagnosis, computer vision, and natural text processing to capture useful features from multimodal data. Since suitable explanation of an AI model is essential in many applications like competitive networks, explainable artificial intelligence, and affective computing, it is essential to have improved attention mechanism-based methods to achieve human-explainability for AI models. [8]

Deep learning models look like a black box and cannot identify exactly what features are learned by neural networks. It is challenging to analyze why these techniques have been conducted prediction. To attenuate this pitfall, attention mechanism is widely utilized in explainable deep learning models to offer the model opacity. This mechanism is capable of weighting input data for AI model; therefore, it can provide importance of each part of input for all output. Such attention-based technique has shown improvement in traditional works, so it seems to be a good practice in multicultural and vulnerable environments. In this direction, substantial attention mechanisms are presented by researchers because they confer deep networks with human-readable results. [3]

6. Case Studies and Applications

Configurations are provided to achieve improved results even in critically adverse environments. Each set of machine learning-based systems aims to detect adversarial attacks. Experimental results showed the HDDIM and attack detection architectures to be environment-robust approaches and provided consistently better detection performance than the state-of-the-art methods in terms of accuracy, the precision-recall curve, and several performance metrics. An extensive discussion and analysis of the limitations and advantages

of the proposed models, as well as an investigation of the use of learning-based methods with varying quantities of limited traffic dataset, are outlined in the discussion section .

The proposed HDDIM offers novel SSL-based intrusion detection models that consist of a hidden deep learning architecture for the process of feature extraction and an interpretable machine learning algorithm for the prediction stage . The first model, Hybrid Deep Learning at both stages (HDL-HD), involves a CNN-based depth-wise separable convolution in the feature extraction stage, which allows for detecting subgraph semantic properties typical in intrusion. The dense layer with a rectified linear unit (ReLu) activation function is used as the second-level classifier. In the second model, Multi-stage Safety-Optimized Input (HDL-MO), clustering with k-prototype is used to generate the first stage hidden inputs using the DUoS algorithm, and a Histogram of Oriented Gradients (HOG) is used to add a safety primitive in the second stage feature extraction. Finally, BPNN with an exponential linear unit (ELU) activation function is used as the second-level classifier. Moreover, an attack detection architecture is implemented that detects the adversarial attacks on the HDL-IDS model .

6.1. Real-world Applications of Explainable Deep Learning in Intrusion Detection

Moreover, Deep learning models seem not to be so black-boxes as envisioned, in particular when specialized techniques have been developed to make interpretability straightforward by design. These approaches can be single-handedly adopted to address the second requisite, namely, the model transparency. It is worth noting that all these proposals have only been considered for the general application of cyber-attacks injection in a vehicle network designed with the specific goal of making embedded security recipe validation. In this perspective, in the field of software development, the employment of these explainable deep learning models and, in general, of the related interpretability explanations, could be particularly applicable from the pilot production stage, close to the final software/vehicle version.

Authors outlined how explanations for models from this space can enable actionability so that defenders can understand and mitigate their impact – an important requirement in real-world intrusion detection applications.In recent research, Deep learning models have been showing interesting results in detecting and diagnosing cyber-attacks aimed at both software and embedded automotive systems, applying different LSTM, Inception-ResNet and CNN architectures. [10] Researchers are developing state-of-the-art models to detect cyber-attacks and accurate prediction of attacker behavior. But, not all the proposed approaches offer a high

level of interpretability and explainability requirements, which are crucial needs in order to make clear how these deep learning models behave during the inference process.

The availability of trustworthy and explainable AI technology for intrusion detection in autonomous vehicles is essential to promote the safe and reliable operation of these advanced technological infrastructures. [17] [19] Advances in deep learning technology have motivated approaches to intrusion detection centered on deep learning models. Meanwhile, the increasingly unsupervised and decentralized operation of autonomous vehicles motivates intrusion detection models which can operate with little or no offline supervision, thereby mirroring the very same characteristics responsible for the unsafe behavior of malicious hosts. In that regard, Indranil Gupta presented an interesting design space that can inspire real-world research and development efforts on deep learning-based intrusion detection systems.

6.2. Performance Comparison with Traditional Methods

Machine learning is applied to network category in the form of decision trees, ensemble learning, K-nearest-neighbors (KNN), naive Bayes, principle component analysis, random forest, and support vector machines. Pieces of evidence from the research community advocate deep learning for detecting network attacks. Xie et al. utilized deep autoencoders for representing the network data and used a softmax classifier to recognize the attack patterns, and the positive results brought attention of the research community to the deep learning methods [29]. Kang et al. presented a deep-learning-based intrusion detection system, named deepIDS, which simultaneously utilized a hybrid model of autoencoders and a recurrent autoencoder to capture the temporal properties among the various features. Deep learning takes constructive advantage from parallel computing and big data analysis for achieving an era of efficient and improved decision-making systems. The training of the network was performed by using a huge dataset i.e., NSL-KDD and unsupervised feature representation was learned. Deep learning-based techniques are CNN, denoising auto-encoder, efficient-learning algorithms, novel framework, and transfer learning techniques. Wu et al. have used VGG16 and ResNet as pre-trained models for transfer learning in the network intrusion detection process [9]. Python's Django, Flask, and Tornado are used for web applications, whereas for mobile and edge devices, developers explore Tensorflow lite, Keras, and Pytorch frameworks for detection of scheme of new type of attacks. Computer networks, web servers,

storage systems, and infrastructure components such as databases and operating systems are equally prone to cyber-attacks as endpoints and applications.

Traditional approaches include anomaly-based techniques like artificial immune systems, fuzzy logic, and machine learning techniques [6]. Machine learning techniques, such as decision trees, support vector machines, and K-nearest-neighbors achieve lower detection rates but consume less time. The major shortcoming of traditional methods is the requirement to have predefined features, exhaustive labeled datasets, and a lack of capability to adapt to newer variation of network attacks. On the other hand, when complexity increases the performance of these methods decreases. Convolutional neural network (CNN) has been proved effective and efficient in big data processing and feature extraction. The efficiency of the CNN model-based methods has been justified over other machine learning techniques for recognizing various kinds of attacks. CNN has good performance and lower detection rate. It works on segmentation of input data and learns the local space structure of the data. Recursive methods like recurrent neural network (RNN) and long short term memory (LSTM) are found good for time series data, and it also contains information about the passage of time, which makes RNN and LSTM more suitable for network intrusion detection. Tensorflow lite and Pytorch are the two popular tools for performing deep learning. While, fully-connected (FC) networks are used for working in a limited area and fine tuning, researchers use lightly different approach using transfer learning, which can reduce the resource consumption and achieve better results. Deep auto-encoders are for the extraction and reconstruction of features, and deep belief networks are for the unsupervised learning of the features.

7. Challenges and Future Directions

[10] Intelligent vehicles equipped with proper sensors are gaining momentum around the world, with step-by-step progression in the research and developmental works for Autonomous and Electric Vehicles (AEV) too. Several types of cyber threats can be launched to attack AEVs and some dexterous defense mechanisms can counteract in the potential cyber-attacks on AEVs as well. Intrusion Detection System (IDS) is one such essential defense mechanism. Therefore, the last few years have observed intensive research efforts to develop innovative IDSs using deep learning (DL), machine learning (ML) or hybrid DL-ML techniques for AEVs. In the coming days, the existing DL based IDSs in AEV will face several challenges and there will be some probable future research directions.[9] A majority of the

existing research works on Autonomous Vehicles (AV) security have focused on developing Intrusion Detection Systems (IDSs) using various deep learning (DL) and Machine Learning (ML) based methodologies. The researchers have proposed and developed several IDSs for both intra-vehicle and inter-vehicle Dense Area Networks (DAN) of AV. As the deep neural networks (annuities) used in these IDS schemes are not explainable, they can only provide a black-box model to make predictions but not an elucidating explanation for the detection decision. The growing need to explain the AV's behavior has raised a concomitant consideration of the trustworthiness of the AVIDM. For years, people have been gaining trust in the Driving Models' prediction through reasoning. Hence, the need arises to make deep learning-based AV IDS Explainable to enhance the trust of the former even in this study, we volition genuine any AVR-I-)MS following some of the extensive previous research's approaches. It can spectrum to the other study direction in the future, to have broad diversity in some feature spectrum observational clusters. As for future work, the lidar features, ladar reduction, prediction, multi-lidar fusion, and its consequences will be examined and hybrid AVIDMS approach for Future research and work is indeed very valuable as we do face quite a few challenges in this piece of research.

7.1. Ethical and Legal Implications of Intrusion Detection in Autonomous Vehicles

One of the most promising ideas regarding the development of autonomous vehicles gives the owner the moral flexibility to set the ethical and legal implementation of these vehicles. Despite its potential advantages, the suggested protective role of driver facilities will ultimately encourage undesirable outputs, such as significant urban transport problems related to increased convictions. Under various moral situations and with concrete quantitative constraints, computer-related behaviour in violent modelling can be measured and advanced. Since a machine's computers are not enveloped by any biological frameworks, their actions fall due to lack of human understanding, and their operators are to comply. An AI system that is entirely rational and totally not conscious is built to explain natural phenomena and generate novel variables in this source. AI behaves according to how it is designed by its creator, as long as it functions [30]. However, there is no possible extension to the world of comparison despite the great success of AI. In reality, machine-learning models can accept highly small targeted modifications of their inputs in various gravitational or stochastic conditions, known as "dangerous" inputs, making misuse of such models in any security-relevant framework much easier.

Interdisciplinary research with an optimal balance of explainability and accuracy in Artificial Intelligence (AI) has significant implications in diverse fields. The focus of this survey is to present an explainable LeNet-based intrusion detection system for autonomous vehicle INTranet (UTOVIN) using a longitudinal dataset. UTOVIN consists of a CAN bus from which features are extracted to capture the vehicle's regular and abnormal activities. The main purpose of this survey is to detect network intrusions in a controlled and benign environment of the continuously evolving Internet-of-Vehicles (IoV) using a feedforward neural network [12]. The model is defendable against repeated adversarial attacks and guarantees signal privacy and integrity. The model is protected against novel and unseen adversarial samples from the same family. The deployment of UTOVIN would enhance the real-time security and safety of IoV deployed systems such as connected autonomous vehicles, electrically-powered vehicles, drones, and Internet-of-Things (IoT) devices such as vehicle infotainment or telematics.

7.2. Future Research Directions in Explainable Deep Learning Models

Future works should also delineate those autonomous vehicles ability to securely share their data. Specifically, distributed learning methods should be utilized that are able to acquire input data for learning deep learning models from multiple units in the ecosystem, and securely consolidate the models obtained over all distributed data at a central server in interconnected vehicle networks. Development of secure, explainable and interpretable machine learning techniques, where occasional threats (attacks and malware data) constantly target any part of the deep learning models or datasets, is another area for future research, that is lacking currently. The developments in this research area can pave the way for an explainable and secure learning technique that are practically more employable in the autonome vehicle networks, thus filling the existing technological gap.

Given the lethal nature of the impacts and the lack of explainability in decisions made by deep learning models, to secure operating conditions of autonomous vehicles against the adversaries better, new methods are required that provide explanations for the decisions made as well as that are robust against potential adversarial attacks. The deep learning models that are deployed in autonomous networks should provide their decisions in such a way that a security analyst, system administrator or operator of the autonomous network could easily understand reasons behind the decisions made by the deep learning models. Further research

directions should consider the massive nature of data and security-critical decision-making in autonomous driving mission. [31] has discussed privacy-preserving solutions for data as future research aspects for machine learning techniques in the connected autonomous vehicle systems.

Deep learning has been frequently used in various autonomous vehicle networking applications, especially in the security domain, such as intrusion detection systems (IDS). The authors in [26] highlighted the importance of deep-learning-based IDS for the automotive sector but without providing methods on how such deep-learning-based approaches are subject to adversarial attacks. Hence, explainability and robustness should be identified and addressed for deep-learning-based IDS for autonomous vehicles. Another avenue for future research is to integrate the machine learning methods to detect whether a vehicle or infrastructure has been infected by possible attacks. [13] discusses the possible implications and future work challenges in automated intrusions detection in power systems namely, attack forecasting methods and methods of network structure interpretations and network state interpretability.

Reference:

1. Tatineni, S., and A. Katari. "Advanced AI-Driven Techniques for Integrating DevOps and MLOps: Enhancing Continuous Integration, Deployment, and Monitoring in Machine Learning Projects". *Journal of Science & Technology*, vol. 2, no. 2, July 2021, pp. 68-98, https://thesciencebrigade.com/jst/article/view/243.
2. Shahane, Vishal. "Evolving Data Durability in Cloud Storage: A Historical Analysis and Future Directions." *Journal of Science & Technology* 1.1 (2020): 108-130.
3. Abouelyazid, Mahmoud. "Comparative Evaluation of VGG-16 and U-Net Architectures for Road Segmentation." *Eigenpub Review of Science and Technology* 6.1 (2022): 75-91.
4. Prabhod, Kummaragunta Joel. "Advanced Techniques in Reinforcement Learning and Deep Learning for Autonomous Vehicle Navigation: Integrating Large Language Models for Real-Time Decision Making." *Journal of AI-Assisted Scientific Discovery* 3.1 (2023): 1-20.

5. Tatineni, Sumanth, and Sandeep Chinamanagonda. "Leveraging Artificial Intelligence for Predictive Analytics in DevOps: Enhancing Continuous Integration and Continuous Deployment Pipelines for Optimal Performance". Journal of Artificial Intelligence Research and Applications, vol. 1, no. 1, Feb. 2021, pp. 103-38, https://aimlstudies.co.uk/index.php/jaira/article/view/104.