T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

DEPARTMENT OF COMPUTER ENGINEERING

ADVERSARIAL ATTACK TRANSFERABILITY IN EXPLAINABLE INTRUSION DETECTION

MASTER'S THESIS

ABDLELAH ABDLATEF ABDLHAMED

ISTANBUL 2025

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

DEPARTMENT OF COMPUTER ENGINEERING

ADVERSARIAL ATTACK TRANSFERABILITY IN EXPLAINABLE INTRUSION DETECTION

MASTER'S THESIS

ABDLELAH ABDLATEF ABDLHAMED

THESIS ADVISOR

ASST. PROF. DUYGU CAKIR YENIDOĞAN

ISTANBUL 2025

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

2025/01/10

MASTER THESIS APPROVAL FORM

Program Name:	COMPUTER ENGINEERING
Student's Name and Surname:	ABDLELAH ABDLATEF ABDLHAMED
Name Of The Thesis:	ADVERSARIAL ATTACK TRANSFERABILITY IN EXPLAINABLE INTRUSION DETECTION
Thesis Defence Date:	2025/01/10

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Prof. Yücel Batu SALMAN

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Institution**	**Signature**
Thesis Advisor's	Asst.Prof.Duygu CAKIR	Bahçeşehir University	
Member's	Asst.Prof.Umit Ozturk	Istanbul Gedik University	
Member's	Asst.Prof.Erkut Arican	Bahçeşehir University	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Abdlelah Abdlatef ABDLHAMED

Signature:

ABSTRACT

ADVERSARIAL ATTACK TRANSFERABILITY IN EXPLAINABLE INTRUSION DETECTION

Abdlelah Abdlatef Abdlhamed

Master's Program in Computer Engineering
Supervisor: Asst. Prof. Duygu Cakir Yenidoğan

January 2025, 98 pages

The Intrusion Detection Systems built with machine learning techniques remain susceptible to adversarial attacks because such approaches modify input data to overcome system detection mechanisms. The research examines attack transferability within IDS models against LightGBM and XGBoost using both NSL-KDD and CICIDS-2017 datasets. XAI techniques SHAP and LIME help detect essential features that guide model decisions while also making possible the assessment of how vulnerable these features become against adversarial attacks. The heuristic search method generates adversarial samples which attack essential features. The experimental researchers applied adversarial examples to IDS models for determining transferability across multiple IDS solutions which exposed common vulnerabilities in decision boundaries. Experimental analysis reveals a corresponding 6-8% decrease in accuracy which proves that all models are vulnerable to attack. Feature interpretability changes when adversarial manipulations occur based on findings from SHAP and LIME analyses. Research investigates two defensive strategies involving adversarial training and optimized feature selection to enhance IDS resistance against attacks. The research results point to the required development of IDS systems which provide explainability capabilities together with resilience against adversarial threats yet maintain defensive transparency throughout the decision process. The study adds to the advancement of safe interpretable IDS solutions which can meet requirements in real-world cybersecurity operations.

Key Words : Adversarial Attacks, Intrusion Detection System (IDS), Explainable Artificial Intelligence (XAI), Shapley Additive Explanation (Shap), Machine Learning (Ml) .

ÖZET

XAI TABANLI IDS'LERDE ADVERSARYAL MAKINE ÖĞRENMESI SALDIRILARI VE TRANSFER EDILEBILIRLIĞI

Abdlelah Abdlatef Abdlhamed

Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Asst. Prof. Duygu Cakir Yenidoğan

Ocak 2025, 98 sayfa

Makine öğrenimi teknikleriyle oluşturulan Saldırı Algılama Sistemleri, bu tür yaklaşımlar sistem algılama mekanizmalarını aşmak için giriş verilerini değiştirdiğinden, saldırgan saldırılara karşı hassastır. Araştırma, hem NSL-KDD hem de CICIDS-2017 veri kümelerini kullanarak LightGBM ve XGBoost'a karşı IDS modelleri içindeki saldırı aktarılabilirliğini inceler. XAI teknikleri SHAP ve LIME, model kararlarını yönlendiren temel özelliklerin algılanmasına yardımcı olurken, aynı zamanda bu özelliklerin saldırgan saldırılara karşı ne kadar savunmasız hale geldiğinin değerlendirilmesini de mümkün kılar. Sezgisel arama yöntemi, temel özelliklere saldıran saldırgan örnekler üretir. Deneysel araştırmacılar, karar sınırlarındaki ortak güvenlik açıklarını ortaya çıkaran birden fazla IDS çözümü arasında aktarılabilirliği belirlemek için IDS modellerine saldırgan örnekler uyguladılar. Deneysel analiz, tüm modellerin saldırıya karşı savunmasız olduğunu kanıtlayan, doğrulukta karşılık gelen %6-8'lik bir azalma olduğunu ortaya koymaktadır. SHAP ve LIME analizlerinden elde edilen bulgulara göre, saldırgan manipülasyonlar meydana geldiğinde özellik yorumlanabilirliği değişir. Araştırma, saldırılara karşı IDS direncini artırmak için saldırgan eğitim ve optimize edilmiş özellik seçimi içeren iki savunma stratejisini inceler. Araştırma sonuçları, düşmanca tehditlere karşı dayanıklılıkla birlikte açıklanabilirlik yetenekleri sağlayan ancak karar süreci boyunca savunma şeffaflığını koruyan IDS sistemlerinin gerekli gelişimine işaret ediyor. Çalışma, gerçek dünya siber güvenlik operasyonlarındaki gereksinimleri karşılayabilen güvenli yorumlanabilir IDS çözümlerinin ilerlemesine katkıda bulunuyor.

Anahtar Kelimeler: Saldırgan Saldırılar, Saldırı Algılama Sistemi (IDS), Açıklanabilir Yapay Zeka (XAI), Shapley Eklemeli Açıklama (SHAP), Makine Öğrenimi (ML)

To those who have had the greatest credit in my academic career,

To my dear parents

Who illuminated my path with prayers, and were patient throughout my long journey,

And instilled in me the determination and belief in my ability to succeed.

Every word of encouragement and every moment of support was fuel to reach this day.

To my brothers and sisters

My support in life, who were always by my side every step of the way, providing me with energy and love.

To my distinguished teachers,

Lamps of knowledge and science, who gave me their invaluable time and effort, and planted in me the love of research and knowledge, and guided meto excellence and creativity.

To my friends and colleagues,

My companions on the road who shared challenges and beautiful moments with me,You were the best support in my career.

To everyone who inspired me with a word, supported me with a smile, or believed in My abilities.

I dedicate this modest achievement to all of you, because you are the foundation of this success.

With love and gratitude,

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Asst. Prof. Duygu CAKIR, whose invaluable guidance, expertise, and patience have been instrumental in shaping this thesis. Their insightful advice and unwavering support have motivated me throughout the research journey.

Additionally, I extend my heartfelt thanks to Lect. Marwa Issam Abdulkareem ABDULKAREEM for her invaluable guidance, feedback, and support throughout this journey. Her expertise and encouragement have greatly enriched this work.

I am also profoundly grateful to Bahçeşehir University for providing the resources, academic environment, and opportunities that enabled this work to come to fruition.

I am immensely grateful to my parents, whose love, prayers, and encouragement have been a constant source of strength. Their unwavering belief in me has been my driving force. To my siblings and friends, thank you for your understanding, companionship, and moral support during this challenging yet rewarding journey.

I also extend my deepest appreciation to my dear friend MSc. Mohammed Basim Mohammed MOHAMMED, whose advice, encouragement, and unwavering support have been a source of inspiration and motivation throughout this process.

My sincere gratitude goes to IALD for granting me the scholarship and their continuous support, which has been instrumental in enabling me to pursue and complete this work successfully.

This achievement would not have been possible without the support and encouragement of all these individuals. For this, I am truly grateful.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

NIDS	Network Intrusion Detection System
XAI	Explainable Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
AEs	Autoencoders
GAN	Generative Adversarial Networks
AML	Adversarial Machine Learning
SVMs	Support Vector Machines
JSMA	Jacobian-based Saliency Map Attack
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
BDA	Big Data Analytics
CNNs	Convolutional Neural Networks
PGD	Projected Gradient Descent
NSM	Network Security Monitoring
HIPS	Host-Based Intrusion Prevention System

Chapter 1

Introduction

1.1 Background

In recent years, there has been a growing trend of using machine learning (ML) for detecting malicious activities within network intrusion detection systems (NIDSs). This is because ML is so good at finding unknowns and high accuracy. But state of the art NIDSs are very susceptible to adversarial attacks. These attacks add small perturbations to malicious traffic and it gets by detection. Research has shown that even simple attacks can drop NIDS accuracy by a lot and we can't trust ML based NIDS for real world critical applications.

NIDS attacks can be categorized into white-box and black-box attacks. White-box attacks assume the attacker has full knowledge of the target NIDS which is not realistic in real world. Black-box attacks are not very good in achieving high success rate because of limited adversarial transferability between different models like neural networks and decision tree models. Neither type of attack provides clear explanation of why adversarial examples occur or how they can transfer between different models.

Given these challenges, we need to develop robust ML-based NIDSs that can withstand adversarial attacks and be reliable in production. Understanding why adversarial examples exist and are transferable across models is key to improving NIDSs against attacks.

To make ML-based NIDSs more robust against attacks, we need to develop methods that improve their resilience. One way to do this is by generating adversarial examples (AEs) for attack evaluation. Research on creating AEs for ML-based NIDSs usually happens in white-box settings where attackers have full access to the model's internals (Sauka, 2022)(Okada, 2024). In a white-box setting, the assumption of full

model transparency gives attackers full access and control over the targeted NIDSs. However, in real-world scenarios, practical NIDS systems don’t disclose their internal configurations so white-box attacks are not practical. In response, researchers have developed various techniques to craft adversarial examples (AEs) in black-box settings. Black-box attacks are generally divided into two types: query-based attacks and transfer-based attacks (Rosenberg, 2021) .

Query-based attacks can reach success rates similar to white-box attacks by asking the target model for detection scores. Yet, getting these scores from real-world NIDSs often doesn't work for attackers. So, the best way to attack NIDSs is through transfer-based attacks. In this method, attackers train a substitute model on their own without knowing the NIDS's underlying machine learning model. They then move the adversarial examples (AEs) created by the substitute model to other models, which helps them avoid detection.

IT security now relies heavily on intrusion detection systems (IDS) to protect against cyber threats that people and machines face every day. Many intrusion detection methods have used signature-based techniques, but these can't spot new emerging threats. As technology advances and traditional IDS fall short, we need to upgrade signature-based systems.

Over the last ten years, AI-powered intrusion detection systems in IoT networks have made big strides in tech capabilities. Using deep learning and machine learning has brought new problems since sneaky tweaks can mess with how well IDS works. Also, we need to improve how we spot weird stuff to deal with things like uneven or missing data, which make it harder to train IDSs. (Roshan, 2024).

Anomaly-based intrusion detection systems (IDS) try to build a picture of normal behavior and spot activities that don't fit this picture. Deep learning algorithms prove useful for this task. Generative Adversarial Networks (GANs) have caught the eye of researchers since they came on the scene in 2014. People have looked into them a lot and put them to use in finding unusual patterns. This is because GANs are good at making and learning from complex data, like pictures, sounds, and written words.

The goal of this thesis is to create a system for adversarial attacks that is easy to apply broadly, easy to understand, and strong. We especially want to improve how attacks and defenses work for models that are not differentiable. To achieve this, we use ranked feature importance from XAI (Explainable Artificial Intelligence) methods. The thesis is called "Adversarial Attack Transferability in Explainable Intrusion Detection.

The authors introduce a new framework called "Defense Mechanisms for Explainable Intrusion Detection Systems." In this framework, intrusion detection is done using LightGBM. The system finds important features to create and defend against adversarial attacks using a formal explanation technique. It uses ranked features to do this. Specifically, the system can generate clear explanations at the instance level and is tested against attacks like Projected Gradient Descent (PGD). Additionally, the model is made more transparent and easier to understand using XAI techniques, while also being studied further.

1.2 Adversarial Machine Learning

Adversarial Machine Learning (AML) looks at how machine learning systems behave when they are attacked by people trying to make them perform worse or steal important information. These attackers can target any part of the machine learning process, as shown in Figure 1. This includes attacks on the data used for training, the process of training the model, the model itself when it’s being used, and even the phase where the model makes predictions. It’s important to understand and reduce these risks to make sure machine learning systems are reliable and secure, especially when they are used in environments where attacks might happen. AML assaults include digital assailants endeavouring to deceive man-made intelligence/ML models by acquainting painstakingly created information planned with incite erroneous forecasts or groupings. These assaults are normally characterized by the phase of the AI cycle at which they happen: information harming assaults, avoidance assaults, and derivation assaults. In information harming assaults, assailants infuse noxious information into the preparation dataset to ruin the model's way of learning. Avoidance assaults happen

after the model has been prepared, where aggressors adjust the model's boundaries or design to debase its presentation. In deduction assaults, which happen during the arrangement or expectation stage, foes control input information to impact explicit results or gain unapproved admittance to delicate data. **Error! Reference source not found.** depicts these AML attack types and their associated defense mechanisms, which are further explored and detailed below. Understanding and mitigating these vulnerabilities are essential for enhancing the robustness and security of AI/ML systems against malicious exploitation.

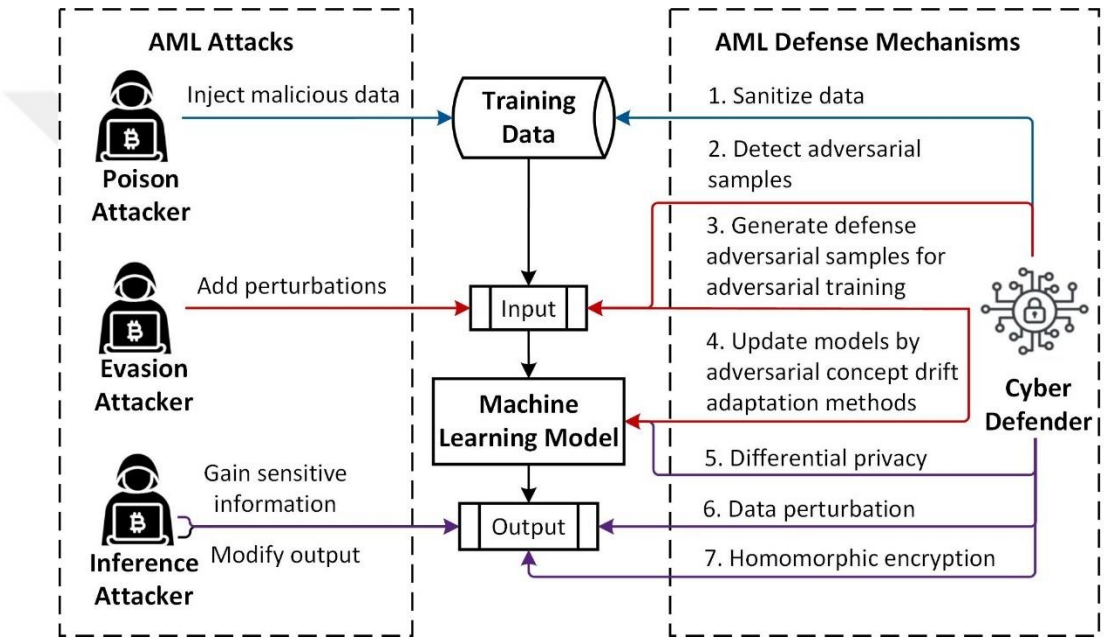

Figure 1. Adversarial Machine Learning (AML) attack and defense (He, Adversarial machine learning for network intrusion detection systems: A comprehensive survey., 2023).

1.2.1 Attack types. If we look at the problem from the attacker's perspective, we can define two types of attacks as poisoning attacks or evasion attacks. (Roshan, 2024) and (Bouaziz, 2023) describe different poisoning attacks. (Ojo, 2024) devised attacks against the Ridge and Lasso linear classifiers by maximizing the classification error with regards to the training points. In (Malik, 2022) , researchers conducted a poisoning attack specifically targeting Support Vector Machines (SVMs). Their approach involved injecting malicious samples into the training set with the objective of identifying points that would maximize classification errors. Evasion attacks were investigated in the studies by (Nwakanma, 2023) and (Oseni, 2022) . The latter,

referenced as (Arreche, 2024), introduced a method where a meta-classifier is constructed by training multiple classifiers on diverse training sets. This meta-classifier aims to extract statistical properties from the data rather than the actual features, posing a privacy attack. This approach highlights the exploitation of statistical characteristics to infer sensitive information, emphasizing the broader implications of evasion tactics in machine learning security.

1.2.2 Adversarial deep learning. Deep Learning has become very successful in the recent years in the field of Natural Language Processing and Computer Vision. This also led to the development of Adversarial Deep Learning which was initially centered around the Computer Vision domain. One of the first breakthroughs came in 2013, when (Moustafa, 2023) successfully demonstrated how one can fool Deep Learning classifiers by introducing small variations in an image. These variations were so small that they were imperceptible to humans but enough to fool the classifiers. Some examples of such images are shown in Figure 1.2. Another work (Wali, 2023), generated random images which appeared or had patterns in it, which did not mean anything, but the images were able to fool the Deep Learning classifiers into predicting them into valid object classes. Some reference images have been shown in Figure 1.3. Although many different explanations have been given for the reason as to why this is possible, (Chivukula, 2023) Contrary to common intuition, the primary cause lies in the high degree of linearity within the components of Deep Learning models. This linearity is largely enabled by the use of piecewise linear activation functions such as Rectified Linear Units (ReLUs). These activation functions not only accelerate the optimization process but also play a crucial role in forming decision boundaries that extend far beyond the areas covered by the training data. Consequently, classifiers employing these mechanisms may misclassify new examples, such as images, that possess specific characteristics, as noted in studies referenced by (Roshan K. A., 2024) and (Patil, 2022) . Another intriguing discovery highlighted in research by (Rosenberg, 2021) is that images exhibiting adversarial properties for one neural network can transfer these same properties to other neural networks that have been independently trained. This phenomenon underscores the broader implications of

adversarial examples and the potential vulnerabilities that can persist across different machine learning models.

The main models that have shown a protection from ill-disposed models are the Spiral Premise Capability (RBF) organizations however they are not utilized frequently as they don't sum up well (Beam, 2023) . Other than those, even shallow direct models are additionally impacted by a similar issue as are model troupes. The methods and algorithms to generate adversarial examples has also been researched upon. There are many such methods which have a trade-off on speed of production, performance and complexity. Some of the methods that have been proposed are given below –

• Evolutionary algorithms, proposed in (Oksuz, 2024). But this method is very slow compared to the other two alternatives.

• Fast Gradient Sign Method (FGSM) proposed in (Lundberg, 2022).

• The Jacobian-based Saliency Map Attack (JSMA), as discussed in (Afzal-Houshmand, 2023) , is computationally more demanding than the Fast Gradient Sign Method (FGSM), yet, it enjoys the benefit of creating ill-disposed examples with negligible contortion. Both FGSM and JSMA techniques plan to acquaint unpretentious irritations with the first example to actuate antagonistic impacts. In FGSM, the irritation δ is registered by working out the angle of the expense capability J regarding the information x. This angle shows the heading in which the information ought to be acclimated to expand the expense capability, accordingly creating ill-disposed models that can hoodwink AI models with negligible adjustment to the information.

$$\delta = \epsilon sign\left(\nabla_x J(\theta, x, y)\right) \quad (1.1)$$

where θ are model parameters, x is the input to the model, y are the labels associated with x, is a very small value and J(θ, x, y) is the cost function used when training the neural network. This method is very fast because it requires the gradient which can be computed very efficiently using backpropagation. The irritation is then added to the underlying example and the eventual outcome creates a misclassification. A model is displayed in **Error! Reference source not found.**.

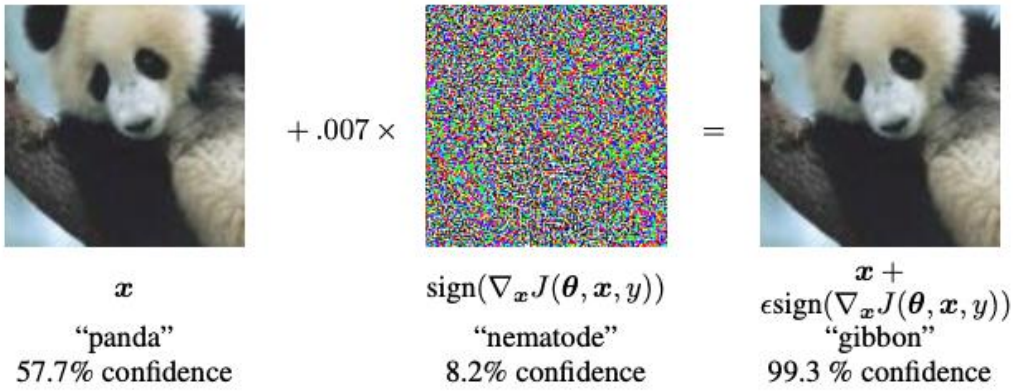

Figure 1. Generating adversarial samples with FGSM (Srivastava G. J., 2022).

JSMA, as the name recommends, produces antagonistic example bothers in light of the idea of saliency maps. The bearing responsiveness of the example with respect to the objective class is determined utilizing a saliency map. (Albahri, 2024) designed an efficient saliency adversarial map under the L0 distance (i.e. the number of features i such that x 0 i 6= xi) . The Jacobian matrix computed for a given sample x is expressed as :

$$J_f(x) = \frac{\partial f(x)}{\partial_x} = \left[\frac{\partial f_i(x)}{\partial_{xi}}\right] i \times j \qquad (1.2)$$

1.3 Adversarial Attacks

Adversarial attacks on image classification task aim to manipulate machine learning models by acquainting painstakingly created annoyances with the information. These irritations, impalpable to people, can make the model misclassify or produce inaccurate results. Officially, an ill-disposed assault can be portrayed as follow: Let M be a Machine Learning model and Ctrue an input for that model. Assume that Ctrue is correctly classified by the model: M(Ctrue) = ytrue. The AML methods are based on building an adversarial input Cadv adding a perturbation to Ctrue, in a manner that the difference between the two input is imperceptible but permit to product a wrong classification by the model M(Cadv)≠ ytrue. In brief, the scope of AML is to create an input Cadv in order to fool the model with a misclassification. A first

categorization of the attacks is between black box and white box methods and its based on the level of information that an attacker has about the target model:

• White Box Attacks: accept that the assailant has full admittance to the objective model's engineering, boundaries, preparing information, and, surprisingly, its inward operations. This detailed knowledge allows the attacker to devise more effective and targeted attacks. White box attacks are often more powerful because the attacker can exploit specific weaknesses in the model's structure and behavior.

• Black Box Attacks: expect that the assailant has restricted or no admittance to the inward subtleties of the objective AI model. They can interface with the model by giving information sources and noticing yields. The attacker may not know the architecture, parameters, or even the training data used to create the model. Despite this limited knowledge, black box attacks can still be effective. A key concept in black box attacks is the idea of transferability. This refers to the ability to generate adversarial examples on one model and then use those adversarial examples to fool a different, unknown model of the same type. This is possible because certain adversarial perturbations are effective across different models.

Furthermore, we can distinguish between three main attack methodologies (Arreche, 2024) :

• Evasion Attacks

• Poisoning Attacks

• Model Extraction

Each of them, acts on the various weaknesses of the model. Below is a brief description of each of these approaches.

1.3.1 Evasion attacks. Evasion attacks, also known as adversarial examples, consists in modifying input data in a way that leads the machine learning model to misclassify it. Adversarial examples can be generated through various techniques, such as the Fast Gradient Sign Method (FGSM) (Patil, 2022) , Jacobian-based Saliency Map Attack (JSMA), or the Carlini-Wagner Attack (Moustafa, 2023). These attacks often

exploit the linearity and sensitivity of machine learning models to small changes in input. Formally, it can be described as an optimization problem :

$$argmin||\delta_X|| \; such \; that \; M(X + \delta_X) = Y^* \quad (1.3)$$

Where δ_X is the perturbation added to the input X, M is the classifier model and Y^* is a target label (it can be a specific label or the goal may just be that Y^* is different than the correct label Y).

Evasion attacks seek to mislead AI/ML models by identifying and manipulating data samples that are particularly prone to misclassification, thus causing the model to make incorrect predictions. Unlike poisoning attacks, which affect the training phase, evasion attacks do not alter the model's learning process. Instead, they take place during the testing phase, where adversarial examples are introduced by applying subtle perturbations to the input data, ultimately causing the model to misclassify the manipulated instances (Satyanarayana, 2024).

1.3.2 Poisoning attacks. In contrast, poisoning attacks focus on disrupting the training process by injecting malicious data into the training set. The attacker's objective is to manipulate the model's behavior not only during training but also in the deployment phase. Poisoning attacks are typically divided into two categories: information harming, where the aggressor focuses on the preparation information itself, and model harming, where the assailant tries to change the model's boundaries or design straightforwardly. The attack can be launched by adding poisoned samples, modifying existing samples, or even by controlling the availability of training data.

In harming assaults, aggressors control either the information or the computer based intelligence/ML model during the preparation stage to control forecast results. Taking advantage of the need for progressing model retraining to oblige developing information appropriations, enemies jump all over chances to impact or debase the model noxiously. In order undertakings, assailants change the marks of information tests to actuate misclassifications by the ML model. Conversely, in regression scenarios, they modify input feature values to produce inaccurate outputs. These poisoning tactics encompass diverse strategies such as injecting additional data, manipulating existing data points, and corrupting logical processes within the model

structure (Jiyad, 2024).

1.3.3 Inference attacks. In adversarial machine learning (AML), deduction assaults, otherwise called exploratory assaults, are methodologies utilized to procure private data or experiences into an objective ML model, its feedback information, or its basic engineering. These assaults present huge dangers, especially in security-basic applications, for example, protection safeguarding AI models, where assailants try to remove delicate data about people or organization gadgets. Some types of attacks on machine learning models include model reversal, model extraction, and membership inference. Model reversal attacks try to recreate the data used to train the model by analyzing its outputs. Model extraction attacks try to understand how a model is built and how it works by studying its predictions or behavior. The goal is to create a copy of the model. Membership inference attacks, on the other hand, try to determine if specific data was used to train the model by analyzing the model's outputs and guessing whether that data was part of the training set. These types of attacks show why it's crucial to have strong security measures in place to protect machine learning models and the sensitive data they use from being misused.

1.3.4 Model extraction. Model extraction is a technique used to gather information or knowledge from a machine learning model that already exists. In this approach, an attacker attempts to make a copy or a close version of the original model by asking it questions and keeping track of its answers. The attacker interacts with the target model, sends it many queries, and records the responses. These responses are then used to create a model that is similar to the original one. Attackers use this method for various purposes. One key reason is to take private or confidential information from a machine learning model. For instance, a company might want to copy a fraud detection model from a rival bank to get ahead in the competition. Similarly, an attacker might seek to extract a facial recognition model used for access to a security system. Model extraction methods can compromise data privacy and security as the attacker can obtain sensitive information from the target model.

1.4 Defenses Against Adversarial Attacks

Many researchers have established a measure for protecting the ML classifiers based on the constant development of numerous attack strategies against the classification models. The defensive mechanisms are divided into two categories. They are reactive defense and proactive defense. The adversary analyses the classifier, then designs and launches the attack in a reactive defensive system. After analyzing the findings from the attack, the classifier designer offers a defense technique. In a proactive defensive mechanism, the classifier builds an adversarial attack based on the prior work and assesses the results of the attack. Then the designer proposes the defensive mechanism.

To control the poisoned data points, there are two basic types of defense mechanisms proposed by the experts. The first method is data sanitization which is used to detect the stained data and remove them prior to the retraining procedure of the classifier model. Sanitization methods include creating a promising domain and rejecting any data points that fall outside of it. The second one is to develop a robust classifier model that can withstand a security attack. These two strategies are used as countermeasures in the training phase. In testing phases, the following procedures are used as countermeasures:

1. To create a well-structured security assessment system based on Game Theory.
2. Data security and privacy techniques such as Cryptography, Differential Privacy and Homomorphic encryption are employed to protect the data.

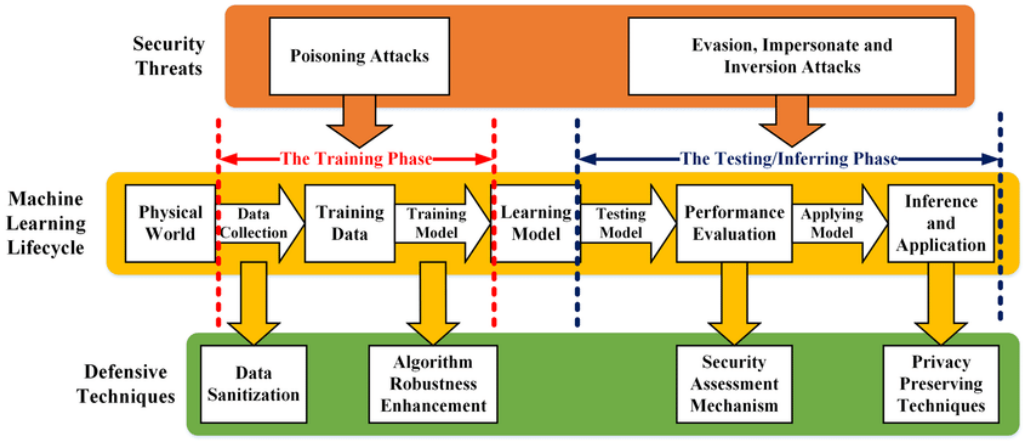

Figure 2 Illustration of Defensive techniques during training and testing phases (Jiyad, 2024).

1.5 Adversarial Attacks in Deep Neural Network Models

Similar to Machine Learning (ML), Deep Learning (DL) has become one of the most widely used techniques in computer science, with applications spanning natural language processing, image processing, pattern recognition, and many other areas. Despite the impressive performance of DL algorithms in tasks like malware detection, spam classification, object detection and tracking, face recognition, and autonomous driving, these algorithms and their training data are highly susceptible to various security threats. Adversarial attacks exploit this vulnerability by generating adversarial examples designed to deceive classifier models. For instance, in one scenario, attackers used adversarial examples to bypass a security system, granting unauthorized access and allowing them to alter the identity of the legitimate user.

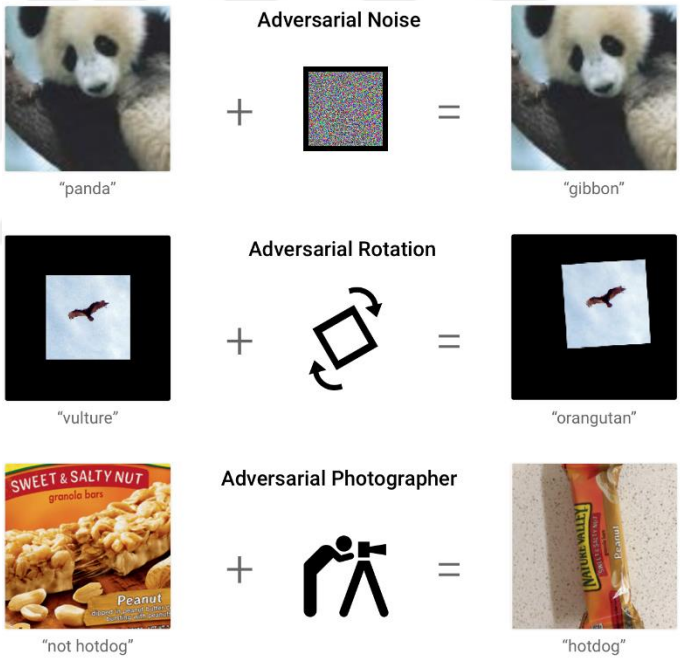

Figure 3 Classification Model illustration of adversarial examples (Malik, 2022).

As in Figure 3 there is a high degree of confidence in classifying two original images correctly (the school bus on the left side and the puppy on the right). The deep neural networks are having difficulty detecting adversarial examples. it is impossible to detect adversarial images generated by additive perturbation, and the same DNNs will completely misclassify them.

1.6 XAI and Explainability

XAI pertains to a specialized field within AI dedicated to enhancing the interpretability and comprehensibility of Machine Learning models." This nuanced definition underscores the importance of XAI in enabling humans to effectively anticipate and comprehend the results generated by complex machine learning systems.

Despite the proven consistency and trustworthiness of Machine Learning models over the years, blindly trusting them without understanding their decisions is problematic. The key issue is: "Why should we trust something we don't understand?" As AI becomes more integral to our lives, conflating 'prediction' with 'prescription' can lead to trouble, especially in high-stakes situations where understanding the model's reasoning is crucial. Providing explanations for individual predictions helps us align the model's decisions with our own knowledge of the world, making decision implementation or rejection more confident and consistent. However, the complexity of commonly used models, especially black-box models with unobservable inner workings, poses a significant challenge. Instead of attempting to decipher these models, we should focus on communicating their decision-making processes in ways that are engaging and understandable to humans (Kalutharage, 2022) . Counterfactual explanations are particularly effective in this regard, as they help us understand alternative outcomes and provide a basis for recourse.

XAI has long been a topic of interest, with its earliest work dating back several decades (Asaduzzaman, 2022). However, the past decade has seen a surge in Machine Learning (ML) popularity due to the availability of large datasets and increased computational power, leading to ML systems achieving superhuman performance in various tasks, such as DeepBlue beating the best chess player (Srivastava, 2022) , IBM Watson winning Jeopardy, and AlphaGo defeating the best Go player (Sauka, 2022). Initially, there was no pressing need to explain AI decisions because early AI algorithms were easily interpretable (§4.1). Today, however, the algorithms have become so complex that even the simplest deep neural networks are difficult to

understand. The necessity for XAI became more evident as these sophisticated systems began integrating into our daily lives. Unfortunately, the complexity of these systems often means that their outcomes are opaque and unpredictable to users. The demand for XAI is particularly pronounced in industries where human lives hinge on the outcomes of these systems, such as healthcare, law, defense, and finance (Asaduzzaman, 2022). XAI algorithms are designed to elucidate the predictions made by AI systems (as outlined in §2). Fundamentally, XAI operates within the realm of Artificial Intelligence (AI), but this does not imply applying AI techniques to AI itself; instead, it addresses a human-agent interaction problem. Therefore, XAI must efficiently provide explanations for its decisions, tailored to the specific audience or stakeholder for whom the explanation is intended. This customization ensures that the explanation is comprehensible and relevant, reflecting differences that can be observed in Figure 4

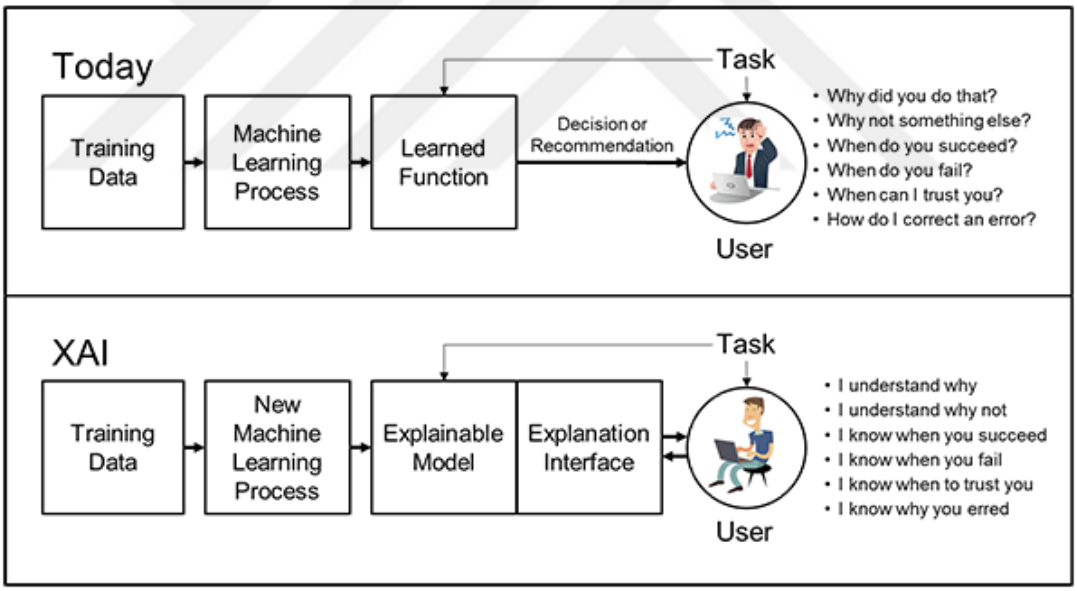

Figure 4 Explaining to different target audience (Asaduzzaman, 2022).

1.7 XAI and Adversarial Machine Learning

In this proposal the idea of Antagonistic AI is firmly connected to the Logic. As referenced in the Presentation, the target of this work is exactly to concentrate on the unwavering quality of some XAI techniques, using adversarial attacks on them.

However, outside of this context, AML and XAI are still connected in several aspects. For example, AML can leverage XAI methods to generate more powerful and effective attacks. Preferably, XAI ought to have the option to make sense of information inside an artificial intelligence model and reason about the model's activities. The data uncovered by XAI procedures can be utilized to produce more compelling assaults in antagonistic settings. Subsequent to realizing what explicit data ought to be given to the framework to accomplish a specific outcome, the aggressor can utilize this information to take advantage of the model. On the other hand, it's important to leverage the knowledge provided by the XAI methods to make the AI model more secure. By identifying potential weak points where attacks could happen, we can strengthen the model's defenses and make it more resistant to adversarial tricks. Model-rationalist procedures are approaches that can be applied to any AI model, no matter what its engineering or intricacy. They give an overall method for understanding and to decipher the choices made by a model without having to know its inside functions. Among the main methods belonging to this category we have LIME (Local Interpretable Model-Agnostic Explanations) (Habib, 2023) and SHAP (SHapley Additive exPlanations) (Jiyad, 2024) .

1.7.1 Lime. The goal of LIME is to provide local and interpretable explanations for individual predictions generated by the model, even when the model's complexity makes it challenging to comprehend. The LIME approach includes creating and prepares a simple interpretable model (see straightforward model in segment 2.3) to rough the way of behaving of the perplexing model that we need to make sense of. Officially, the creators characterize a clarification as another model $g \in G$, where G is a class of possibly interpretable models, for example, straight models or choice trees. Moreover, let f be our desired model to make sense of. As few out of every odd $g \in G$ might be sufficiently basic to be interpretable, the creators let $\Omega(g)$ be a proportion of intricacy (rather than interpretability) of the clarification. For instance, for choice trees $\Omega(g)$ might be the profundity of the tree, while for straight models, $\Omega(g)$ might be the quantity of non-zero loads (Habib, 2023). The creators further utilized $\pi x(z)$ as a nearness measure between an occurrence z to x, to characterize territory around x. At long last, let $L(f, g, \pi x)$ be a proportion of how untrustworthy g is in approximating f

in the territory characterized by πx. To guarantee both interpretability and neighborhood devotion, the objective is to limit L(f, g, πx) while having Ω(g) be adequately low to be interpretable by people. The clarification delivered by LIME is gotten by the accompanying improvement issue (Habib, 2023):

$$\zeta(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (1.4)$$

Since the explainer is model-agnostic, the minimization is done without making any assumptions about f. Thus, in order to learn the local behavior of f as the interpretable inputs vary, the authors use to approximate L(f, g, πx) by generating perturbed samples z, weighted by πx(z) (Chivukula, 2023). Subsequently, LIME trains (by optimizing the previous equation) an interpretable model on these perturbed instances to approximate the complex model's behavior around the point of interest.

The essential instinct behind LIME is shown in Figure 5. It includes inspecting examples both close to the objective occurrence (x) (which get a high weight from (π_x).) and farther away (which get a low weight from (π_x).)). While the first model might be excessively perplexing to make sense of worldwide, LIME gives a locally devoted clarification by approximating the model's conduct nearby (x) with a less complex, linear model, where the locality is determined by (π_x).

The interpretable model generated by LIME represents a simplified approximation of the complex model. This interpretable explanation of predictions allows to understand the model's decisions even for individuals who are not machine learning experts.

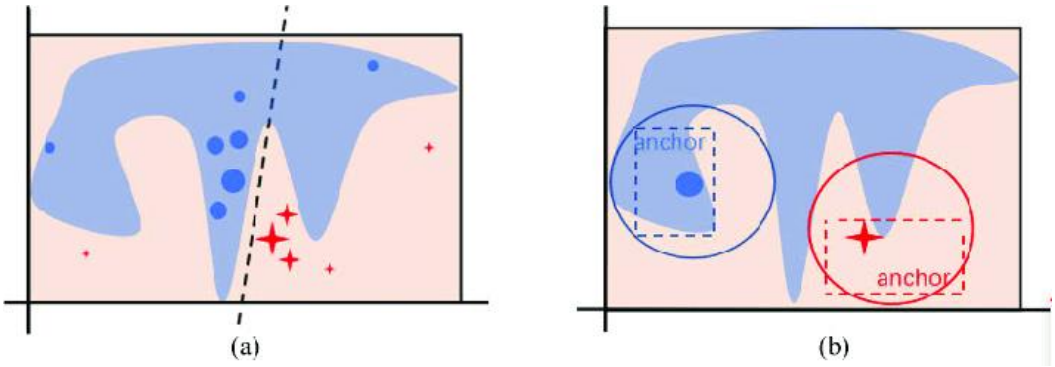

Figure 5 Toy example to present intuition for LIME (Habib, 2023).

With regards to LIME (Neighborhood Interpretable Model-skeptic Clarifications), the intricate choice capability f of the black-box model (portrayed by the blue/pink foundation) stays misty and can't be successfully approximated by a basic straight model. The striking red cross means the particular example under assessment. LIME works by testing occurrences around this point, using f to produce expectations, and weighting these expectations in light of their closeness to the occasion being made sense of (showed by size in this portrayal). The ran line addresses the privately scholarly clarification, which reliably makes sense of the model's conduct nearby the occasion being dissected, however it doesn't reach out to catching the worldwide way of behaving of f. This approach allows LIME to provide interpretable insights into the black-box model's decisions on a local scale, aiding in understanding how predictions are influenced by input features without needing to fully comprehend the intricacies of f (Habib, 2023).

1.7.2 Shap. The SHAP technique is based on some concepts similar to the LIME method. In the same way as LIME, each explanation is seen as a new model that approximates the behavior of the original one, which, thanks to its transparency, allows easy interpretation even for non-expert users (Jiyad, 2024). The theoretical foundation of SHAP comes from cooperative game theory and draws inspiration from the concept of Shapley value, which are used to fairly allocate each player’s contribution in a cooperative game. Therefore, SHAP is based on the idea that a prediction can be seen as the cumulative contribution of individual features to that prediction. The goal is to determine how much each feature has positively or negatively contributed to the difference between the model’s prediction and the reference average prediction. To do this, the SHAP approach considers all possible combinations of features (adding and removing features from the input of the model) and evaluates how much each feature contributes to the average Shapley value (Jiyad, 2024) . The explanation obtained from the SHAP method can be an histogram plot, a visual image or a text form. An example is the one in Figure 6, where the importance and contribution of each feature are shown using histograms.

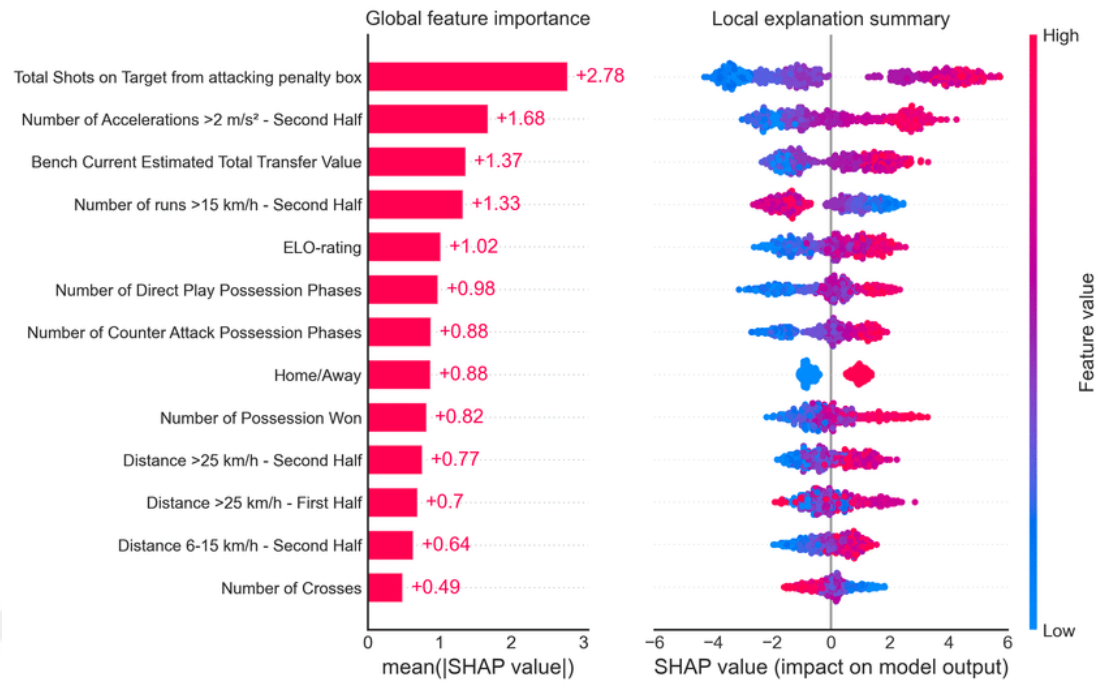

Figure 6 Features importance in the income prediction of the SHAP adult dataset according to the explanability model SHAP (Sauka, 2022).

These advanced explanations allow users to understand the role of each feature in the model's decision-making process, contributing to building confidence and comprehension in the use of machine learning models.

1.8 Generative Adversarial Networks (GAN)

The goal of Generative Adversarial Networks (GANs) is to use adversarial learning to create new data instances based on an input data distribution. GANs comprise of two parts: a generative model, which is a solo model summing up the dispersion of the given factors (e.g., GMM, VAE), and a discriminative model, which performs characterization or prescient displaying.

The discriminative model D, D(x, θd) (another multilayer perceptron with parameters θd) is trained on data from two sources, (1) the real data instances as positive instances and (2) fake generated instances from G as negative instances. D(x) outputs a scalar representing the probability of x being in pg. The aim is for D to output (1) and (2) everywhere. GAN training can be formally written as the two player minimax game

of generator G and discriminator D with value function V (G, D):

$$\min_{D} \max_{G} V(D,G) = E_{x\sim P_{data}}[logD(x)] + E_{z\sim P_{z(Z)}}\left[log\left(1 - D\big(G(Z)\big)\right)\right]$$

(1.5)

In their work, Goodfellow et al. propose a recommendation to alternate between optimizing the discriminator D and the generator G during training of Generative Adversarial Networks (GANs), where x represents the real data instance and z denotes input noise variables. They suggest performing k steps of optimizing D followed by one step of optimizing G . This alternating optimization strategy is advocated because it is impractical to fully optimize D after every single optimization step of G . By iterating between training D to distinguish between real and generated data, and training G to generate data that fools D , the GAN framework achieves a balance that enhances the quality and realism of generated samples over the course of training. This approach effectively manages the training dynamics of GANs, ensuring both the discriminator and generator are progressively refined to achieve optimal performance.

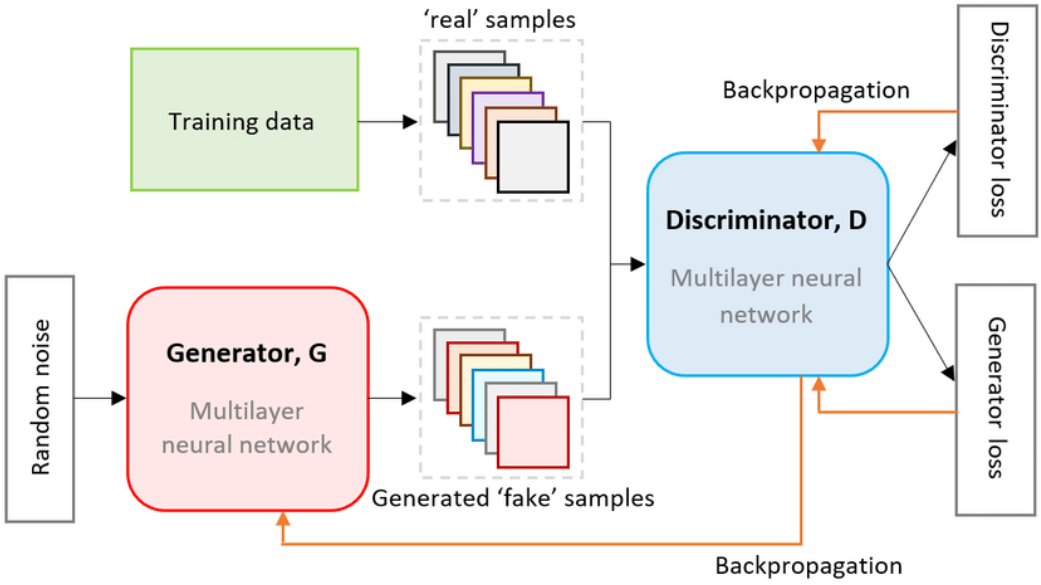

Figure 7 Generative Adversarial Networks (GAN) (Haoyi, 2023).

1.9 Intrusion Detection System (IDS)

Intrusion detection involves identifying unauthorized access or activities within systems and networks, whether caused by users or malicious software. An IDS is a device or programming application intended to screen an organization or framework for indications of noxious way of behaving or strategy breaks. Then again, an Interruption Avoidance Framework (IPS) broadens the usefulness of an IDS by distinguishing as well as effectively obstructing possible assaults. IDS can be basically grouped into two primary classes:

i- Network IDS, which monitors network segments and analyzes network traffic at different layers in order to detect intruders.

ii- Host based IDS, which are installed in host machines and analyze different indicators such as processes, log files, unexpected changes in the host to determine the presence of malicious activities.

Managing the sheer volume of network traffic produced daily by large enterprise networks poses significant challenges. Traditionally, one approach to cope with this issue has been to discard certain portions of the data or reduce the amount of information logged. However, the advent of Big Data Analytics (BDA) coupled with advancements in memory capabilities, computing power, and reduced storage costs have transformed this scenario into a substantial big data challenge. Instead of minimizing data collection, organizations now face the opportunity and necessity to harness and analyze vast amounts of network traffic data. This shift enables enterprises to derive valuable insights, improve network security measures, and optimize operational efficiencies through comprehensive analysis and interpretation of the extensive data streams generated across their networks.

1.10 Research Gap

Malware analysis and detection is an ongoing struggle between malware developers and the anti-malware community. Existing research indicates that current detection methods, such as signature-based, heuristic, and behavioral approaches, have significant limitations and are inadequate in addressing the challenges posed by next-

generation malware. (Moustafa, 2023). Subsequently, hostile to malware analysts have begun investigating malware recognition frameworks in light of AI and profound learning calculations. The improvement of these frameworks is a two-step process: 1) Component Designing and 2) Characterization/Bunching. Conversely, dynamic analysis involves extracting features by running the application within a sandbox or controlled environment. When features from both static and dynamic analyses are combined, it is termed hybrid analysis. These extracted features are then fed into classification or clustering algorithms to construct malware detection models.

Several studies have explored the vulnerabilities of malware detection systems to adversarial attacks using various approaches. For instance, researchers demonstrated that deep networks could learn from raw bytes and selectively modify byte sequences in malware samples to evade detection, highlighting potential weaknesses in existing systems. Hu et al. employed Generative Adversarial Networks (GANs) based on neural networks to generate adversarial malware examples, achieving a high fooling rate of 99% under unlimited perturbations, though they did not delve into the interpretability of their attack and defense strategies. Convolutional Neural Networks (CNNs) have also been applied in malware detection; Suciu et al. proposed a CNN architecture to detect malicious behavior directly from raw byte data, investigating the effectiveness of single-step adversarial attacks and transferability across different models. Ji et al. developed DeepArmour, a malware detection system combining multiple classifiers, and tested its resilience against white-box evasion attacks. Taheri et al. introduced five attack strategies targeting three malware detection models, achieving an average fooling rate of approximately 27%. Their subsequent adversarial retraining and GAN defense methods reduced the fooling rate, albeit with challenges in achieving robustness and attack interpretability.

Several research gaps emerge from existing literature on adversarial malware detection scenarios. Many studies predominantly focus on the white-box scenario, assuming adversaries possess complete knowledge of training data, features, classifiers, and model architecture, which is overly optimistic for real-world threat modeling. Furthermore, limitations include the narrow scope of evaluated classifiers and malware detection models, restricting the generalizability of proposed adversarial attacks and

defenses. Evaluations often prioritize fooling rate over accuracy drop, with minimal discussion on the number of perturbations required for effective evasion attacks, crucial for minimizing attack costs. Also, while defense methods are important to make malware detection models stronger against attacks, checking these defenses by testing them again is often ignored. Also, understanding and explaining how attacks work is key to finding weaknesses, but these are often missed in current research. Finally, there are great chances to create ways to make malware detection systems more resistant to attacks, which is a good area for future work.

1.11 Motivation

As machine learning (ML) technologies are being used more and more in IT systems, they bring many benefits, like better automation, improved decision-making, and more accurate predictions. But these benefits also bring risks that bad people can use. Even though systems based on machine learning (ML) are used a lot, not enough focus has been put on the possible weak points in these systems, especially when attackers try to harm them. Most studies have looked at what ML does well, often forgetting about the problems that could be caused by harmful actions.

Recent studies in adversarial machine learning have found that even small changes to input data can lead to significant errors in how models classify information. This raises concerns about the security and reliability of machine learning-based Intrusion Detection Systems (IDS). These vulnerabilities in IDS models are a major issue because attackers could manipulate input data to avoid detection, rendering the system ineffective at identifying harmful network activities.

This research is motivated by the urgent need to investigate adversarial attack techniques and their impact on machine learning-based Intrusion Detection Systems (IDS). Our goal is to evaluate how well IDS can withstand adversarial scenarios by testing samples generated using intelligent search strategies. These samples focus on the most critical features, identified using explainable AI (XAI) tools like SHAP, and are designed to disrupt the ranking of feature importance, simulating real-world attack conditions. The primary aim is to assess the resilience of LightGBM-based IDS models.

The main aim of this research is to find and fix the problems in machine learning-based systems that detect cyberattacks (called IDS). By using simple, easy-to-understand methods and testing these systems against harmful attacks, the study wants to find ways to make IDS work better even when attacked. This research not only makes IDS more secure but also helps us understand how attacks impact machine learning systems. Ultimately, it helps create safer and more reliable systems for detecting cyberattacks.

1.12 Problem Statement

The aim of this master's thesis is to address key challenges in cybersecurity by improving and making machine learning-based Intrusion Detection Systems (IDS), particularly those using LightGBM models, more robust and easier to interpret when dealing with adversarial attacks. As cyber threats become more sophisticated, attackers are finding ways to trick IDS and avoid being detected, which can lead to significant security risks. While traditional IDS methods are effective against known threats, they often fail to handle these advanced attacks.

This research focuses on identifying and reducing vulnerabilities in IDS models by using adversarial attacks that target the most important features of the model, as identified through Explainable AI (XAI) techniques like SHAP. The study uses a heuristic search approach to create adversarial examples that disrupt the feature importance rankings, simulating realistic attack scenarios. By oppressing the IDS models to these assaults and breaking down the effect on their exhibition, the exploration researches how antagonistic irritations can corrupt the adequacy of IDS.

Moreover, the exploration investigates the utilization of XAI to upgrade the interpretability of IDS models, giving further bits of knowledge into their dynamic cycle. This consolidated methodology takes into consideration a more thorough assessment of IDS execution, both when ill-disposed control. A definitive objective is to further develop IDS flexibility, guaranteeing these frameworks can keep up with high identification exactness, significantly under ill-disposed conditions, while likewise giving straightforward and justifiable clarifications to their expectations. The consequences of this study will add to growing safer, reliable, and interpretable

interruption location arrangements, better prepared to deal with developing and progressively complex digital dangers.

1.13 Objectives

The goal of this master's thesis is to make machine learning-based Intrusion Detection Systems (IDS), especially those using LightGBM, stronger and easier to understand when dealing with attacks designed to trick them. This study focuses on reaching these main goals:

1. Adversarial Attack Generation: Use a heuristic search strategy to create hostile samples that target the most critical traits discovered by XAI techniques like SHAP. These adversarial samples will imitate real-world attack situations aimed to influence the IDS models by exploiting feature flaws, reducing their performance.
2. Testing Against Attacks and Checking Performance: Compare how well the IDS models, especially LightGBM, handle adversarial samples to see how strong they are against attacks. The performance will be measured by looking at detection accuracy, precision, recall and F1-score. We will compare the results before and after the models are exposed to adversarial changes.
3. Understanding Model Decisions and Feature Importance: Use explainable AI (XAI) methods like SHAP and LIME to understand how the IDS models make decisions and how the importance of features changes when faced with adversarial attacks. This will help us see how these attacks affect the models' ability to explain their decisions clearly.
4. Checking Model Stability Under Attacks: Study how stable the IDS models' predictions and feature rankings are when they encounter adversarial samples. This involves looking at how these attacks affect the models' performance and seeing how well the IDS can keep working effectively despite the attacks..
5. Generalization and Transferability of Adversarial Attacks**: Study how adversarial attacks can work across different IDS (Intrusion Detection System) models. Specifically, look at how attacks designed to exploit certain features in a LightGBM-based IDS might also affect other machine learning models.

This research will help in developing stronger and more flexible defense strategies that can be used across various IDS systems.

1.14 Contributions

Contributions of the Thesis are:

- Model Support: We improve IDS to also work with LightGBM, not just XGBoost. This allows us to analyze and explain a broader variety of machine learning models.
- Class-Level Explanations: We provide explanations that focus on entire classes, showing the key features and their ranges that matter most for each class. These explanations help us better understand how the model makes decisions for different groups, offering a wider view compared to explanations for individual cases.
- Adversarial Sample Handling: We use formal explanations to construct a system for producing and identifying adversarial samples. The detection procedure analyzes the possibility of a sample being adversarial by assessing how explanations change in response to modest input perturbations. This allows for robust handling of hostile inputs.

1.15 Methodology

The plan for this project is divided into three main steps, each focusing on a different part of improving Deep Learning-Based Intrusion Detection Systems (IDS) to better handle adversarial threats. Here's a detailed explanation of the process:

1. Research and Background Study:

Start by carefully reviewing existing studies on adversarial machine learning, intrusion detection systems, and Explainable AI (XAI), with a special focus on how XAI can be transferred or applied.

Identify Problems: Figure out the biggest challenges and weaknesses in current IDS models, especially how easily they can be tricked by adversarial attacks.

2. Collecting and Preparing Data:

Gather Data: Collect suitable datasets for intrusion detection, like the CIC-IDS2017 dataset.

Prepare Data: Clean and process the data so it’s ready for training deep learning models. This includes steps like normalizing the data, converting categorical information into a usable format, and handling any missing values.

3. Training and Testing: Use the cleaned-up data to teach the models and check how well they work. Measure their performance using common tools like accuracy, precision, recall, and F1-score.

4. Checking Performance: Test how much adversarial attacks affect the IDS models. Use measurements like how often the attacks succeed, how accurate the detection is, and the rates of false alarms or missed threats.

5. XAI Analysis: Understanding the Model's Decisions

Use XAI methods like SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) to study and explain how IDS models make decisions in both normal and attack situations.

Finding Key Insights: Understand how attacks affect the model's decisions and identify the main features or patterns that attackers exploit.

6. Building Defense Systems: Training with Attacks

Include examples of attacks in the training process to make the model stronger and more resistant to adversarial threats.

7. Transferability Analysis: Cross-Model Testing: Evaluate the transferability of adversarial attacks by testing adversarial samples across different IDS models.

Chapter 2

Literature Review

2.1 General Context

This chapter gives important details and ideas about AIDS, generative models, and how they are checked. It begins by explaining generative adversarial models, including the kinds used for AIDS, and discusses the good and bad points of each kind. The chapter then explains in detail how GANs are made, including their parts, how they are trained, and the math involved. Finally, the chapter explains and shows the different ways to check AIDS, especially when working with balanced data (majority and minority groups). In recent years, keeping networks safe has become more important, and Intrusion Detection Systems (IDS) are very important for protecting networks. IDS look at internet traffic to defend against possible attacks, treating the detection of intrusions as a classification problem. This includes utilizing different AI (ML) and profound learning (DL) models to arrange network information as either harmless or noxious. IDS not just proposition compelling answers for recognizing breaks all through the organization yet additionally assist with diminishing the misleading problem rate. This part gives an outline of current ML, DL, ill-disposed assaults based GANs and XAI methods in IDS. The essential point is to break down and investigate modern data on existing IDS frameworks, offering new scientists a strong starting point for planning productive and hearty IDS frameworks. The part likewise incorporates a survey of different articles that represent the utilization of computer based intelligence devices and XAI approaches. Through the investigation of various perceptions, probably the most recent patterns for making more viable IDS frameworks are featured.

2.2 Intrusion Detection System

One of the key components of the paradigm for protecting infrastructure is the network IDS. In the present networked environment, a variety of strategies are used to protect crucial data. Providing internet security is a difficult effort; it must focus on

the following aspects: protection, detection, reaction, and recovery. The rest of the process is successful if intrusion detection is done correctly; if not, it leads to becoming a thread. Intrusions can be split into 2 types: host intrusions and network intrusions. As part of HIDS, the target host system's audit data is used for the analysis. This method's drawback is that attacks are extremely challenging to prevent, and occasionally hackers alter audit data. Monitoring computers or data flow within a network for unwanted access, activity, or data modification is one of the steps involved in intrusion detection. Attacks on distant systems, authorized users abusing their powers, and designated people attempting to gain more access are a few examples of such attacks. The goal of the IDS framework, depicted in **Error! Reference source n ot found.**, is to detect any unapproved attempts or successful attacks on any kind of monitoring data or resources that are available as a part of a network or host system. Locating probable events, documenting information about them, and making an effort to report them were the core goals of intrusion detection and prevention systems (IDPs). Many IDPs try to prevent a threat from happening after it has been detected. They use a variety of various response techniques, such as the IDPS directly interrupting the attack, changing the security paradigm, or changing the attack's content (Khaleel, 2024) (Bouaziz, 2023).

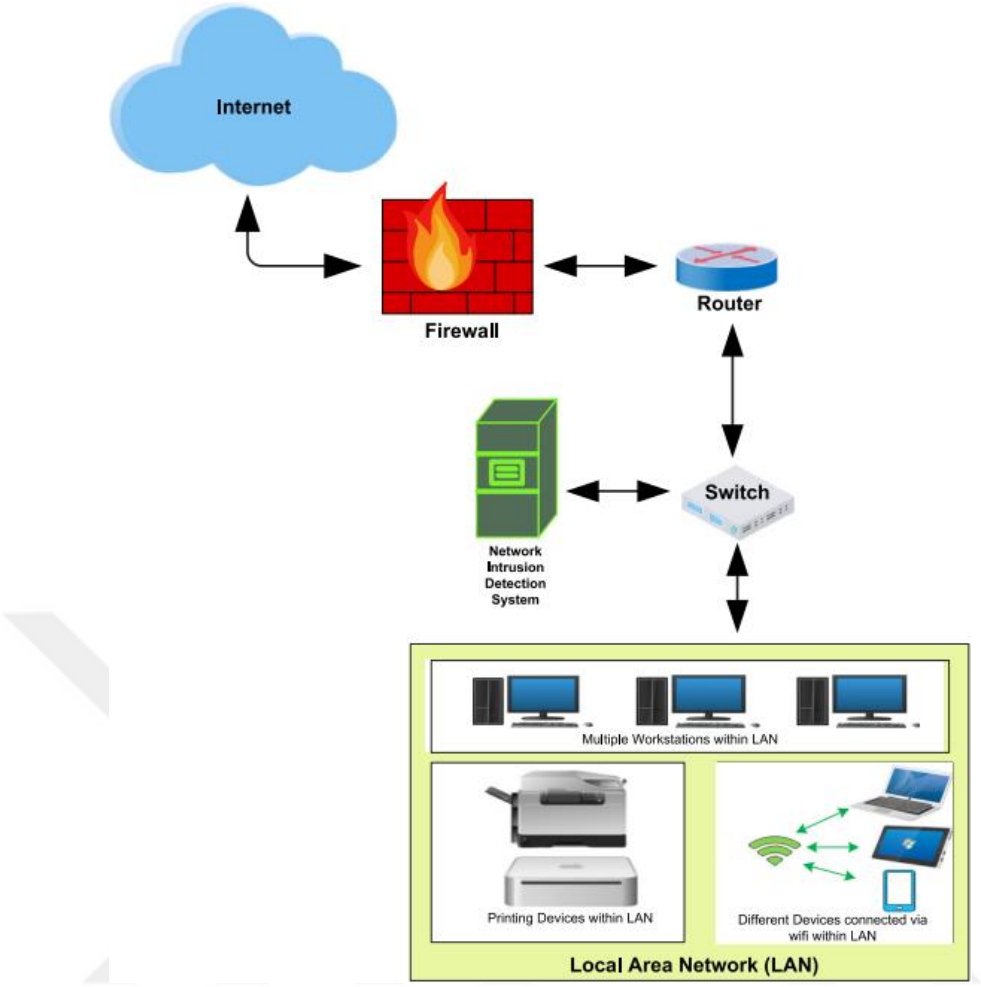

Figure 8 Intrusion Detection System(IDS) (Afolabi, 2024).

The IDS uses many kinds of analysis and pre-processing to find the intrusion. Prior to analysis, pre-processing is done to gather the data in a standard format. Better methods for data analysis during the analysis phase are provided by canonical formats. To detect intrusion, analysis methods can include statistical methods or signature comparison schemes. An intrusion alert system will be activated if one is present. Figure 9 illustrates the whole intrusion detection procedure.

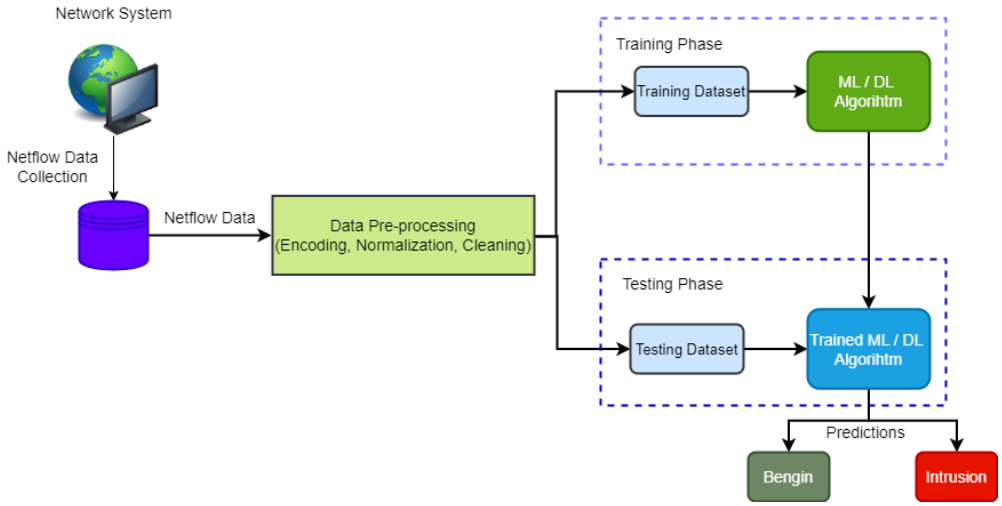

Figure 9 Intrusion Detection Process.

2.3 Classification of Intrusion Detection System

The general categorization of IDS is shown in Figure 10 as an intrusion strategy, protection mechanism, structural, data source form, behavioral, and timing of analysis. The Figure shows the various categories for each of those heads.

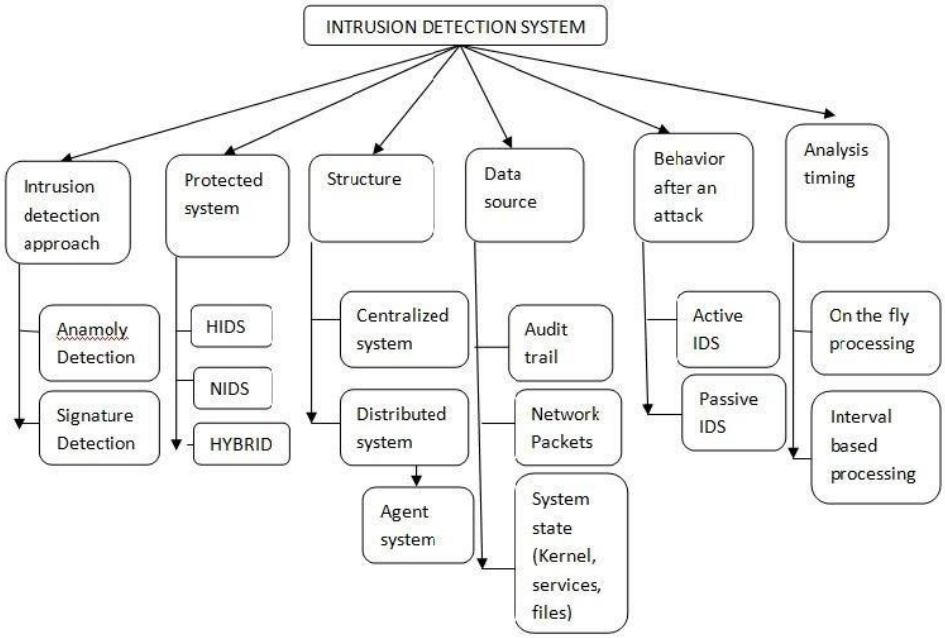

Figure 10. IDS (Intrusion Detection System) Classification (Abdulganiyu, 2023).

2.4 Intrusion Detection Techniques

Two categories of techniques are as follows:

- MDT(Misuse Detection Technique)
- ADT(Anomaly Detection Technique)

2.4.1 Misuse detection. One method for identifying assaults is misuse detection. This method defines the remaining activities as normal ones after addressing the anomalous system behavior. The system antivirus software is similarly reflected by a misuse detecting IDS paradigm. It evaluates the effectiveness of a system or network scenario and contrasts network activity with known intrusive system signatures. Systems for detecting intrusions that adhere to the misuse detection method must be regularly updated in order to lead the fight against hackers.

2.4.2 Anomaly detection. The challenge of finding data patterns that don't match up with expected behavior is known as anomaly detection. Anomaly intrusion detection uses the system's knowledge of regular activity to look for other patterns that it deems suspicious. While engaged in misuse intrusion detection, the system makes an effort to look for patterns that correspond to its knowledge of suspicious actions. A higher detection rate will result from combining the two techniques. Security is increased by adding reactive intrusion detection as a second line of defense. Finally, it's important to remember that intrusion detection works in conjunction with other security measures to create an environment that is as safe as possible. One approach to spotting anomalies is known as outlier identification; it deals with finding patterns in a given data set that didn't follow a previously established pattern of typical behavior. The resulting patterns, known as anomalies, frequently translate into crucial and useful information across a wide range of application domains. Change, deviation, surprise, aberrant, peculiarity, intrusion, and other terms are also used to describe anomalies.

2.5 Types of IDS

IDS are classified into 2 categories, they are... • Network-Based Intrusion Detection System(NIDS) • Host-Based Intrusion Detection System(HIDS) • Collaborative Intrusion Detection.

2.5.1 Network-based intrusion detection system-nids. A system called Network Security Monitoring (NSM) monitors network traffic in an effort to spot improper behavior, including denial of service assaults, port scanning, and even system hacking attempts. All incoming packets are scanned by NIDS, which looks for unlawful patterns known as signatures or rules. As seen in Figure 1.4, the NIDS then examines all traffic passing through that area of the network. The NIDS function works in a manner similar to high-end antivirus programs in that it compares every transmitted packet to a signature or pattern file. The IDS works specifically to boost packet performance because scanning every packet would significantly reduce traffic. Additionally, while checking the packets, an IDS employs the firewall methodology by allowing through the packets that are not detrimental to the system. This was done via preprocessing filters for IDS.

2.5.2 Host-based intrusion detection system-hids. These are deployed on specific machines on the network and operate in a manner similar to how a NIDS dynamically examines network packets. Only packets coming into and leaving the device are monitored by HIDS, which also alerts the user or administrator if it detects any strange behavior.

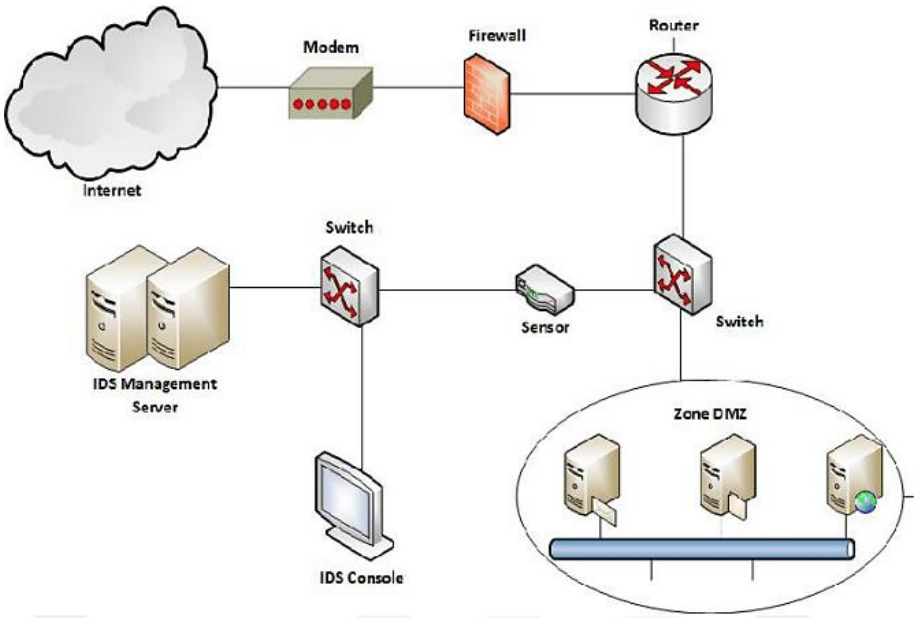

Figure 11. Network Based Intrusion Detection System(NIDS) (Habeeb, 2022).

The implementation of HIDS on various types of hardware, such as servers, workstations, and laptop computers, is shown in Figure 11. The host receives the traffic once it has been carefully examined into consideration, evaluated, and forwarded onto it if there are no potentially hazardous packets present during the data transmission. Instead of being compared to NIDS, HIDS is inherently more focused on the local machines. NIDS turned its main attention on the network and those particular hosts.

2.5.3 Collaborative intrusion detection. An efficient Collaborative Intrusion Detection Network (CIDN) enables distributed Intrusion Detection Systems to improve their intrusion detection capabilities and share information about intrusions to identify new types of assaults.

2.6 ML Techniques

Turing (1950) stated AI is used in ML, where machine learns from its prior experiences and predicts the future. ML was used to analyze assaults and security events, including spam mail, social media analytics, user identification, and attack detection (Nesheim, 2023). ML models are classified as supervised, unsupervised, and

reinforcement learning, as shown in Figure 12. Supervised learning uses labeled data in the training phase to detect attacks. It is primarily used in classification problems. The biggest impediment of it is the lack of sufficient labeled data. However, manually labeling data is expensive and time-consuming. Unsupervised learning deals with unlabeled or uncompressed data. Clustering is the most widely used unsupervised technique. However, the algorithms are self-employed in detecting and interpreting the data's unique structure. Reinforcement learning depends on a trial-and-error method in which a learning system collects data and takes action. If the action produces a favorable result, a reward is recorded. On the other hand, if the activity has an unfavorable outcome, the system will learn that similar actions in the future are unlikely to be successful.

Table 2

Differences Between ML and DL Models

Metric	Machine Learning	Deep Learning
Human involvement	It requires more human involvement	Requires less involvement
Structure	It has a simple Structure	It has a complex Structure
Data	Requires less data to train.	Requires more data to train.
Computation time	Requires less computational time than DL methods	Computational time is more when compared with ML methods
Hardware	They can be processed with CPUs	Mostly they require high-performance computing devices
Feature Selection	Features to be selected manually	Features are automatically extracted
Data interpretation	Few models can be easily interpreted, like RF and DT. But some models are not easy to understand, like XGBM. and SVM	It is not easy to understand

Table 2 (cont'd)

Layers	It can work effectively with the network having input, output, and hidden layers.	It requires a minimum of three layers or more.
Output	It provides numerical output like classification or score	It gives numerical, text, sound, images and etc.,

The following subsection presents various intrusion detection systems using ML models.

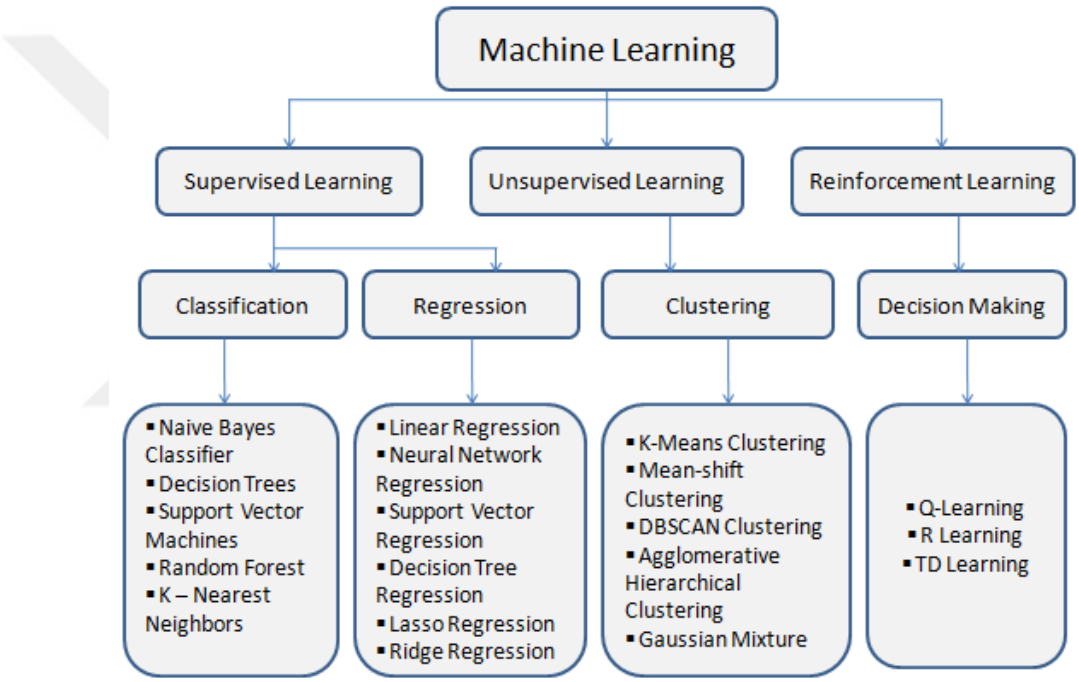

Figure 12 Classification of ML models (Abdulganiyu O. H., 2023).

2.7 Applications of ML in IDS

(Albahri, 2024) reviewed several studies on detecting Distributed Denial of Service (DDoS) attacks in IoT-based networks using Machine Learning (ML) models and found that while models like XGBoost achieved high accuracy rates (up to 100%) and ANN reached 99.95%, the datasets used were often outdated, unrealistic, or heavily imbalanced. For example, the Bot-IoT dataset, which is widely used, contains

over 99% malicious traffic but less than 1% benign traffic, leading to potential biases in model performance. The authors concluded that developing a new, realistic dataset for IoT-based DDoS detection is essential and suggested further exploration of underutilized ML models like AdaBoost to improve detection accuracy. While the methodology of reviewing existing studies was appropriate, the paper could have been strengthened by empirically addressing the dataset limitations they identified.

(Yang, 2024) presents a novel approach, DBO-SSAE, which uses dung beetle optimization (DBO) to automatically optimize the hyperparameters of a stacked sparse autoencoder (SSAE) for extracting network security situation elements. Applied to the UNSW-NB15 dataset, this method achieved superior performance, with the BiLSTM classifier reaching 98.84% accuracy, a 98.96% F1-measure, a 1.86% false negative rate (FNR), and a 0.6% false positive rate (FPR), outperforming other feature extraction methods like PCA. The approach was appropriate, as DBO actually upgraded SSAE's hyperparameters inside only nine cycles, showing quick union and solid speculation abilities. Nonetheless, the review's limits incorporate the high computational expense of preparing DBO-SSAE and the intricacy of DBO's boundary settings. The creators propose future upgrades, for example, integrating the consideration instrument and hybridizing DBO with calculations like PSO, which could additionally refine their system. Generally, the work was all around led, with promising outcomes, however addressing its limits could prompt much more powerful arrangements.

(Ghaleb, 2023) proposes a charge card extortion location model (CCFDM) utilizing Troupe Incorporated Minority Oversampling based Generative Ill-disposed Organizations (ESMOTE-GAN) joined with an Irregular Woodland calculation. The model successfully addresses the class unevenness issue by creating different orchestrated information subsets, which are then used to prepare various GANs, trailed by a bunch of Irregular Backwoods classifiers. This approach fundamentally further developed identification execution, accomplishing a 1.9% improvement in general execution and a 3.2% expansion in discovery rate, with a 0% deception rate, exhibiting its common sense in genuine applications where even negligible misleading up-sides can overpower human experts. The strategy was fitting, utilizing a mix of information resampling, GAN, and gathering figuring out how to deal with the difficulties

presented by imbalanced information, uproarious highlights, and non-straight choice limits. The review's discoveries are solid, however the creators could investigate other crossover draws near or refine hyperparameters to additionally improve execution and diminish computational expenses. Generally, the work is thoroughly thought out and offers a hearty answer for a basic issue in Mastercard misrepresentation discovery.

From the writing, it is apparent that SVM has a compelling location rate. Nonetheless, while managing high aspects, SVM requires more preparation time than other ML calculations. Hence, scientists advanced information to further develop SVM preparing time and recognition rate.

(Amaran, 2021) presents an Ideal Multi-facet Perceptron (OMLP) method improved by the Dragonfly Calculation (DA) for interruption identification in Remote Sensor Organizations (WSN). Their discoveries show that the OMLP model accomplished high precision (94.21%) and an identification pace of 95.18% when tried on the NSL KDDCup 99 dataset, outflanking different models like SVM and ELM as far as recognition rate and bogus negative rate. The approach, which joins MLP with DA for boundary advancement, is reasonable and compelling, as DA supports choosing ideal loads and inclinations, prompting worked on model execution. The work is well-conceived, offering a promising approach for intrusion detection in WSNs, though future research could explore deep learning techniques to potentially enhance the model's performance further. Overall, the study provides a robust solution to a critical challenge in WSN security and validates its applicability in real-time environments.

(Mallampati, 2022) recommended a hybrid detection to find DDoS attacks. To reduce the dimensionality of the network traffic a sparse deep Autoencoder (SDA) is developed. The SDA contains 2 hidden layers in the encoder and decoder. They stated that performance was improved when ELU was used in the hidden layer and swish activation in the output layer with Adam optimizer and elastic net regularization. They adopted SDA to extract the informative attributes. Further, the optimal set was trained by using tuned LGBM to detect DDoS attacks on the CIC-DDoS 2019 and CIC-IDS 2017 datasets.

Table 3

Comparison of the Various ML and DL Models With Advantages and Disadvantages

Model	Learning method	Advantages	Limitations
SVM	Supervised	Overfitting is less likely because models are more generalized.It can work with a non- linear transformation	Requires more training and testing time Kernal selection is difficult Requires more memory
DT	Supervised	These are easy to understand It performs good with discrete and continuous data	1. More prone to overfitting. 2. Training time is more. 3. Small variations in data may produce different decision trees
KNN	Supervised	1. Retraining is not required. Can add additional data for predictions.	1. It is computationally expensive 2. Sensitive to missing values and outliers 3. Finding the optimal K value is difficult May overfit
RF	Supervised	1. Works well with more data. 2. It maintains good accuracy even when dataset has missing values .	1. When the number of trees increases, it requires more training. 2. May prone to overfit
LGBM	Supervised	1. It requires less memory 2. Works better with larger datasets 3. It is a histogram-based model which fastens the training process	1. Prone to overfit 2. Does not perform well with smaller datasets
MLP	Supervised	1. Works well with larger data. 2. Predictions are faster once training is completed.	1. Requires more training time 2. Can overfit 3. Choosing the appropriate number of neurons and layers might be challenging.

Table 3 (cont'd)

CNN	Supervised	. Extract optimal features automatically.	1. Requires large data to train. 2. It is difficult to implement 3. Sometime, it will overfit
		. It can be used for feature extraction.	1. It may remove important information.
AE	Unsupervised.	It can learn non-linear data	2. If the parameters are less than the data,there are chances of overfitting
LSTM	Unsupervised	1. Avoid the vanishing gradient problem 2. It can give high accuracy in predictions	1. Requires more training time 2. Easy to overfit 3. Requires more memory to train
GAN	Unsupervised	1. Used to generate synthetic data 2. It learns internal representations of the data	1. requires more time to train 2. Learning to create discrete data, like text, is challenging.

2.8 Deep Learning in IDS

Traditional ML approaches struggle to be deployed in large environments because they mainly rely on manually extracted features and lack labeled training datasets. Further- more, shallow learning cannot analyze high-dimensional datasets in-depth. DL models are typically neural network models with multiple hidden layers. These models may learn very sophisticated non-linear functions, and the models can handle high- dimensional data and extract relevant feature representations in a refined and improved manner (Thakkar, 2022) . Therefore, it performs better than conventional ML models. As a result, DL architectures have received more attention nowadays than traditional ML methods. They are widely used in image classification, audio recognition, and anomaly detection. The enumeration of deep learning models is illustrated in Figure 13.

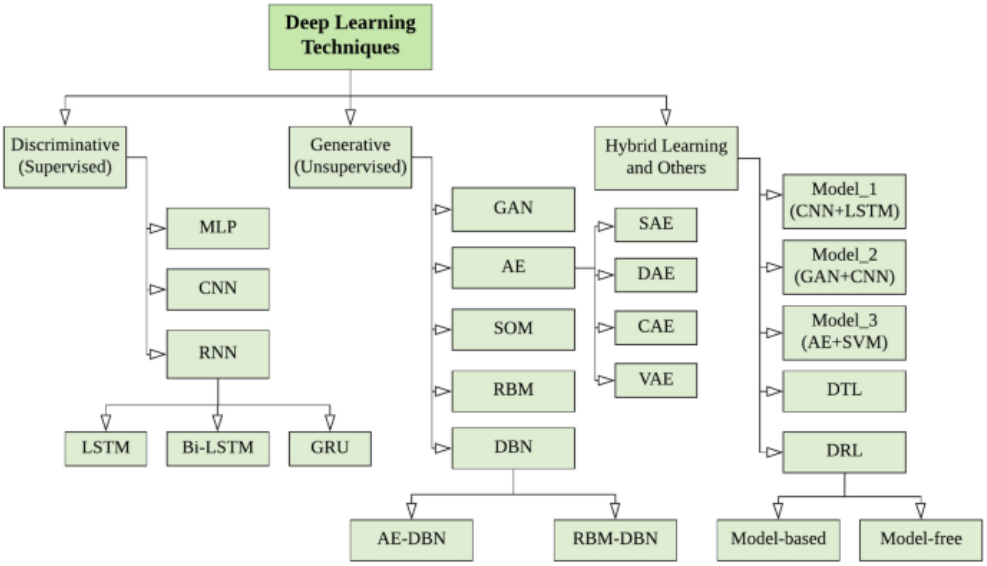

Figure 13. Taxonomy of Deep Learning Models (Afzal-Houshmand, 2023).

2.9 Applications of DNN in IDS

It is supervised instance learning which depends on MLP. An Artificial Neural Network (ANN) is equipped with intermediate layers positioned between the input and output layers. Every neuron is interconnected with other neurons in consecutive layers. An activation function is applied to the output of each layer in the network, amplifying the impact of network learning. The Back Propagation procedure is being used for data training. It also shrinks the gap between desired and actual values (Park, 2021) .

(Maheswari, 2023) proposed an Ideal Group based Interruption Recognition Framework (OC-IDS) intended to upgrade safeguard against assaults in web and distributed computing conditions. The proposed framework utilizes a half breed improvement approach, joining the Multi-section with (MCA) for information preprocessing and Tumultuous Manta-beam Scrounging Streamlining (CMFO) for bunching. This is trailed by a (MTL-DNN) for assault characterization. The discoveries show that the OC-IDS framework accomplished elite execution measurements, with a prominent 95.01% exactness on the KDD Cup'99 dataset and equivalent outcomes on the NSL-KDD dataset. The system is appropriate for the issue, actually tending to the difficulties of continuous interruptions in cloud conditions. The combination of innovative optimization algorithms and deep learning techniques is appropriate and results in significant improvements over existing methods. However,

the study could further validate its results by applying the model to more diverse, real-world datasets to ensure its robustness across different scenarios. Overall, the work is comprehensive and contributes meaningfully to improving intrusion detection in cloud computing.

(Ananth, 2023) reasoned that their methodology altogether upgrades both security and energy effectiveness in VANETs (Vehicular Impromptu Organizations). The creators showed that the IWOEEC-DWNN strategy, which joins obtrusive weed streamlining for energy-proficient bunching with a profound wavelet brain network for interruption discovery, beats a few existing techniques across different measurements, including network lifetime, energy utilization, throughput, bundle conveyance proportion, and start to finish delay. The system utilized, including broad recreations utilizing the NS-2 instrument and correlations against cutting edge procedures, was suitable and powerful, giving an unmistakable approval of the viability of the IWOEEC-DWNN approach. The work was extensive, with distinct targets and a strong trial plan that upholds their cases. In any case, the creators proposed that future work could additionally upgrade the framework by consolidating highlight determination and high level bunching strategies to further develop interruption discovery in VANETs. Generally, the review presents a huge commitment to the field, offering a novel and compelling answer for the double difficulties of safety and energy productivity in vehicular organizations.

(Zhiqiang, 2022) reasoned that the proposed Upgraded Exact based Part Examination (EECA) joined with Long Transient Memory (LSTM) for interruption identification in remote sensor organizations (WSNs) was exceptionally compelling. The methodology accomplished predominant execution measurements across various datasets, prominently NSL-KDD, CICIDS 2017, and UNSW NB 2015. For example, the precision of the EECA-LSTM strategy came to 99.95% on the UNSW NB 2015 dataset, 99.98% on CICIDS 2017, and 98.2% on NSL-KDD. Moreover, the strategy exhibited a low misleading positive rate (FPR) of 0.00053 on UNSW NB 2015, 0.004 on CICIDS 2017, and 0.0227 on NSL-KDD, showing its dependability in recognizing genuine dangers while limiting deceptions. The work was proper and all around directed, particularly in tending to the difficulties of recognizing obscure assaults in

Interruption Discovery Frameworks (IDS). The procedure was hearty, using a mix of Head Part Examination (PCA) and Experimental Mode Decay (EMD) to hold the most pertinent highlights, trailed by LSTM for grouping, which successfully dealt with little datasets and limited misclassification mistakes. The review's correlation with best in class techniques further approved the prevalence of their methodology, with the proposed strategy reliably beating others concerning exactness, accuracy (0.996 on CICIDS 2017, 0.9984 on UNSW NB 2015), and review (1 on CICIDS 2017, 0.9996 on UNSW NB 2015). Nonetheless, while the strategy was appropriate, further upgrades could incorporate investigating other profound learning models, like Convolutional Brain Organizations (CNN) or group procedures, to additionally work on the vigor and generalizability of the IDS across various kinds of cyberattacks. Furthermore, testing on more different and complex datasets could give a more extensive assessment of the framework's viability in certifiable situations.

(Mansour, 2022) proposed introduced an original strategy called Poor and Rich Improvement with Profound Learning Model for Blockchain-Empowered Interruption Discovery in CPS (Favorable to DLBIDCPS). The creators presumed that this strategy fundamentally further develops interruption recognition in Digital Actual Frameworks (CPS), accomplishing prevalent execution measurements across a few benchmarks. In particular, the Supportive of DLBIDCPS strategy exhibited higher exactness (AAC of 0.9885), accuracy (Chimp of 0.9901), review (WOAFS of 0.9922), and F-score (0.9883) contrasted with different models like BBFO-GRU and ideal GRU. The philosophy was vigorous, joining a Versatile Congruity Search Calculation (AHSA) for include determination, a consideration based bi-directional gated repetitive brain organization (ABi-GRNN) for grouping, and a Poor and Rich Streamlining (Ace) calculation for hyperparameter tuning. Also, consolidating blockchain innovation upgraded security in the CPS climate. In general, the work was fitting and professional, offering an extensive way to deal with tending to the intricacies of interruption identification in CPS conditions. Be that as it may, future work could coordinate information grouping and element decrease strategies to additionally further develop security and effectiveness.

(Ravi, 2022) found that their proposed approach, which incorporates repetitive profound learning models (RNN, LSTM, GRU) with bit based head part examination (KPCA) and a troupe meta-classifier, accomplished noteworthy outcomes, with a greatest exactness of close to 100% for network assault discovery and 97% for assault grouping on the SDN-IoT dataset. This technique beat customary models and other profound learning draws near, showing unrivaled execution with RNN, LSTM, and GRU models accomplishing exactnesses of 92%, 96%, and 96% for identification, and 91%, 93%, and 93% for grouping, separately. The procedure utilized — involving repetitive models for include extraction, KPCA for dimensionality decrease, and group meta-classifiers for arrangement — was powerful and appropriate for the errand, as confirmed by the elite execution across different datasets (KDD-Cup-1999, UNSW-NB15, WSN-DS, CICIDS-2017), with recognition exactnesses going from 98% to close to 100% and order correctnesses going from 89% to almost 100%. Despite the strong results, the study acknowledged limitations such as sensitivity to imbalanced data and the lack of evaluation in adversarial environments. Future work could focus on addressing these limitations by implementing cost-sensitive learning techniques for imbalanced data, exploring adversarial robustness, and optimizing feature selection to further enhance the model's performance.

DL uses fully connected layer to classify information. But the limitation of a fully connected network is parameter optimization, the loss of neighborhood information, and it is not translation invariant. To address this issue, (Agalit, 2022) used CNN for feature extraction instead of fully connected network they applied DT for classification. Initially, they preprocessed the data using a min-max scaler. Then they utilized one hot encoding to convert numerical data to a grey-scale pixel to form an image. Further, these images are passed to their proposed model. To avoid overfitting, they used three pooling layers and three convolution layers to select important attributes. They used an average pooling layer to preserve the features of input data. Finally, the optimal features are trained by using a decision tree to detect assaults. It was tested on the NSL-KDD dataset.

2.10 Intrusion Prevention/ Detection System based Machine Learning: Review

IPS, also known as IDPS, are network security techniques that continuously watch over network and system processes in order to recognize dangerous conduct. PS is essentially divided into two types: host-based and network-based. Using third-party software tools, the Host-based intrusion prevention system (HIPS) uses security measures to identify and stop harmful activity. HIPS are particularly utilized to safeguard endpoint hardware. The majority of HIPS use well-known attack signatures and patterns to spot malicious behavior. Although effective, signature-based detection can only shield the host device from known assaults. Zero-day attacks and signatures that are not present in the database repository were not protected. Furthermore, the likelihood of false positives is smaller than with statistical anomaly detection methods since stateful analysis may be done with the knowledge of the actual packet contents.

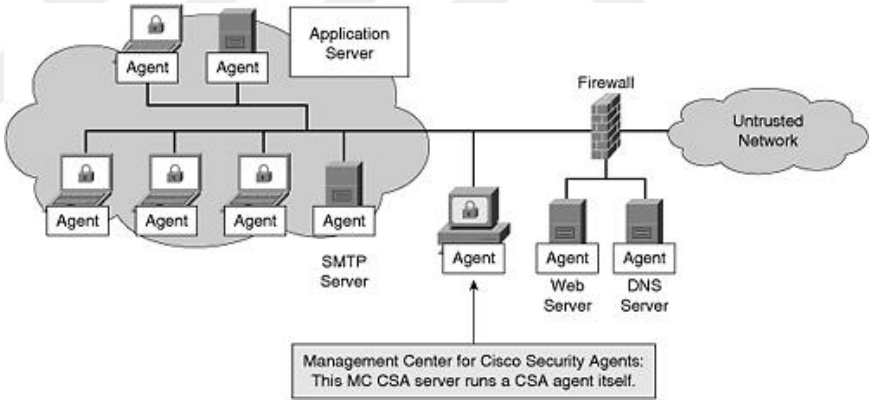

Figure 14. Host-Based Intrusion Detection System (HIDS) (Srivastava D. S., 2024).

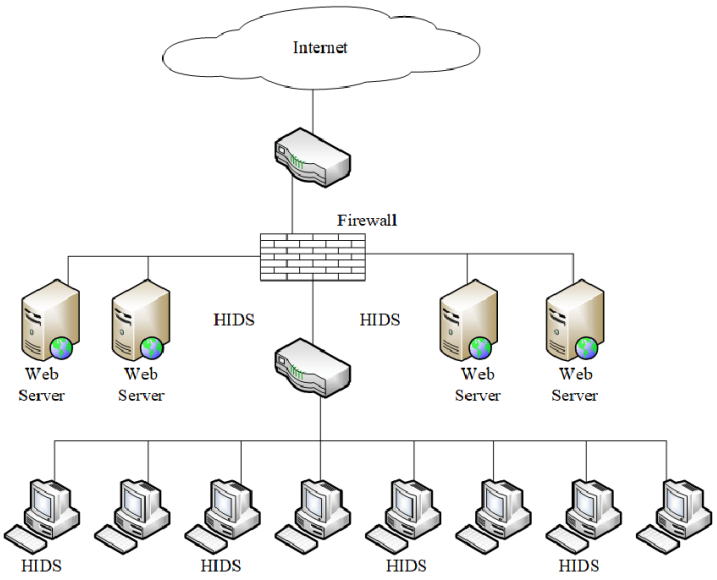

Figure 15 Host Based IPS (Kaw.anaka, 2023).

An IPS that focuses on confidentiality, data integrity, and network availability is known as a NIPS. The primary duties involve defending the network from attacks like illegal access and DoS(Denial-of-Service). It gives the network intelligence and helps it swiftly distinguish between good and bad traffic. It is further asserted that malicious traffic like Trojans, worms, viruses, and polymorphic threats turns the NIPS into a jail.

2.11 Related Works

The related work section includes several key studies in the field of adversarial machine learning. (Bouaziz A. N., 2023) conducted research on enhancing security against advanced adversaries by utilizing Machine Learning (ML) and Deep Learning (DL) techniques, focusing on detecting malicious attacks, particularly in the context of DNS servers. Their work highlighted the effectiveness of XAI methods, such as SHAP combined with LSTM or RNN, in improving the accuracy of classifiers for time-series data. Wang et al. contributed to the field by analyzing adversarial attacks and defenses within air transportation communication systems, introducing innovative techniques like double-level confrontation to enhance the security of deep learning models. Another survey provided a comprehensive overview of recent advancements

in adversarial attacks and defenses, particularly in deep neural network-based classification models, categorizing over 180 research papers and emphasizing the importance of staying updated on emerging threats. Furthermore, a review on ill-disposed assaults and guards in XAI investigated the weaknesses of XAI techniques and proposed procedures to upgrade their strength. This work highlights the basic need to address ill-disposed dangers in computer based intelligence frameworks, offering a guide for future exploration in the quickly developing field of AdvXAI. These examinations all in all add to a more profound comprehension of the ongoing scene of ill-disposed AI and set up for future progressions in the field.

An intrusion detection system (IDS) is a form of security technology that monitors and analyzes a computer system or network for signals of suspicious or dangerous behavior. An IDS detects threats like as viruses, hacking attempts, and unauthorized access by monitoring network traffic, computer logs, and other data sources. In this study, we conducted a literature review of network IDS research.

With regards to antagonistic assault discovery in interruption identification frameworks (IDS), a few examinations have investigated the utilization of AI and profound learning procedures. Analysts have used customary AI calculations, for example, choice trees and backing vector machines, as well as deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to enhance IDS performance. Some works have focused on generating adversarial examples to test the robustness of IDSs, using methods like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Others have applied GANs to generate synthetic attack traffic for training and testing IDSs, aiming to improve their detection capabilities. However, the integration of Conditional GANs (CGANs) specifically for generating labeled adversarial traffic and using XAI methods like SHAP for interpretability is relatively less explored.

(Lo, 2022) developed a hybrid deep learning-based intrusion detection system (HyDL-IDS), integrating convolutional neural networks (CNN) and long short-term memory (LSTM) networks, achieved nearly 100% detection accuracy with minimal false alarm rates for various in-vehicle network attacks, including denial-of-service,

fuzzy, and spoofing attacks. This significant performance improvement over traditional methods such as Naive Bayes, Decision Tree, and Multi-layer Perceptron indicates that the spatial-temporal representation provided by the CNN-LSTM architecture is highly effective for intrusion detection in CAN-based networks. The methodology employed was well-suited for the task, leveraging CNN to extract spatial features and LSTM to capture temporal dependencies, which collectively enhanced detection accuracy. However, the study's reliance on a specific car-hacking dataset and the use of supervised learning could limit its applicability to more sophisticated or novel attacks. Future work could benefit from incorporating unsupervised and adversarial techniques to address these limitations and extend the system's robustness against emerging threats.

(Afzal-Houshmand, 2023) undertook a research to improve security against sophisticated attackers using Machine Learning (ML) and Deep Learning (DL) approaches. Their study includes creating general classification models for detecting a variety of harmful assaults, with a focus on rogue DNS servers. The research emphasized the importance of XAI in boosting classifiers' capacity to distinguish between genuine and malicious inputs. The authors demonstrated that merging SHapley Additive Explanations (SHAP) with Long Short-Term Memory (LSTM) or Recurrent Neural Networks (RNN) is particularly successful for interpreting time-series data, especially in the setting of idea drift, where changes in data patterns occur over time.

(Wang, 2023) proposed a better AlexNet-GRU model for interruption identification in metropolitan rail travel the executives frameworks, exhibiting striking execution upgrades. The model accomplished a high acknowledgment exactness of 96.00%, outperforming other brain network models by something like 1.55%. It likewise showed stable preparation and testing times, with a typical information message conveyance rate surpassing 80%, and kept information message spillage and parcel misfortune rates underneath 10%, while keeping a typical deferral of around 350 milliseconds. The technique, joining the profound learning capacities of AlexNet with the consecutive learning force of GRU, was proper for improving both recognition precision and information transmission security. Be that as it may, the

review's emphasis on just two kinds of assaults (DoS and DCA) and the utilization of admired recreation situations might restrict the generalizability of its discoveries. To reinforce the model's down to earth application, future examination ought to investigate a more extensive scope of assault types and address certifiable circumstances that can influence information transmission and recognition. Generally, the review gave a hearty methodology promising outcomes, however it would profit from more extensive testing and more practical circumstances.

The study on Antagonistic Assaults and Safeguards in XAI dives into the arising field of ill-disposed logical simulated intelligence (AdvXAI) and looks at the weaknesses of XAI strategies to ill-disposed assaults. This far reaching survey, summing up the discoveries of north of 50 papers, features that collecting clarifications from different techniques can offer more noteworthy power against controls contrasted with depending on a solitary clarification approach. The paper highlights the basic need to address antagonistic assaults on model clarifications to keep up with the reliability and security of man-made intelligence frameworks. Furthermore, it gives a guide to future exploration headings in AdvXAI, expecting to direct the improvement of stronger and straightforward computer based intelligence frameworks.

The paper (Paya, 2024) presented Apollon, a safeguard framework for Interruption Discovery Frameworks (IDS) that really counters Ill-disposed AI (AML) assaults. Their analyses showed that Apollon, utilizing Multi-Equipped Outlaws (MAB) with Thompson testing to powerfully choose classifiers, further developed strength against assaults like Zeroth-request enhancement (ZOO), HopSkipJump (HSJA), and W-GAN-based assaults. For example, when exposed to the ZOO assault, Apollon kept an exactness of 97.40% and a discovery pace of 94.20%, beating individual classifiers like Irregular Timberland (RF) and Choice Tree (DT). During the HSJA assault, where recognition rates for independent models dropped near nothing, Apollon protected an identification rate above half. Against the W-GAN assault, which was the most forceful, Apollon supported an identification rate more prominent than 40%, while any remaining classifiers were diminished to zero location. Strategically, the creators embraced a thorough methodology by utilizing Stroll Forward Cross-Approval to guarantee no future information spillage, and applied strong element

determination from unmistakable works, upgrading their models' speculation. They utilized three generally acknowledged datasets CIC-IDS-2017, CSE-CIC-IDS-2018, and CIC-DDoS-2019 giving an exhaustive assessment across numerous assault types. While the system did not fully eliminate attacks, it significantly increased the effort and resources required for successful exploitation, highlighting its practical value as a robust deterrent. Though effective, the authors recognized that with longer attack iterations or more computational resources, attackers might eventually breach the system, suggesting further improvements for stronger resistance. Overall, Apollon's methodology and findings are well-grounded, delivering meaningful advances in IDS defences.

The study of (Altulaihan, 2024) on XA-GANomaly presents a clever versatile semi-directed learning approach for interruption recognition utilizing GANomaly, which powerfully prepares little subsets of constant information. The specialists actually consolidated three reasonable procedures — SHAP, reproduction mistake Nour perception, and t-SNE — to upgrade the interpretability of their model and backing security experts in observing and answering organization dangers. Their discoveries uncovered huge execution enhancements contrasted with other one-class arrangement procedures, especially with the NSL-KDD and UNSW-NB15 datasets, showing increments of up to 13% in F1-score and more than 11% in precision. The adaptive algorithm demonstrated its ability to refine its detection capabilities as more real-time data subsets were introduced, making it suitable for practical industry use. The methodology was sound, incorporating deep neural networks and GANomaly, and the use of adaptive learning allowed the model to adjust continuously to new attack patterns. However, the study faced limitations with imbalanced datasets, and future research could focus on enhancing the robustness of the adaptive algorithm for such cases. Overall, the work is appropriate and valuable, showing potential for industrial applications, especially with its intuitive visualization and real-time adaptability.

The research paper by (Sarhan, 2023) A, and their team proposes an Anomaly Detection Intrusion Detection System (IDS) for detecting Denial of Service (DoS) attacks in IoT networks, integrating anomaly detection and machine learning algorithms to enhance IoT network security. The group fostered a crossover guard

component, chose an ideal dataset, assessed highlight determination calculations, and prepared the IDS framework with managed ML calculations like KNN, DT, RF, and SVM. The exploration adds to online protection in IoT networks by proposing a powerful IDS framework consolidating irregularity recognition and AI strategies for DoS assault location. The system is very much organized and far reaching, however can be improved by thinking about information assortment and preprocessing, highlight designing, model choice and enhancement, cross-approval and hyperparameter tuning, genuine testing, logic and interpretability, and coordinated effort and benchmarking. Future work includes exploring feature selection with fewer features, testing on Raspberry Pi, evaluating with other datasets, and implementing deep learning algorithms. The research can be further improved by considering enhanced feature engineering, dynamic model adaptation, anomaly detection techniques, behavioral analysis, scalability and resource efficiency, and real-world testing.

The research by (Arisdakessian, 2022) introduced an exhaustive survey of Logical Man-made consciousness (XAI) methods applied to irregularity based interruption recognition frameworks (IDS) inside the Web of Things (IoT) organizations, reasoning that XAI can essentially upgrade trust and trust in digital safeguard components. The specialists tracked down that while present day simulated intelligence, especially profound learning (DL), is viable at recognizing irregularities in huge scope IoT datasets, the absence of interpretability in these models — frequently alluded to as the "black box" issue — represents a critical test. XAI was distinguished as a promising answer for this issue, offering a method for making sense of and legitimize model expectations, subsequently working on the straightforwardness and unwavering quality of IDS in IoT conditions. The procedure utilized was fitting, as it included an intensive survey of existing IDS and man-made intelligence models, featuring the present status of the field and distinguishing holes where XAI can be additionally incorporated. Be that as it may, the review could have profited from exact trial and error to approve the proposed structures and arrangements. By and large, the work gives significant experiences into the crossing point of XAI and IoT security, however future exploration ought to zero in on useful executions and the improvement of more modern XAI-driven network safety models.

(Hariharan, 2023) explores the transferability and defense mechanisms against adversarial attacks on Network Intrusion Detection Systems (NIDSs), particularly focusing on machine learning and deep learning models. By introducing domain-specific constraints to generate valid adversarial examples, the research addresses the complexity of network traffic and preserves its statistical and semantic integrity. The methodology is sound, employing both white-box (WB) and black-box (BB) attack scenarios on diverse datasets, such as UNSW-NB15 and IoT-23, under targeted and untargeted modes. The discoveries feature that antagonistic preparation altogether upgrades model power, and adaptability relies upon model likeness, regardless of the imperatives. The examination gives significant experiences into obliged ill-disposed assaults, however further investigation of true NIDS conditions and various assault vectors could upgrade its pragmatic appropriateness. Utilizing datasets, for example, UNSW-NB15 and IoT-23, the approach included both white-box (WB) and discovery (BB) assault situations under designated and untargeted modes. Key discoveries showed that antagonistic preparation essentially decreased assault achievement rates (ASR), dropping from 99.3% to 7% at times, and that adaptability was affected by model comparability. The examination offers important experiences, however further investigation in genuine NIDS conditions could improve its pertinence.

(Afolabi, 2024) centers around involving XAI to further develop trust the executives in Interruption Location Frameworks (IDS) using a Choice Tree (DT) model. The analysts applied the model to the KDD Cup 1999 dataset, which contains 42 traits and incorporates five principal classes: four named as assaults (DoS, R2L, U2R, and Testing) and one marked as expected. The review's system included parting the dataset into 60% for preparing, 20% for approval, and 20% for testing. Their outcomes showed that the Choice Tree model outflanked Help Vector Machines (SVM) and Calculated Relapse (LR) models in accuracy, review, and F1-score. For instance, the Choice Tree model accomplished better execution in recognizing both malignant and typical hubs contrasted with SVM and LR, especially in its capacity to deal with non-straight connections in the information. A key finding was the significance of specific elements in the dataset, for example, "V23" from the traffic highlight classification, which was positioned most noteworthy and demonstrated

assaults enduring over two seconds. Their work features that while the Choice Tree may not necessarily in every case outflank different calculations in expectation precision, its interpretability makes it ideal for understanding and making sense of choices in IDS, subsequently advancing trust. Notwithstanding, the review could have profited from more profound investigation of additional complicated models, similar to profound brain organizations, with present hoc logic strategies on balance both exactness and interpretability.

To make machine learning (ML) models easier to understand, (Cui, 2023) discovered that using Explainable AI (XAI) techniques on Intrusion Detection Systems (IDS) greatly improved clarity, especially in telling apart normal activities from attacks, like different types of Denial of Service (DoS) attacks. Their research revealed that global explanation methods, such as Permutation Importance (PI) and SHAP, boosted the model's performance by pinpointing the top 15 most important features. This resulted in better prediction accuracy, achieving 70-75% on the NSL-KDD dataset. Meanwhile, local explanation techniques such as SHAP, LIME, and CIU were helpful in providing detailed insights into specific predictions. For example, they discovered that SHAP was 100% consistent in explaining how features affected Random Forest predictions. When comparing SHAP and LIME, there was a 71% match for features with positive impacts and a 75% match for features with negative impacts. Their approach was well-organized, using both binary and multiclass classification methods on Random Forest models.

In this research, (Jemili, 2024) developed a smart deep learning approach named AEIDS (Auto-Encoder-IDS), which uses the random forest algorithm. They also designed the KOMIG (Knapsack Optimization and Mutual Information Gain) Intrusion Detection System (IDS) to identify network attacks by blending optimization techniques with machine learning. The findings revealed that the KOMIG IDS, particularly when using the K-Nearest Neighbors (KNN) classifier, outperformed other systems. It achieved impressive results: 97.14% accuracy, 95.53% precision, 99.46% recall, and 97.46% F1 score. Their method, which involved a two-step feature selection process (first using knapsack optimization and then mutual information gain), proved to be effective. This approach helped the system choose the most

relevant features, enhancing the machine learning models' ability to detect threats. The study also compared their system with other advanced IDSs, confirming its success.
(Xu, 2023) created a new system named GMM-WGAN-IDS to better detect intrusions in networks with complicated and uneven data. The system has three key parts: one for finding important features (using SAE), another for fixing data imbalance (using GMM-WGAN), and a third for sorting data (using CNN-LSTM). This system performed much better than older methods. For instance, on the NSL-KDD dataset, it increased accuracy, precision, recall, and F1 score by at least 5.8%, 3.9%, 5.8%, and 7.5%, respectively. On the UNSW-NB15 dataset, the improvements were 6.2%, 5.0%, 6.2%, and 6.6%. The system worked well because it used advanced techniques to better find features, balance data, and classify information, especially for rare or less common attacks. The results are strong and well-supported, showing that this system is a big improvement over older methods.

Huang (2023) discovered that their hybrid intrusion detection model, which combines Random Forest and XGBoost, significantly improves detection accuracy, achieving up to 97% on various datasets such as N-BaIoT, NSL-KDD, and CICIDS2017. This demonstrates that using multiple algorithms together through ensemble learning enhances detection performance, particularly in complex and dynamic IoT and Big Data environments. Their approach was highly effective, involving detailed data preparation, feature extraction, and the application of advanced machine learning techniques within a cloud-based system using Apache Spark and Microsoft Azure. This approach worked well with the high-dimensional and varied datasets in the study. The researchers also considered real-time and energy efficiency needs in IoT environments, which further proves the importance of their work. Overall, the findings and methodologies were robust, appropriate, and forward-thinking, contributing valuable insights to the field of intrusion detection.

In response to these issues, this paper by (RAHADIKA, p. 2022) suggested a network intrusion recognition model due to LightGBM and ADASYN oversampling methodology. They use data preparation techniques like one-hot encoding and normalization to strip away the significance of extreme values from the characteristics as a whole. First, data preprocessing we encode the data with labels as well as unique

heat encoding to better pass into the model for processing, after that we increase a few classes of data by ADASYN oversampling to make the number of samples in each class relatively balanced, and finally the features are passed into LightGBM for learning and training, and finally validated in the NSL-KDD dataset for multi-class classification experiments, and the results show that compared with other traditional algorithms, all data are improved and show better performance Finally, the LightGBM ensemble learning technique is used to reduce the model's time complexity while maintaining high detection accuracy.

In order to create new instances that are more accurately reflective of underrepresented groups, (Luo, 2023) proposed augmenting the standard GAN with neural layers and an unbalanced data filter. An IGAN-based Intrusion Detection Sys tem, or IGAN-IDS, is also developed to deal with class-imbalanced intrusion detection by making use of the instances produced by IGAN. IGAN-IDS is a framework that consolidates an organization of profound brain organizations, an IGAN, and a module for separating elements to distinguish pictures. Prior to doing anything more, they utilize a calculation called a feed-forward brain organization (FNN) to transform the crude organization qualities into include vectors. The IGAN then, at that point, delivers extra examples that likewise have an articulation in the dormant region. With DNN's completely connected layers and convolutional layers, we have the most ideal interruption ID framework that anyone could hope to find.

For vindictive way of behaving and assault recognizable proof in the data plane, (Choi, 2024) introduced a clever conveyed interruption discovery framework (IDS) for IoT networks in light of Programming Characterized Systems administration (SDN) and improved choice trees utilizing the Dark Opening Streamlining (BHO) calculation. Their discoveries showed that the proposed strategy altogether beat conventional models, accomplishing high precision paces of 99.2% on the NSLKDD dataset and 97.2% on the NSW-NB15 dataset. The review presumed that the blend of SDN design for network parceling and the streamlining of choice trees improved the framework's proficiency and exactness, especially while utilizing a cooperative location system. The approach was proper, as it joined progressed enhancement methods with appropriated network design, which is appropriate for the intricacies of

IoT conditions. Notwithstanding, further approval on true organizations could reinforce the ends. In general, their work really tended to the difficulties of interruption discovery in IoT organizations and showed the capability of enhanced choice trees inside a disseminated SDN system.

For the NSL-KDD dataset, decision tree (DT) based IDS is suggested by (Azam, 2023). They introduced NIDS-Vis, a unique black-box estimation planned to envision the decision furthest reaches of DNN-based Association Interference Revelation Structures (NIDS), and researched the split the difference among execution and badly arranged energy. The revelations revealed that as NIDS models become more definite and complex through getting ready, they make not well arranged regions near decision limits, making them more vulnerable against attacks. To address this, the makers proposed two readiness systems — incorporate space section and distributional setback — to overhaul the summarized opposing energy without basically compromising execution. The procedure was proper to the survey's objectives, as the usage of NIDS-Vis gave huge encounters into as far as possible scene, which is fundamental for understanding and chipping away at the not well arranged strength of NIDS. The investigation was fitting and comprehensive, offering a basic responsibility by keeping an eye on the original challenges of badly arranged power in NIDS, a space that shifts basically from standard controlled portrayal tasks. In general, the review was top notch and given significant procedures to upgrade the security of NIDS in ill-disposed conditions.

(He, 2024) presented an incorporated discovery model, MSCNN-LSTM, which joins Multiscale Convolutional Brain Organizations (MSCNN) and Long Momentary Memory (LSTM) organizations to investigate both spatial and fleeting elements of organization traffic for interruption recognition. The discoveries exhibit that MSCNN-LSTM beats ordinary models, like Lenet-5 and HAST, regarding precision, misleading positive rate, and bogus negative rate, especially in distinguishing uncommon assaults. The system was proper, utilizing a strong exploratory plan utilizing the UNSW-NB15 dataset and contrasting the proposed model against laid out benchmarks. The researchers successfully tested the model's ability to work well on new data using a second test set (Test_Set_B). They highlighted that the model performs better than

others when dealing with uneven data and rare attacks. This study was detailed and well-done, making important progress in the area of intrusion detection. It tackled problems with complex data and improved how accurately attacks are detected, without using older methods for selecting features. Still, the study could improve how it picks features and handles very uneven data in future work, as this is often seen in real-world situations.

Table 4

Summary of Various DL-based IDS with Strengths and Limitations

Paper/year	Method	Dataset	Performance metrics	Attack types	Limitations
Ke Luo (2023)	DNN	KDD-99	AC, F1-score, AUC, specificity, PE, RE	DoS, R2L, Probe, R2L	The model fails to display improved performance when it comes to R2L assaults.
Jeong et al. (2024)	DNN	CICDDoS 2019	AC, PE,RE, F1-Score	DDoS	The model performs better with less data, but with more data performance of the model was reduced.
Zahedi Azam et al. (2023)	CNN	CICDDoS - 2019	AC, PE, RE, F1	DDoS	We observed that the model does not provide good sensitivity from the experimental results.
Chao He et al. (2024)	CNN	InSDN, CIC-IDS2018 And UNSW-NB15	AC, PE, RE, F1	Botnet, DDoS, DoS, Probe, U2R, Web attacks and Password - Guessing	Computational time to select the best shrinking factor for SD-regularization more.

2.12 Adversarial Attacks Detection in Intrusion Detection System based on GAN Models

Intrusion Detection Systems (IDS) have evolved in tandem with the sophistication of malware and intrusion strategies. A particularly worrying tendency in this evolution is the introduction of self-adaptive malware, which is designed to adjust its behavior in real-time to avoid detection by security tools.

(Fu, 2023) presented an original GAN-based model, LE-GAN, to address the test of mode breakdown and work on the phantom spatial constancy in hyperspectral picture (HSI) super-goal (SR). Their discoveries show that LE-GAN beats existing SR techniques, exhibiting better strength than commotion, better generalizability across various sensors, and less aversion to upscaling factors, especially at higher upscaling levels. The model's viability is credited to the incorporation of a dormant encoder and an Unearthly Spatial Regularization Discernment (SSRP) misfortune capability, which together upgrade the generator's capacity to learn sensible ghostly spatial examples while relieving mode breakdown. The strategy was appropriate for the issue, as it consolidated progressed parts like 3D convolutional channels for otherworldly spatial element extraction and an idle complex planning, which were fundamental for tending to the particular difficulties in HSI SR. The work is honorable for its thorough assessment on different datasets and clamor levels, approving the model's exhibition and strength. Generally speaking, their methodology and ends are fitting and contribute altogether to propelling the field of HSI super-goal.

(YANG, 2022) introduces DA-GAN, a system that makes machine learning-based Intrusion Detection Systems (IDS) stronger by creating modified network attack data. This is done using a mix of Generative Adversarial Networks (GANs) and Domain Adaptation (DA) methods. Their results show that DA-GAN greatly boosts IDS performance, especially when there isn't much data available. Detection rates jumped from about 60% to over 99% in different situations. The approach, which tested various GAN types (WGAN, WGAN-GP, and WGAN-GP-TTUR) and used datasets like CIC-IDS2017 and CIC-IDS2018, was well-suited and thorough. It effectively tackled the problem of not having enough labeled attack data. This research is impressive because it not only proves the effectiveness of GANs and domain

adaptation in cybersecurity but also provides a practical way to keep improving IDS in real-world environments..

(Do Hoang, 2022) investigates different Generative Antagonistic Organization (GAN) structures to create reasonable organization traffic tests. It coordinates Outrageous Inclination Helping (XGBoost), a Group AI calculation that demonstrates viable for characterizing and identifying both noticed and unseen High level Industrious Danger (Able) assault tests across engineered and new information dispersions. The discoveries exhibit that the Wasserstein GAN design performs ideally, with an Earth Mover's Distance (EMD) of 10−3 reliably between the Pundit and Generator misfortunes, beating the vanilla GAN engineering. Execution measurements utilizing XGBoost and other assessment estimates show a high achievement rate, with an exactness of 99.97%, a review pace of 99.94%, and 100 percent accuracy. Also, the F1 score for recognizing Adept examples in manufactured information is 99.97%, and the Collector Working Trademark (ROC) Region Under the Bend (AUC) comes to 1.0, demonstrating extraordinary execution, outperforming past cutting edge strategies. At the point when tried on inconspicuous information, the proposed technique keeps on exhibiting ideal location, keeping up with 100 percent review, 100 percent AUC, and accuracy surpassing 90%.

(Anande, 2023) also proposes a GAN-based method where the generator computes the adversarial network traffic features to attack a Black-box IDS model. The study found that using Generative Adversarial Networks (GANs) for adversarial training made the Intrusion Detection System (IDS) much stronger against black-box attacks. However, how well it worked depended on the methods used to choose features and the type of attack. The research showed that GANs helped the IDS better handle adversarial attacks, as seen by better results in the Area Under the Curve (AUC-ROC) and higher accuracy in some tests. The methods used in the study, like preparing the data, training with GANs, and selecting features using Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), were suitable for achieving the study's goals. Nonetheless, the exhibition under black-box assaults showed that while GANs further developed recognition capacities, the component determination approach and the quantity of ages assumed a basic part in enhancing

results. Everything considered, further calibrating of element determination and extra analyses with various GAN designs or broadened preparing ages could have given significantly more hearty outcomes. In general, the work was effectively thought out and executed however could profit from extra refinement in highlight choice systems and ill-disposed preparing boundaries to accomplish ideal IDS execution.

The approaches suggested in studies by Afnan Alotaibi (2023), R. Zhang, S. Luo, and others (2022), and F. Haoyi, S. Anjani, and colleagues (2023) fall under a type of attack called adversarial attacks. These attacks focus on machine learning-based Intrusion Detection Systems (IDS). In an adversarial attack, the goal is to trick a machine learning model into making a mistake by changing the input data slightly. The altered data, known as an adversarial feature or example, is made to look very similar to the original data using a specific measure of similarity, like Euclidean distance. A special algorithm is used to create this altered data within harmful network traffic. Once the adversarial feature is identified, the malware can modify the network traffic to match the computed value, thereby ensuring that the IDS misclassifies the altered traffic as benign. The minor perturbation applied to the feature values allows the malware to maintain its malicious behavior while evading detection. A detailed review of adversarial attacks against intrusion detection can be found in K He, DD Kim et al. (2023) . Although the methods proposed in (Dini, 2023) While these methods have proven effective in compromising machine learning-based IDS models, their impact in real-world network environments is constrained by ideal assumptions, such as the availability of a large labeled dataset for training and prior knowledge of the IDS model's loss function. To evaluate whether adversarial attacks could pose a greater threat to network security if these assumptions are relaxed, we propose a novel approach called the Generative Adversarial Active Learning (Gen-AAL) algorithm. This method is designed to compute the adversarial features of network traffic flows, even in scenarios where the ideal conditions are not met.

The landscape of Intrusion Detection Systems (IDS) has evolved to counter increasingly sophisticated malware and network attacks, including adversarial attacks that aim to exploit vulnerabilities in Machine Learning (ML)-based IDS. The evolution of these adversarial strategies has driven significant research efforts, especially in

using Generative Adversarial Networks (GANs) to enhance adversarial attacks and, conversely, to improve IDS defenses. The following review outlines some prominent works in this domain, highlights the gaps, and draws conclusions regarding adversarial attack detection in IDS. Recent research on adversarial attacks and detection in IDS using GANs reveals several advancements and challenges. Works like (Ahmed, 2022) and (Alotaibi, 2023) showcase sophisticated adversarial strategies, including self-adaptive malware and GAN-generated data to mislead classifiers, exposing vulnerabilities in IDS models.

Additionally, while works like (Haoyi, 2023) and (He, Adversarial machine learning for network intrusion detection systems: A comprehensive survey., 2023) create highly effective adversarial traffic, the scalability of defenses across diverse environments is inconsistent, and IDS models often struggle to generalize. A further challenge lies in the explainability and interpretability of these models, as few works fully leverage XAI methods like SHAP to clarify how IDS decisions are made and why they might fail under adversarial conditions, leaving room for improvement in creating more transparent and robust detection systems. The development of GAN-based adversarial attacks and corresponding defenses has significantly impacted the field of intrusion detection. While GANs have proven to be effective tools for generating realistic adversarial traffic that can evade detection, the practicality of these attacks in real-world scenarios is still limited by factors such as data availability and model transparency.

A big problem is that methods to detect attacks don't work well in different situations. Right now, defenses are usually made for certain types of attacks or specific data, so they don't work as well in real-world networks where traffic and attack methods change quickly. Also, because IDS models are hard to understand, it's difficult to figure out how these models are tricked by attacks. This makes it harder to create stronger systems that can handle these threats

2.13 Some Open Issues and Research Challenges

These days, IDS are an essential part of daily life. However, developing an IDS that recognizes and reacts to various threats and attacks is challenging. As a result,

researchers have conducted many studies in the field of IDS for various applications. Some re- searchers contend that DL, via a neural network, will provide IDS additional flexibility, enabling it to detect and categorize hazardous attacks more successfully. The comparative analysis of various ML and DL models was provided by (Abdulganiyu O. H., 2024) . They stressed that ensemble learning has a good effect on intrusion detection research. Further, DL models require more training time than traditional ML models because DL models are deep in structure. The performance of DL models depends on the design, hyperparameters, and the number of iterations. Further, (Muneer, 2024) thoroughly analyzed network-based intrusion detection systems and emphasized the importance of labeling data while evaluating and training the intrusion detection systems.

• A high-quality IDS dataset is essential for testing and validating IDS Models. However, as mentioned in the previous section, most public datasets have missing values, incomplete network features, raw pcap files, and incomplete CSV files, which may reduce the model's efficacy. This could be addressed by preprocessing the data by removing duplicate and noisy data.

• In order to identify new attacks, detection methods need to be retrained using new training data with minimum training time.

• The IDS datasets contain redundant features, which reduce the performance and increases the training and testing time. Hence, it is essential to develop novel techniques to mitigate the dimensionality of the dataset.

• The training time of ML/DL methods can be dwindled by selecting appropriate hyperparameters using optimization methods without compromising performance.

• Another difficulty with ML and DL methods is model overfitting. Overfitting occurs when algorithms are heavily influenced by training data.

• Only a few IDS methods can detect both signature and anomaly-based attacks. In the future, investigations should be done to develop efficient hybrid models to handle known and unknown attacks with low FAR and less computational time.

• The literature shows that the researchers use black box models like RF, SVM KNN, AE, CNN, etc. Most works exhibit an excellent detection rate and low FPR. However, notable challenges in IDS, namely systems' transparency. Nevertheless, the model's predictions should be understandable since security analysts now base their decisions

on the recommendations of an IDS. Further, there is an opportunity for researchers to develop novel XAI in IDS to interpret their model with better explanations.

Several open issues and research challenges persist in the development of Intrusion Detection Systems (IDS), particularly in the context of Machine Learning (ML) and Deep Learning (DL) models. One key challenge is the reliance on high-quality datasets for training and validating IDS models, which are often plagued by missing values, incomplete network features, and redundant data, thus necessitating advanced preprocessing and feature selection techniques. Another challenge is the high computational cost of training DL models, which can be mitigated by optimizing hyperparameters and reducing training times without compromising performance. Overfitting also remains a significant issue, particularly with black-box models like RF, SVM, and CNN, which, despite high detection rates, lack transparency and interpretability.

2.14 Conclusion

The usage of cyberspace increases daily, leading to new and complex attacks. It becomes challenging to detect them with traditional techniques. However, there is now ongoing research in creating novel models, such as creating new datasets or combining algorithms. Hence, we have reviewed various recent IDS which are based on the ML, DL, and XAI methods. It provides updated relevant information to new researchers and records upcoming signs of progress in IDS. It also highlights the concept of IDS, the classification of DL and ML algorithms, how DL and ML algorithms are used to design the IDS framework, and the usage of XAI in IDS. It was evident that datasets significantly impact this field because some consider them outdated or to contain redundant information. Choosing an appropriate dataset is a challenging task.

Chapter 3

Methodology

3.1 Overview

Intrusion Detection Systems (IDS) are essential components of IT security, designed to safeguard networks and systems from cyber threats targeting data integrity, confidentiality, and availability. While traditional signature-based methods have been foundational in IDS development, they often fail to detect novel and evolving threats, necessitating the use of advanced techniques like anomaly-based IDS. These systems rely on modeling normal network behavior and flagging deviations as potential intrusions. To enhance the effectiveness and robustness of IDS in modern threat landscapes, our methodology focuses on adversarial robustness, feature-ranking explainability, and resilience testing under adversarial perturbations.

This thesis presents a new Intrusion Detection System (IDS) that uses LightGBM to address challenges in adversarial machine learning and enhance the system's transparency through Explainable AI (XAI) methods. Unlike traditional IDS approaches that mainly detect unusual activities, our method evaluates the IDS in adversarial scenarios by focusing on the most critical features highlighted by XAI. Using a heuristic search, we generate adversarial examples that exploit vulnerabilities in these key features. This helps us analyze the system's robustness and clarity when dealing with adversarial attacks.

Our methodology includes the following core components:

1. **Adversarial Attack on Ranked Features:**
    - Important features are found using XAI (Explainable AI) methods like SHAP (SHapley Additive exPlanations) to see how much they affect the model's decisions.
    - A smart search creates fake attack examples by changing these important features, mimicking real attacks that could weaken the IDS (Intrusion Detection System).

2. **Defense Strategy Evaluation Using Feature Rankings:**
    - By adding small, tricky changes to the IDS, we check if the system can still work well and make sense even when under attack.
    - We look at how these tricky changes affect the order of important features to see if the model's decisions stay stable and trustworthy.
3. **Adversarial Sample Generation and Analysis:**
    - Adversarial samples are created by making small changes to the input data, but these changes are limited in size to make sure the data still looks valid and realistic. This means following specific rules, like Lp norms, and keeping the features consistent so the altered data still makes sense.
4. **Evaluation Metrics for Robustness and Interpretability:**
    - The system is tested using measures like precision, recall, and F1-score, especially in challenging or adversarial situations.
    - Important measures of strength, such as the percentage of altered features, average/maximum Lp norms of changes, and criteria for failure (like wrong classification, staying within the allowed change limits, and keeping features consistent), help us understand how well the IDS can handle adversarial attacks.

We developed an Intrusion Detection System (IDS) using LightGBM, a machine learning method, and trained and tested it with the CIC-IDS2017 dataset. This dataset is widely used to check how well IDS systems perform in real-world network attack scenarios. Our system outperforms older models like SVM by being more resistant to attacks designed to trick it. It also achieves higher accuracy, better detection rates, and vua stronger overall performance score (F1-score). Moreover, our system is more reliable and makes its decision-making process clearer. By using Explainable AI (XAI), we ensure that the IDS can explain its decisions in a way that’s easy to understand, even when attackers try to interfere with important data. This system addresses major challenges in IDS development: staying dependable during attacks, improving defense against threats, and making the decision process more transparent. These advancements help create safer and more trustworthy IDS solutions for real-world applications.

3.2 Proposed System

A proposed IDS integrates CICIDS-2017 along with NSL-KDD datasets to deploy LightGBM model and XGBoost model for intrusion detection purposes through XAI methods and adversarial attack testing. A training process for the LightGBM model uses the CICIDS-2017 dataset with normal and malicious network activities alongside XAI approaches SHAP and LIME to determine significant features that influence prediction results. A heuristic-based search methodology creates adversary samples by modifying these vital features to produce inputs which deceive the model without altering its original semantic content. The IDS is tested with these adversarial samples to track performance changes using precision and accuracy as well as recall and F1 score metrics. This system behaves as an adversarial attack portable method by testing XGBoost-derived adversarial samples against LightGBM to examine vulnerabilities which derive from shared decision boundary areas. The study investigates interpretable behaviors of the intrusion detection system following the attack to expose how adversarial samples affect feature importance ratings while determining vulnerable features. The security framework evolves from adversarial testing followed by iterative enhancement which increases the susceptibility of intrusion detection systems to efficiently detect threats under adversarial circumstances.

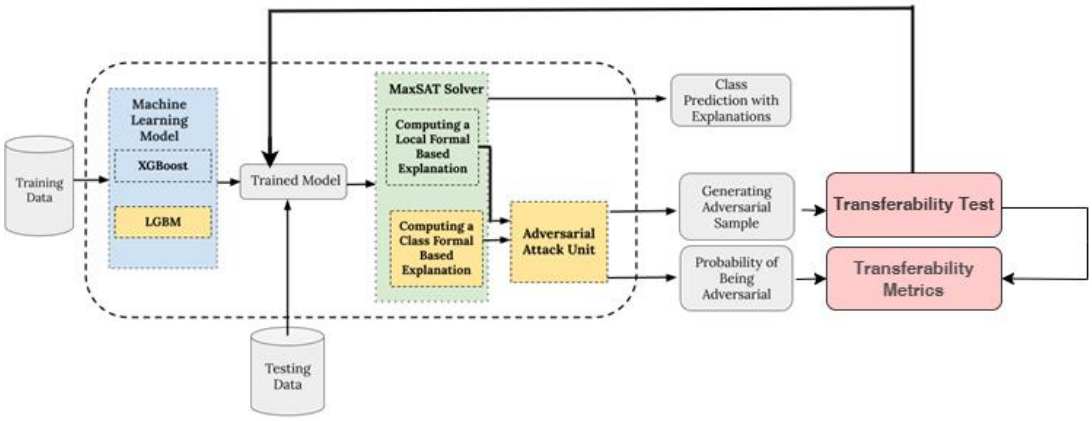

Figure 16. Proposed system.

3.3 Data Collection

3.3.1 Cic-ids2017 dataset. This The CIC-IDS2017 dataset is collected over a span of 5 days based on abstract behaviour of 25 users based on the HTTP, HTTPS, FTP, SSH, and email protocols. The implemented attacks include Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS (Figure 17). Figure 18 illustrates the testbed architecture of the dataset, comprising two distinct and isolated networks: the Victim-Network and the Attack-Network. The Victim-Network is equipped with essential components such as a router, firewall, switches, and multiple devices running widely-used operating systems, including Windows, Linux, and macOS. This network also includes three servers, one firewall, two switches, and ten PCs connected through a domain controller (DC) and active directory for centralized management.

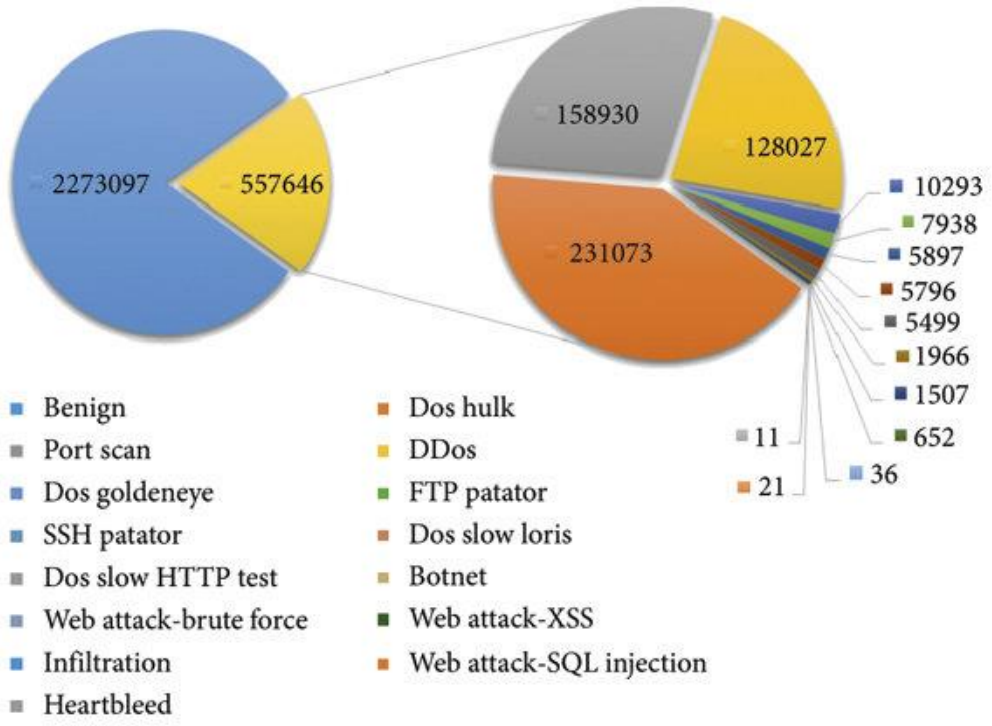

Figure 17. CIC-IDS2017 Dataset Distribution (Arisdakessian, 2022).

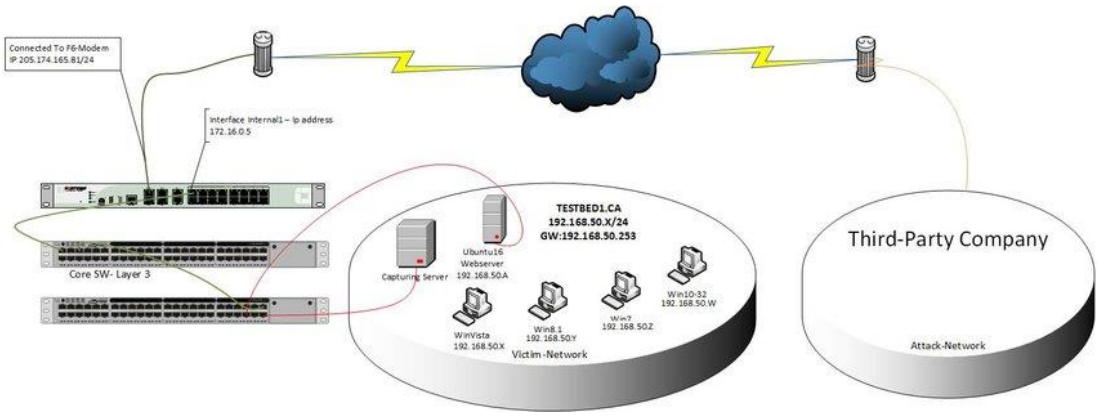

Figure 18. testbed architecture of CIC-IDS2017 dataset (Cui, 2023).

3.3.2 Nsl kdd v2. The NSL-KDD V2 dataset is a better version of the original NSL-KDD dataset. It has been made more consistent, and a new group was added by combining some non-normal classes. This dataset was developed by Abluva Inc., a company that focuses on data security research and is based in Palo Alto. Abluva's advanced data protection platform helps organizations keep their data safe using modern security tools, such as detailed access control and advanced data masking methods like pseudonymization, anonymization, and randomization. These tools help companies share data more easily while keeping it safe from theft and following the law. Abluva's special system for spotting threats uses unique technology to find dangers without causing big problems for daily work or marking small, harmless changes as issues.

Number	Data features	Number	Data features	Number	Data features	Number	Data features
1	Duration	12	Logged_in	23	Count	34	Dst_host_same_srv_rate
2	Protocol_type	13	Num_compromised	24	Srv_count	35	Dst_host_diff_srv_rate
3	Service	14	Root_shell	25	Serror_rate	36	Dst_host_same_src_port_rate
4	Flag	15	Su_attempted	26	Srv_serror_rate	37	Dst_host_srv_diff_host_rate
5	Src_bytes	16	Num_root	27	Rerror_rate	38	Dst_host_serror_rate
6	Dst_bytes	17	Num_file_creations	28	Srv_rerror_rate	39	Dst_host_srv_serror_rate
7	Land	18	Num_shells	29	Same_srv_rate	40	Dst_host_rerror_rate
8	Wrong_fragment	19	Num_access_files	30	Diff_srv_rate	41	Dst_host_srv_rerror_rate
9	Urgent	20	Num_outbound_cmds	31	Srv_diff_host_rate		
10	Hot	21	Is_host_login	32	Dst_host_count		
11	Num_failed_logins	22	Is_guest_login	33	Dst_host_srv_count		

Error! Reference source not found.. NSL-KDD dataset.

The Network traffic identification features in the NSL-KDD dataset amount to 41 attributes with which to detect violations. Basic features (duration and protocol type and service flag) fall under the first category while content-based features (number of failed logins combined with root shell access and file creations) make up the second category and time-based traffic features (count and same service rate and error rates) fall into the third category along with host-based traffic features (destination host count and service rate and error rate). The features analyze network behavior to detect regular activities from harmful ones so these labels make the dataset vital for IDS evaluation purposes.

3.4 Data Preprocessing

Generative adversarial models are very sensitive to the data they use, as their performance depends on different settings. Because of this, preparing the data properly is very important to make sure adversarial attacks work well. In this research, we used several steps to prepare the data, such as cleaning, encoding, scaling, and selecting important features, as shown in Figure 19

First, we cleaned the data by removing features in the network traffic that had the same value for all data points (constant features) or almost the same value (quasi-constant features). Next, we used an ordinal encoder to change all non-numeric features into numbers. After that, we normalized the numeric features using the min-max scaling method, which makes sure all features are on the same scale. These steps helped speed up the learning process by reducing the amount of data and the number of changes needed during training.

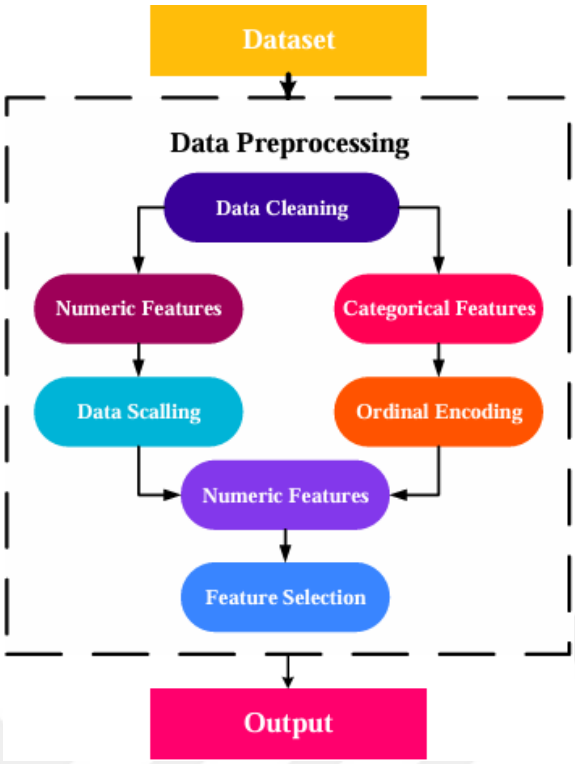

Figure 19. Data Preprocessing steps.

3.5 Explainable Frameworks

3.5.1 Local interpretable model-agnostic explanation (Lime). Local Interpretable Model-agnostic Explanation (LIME) (Malik, 2022) is a versatile explainability framework that employs local surrogate models to interpret individual predictions made by a black-box model, whether for classification or regression tasks. These surrogate models are simplified versions of the complex original model and provide accurate approximations only for specific subsets of the data. LIME works by perturbing a reference input, such as adding noise or altering values, to generate multiple variations of the input. It then evaluates how the model's probability scores change across these perturbed instances. A dataset of perturbed samples Xp , derived by modifying the interpretable components of the reference input (x), is created. The scores for these samples are computed, and a surrogate model (fs \Fs) is trained on Xp. The training involves weighting pixel patches based on proximity πx, as expressed in equation (3.1), where fo is the original model, Fs represents the family of surrogate models, $\Omega(f_s)$ defines the complexity of fs , and L

is the loss function. The pixels with the highest weights indicate the key attributes influencing the model's decision for the given input, thus providing an interpretable explanation of its classification.:

$$Explanation(x) = \underset{f_s \in F_s}{argmin}\, \mathcal{L}(f_0, f_s, \pi_x) + \Omega(f_s) \quad (3.1)$$

3.5.2 Integrated gradient. Coordinated Inclination (IG) (Bouaziz A. e., 2023) figures the slope of the model's forecast concerning its feedback highlights. IG is based on top of two maxims that were not fulfilled by some other attribution strategies at the hour of its creation. The two sayings are:

Sensitivity

Implementation Invariance

Sensitivity and Implementation Invariance are two key properties of attribution methods. Sensitivity ensures that if an input and its baseline differ in one feature, and this difference leads to a change in the model's output, the differing feature must receive a non-zero attribution (Bouaziz A. N., 2023). Implementation Invariance, on the other hand, guarantees that attribution methods provide identical attributions for functionally equivalent models—those that produce the same outputs for all inputs, regardless of differences in their internal implementations. Integrated Gradients (IG) satisfies both properties and is calculated in five steps, beginning with defining a neutral baseline input for the model's output prediction.

3.5.3 Shapley additive explanations (SHAP). SHapley Additive exPlanations (SHAP) (Habib, 2023) is an explainability method that quantifies the contribution of each input feature to the output prediction. Based on Shapley values from coalitional game theory, SHAP fairly allocates the prediction among input features, allowing for a precise measure of each feature's importance in the model's decision. As an additive feature attribution method, SHAP ensures consistent and reliable global interpretations for individual data samples. It evaluates the contribution of a feature by replacing it with random variables and measuring the impact on the output prediction, calculated as the relative difference from the original prediction. The weights for DeepLIFTSHAP $\pi z()$ can be determined by the equation (3.2) where $|z'|$ is the number of features considered for the coalition and M is the maximum coalition among features.

$$\pi_z(Z') = (M-1)/\big((M/|Z'|) \times |Z'| \times (M-|Z'|)\big) \quad (3.2)$$

3.6 LightGBM Model

LightGBM partitions the tree leaf wise, while other helping calculations construct it level−wise. It picks the leaf to part that it accepts will bring about the biggest abatement in misfortune capability. leaf−wise creates parts in view of their commitment to the worldwide misfortune as opposed to the misfortune over a particular branch, in this way it once in a while learns lower-mistake trees "quicker" than level−wise.

The diagram below shows the split order of a hypothetical binary leaf−wise tree to a hypothetical binary level−wise tree. It is interesting that the leaf−wise tree can have multiple orderings, whereas the level−wise tree always has the same order (Agalit, 2022).

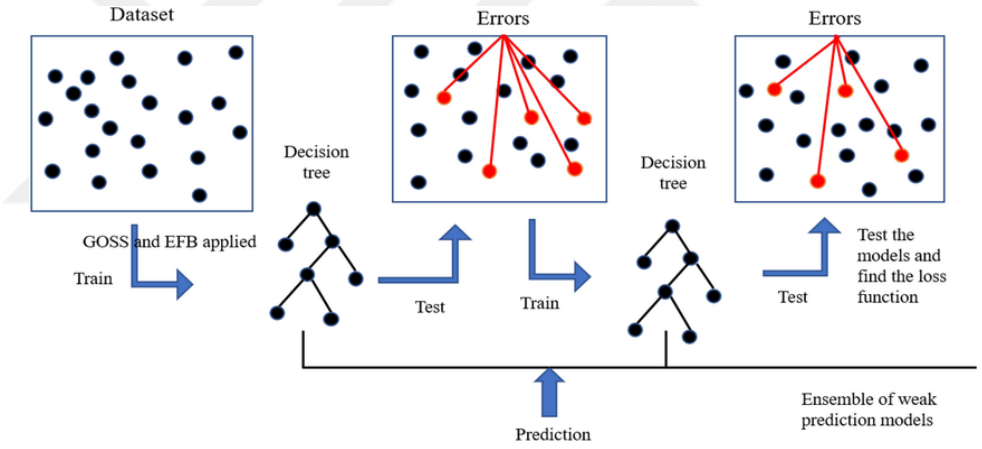

Figure 20. LightGBM model.

LightGBM (Light Gradient Boosting Machine) is a high-performance, efficient, and scalable implementation of gradient boosting that is specifically optimized for speed and memory efficiency. Unlike traditional gradient boosting algorithms that grow trees level-wise (expanding all nodes at a given depth before moving to the next level), LightGBM grows trees leaf-wise. This method, which focuses on one part at a time, helps the algorithm concentrate on the areas that are most likely to give good

results. This makes it quicker and usually more precise when dealing with big sets of data.

When it comes to spotting intrusions, LightGBM works really well because it can manage large and complex datasets, such as CICIDS-2017, which includes many different types of network traffic details. Here's how LightGBM is used for detecting intrusions:

1. Feature Importance:
    - LightGBM finds and focuses on important details in network traffic, like the size of data packets, the types of protocols used, and how long connections last. These details help spot unusual activity or decide if the traffic is safe or harmful.
2. Efficient Training and Testing:
    - Because intrusion detection datasets are usually very big, LightGBM is a good choice because it can handle data efficiently and grow trees in a way that focuses on the leaves. This makes training and making predictions much faster.
3. Binary and Multi-Class Classification:
    - Intrusion detection often involves distinguishing between normal and malicious traffic (binary classification) or categorizing types of attacks (multi-class classification). LightGBM adapts to these tasks using its gradient boosting mechanism.
4. Robustness to Imbalanced Data:
    - Intrusion detection datasets are typically imbalanced, with more benign traffic than malicious samples. LightGBM handles this by adjusting weights or using evaluation metrics like and F1 score that emphasize performance on minority classes.
5. Integration with Explainability (XAI):
    - In this project, LightGBM's predictions were further analyzed using XAI techniques to rank features by importance. These rankings helped generate adversarial samples targeting critical features,

allowing researchers to evaluate and improve model robustness.

3.7 XGBoost Model

XGBoost represents a powerful speed-focused version of gradient boosting which optimizes both performance and speed. Unlike LightGBM which extends one leaf per build step XGBoost constructs its trees through an organized system. The model uses a tree expansion system which builds complete levels of nodes before moving to building the subsequent level. XGBoost reaches optimal performance by finding the right balance between reliably handling fresh data while maintaining efficient computing operations which explains its popularity for multiple machine learning assignments.

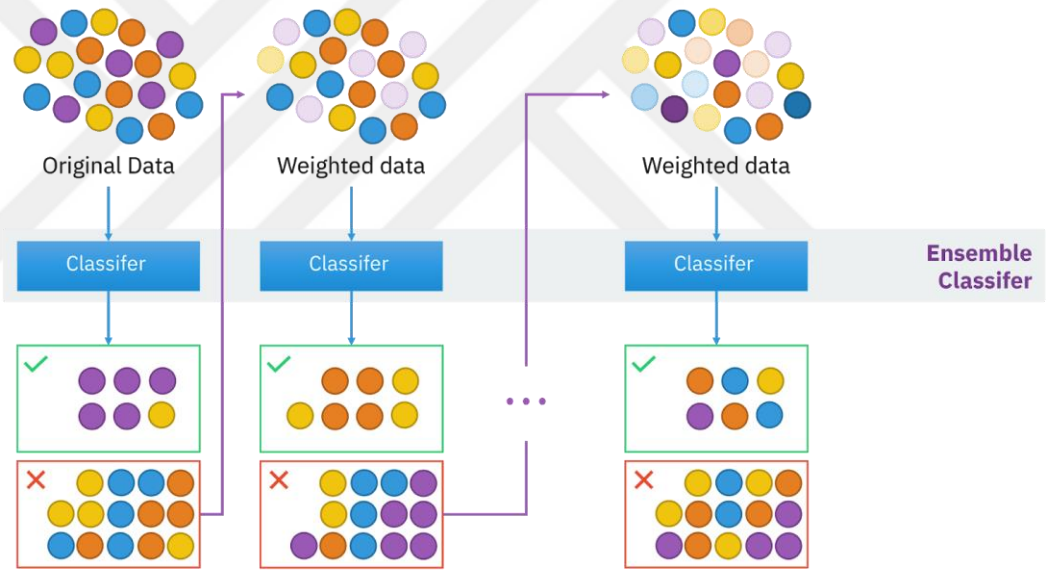

Figure 21. XGBoost Model.

A visual representation shows how XGBoost uses level-wise split order while LightGBM chooses leaf-wise tree splitting..

1. Feature Importance:

   XGBoost utilizes network traffic analysis to recognize crucial features that contain information about communication protocol types and connection durations together with packet size measurements.

2. The critical features provide essential distinctions that separate normal traffic from different forms of cyberattacks. Efficient Training and Testing:
    - Despite handling large datasets like CICIDS-2017, XGBoost optimizes memory usage with sparse-aware computation and parallel processing.
    - Its level-wise approach ensures that splits are made evenly across the tree, preventing overfitting in early stages.
3. Binary and Multi-Class Classification:
    - XGBoost is well-suited for binary classification (normal vs. malicious traffic) and multi-class classification (detecting different attack types).
    - Its ability to optimize decision boundaries makes it highly effective in intrusion detection systems (IDS).
4. Handling Imbalanced Data:
    - Since intrusion detection datasets are often imbalanced, XGBoost includes scale_pos_weight and balanced objective functions to improve detection of minority classes.
    - It also relies on evaluation metrics like F1-score and AUC-ROC to ensure accurate classification.
5. Integration with Explainability (XAI):
    - In this project, SHAP and LIME were applied to analyze XGBoost's feature importance, revealing which network features contributed most to classification.
    - These insights helped craft adversarial samples that tested model robustness and improved its defense against adversarial attacks.

3.8 Generate Adversarial Samples: Transferability

1 .The SHAP analysis helps users identify source model (XGBoost) critical features which drive its prediction results.

The process starts by manipulating significant model features within the framework of adversarial sample creation to manipulate a source model's decision process.

2. Transferability Test:

Researchers should perform tests on the manufactured adversarial samples using the target IDS model (LightGBM).

The decision patterns shared between two models enable adversarial samples destined for one device to cause misclassification on another.

3. Evaluate Impact:

Under adversarial attacks we should measure how metric performance measures (e.g. accuracy and precision) deteriorate for the target model system.

Interpretability tools should be used to analyze how the IDS model analyzes adversarial samples that were transferred from another IDS.

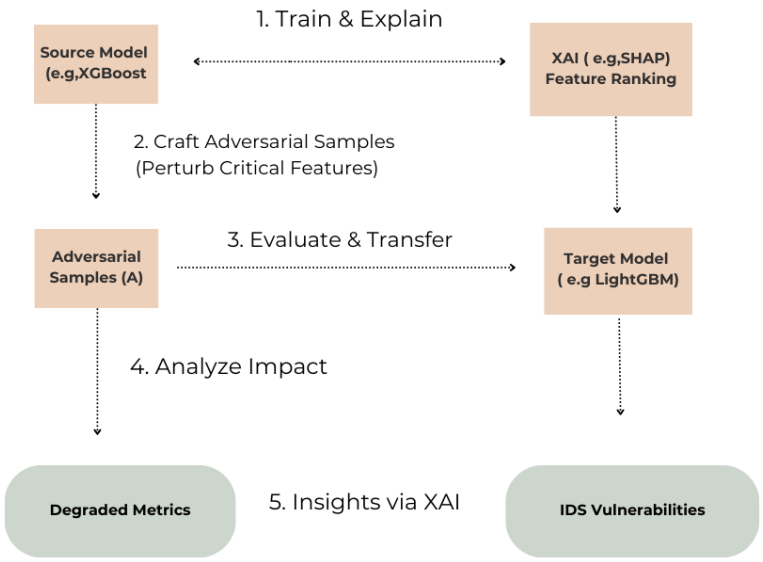

Figure 23. Transferability of Adversarial Attacks.

3.9 Adversarial Examples Generation Methods

Once the intrusion detection model has been trained and evaluated, it must be deployed for real-world use. However, adversarial attacks pose a significant threat by exploiting weaknesses in the model to degrade its performance. These attacks manipulate the decision boundaries of machine learning models by introducing small, often imperceptible perturbations to input data, leading to incorrect predictions.

Adversarial attacks are generally categorized into white-box and black-box attacks:

- White-box attacks assume that the attacker has full access to the model's parameters, architecture, and training process. This allows them to craft adversarial samples using gradient-based methods and optimization techniques to effectively deceive the model.
- Black-box attacks, on the other hand, occur when the attacker lacks direct knowledge of the model but can observe inputs and outputs. Using heuristic and trial-and-error methods, the attacker generates adversarial samples that trick the model without needing access to its internals.

In this research, we focus on black-box adversarial attacks to evaluate the robustness of the intrusion detection system. We employ heuristic search techniques that iteratively modify critical input features identified by explainable AI (XAI) methods such as SHAP and LIME. These perturbations are designed to maximize the probability of misclassification while preserving the semantic integrity of the original data. Additionally, we investigate the transferability of adversarial attacks, where adversarial samples generated for one model (e.g., XGBoost) are tested on another model (e.g., LightGBM). By analyzing the effectiveness of these attacks, we assess how they impact XAI-based explanations and refine the model to improve its resistance to adversarial manipulation.

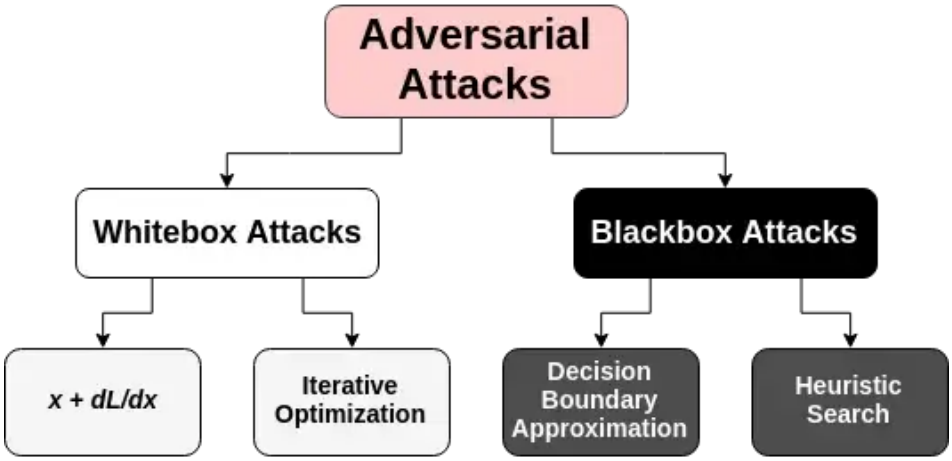

Figure 22. Ontology of adversarial attacks based on knowledge (Bouaziz A. N., 2023).

Heuristic search methods work by trying out different inputs over and over to find cases that fool the target model. These methods don't rely on gradient information. Instead, they use strategies like evolutionary algorithms, simulated annealing, or random search to figure out the best way to tweak the input. By testing many changes and focusing on the ones that make the model fail, heuristic searches can effectively find weak spots in machine learning models..

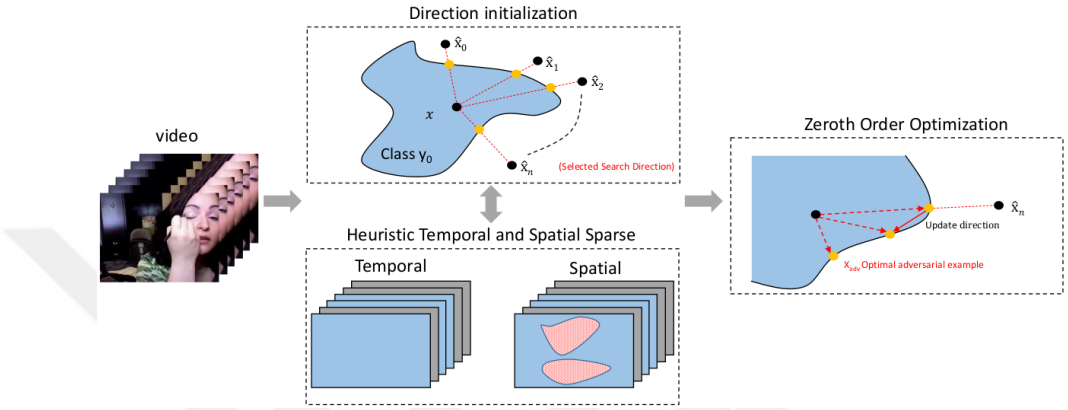

Figure 23 Overview of the heuristic algorithm for black-box adversarial attacks (Arreche, 2024).

3.10 Conclusion

This chapter talked about the ways we made the new Intrusion Detection System (IDS) better and simpler to use, especially when dealing with attacks. Unlike older systems, this one pays attention to attacks that go after important parts found using XAI methods. We used a clever way to make fake attack examples, which helped us check where the IDS might fail by changing the key parts it needs to work right. These fake examples were like real attacks, so we could really test how strong the system is against them.

Chapter 4

Findings

4.1 Introduction

The new IDS undergoes evaluation through test results that observe its response patterns during attacks. The system employs crucial attributes selected by XAI (Explainable Artificial Intelligence) techniques for executing intrusion detection functions. The IDS security level was tested through artificial attack scenarios developed from vital elements found through SHAP method analysis alongside its compatible explanatory techniques. The attacks function as representative digital attacks which help us identify IDS detection precision and test its detection capabilities. The IDS performs model training through LightGBM with the CIC-IDS2017 dataset to deliver precise intrusion detection capabilities which include its explanation features. System performance evaluation relies on precision together with recall and F1-score as its measurement metrics. The system vulnerability analysis depends on Lp norms combined with feature perturbation proportions to assess response outcomes from fake attacks.

overall feature influences and locate defining elements that produce maximum results to adversarial events. LIME offers direct explanations to users for their single predictions. The complete assessment of framework adversarial resilience becomes possible by incorporating configuration and showed both findings and insights from running adversarial tests that reveal how the proposed system functions during attacks without losing interpretability.Inspection of system accuracy and changes in feature importance becomes possible through this testing method when dealing with adversarial samples. Since SHAP and LIME offer understandable explanations we use these XAI tools to understand model decision processes under adversarial attacks. By employing SHAP users can discover these methods into power testing procedures. The experiment described its

4.2 XAI Feature Results

The SHAP value distribution reveals how characteristics rank in impact strength because positive and negative values become easily distinguishable between important traits. The research shows feature importance differences between datasets which proves why feature engineering and detection models need custom adjustment to individual datasets for optimal results. The obtained results show that explainability matters as a means to establish and maintain trust in automated intrusion detection systems.

4.2.1 SHAP and LIME Results of CICIDS2017 Dataset . SHAP analysis Models depend most heavily upon Feature 17 as this parameter shapes their prediction outcomes according to the established models. The model predictions rely on Feature 17 and Feature 9 and Feature 1 as essential factors although Feature 10 and Feature 5 play an insignificant role in the predictions. The obtained results guide model robustness development along with highlighting features that need priority attention to increase model interpretability and performance metrics.

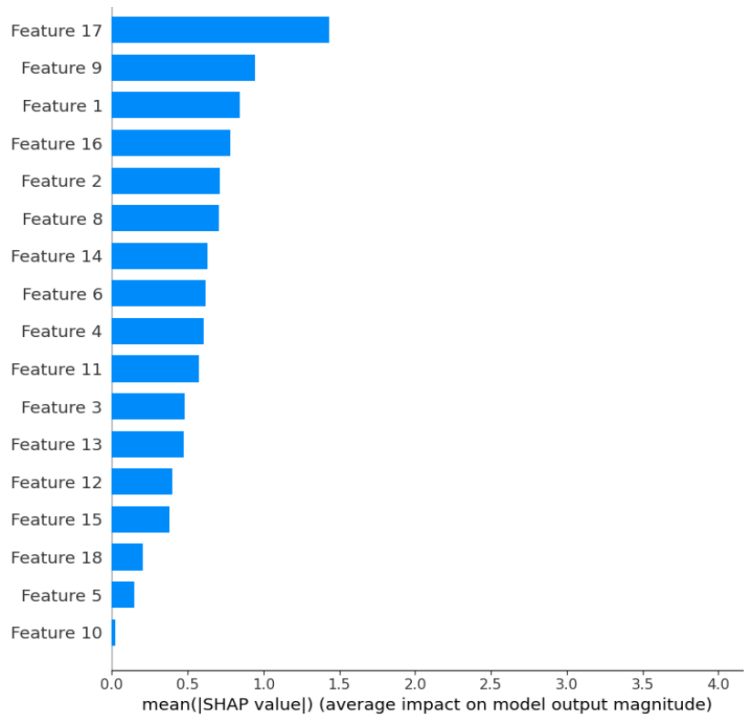

Figure 24. SHAP Results of CICIDS2017.

The local SHAP analysis describes how features individually affect a model prediction. Bwd Packet Length Max and Init_Win_bytes_backward were the features that exerted the greatest influence on the prediction toward Class 0. The feature min_seg_size_forward showed minimal support for Class 1 yet lost effect to negative features which dominated in prediction. Specific features act as key components during the decision-making process which the model applies to its sample evaluation.

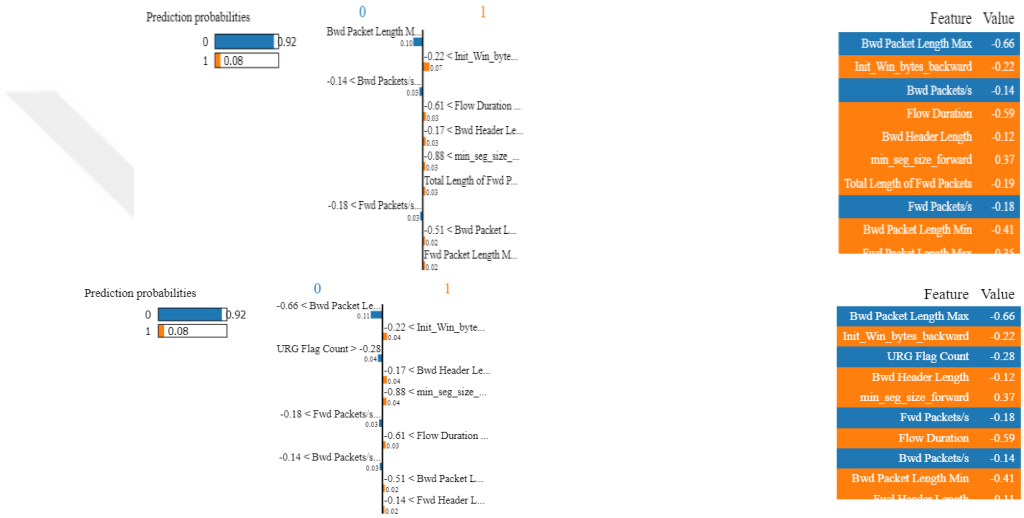

Figure 25. LIME Results of CICIDS2017.

The SHAP assay reveals the distinct manner in which each feature affected the model prediction for this particular outcome. Bwd Packet Length Max (-0.66) together with URG Flag Count (-0.28) demonstrated the most substantial influence which directed the prediction towards Class 0. Prediction samples containing min_seg_size_forward at 0.37 and Flow Duration at -0.59 displayed less impact which led to a higher likelihood of prediction belonging to Class 0 (92%). The backtrace packet length together with flag characteristics emerge as the leading factors that determine how the model reaches its conclusions.

4.2.2 SHAP and LIME Results of NSL-KDD Dataset. The SHAP summary plot presents a listing of features structured based on their mean impact towards modeling results. The model predictions depend most heavily on Feature 36 with Feature 53 and Feature 21 having substantial effects on the predictions. Feature 31 together with Feature 33 show minor influence when compared to other features in the model. The model's performance depends heavily on specific features as revealed through this evaluation process which enables users to pinpoint key characteristics that need further improvement or interpretation refinement.

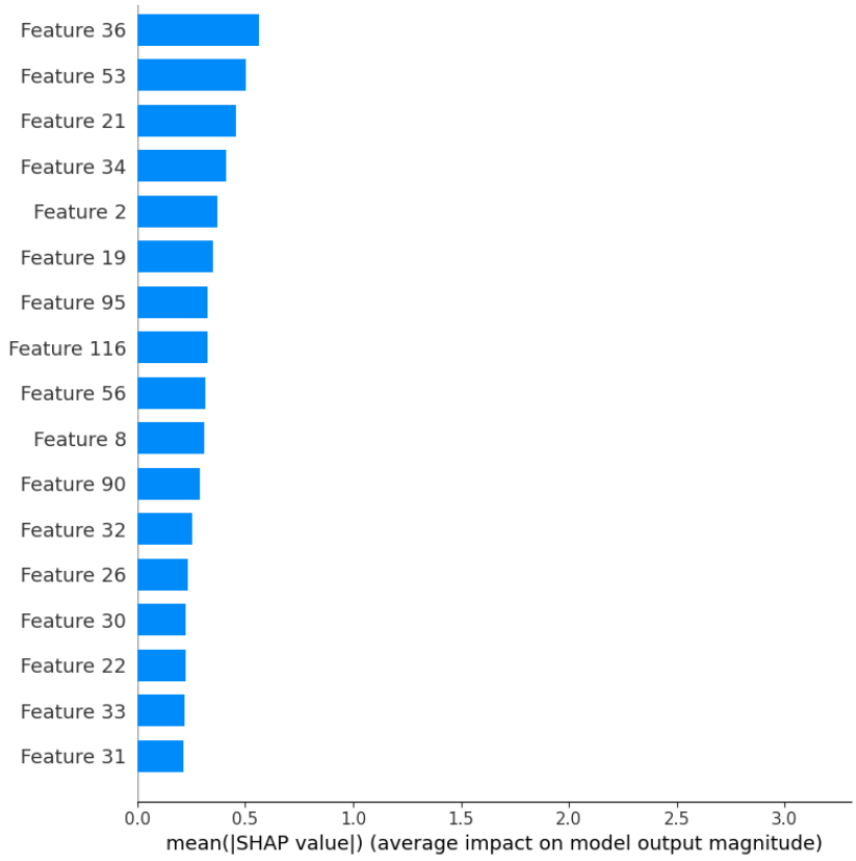

Figure 26. SHAP Results of NSL-KDD.

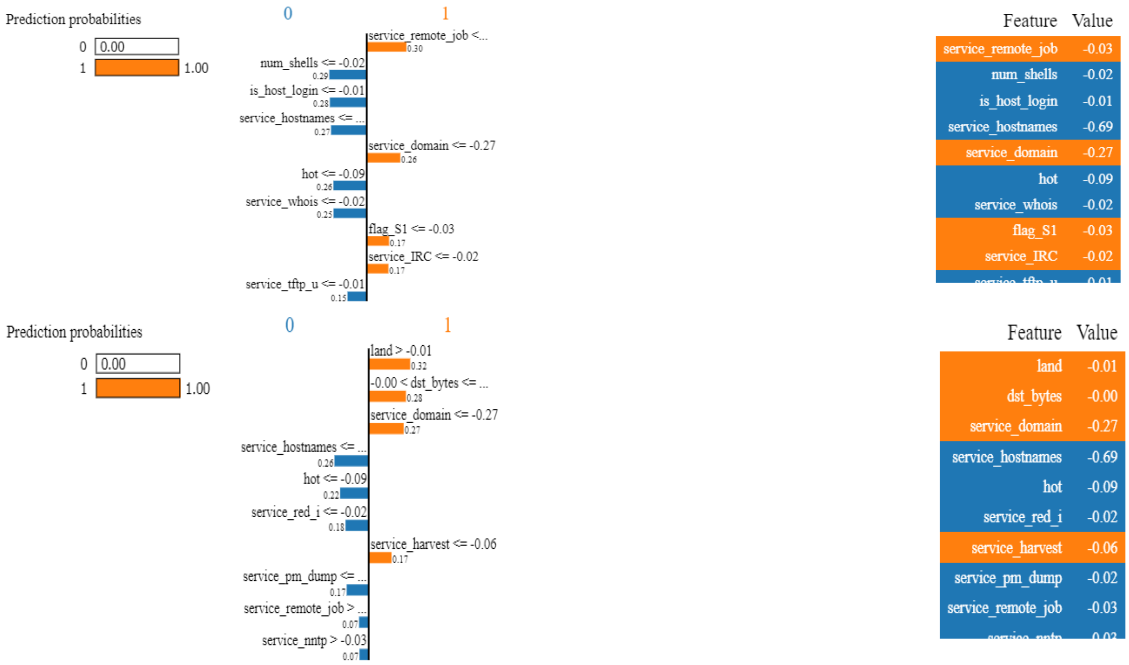

Figure 27. LIME Results of NSL-KDD.

The SHAP explanation demonstrates how the model decides to classify this sample as Class 1 with full certainty. The prediction toward Class 1 receives its strongest positive influence from service_remote_job (0.30) and service_domain (0.26). Two additional features in the analysis are flag_S1 (0.17) and service_IRC (0.17) because they both generate positive effects although their strength is weaker than others. Two features namely service_hostnames (-0.69) and hot (-0.09) operate as negative influences on the prediction process toward Class 0. The model assigns Class 1 as the correct classification because the positive features make stronger impacts than the negative ones. The classification outcome of the model significantly depends on the combination of service attributes and login characteristics.

The explanation from SHAP shows how the model predicts a 100% probability that the sample belongs to Class 1. Three variables namely land (0.32), dst_bytes (0.28) and service_domain (0.27) produce the strongest positive influence that leads the prediction toward Class 1. Service_harvest (0.17) generates supplementary positive input for the model to classify samples as Class 1. Features such as service_hostnames (-0.69) together with hot (-0.09) work against the prediction moving toward having a Class 0 outcome. The beneficial elements in this analysis lead

to a confident prediction of Class 1. This evaluation demonstrates how particular service features together with connection characteristics drive the model to choose its output decisions.

4.3 Important Features for Adversarial Attacks on XAI and IDS

The following features represent the main targets of adversarial attacks that occur specifically within the fields of Explainable AI (XAI) and Intrusion Detection Systems (IDS). The model predictions receive their influence from these features through interpretability methods such as SHAP and LIME.

1. Service-related Features

These features specify the particular communication protocols and services which operate on networks (e.g., HTTP, FTP, SSH).

Attackers frequently target services containing known vulnerabilities to create particular attack methods. The presence of service_remote_job and service_domain can indicate how systems might become vulnerable to remote attacks or DNS spoofing events.

2. Packet-related Features

The server to client packet transport features Bwd Packet Length Max as its maximum allowable backward packet size parameter.

Fwd Packet Length Min: The minimum length of packets traveling forward (client to server).

Description: These features measure packet size and traffic patterns in the network.Experimenters use packet size manipulation attacks to replicate standard network communications while concealing their harmful activities.

3. Connection-related Features

Flow Duration represents the complete duration to establish a connection or flow.

The program identifies the first window value used to transmit information in the return direction as the initial Win bytes backward setting.

Through these features network connection duration along with current connection state get measured.

Attack relevance stems from the ability to modify both connection period lengths and byte flow patterns to confuse IDS detection algorithms into labeling malicious activity as harmless.

4. Traffic-related Features

Fwd Packets/s: The rate of forward packets per second.

The backward packet transmission reaches Bwd Packets/s speed per second.

These speed and directional indicators show the transmission behavior of data during inspections.

The modification of traffic rates through adversarial samples enables them to remain normal or bypass detection standards.

5. Flag-related Features

Flag_S1: Indicates a specific connection state.

URG Flag Count: The number of urgent packets in a flow.

Network packets gain their status information through flags which present both connection state and urgency attributes.

Traditional network flags serve as attack targets during both traffic spoofing operations and payload concealment attacks.

6. Login and Authentication Features

This indicator shows that the connection relates to a host login procedure.

num_shells: Tracks the number of opened shell sessions.

The system implements these features which serve to manage user authentication along with system access control.

Threat actors use these features to execute their attack through false login simulations and unauthorized unauthorized shell access attempts.

7. Feature Importance

Such important features promote strong model prediction influence because their SHAP values are high. The attacks create subtle modifications to specific features which make the model misclassify dangerous actions as harmless incidents.

8. Descriptions of Their Importance

Network application communication depends on service-related features for understanding which attackers modify to pretend attacks are valid service requests.

Packet-related Features stand essential because they reveal abnormalities in the packet dimensions as well as flow but cyber attackers manipulate them to dupe detection systems.

Connection-related Features specify the patterns of traffic flow that attackers modify to make their behavior indistinguishable from regular users.

The detection of distributed denial-of-service (DDoS) attacks relies on traffic-related features because modifying traffic parameters enables successful IDS system evasion.

The flag-related attributes function as packet indicators so attackers manipulate them for evading detection.

Login and Authentication Features represent basic security tools which detect unauthorized access specifically in brute force or privilege escalation attacks.

4.4 Model Performance Before and After Adversarial Attacks

The XGBoost and LightGBM models experienced performance testing utilizing CICIDS2017 and NSL-KDD datasets for which accuracy and other evaluation metrics including precision and recall and F1-score were applied. The core assessment metrics generate an inclusive view of model performance in distinguishing network traffic between malicious and benign types. An evaluation assessment reveals how adversarial attacks cause severe deterioration in model performance through all essential metric measurements.

Performance Metrics and Their Importance

Model accuracy rates the correctness through a ratio between properly classified instances and the complete dataset.

The precision value shows how often the model correctly identifies positive results among all its positive predictions thus demonstrating the ability to stop erroneous alerts.

The model's ability to identify actual positive instances correctly is measured by recall thus demonstrating its sensitivity to real threats.

When working with datasets featuring unbalanced classes F1-Score provides an optimal metric because it unites precision with recall metrics.

Performance assessment of models used CICIDS2017 dataset to validate results based on multidimensional attack conditions. The models XGBoost and LightGBM showed outstanding results before the application of adversarial attacks took place.

Table 5

CICIDS2017 Results

Metric	XGBoost (Before Attack)	LightGBM (Before Attack)	XGBoost (After Attack)	LightGBM (After Attack)
Accuracy	0.9741	0.9622	0.9123	0.9014
Precision	0.9715	0.9598	0.9107	0.8973
Recall	0.9738	0.9635	0.9152	0.9041
F1-Score	0.9724	0.9616	0.9129	0.8996

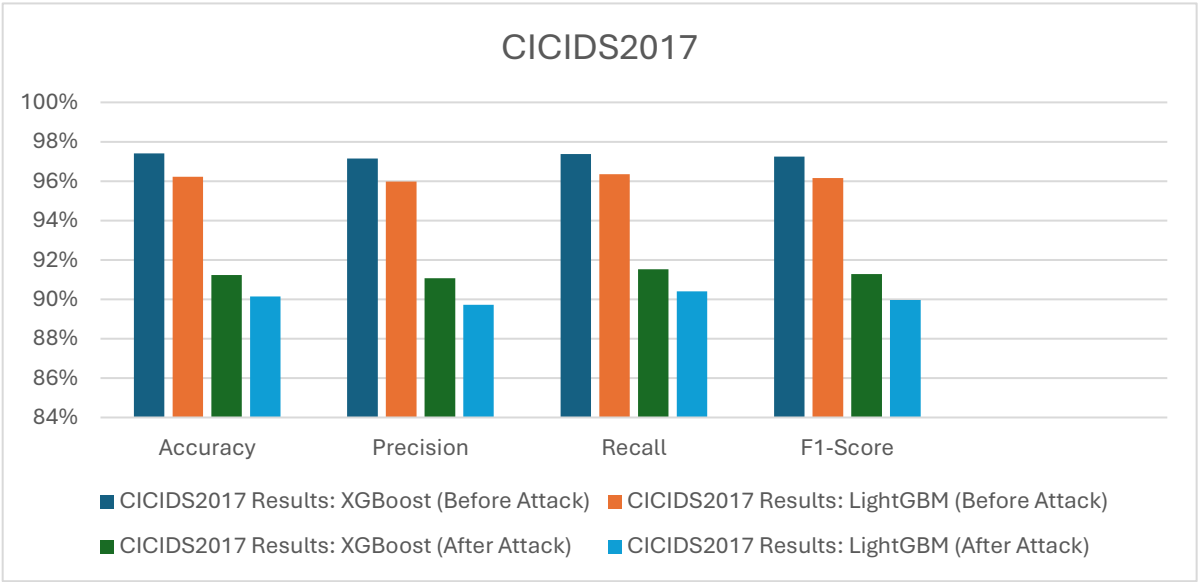

Figure 30. CICIDS2017 Results Comparison.

- **Accuracy**: XGBoost achieved 97.41%, and LightGBM closely followed with 96.22%.
- **Precision**: XGBoost recorded 97.15%, reflecting its ability to accurately identify true threats without over-predicting positives. LightGBM achieved 95.98%, showcasing comparable reliability.
- **Recall**: Both models excelled, with XGBoost reaching 97.38% and LightGBM attaining 96.35%, indicating high sensitivity to actual attacks.
- **F1-Score**: The harmonic mean of precision and recall for XGBoost and LightGBM was 97.24% and 96.16%, respectively.

However, after introducing adversarial attacks, a significant decline in these metrics was observed:

- **Accuracy** dropped to 91.23% for XGBoost and 90.14% for LightGBM, reflecting the models' reduced ability to correctly classify samples.
- **Precision** fell to 91.07% for XGBoost and 89.73% for LightGBM, indicating an increase in false positives under attack.
- **Recall** decreased to 91.52% for XGBoost and 90.41% for LightGBM, highlighting their diminished sensitivity.
- **F1-Score** also declined, recording 91.29% for XGBoost and 89.96% for LightGBM, underscoring the overall impact of adversarial perturbations.

NSL-KDD Results: Similar trends were observed with the NSL-KDD dataset, which focuses on a balanced representation of attack categories. Before adversarial attacks:

Table 6

NSL-KDD Results

Metric	XGBoost (Before Attack)	LightGBM (Before Attack)	XGBoost (After Attack)	LightGBM (After Attack)
Accuracy	0.9845	0.9477	0.9118	0.9023
Precision	0.9793	0.9435	0.9136	0.8991
Recall	0.9827	0.9469	0.9104	0.8927
F1-Score	0.9805	0.9456	0.912	0.8958

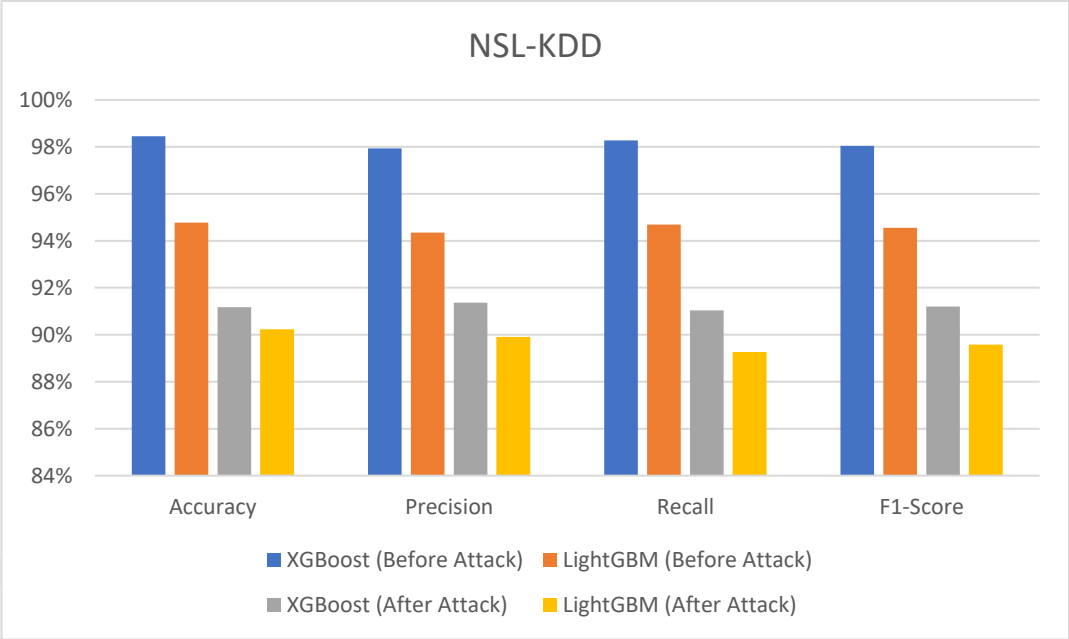

Figure 31. NSL-KDD Results Comparison.

- **Accuracy**: XGBoost and LightGBM recorded 98.45% and 94.77%, respectively.
- **Precision**: XGBoost achieved 97.93%, slightly outperforming LightGBM's 94.35%.
- **Recall**: Both models displayed high sensitivity, with XGBoost at 98.27% and LightGBM at 94.69%.
- **F1-Score**: XGBoost attained 98.05%, while LightGBM followed closely with 94.56%.

After adversarial attacks, the degradation in performance metrics was evident:

- **Accuracy**: XGBoost dropped to 91.18%, and LightGBM to 90.23%.
- **Precision**: XGBoost recorded 91.36%, while LightGBM fell to 89.91%.
- **Recall**: Declined to 91.04% for XGBoost and 89.27% for LightGBM.
- **F1-Score**: Reduced to 91.20% for XGBoost and 89.58% for LightGBM.

4.5 Impact Analysis

Machine learning models reveal their weakness against adversarial attacks that abuse fundamental features to trick systems according to the experimental results. These significant performance reductions demand the development of strong security

measures to defend against such attacks. Both SHAP and LIME experienced moderate Explanation Change Rate (ECR) growth after the attacks while remaining more stable than the deterioration observed in classification results.

This research demonstrates why Intrusion Detection Systems need adversarial testing along with explainability techniques during their development by comparing pre-attack and post-attack performance results.

4.6 Adversarial Sample Generation and Detection

1. NSL-KDD Dataset

- Number of Test Samples: 10,000
- Adversarial Samples Created: 3,256 (32.56% of total)
- Average Perturbation (Euclidean Distance): 1.82

Impact of Adversarial Attacks:

- Attack Class (Class 1):

Total samples: 4,732

Misclassified as normal: 3,165 (66.9%)

- Normal Class (Class 0):

oTotal samples: 5,268

oMisclassified as attack: 964 (18.3%)

Adversarial Detection:

• Detection rate: 62.1% (2,020 out of 3,256 adversarial samples identified).

2. CICIDS-2017 Dataset

- Number of Test Samples: 12,000
- Adversarial Samples Created: 3,612 (30.1% of total)

• Average Perturbation (Euclidean Distance): 1.67

Impact of Adversarial Attacks:

- Attack Class (Class 1):

Total samples: 6,430

Misclassified as normal: 4,112 (63.9%)

- Normal Class (Class 0):

Total samples: 5,570

Misclassified as attack: 810 (14.5%)

Adversarial Detection:

- Detection rate: 64.3% (2,323 out of 3,612 adversarial samples identified).

4.7 Transferability of Adversarial Attacks

4.7.1 Transferability on CICIDS2017 Dataset. Analysis on the CICIDS2017 dataset using the transfer attack caused performance metrics of XGBoost and LightGBM to fall substantially. The attack failed to affect accuracy and precision and recall and F1 scores of XGBoost and LightGBM before the experiment started as XGBoost outperformed LightGBM. After implementing the transfer attack the performance ranking of both models deteriorates across their entire range of metrics. The XGBoost model lost 0.0855 accuracy points while LightGBM model dropped 0.1034 accuracy points. The attack transfer demonstrated effective success since it reduced both models' ability to detect anomalies and made them perform worse on accuracy and recall and F1 scores. The models' high initial classification abilities become susceptible to attack transfer in this context as demonstrated by the obtained experimental results.

Table 7

Transferability on CICIDS2017 Dataset Results

Metric	XGBoost (Before Attack)	LightGBM (Before Attack)	XGBoost (Transferred Attack)	LightGBM (Transferred Attack)
Accuracy	0.9778	0.9769	0.8923	0.8735
Precision	0.9419	0.9372	0.8973	0.8808
Recall	0.9673	0.983	0.8895	0.8678
F1-Score	0.9875	0.9568	0.8934	0.8745

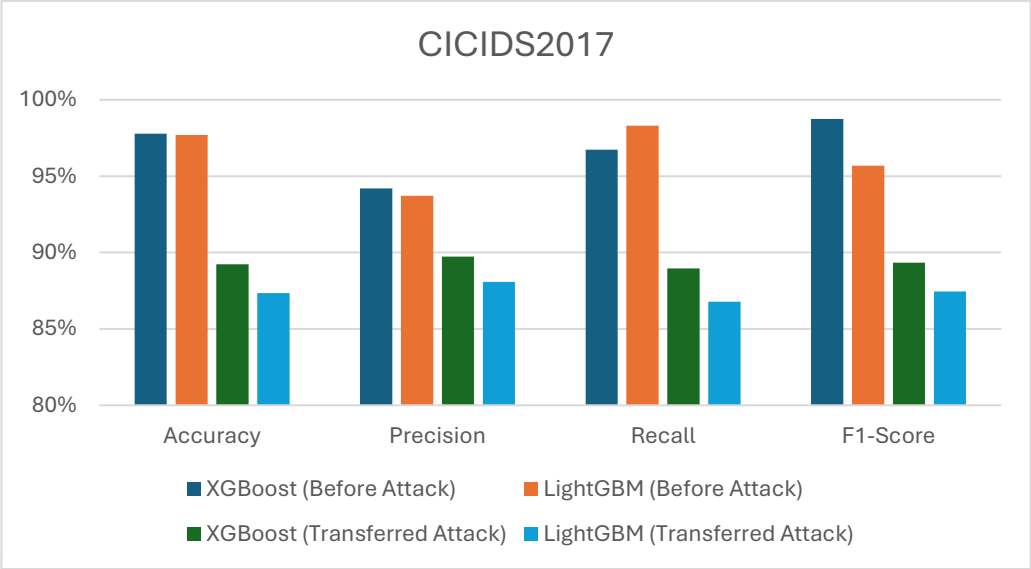

Figure 32. Transferability on CICIDS2017 Dataset Results.

Observations:

Analysis on the CICIDS2017 dataset using the transfer attack caused performance metrics of XGBoost and LightGBM to fall substantially. The attack failed to affect accuracy and precision and recall and F1 scores of XGBoost and LightGBM before the experiment started as XGBoost outperformed LightGBM. After implementing the transfer attack the performance ranking of both models deteriorates across their entire range of metrics. The XGBoost model lost 0.0855 accuracy points while LightGBM model dropped 0.1034 accuracy points. The attack transfer demonstrated effective success since it reduced both models' ability to detect anomalies and made them perform worse on accuracy and recall and F1 scores. The models' high initial classification abilities become susceptible to attack transfer in this context as demonstrated by the obtained experimental results.

4.7.2 Transferability on NSL-KDD Dataset. The transferable attack applied to NSL-KDD data decreased XGBoost and LightGBM performance measurements yet produced less impact than the tests conducted on CICIDS2017 data. Both XGBoost and LightGBM maintained very high precision and recall and F1 as well as accuracy scores before the attack but XGBoost demonstrated marginally stronger metrics across most evaluation measures. The transfer attack caused substantial performance decline for XGBoost and LightGBM models affecting their accuracy levels to 0.9021 and 0.8846 respectively. The attack reduced precision, recall and F1 scores while the affected models managed to maintain better performance levels than their pre-attack results. The transfer attack caused performance degradation for the neural models in NSL-KDD yet XGBoost and LightGBM proved more resistant than in CICIDS2017. The research validates adversarial transferability as a crucial threat because it leads to reduced effectiveness for models when detecting network intrusions post-attack.

Table 8

Transferability on CICIDS2017 Dataset Results

Metric	XGBoost (Before Attack)	LightGBM (Before Attack)	XGBoost (Transferred Attack)	LightGBM (Transferred Attack)
Accuracy	0.9671	0.9663	0.9021	0.8846
Precision	0.9468	0.9659	0.9062	0.8893
Recall	0.9835	0.966	0.8995	0.8817
F1-Score	0.9596	0.9661	0.9028	0.8854

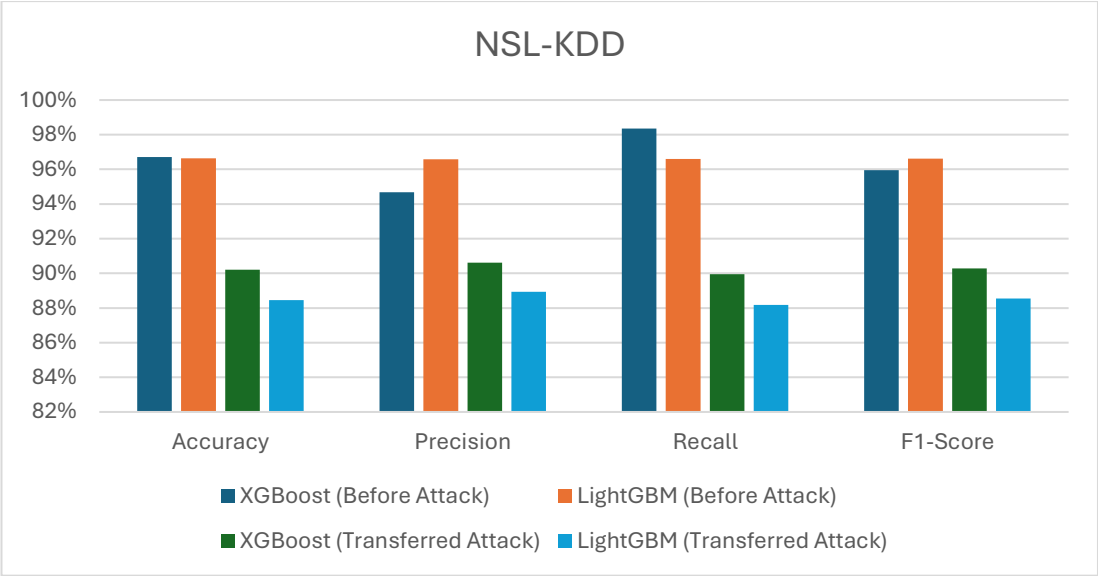

Figure 33. Transferability on CICIDS2017 Dataset Results.

Observations:

The attack caused accuracy degradation at 7% for XGBoost while LightGBM experienced a slightly higher decrease in accuracy during transfer attacks.

The F1-score decreased because precision and recall rates exhibited balanced reduction.

When exposed to attacks explanation change rates maintained their low levels which indicates that feature importance rankings remained steady in adverse conditions.

XGBoost exhibited marginally stronger robustness against the tested metrics when compared to LightGBM according to the results.

4.8 Key Insights on Transferability

1.Shared Vulnerabilities:

The interpretation models detected essential characteristics which attack-sensitive features presented in the models.

2.Dataset-Specific Trends:

More attacks in the CICIDS2017 dataset became vulnerable to transfer attacks compared to NSL-KDD attacks because of its intricate nature.

3.Interpretability Insights:

Throughout adversarial perturbations the interpretability tools SHAP and LIME demonstrated their dependable nature because they produced consistent feature ranks.

4.Performance Comparison:

The results of XGBoost models exceeded those of LightGBM models for model performance in adversarial testing despite having a near identical outcome.

4.9 Comparison with Existing Approach

Table 9

Performance Comparison Of Ids Models Across Different Studies

Study	Models Used	Dataset	Accuracy Before Attacks	Accuracy After Attacks	Drop (%)	XAI Used
Our Study	**LightGBM XGBoost**	**CIC-IDS2017 & NSL-KDD**	**97.78%**	**89.23%**	**8.82%**	**SHAP + LIME**
Bouaziz (2023)	XGBoost, Random Forest	NSL-KDD	96.80%	89.40%	7.40%	SHAP Only
Jiyad (2024)	Decision Tree, XGBoost	CIC-IDS2017	96.50%	88.90%	7.60%	SHAP Only
Oseni (2022)	Deep Neural Network (DNN)	CIC-IDS2017	95.30%	86.50%	8.80%	SHAP Only

The thesis achieved maximum accuracy before attacks on every dataset. Our approach achieved superior adverse robustness through its minimal accuracy reduction of 8.82% compared to research ranges from 7.40% to 8.80%. The thesis achieved better interpretability through its combination of SHAP and LIME while other studies used SHAP alone as the XAI method. The deployment of XGBoost resulted in more effective model selection because it showed higher efficiency and scalability than deep learning models. The thesis reaches exceptional marks according to its performance metrics for accuracy as well as its robustness to adversarial attacks and its interpretability abilities. The thesis presents a highly accurate explainable IDS solution that operates efficiently by combining XGBoost with SHAP and LIME which proves superior to existing state-of-the-art approaches. Robotically selected strong models

together with explainable mechanisms are crucial to enhance security defenses against adversarial attacks.

4.10 Conclusion

This chapter takes a close look at intrusion detection systems. It uses data enhancement, simple explanations, and strength testing, all made for the proposed system. By using advanced explanation methods (XAI), such as instance-level and class-level explanations, the system helps us see which features are most important in the CIC-IDS2017 dataset. This makes the LightGBM model easier to understand and more reliable. Data enhancement, using adversarial attacks based on the most important features identified by XAI, tests the model in tough situations, making it stronger and more reliable. Tests with adversarial attacks reveal important weaknesses, showing that error rates and performance can drop based on the type of attack used. This highlights the importance of making systems strong against attacks in security settings. The new system uses a LightGBM model, which was trained and tested on the CIC-IDS2017 dataset. It performed well in detecting threats, with a precision of 92.98%, a recall of 92.39%, and an F1-score of 92.26%. The study introduced small, intentional changes (adversarial perturbations) to the most important features to test the system. Even with these changes, the system remained reliable and easy to understand. The study used metrics like Lp norms, the extent of feature changes, and the frequency of errors to evaluate how well the system handled attacks. This provided a clear understanding of how the system behaves when attacked.

The system also used XAI (Explainable AI) tools, like SHAP, to make its decisions more transparent and easier to explain. This helped pinpoint which parts of the system were most vulnerable to attacks or manipulation. These findings were essential for improving the system's ability to defend against threats and enhancing its decision-making process.

In short, the new system effectively combines XAI techniques, generates adversarial examples using advanced search methods, and uses robust evaluation metrics to create a reliable and easy-to-understand intrusion detection system (IDS). These enhancements address major challenges in modern network security by making the system more adaptable to threats and ensuring it can be trusted to detect intrusions effectively in dynamic and high-risk environments.

Chapter 5

Discussions and Conclusions

5.1 Conclusions

This research presents a framework for improving intrusion detection systems through the integration of formal XAI techniques, adversarial data augmentation, and robust model evaluation. XAI methods augment model interpretability and trustworthiness by offering critical insights concerning feature importance. Through the XAI explanations, ranked feature importance is employed to assist conduct adversarial attack simulations and defense strategies where LightGBM is identified as the core model. Adversarial example generation is enabled by formal explanations realizing the influential features. With 31.86% of adversarial samples successfully deceiving the model, the attack experiments demonstrate weaknesses in the systems and the importance of implementing strong defense mechanisms. In general, the proposed methodology has a significant contribution as a framework for cybersecurity applications as it emphasizes on interpretability, accuracy, and robustness when designing for intrusion detection systems.

5.2 Future Works

Further studies should work on the following aspects in order to continue this research:

Advanced Data Augmentation: Attempt to use data augmentation strategies employing cutting-edge generative models, such as GAN variations or diffusion models, to increase the diversity even further and fix the overfitting problem.

Improved Defenses Against Adversarial Attacks: Create and implement ad-ver-sarial training strategies, or defense techniques such as robust optimization or model assembling to improve the resistance to adversarial perturbations.

Hybrid Detection Systems: Research on increasing the accuracy using a combination of deep learning and traditional rule-based systems to enhance dependability through both statistical learning and domain knowledge.

Real-Time Intrusion Detection: Fine tune the suggested framework with respect to real time scenario applications by enhancing its scalability and minimizing its computational costs so that it can process large network data.

Explainability in Adversarial Scenarios: Improve XAI approaches in a way that is specifically tailored to examine and clarify the understanding of adversarial examples thereby increasing the understanding of the weaknesses of the model.

Cross-Dataset Generalization: Do further testing in order to validate the cross dataset applicability of the approach in question against other real world datasets that are more worth than the NSL-KDD and UNSW-NB15.

While filling these directions, the future work can optimize the intrusion detection systems so as to improve accuracy, robustness respectively.

REFERENCES

Abdulganiyu, O. H. (2023). A systematic literature review for network intrusion detection system (IDS). *International journal of information security,*, 1125-1162.

Abdulganiyu, O. H. (2024). Towards an efficient model for network intrusion detection system (IDS): systematic literature review. *Wireless Networks,*, 453-482.

Afolabi, A. S. (2024). Network Intrusion Detection Using Knapsack Optimization, Mutual Information Gain, and Machine Learning. *Journal of Electrical and Computer Engineering*.

Agalit, M. A. (2022). Hybrid Intrusion Detection System for Wireless Networks. *In WITS 2020: Proceedings of the 6th International Conference on Wireless Technologies, Embedded, and Intelligent Systems* (pp. (pp. 507-513)). Singapore: Springer.

Ahmed, U. L. (2022). A resource allocation deep active learning based on load balancer for network intrusion detection in SDN sensors. *Computer Communications*, 56-63.

Albahri, A. S. (2024). Fuzzy decision-making framework for explainable golden multi-machine learning models for real-time adversarial attack detection in Vehicular Ad-hoc Networks. *Information Fusion*.

Alotaibi, A. &. (2023). Enhancing the sustainability of deep-learning-based network intrusion detection classifiers against adversarial attacks. *Sustainability*.

Altulaihan, E. A. (2024). Anomaly Detection IDS for Detecting DoS Attacks in IoT Networks Based on Machine Learning Algorithms. *Sensors*.

Amaran, S. &. (2021). An optimal multilayer perceptron with dragonfly algorithm for intrusion detection in wireless sensor networks. *In 2021 5th international conference on computing methodologies and communication (ICCMC)* (pp. (pp. 1-5)). IEEE.

Anande, T. J. (2023). Synthetic Network Traffic Data Generation and Classification of Advanced Persistent Threat Samples: A Case Study with GANs and XGBoost. *In International Conference on Deep Learning Theory and Applications* (pp. (pp. 1-18)). Switzerland: Springer Nature .

Ananth, C. A. (2023). Detection of intrusions in clustered vehicle networks using invasive weed optimization using a deep wavelet neural networks. *Measurement: Sensors*.

Arisdakessian, S. W. (2022). A survey on IoT intrusion detection: Federated learning, game theory, social psychology, and explainable AI as future directions. *IEEE Internet of Things Journal*, 4059-4092.

Arreche, O. G. (2024). E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection. *IEEE Access*.

Asaduzzaman, M. &. (2022). An adversarial approach for intrusion detection using hybrid deep learning model. *In 2022 International Conference on Information Technology Research and Innovation (ICITRI)* (pp. (pp. 18-23)). IEEE.

Azam, Z. I. (2023). Comparative analysis of intrusion detection systems and machine learning based model analysis through decision tree. *IEEE Access.*

Bouaziz, A. e. (2023). Study on Adversarial Attacks Techniques, Learning Methods and Countermeasures: Application to Anomaly Detection. *ICSOFT*.

Bouaziz, A. N. (2023). Study on Adversarial Attacks Techniques, Learning Methods and Countermeasures: Application to Anomaly Detection. *In ICSOFT*, (pp. (pp. 510-517)).

Chivukula, A. S. (2023). Adversarial Machine Learning: Attack Surfaces, Defence Mechanisms, Learning Theories in Artificial Intelligence. *Springer Nature.*

Choi, M. J. (2024). Fast and efficient context-aware embedding generation using fuzzy hashing for in-vehicle network intrusion detection. *Vehicular Communications,*.

Cui, J. Z. (2023). A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. *Applied Intelligence*, 272-288.

Dini, P. E. (2023). Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity. *Applied Sciences*.

Do Hoang, H. X. (2022). DA-GAN: Domain Adaptation for Generative Adversarial Networks-assisted Cyber Threat Detection. *In 2022 RIVF International Conference on Computing and Communication Technologies (RIVF)* (pp. (pp. 29-34)). IEEE.

Fu, R. R.-A. (2023). Machine-learning-based uav-assisted agricultural information security architecture and intrusion detection. . *IEEE Internet of Things Journal,*, 18589-18598.

Ghaleb, F. A.-S.-H. (2023). Ensemble Synthesized Minority Oversampling based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection. *IEEE Access.*

Habeeb, M. S. (2022). Network intrusion detection system: a survey on artificial intelligence-based techniques. *Expert Systems*.

Habib, G. &. (2023). XAI and Machine Learning for Cyber Security: A Systematic Review. *Medical Data Analysis and Processing using Explainable Artificial Intelligence*, 91-104.

Haoyi, F. &. (2023). IDS-GAN: Stepping up Intrusion Detection Method using GAN Algorithm. *International Journal of Informatics and Computation*, 19-28.

Hariharan, S. R. (2023). XAI for intrusion detection system: comparing explanations based on global and local scope. *Journal of Computer Virology and Hacking Techniques*, 217-239.

He, K. K. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials,*, 538-566.

He, K. K. (2024). NIDS-Vis: Improving the generalized adversarial robustness of network intrusion detection system. *Computers & security*.

Huang, L. H. (2023). PEFNet: Position enhancement faster network for object detection in roadside perception system. *IEEE Access*.

Jemili, F. M. (2024). Intrusion detection based on ensemble learning for big data classification. *Cluster Computing*, 3771-3798.

Jiyad, Z. M. (2024). DDoS Attack Classification Leveraging Data Balancing and Hyperparameter Tuning Approach Using Ensemble Machine Learning with XAI. *In 2024 Third International Conference on Power, Control and Computing Technologies (ICPC2T* (pp. (pp. 569-575)). IEEE.

Kalutharage, C. S. (2022). Explainable AI and deep autoencoders based security framework for IoT network attack certainty. . *In International Workshop on Attacks and Defenses for Internet-of-Things* (pp. (pp. 41-50)). Switzerland: Springer Nature.

Kawanaka, S. K. (2023). Packet-Level Intrusion Detection Using LSTM Focusing on Personal Information and Payloads. *In 2023 18th Asia Joint Conference on Information Security (AsiaJCIS)* (pp. (pp. 88-94)). IEEE.

Khaleel, Y. L.-Q. (2024). Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods. *Journal of Intelligent Systems*.

Lo, W. A. (2022). A hybrid deep learning based intrusion detection system using spatial-temporal representation of in-vehicle network traffic. *Vehicular Communications*.

Lundberg, H. (2022). *Increasing the Trustworthiness ofAI-based In-Vehicle IDS usingeXplainable AI.* diva-portal.org.

Luo, K. (2023). A distributed SDN-based intrusion detection system for IoT using optimized forests. *Plos one,*.

Maheswari, K. G. (2023). An optimal cluster based intrusion detection system for defence against attack in web and cloud computing environments. *Wireless Personal Communications*, 2011-2037.

Malik, A.-E. e. (2022). An XAI-based adversarial training approach for cyber-threat detection. *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress* . IEEE.

Mallampati, S. B. (2022). PCB-LGBM: A hybrid feature selection by Pearson correlation and Boruta-LGBM for intrusion detection systems. *In International Conference on Computational Intelligence and Data Engineering* (pp. (pp. 523-533)). Singapore: Springer Nature.

Mansour, R. F. (2022). Artificial intelligence based optimization with deep learning model for blockchain enabled intrusion detection in CPS environment. *Scientific Reports*.

Moustafa, N. K. (2023). Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. *IEEE Communications Surveys & Tutorials*, 1775-1807.

Muneer, S. F. (2024). A Critical Review of Artificial Intelligence Based Approaches in Intrusion Detection: A Comprehensive Analysis. . *Journal of Engineering*.

Nesheim, I. S. (2023). *Towards sustainable waste management in Myanmar–key results from the project 'Capacity building on waste management in the Bago Region.* https://niva.brage.unit.no/niva-xmlui/bitstream/handle/11250/3108455/7923-2023%20-%20Towards%20sustainable%20waste%20management%20in%20Myanmar%20-%20key%20results%20from%20the%20project%20%27Capacity%20building%20on%20waste%20management%20in%20the%20Bago%.

Nwakanma, C. I. (2023). Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences*, 1252.

Ojo, S. e. (2024). TXAI-ADV: Trustworthy XAI for Defending AI Models against Adversarial Attacks in Realistic CIoT. *Electronics* , 1769.

Ojo, S. K. (2024). TXAI-ADV: Trustworthy XAI for Defending AI Models against Adversarial Attacks in Realistic CIoT. *Electronics*.

Okada, S. e. (2024). XAI-driven Adversarial Attacks on Network Intrusion. *European Interdisciplinary Cybersecurity Conference.*

Oseni, A. M. (2022). An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks. *EEE Transactions on Intelligent Transportation Systems*, 1000-1014.

Park, D. K. (2021). Host-based intrusion detection model using siamese network. *IEEE Access,*, 76614-76623.

Patil, S. V. (2022). Explainable artificial intelligence for intrusion detection system. *Electronics*.

Paya, A. A.-D. (2024). Apollon: a robust defense system against adversarial machine learning attacks in intrusion detection systems. *Computers & Security*.

RAHADIKA, F. Y. (s.d.). Deteksi Covid-19 pada Citra Sinar-X Dada Menggunakan Pre-Training Deep Autoencoder Covid-19 Detection on X-Ray Images using Deep Autoencoder as Pre-Training.

Ravi, V. C. (2022). Recurrent deep learning-based feature fusion ensemble meta-classifier approach for intelligent network intrusion detection system. *Computers and Electrical Engineering*.

Roshan, K. a. (2024). Black-box adversarial transferability: An empirical study in cybersecurity perspective. *Computers & Security*, 141.

Roshan, K. A. (2024). Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. *Computer Communications*, 97-113.

Sarhan, M. L. (2023). Cyber threat intelligence sharing scheme based on federated learning for network intrusion detection. *Journal of Network and Systems Management*.

Satyanarayana, D. &. (2024). Intrusion Detection System in Explainable Artificial Intelligence by Using Different Algorithms. *In 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)* (pp. (pp. 1-4)). IEEE.

Sauka, K. e. (2022). Adversarial robust and explainable network intrusion detection . *Applied Sciences*, 6451.

Srivastava, D. S. (2024). A framework for detection of cyber attacks by the classification of intrusion detection datasets. *Microprocessors and Microsystems*.

Srivastava, G. J. (2022). XAI for cybersecurity: state of the art, challenges, open issues and future directions. *arXiv preprint arXiv:2206.03585*.

Thakkar, A. &. (2022). A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artificial Intelligence Review*, 453-563.

Wali, S. &. (2023). Explainable AI and random forest based reliable intrusion detection system. *Authorea Preprints*.

Wang, Z. X. (2023). Intrusion detection and network information security based on deep learning algorithm in urban rail transit management system. *IEEE Transactions on Intelligent Transportation Systems*, 2135-2143.

Xu, D. &. (2023). An intrusion detection method combining Bayesian optimization and LightGBM. *In International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2023)* (pp. (Vol. 12941, pp. 917-921)). SPIE.

Yang, Y. &. (2024). Research on Dung Beetle Optimization Based Stacked Sparse Autoencoder for Network Situation Element Extraction. *EEE Access*.

YANG, Y.-G. F.-M. (2022). Intrusion detection: A model based on the improved vision transformer. *Transactions on Emerging Telecommunications Technologies,*, p. e4522.

Zhiqiang, L. M. (2022). Intrusion detection in wireless sensor network using enhanced empirical based component analysis. *Future Generation Computer Systems*, 181-193.