

Chandan Kumar

Chandan_Thesis

 M.TECH MAJOR PROJECT II

Document Details

Submission ID

trn:oid:::27535:140317378

Submission Date

May 25, 2026, 9:42 AM GMT+5:30

Download Date

May 25, 2026, 9:50 AM GMT+5:30

File Name

Chandan_Thesis.pdf

File Size

997.9 KB

55 Pages

13,374 Words

77,008 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 8 words)

Match Groups

-  **84 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **9 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 3%  Internet sources
- 2%  Publications
- 5%  Submitted works (Student Papers)

Match Groups

- 84 Not Cited or Quoted** 6%
Matches with neither in-text citation nor quotation marks
- 9 Missing Quotations** 1%
Matches that are still very similar to source material
- 0 Missing Citation** 0%
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted** 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 3% Internet sources
- 2% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1

Student papers

University of Ulster on 2026-04-19

<1%
- 2

Student papers

University of Hertfordshire on 2025-01-06

<1%
- 3

Internet

www.diva-portal.org

<1%
- 4

Internet

dspace.library.uvic.ca

<1%
- 5

Student papers

Alliance College of Engineering and Design on 2026-04-29

<1%
- 6

Publication

Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dharendra Kumar Shukla. "Re...

<1%
- 7

Student papers

University of Hertfordshire on 2026-04-27

<1%
- 8

Internet

dspace.daffodilvarsity.edu.bd:8080

<1%
- 9

Student papers

University of Hong Kong on 2020-08-14

<1%
- 10

Student papers

Napier University on 2025-04-29

<1%

11	Publication	Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniqu...	<1%
12	Publication	Hesham Kamal, Maggie Mashaly. "Enhanced Hybrid Deep Learning Models-Based...	<1%
13	Publication	Mourade Azrour, Jamal Mabrouki, Sultan Ahmad. "IoT and Advanced Intelligence ...	<1%
14	Internet	archive.org	<1%
15	Student papers	American National University on 2026-02-02	<1%
16	Student papers	LTI 202223 on 2026-05-04	<1%
17	Student papers	University of Sunderland on 2026-05-09	<1%
18	Internet	export.arxiv.org	<1%
19	Internet	www.securitycn.net	<1%
20	Student papers	Moodle2025 on 2026-05-01	<1%
21	Student papers	Coventry University on 2025-08-15	<1%
22	Publication	Joy Buuri, Samah Mansour, Mostafa El-Said, Xinli Wang. "An Empirical Study Usin...	<1%
23	Internet	www.mdpi.com	<1%
24	Internet	www.newma.co.uk	<1%

25	Internet	www.preprints.org	<1%
26	Student papers	University of Wales Swansea on 2020-09-30	<1%
27	Internet	depo.btu.edu.tr	<1%
28	Internet	www.researchsquare.com	<1%
29	Student papers	Brunel University on 2026-03-27	<1%
30	Student papers	National Institute of Technology Delhi on 2020-05-27	<1%
31	Student papers	University of Hertfordshire on 2023-12-03	<1%
32	Student papers	University of Western Ontario on 2022-07-16	<1%
33	Student papers	Vrije Universiteit Amsterdam on 2026-03-17	<1%
34	Internet	arxiv.org	<1%
35	Internet	theses.whiterose.ac.uk	<1%
36	Internet	koreascience.kr	<1%
37	Internet	trepo.tuni.fi	<1%
38	Internet	www.infosec.gov.hk	<1%

39	Publication	Imtiaz Ullah, Qusay H. Mahmoud. "A Two-Level Hybrid Model for Anomalous Acti...	<1%
40	Student papers	Middlesex University on 2021-05-04	<1%
41	Student papers	Richfield Graduate Institute of Technology on 2025-12-15	<1%
42	Publication	Swati Sah, Rejwan Bin Sulaiman, Aditya Dayal Tyagi. "Federated Learning in Finan...	<1%
43	Student papers	University of East London on 2020-09-09	<1%
44	Student papers	University of Queensland on 2019-11-04	<1%
45	Student papers	University of Strathclyde on 2022-04-01	<1%
46	Student papers	University of Sunderland on 2019-01-18	<1%
47	Student papers	University of West London on 2025-07-31	<1%
48	Student papers	University of York on 2022-03-07	<1%
49	Internet	ebin.pub	<1%
50	Internet	etd.aau.edu.et	<1%
51	Internet	thesesjournal.com	<1%
52	Internet	www.frontiersin.org	<1%

53	Student papers	Islington College,Nepal on 2026-05-10	<1%
54	Publication	J. Dhilipan, V. Saravanan, R. Agusthiyar. "Human Machine Interaction in the Digit...	<1%
55	Student papers	Loughborough University on 2026-04-20	<1%
56	Student papers	Macquarie University on 2021-06-06	<1%
57	Student papers	Manchester Metropolitan University on 2024-10-04	<1%
58	Student papers	Multiversity on 2025-10-21	<1%
59	Student papers	National Forensic Sciences University on 2024-11-28	<1%
60	Publication	Nidhi Sehgal, Balamurugan Balusamy, Rachel Sheffield, Umer Zaman. "The Futur...	<1%
61	Publication	S. Nikkath Bushra, Nalini Subramanian, A. Chandrasekar. "An optimal and secure ...	<1%
62	Student papers	University of Birmingham on 2022-09-10	<1%
63	Student papers	University of East London on 2025-09-08	<1%
64	Student papers	University of Maine on 2026-05-09	<1%
65	Internet	ais.khpi.edu.ua	<1%
66	Internet	file.techscience.com	<1%

67

Internet

journals.unesco.go.ke

<1%

68

Internet

www.ijert.org

<1%

CHAPTER-1

INTRODUCTION

13

The global amount of data is growing exponentially as a result of digitalization, increased use of the Internet of Things (IoT), cloud services and advances in artificial intelligence (Bartley 2024). Forecasts show that the data volume could exceed 181 zettabytes by 2025, which places increased demands on storage, processing and data security (Bartley 2024). In parallel, the vulnerability of IT infrastructures is increasing as organizations' dependence on network-based communication intensifies.

Networks generate large amounts of raw data in the form of logs from firewalls, servers, applications and other network components. This type of network traffic is a concrete example of so-called big data, i.e. large, complex and rapidly generated data sets that are difficult to handle with traditional methods. This further highlights the requirement for scalable and automated methods for processing data, particularly in cybersecurity. Previously, analysis of such data was done through security tools like SIEM (Security Information and Event Management), although a very small percentage of this data has ever been analyzed in real-time (Joshi, Ghuge, Anjarlekar and Chavan, 2025). In essence, this means that a lot of data that could help in predicting future threats to cybersecurity remains unused.

This could be because of the fact that the volume and complexity of information have become overwhelming for human beings to process and react to in a timely manner. According to Mahboubi et al. (2024), the major challenge associated with threat hunting includes the challenge of lack of enough time and resources to detect threats in huge amounts of data generated by security systems. The larger the amount of data, the higher the likelihood that the data will not be detected manually.

According to Joshi et al. (2025), manual log analysis is a lengthy and expensive activity that inevitably leads to mistakes in the procedure. With the fast-changing nature of cyberattacks and their automation, traditional techniques of threat detection appear to be unable to provide a proactive approach to cybersecurity defense. Therefore, it seems necessary to apply

advanced instruments able to analyze large amounts of information quickly and effectively, according to Joshi et al. (2025).

It is the rising complexity of attacks that drives experts to create new methods of pattern recognition and anomaly detection. As stated by Ortiz-Garcés, Govea, Sánchez-Viteri, and Villegas-Ch. (2024), pattern recognition and anomaly detection become crucial because these deviations indicate the existence of attacks. It is important to note that these deviations are key in recognizing attacks since signature-based detection cannot identify them all.

Machine learning and artificial intelligence provide new ways of automating the process of recognizing anomalies and attack patterns in network traffic (Han, Guo, Wang, and Zhang, 2023). Through the analysis of raw data in logs and discovering of hidden relations and anomalies, security work becomes more effective and efficient. Automation of the process is likely to decrease the workload of security specialists while making threat detection and management possible.

Against this background, the study aims to investigate how machine learning and artificial intelligence techniques can be applied to analyze raw data from network logs in order to identify patterns and anomalies related to cyberattacks. The ability to address the imbalances in the training data is one of the key factors to be considered when assessing the performance of the proposed models, since it could influence their capability to detect attacks. In the current research, the attention will be paid to analyzing structured network traffic data from the CICIDS2017 database.

1.1 BACKGROUND

The growth of data volume in IT infrastructure has given rise to both new opportunities for business growth and security challenges (Zuech, Khoshgoftaar & Wald, 2015). Corporations have an ever-increasing amount of raw data coming from corporate networks in the form of firewall, server, application, and other network component logs. Often, such logs may contain information about the behavior and security state of the system (Joshi et al., 2025). Conventionally, processing of this kind of data occurs in Security Information and Event Management (SIEM) systems, which specialize in analyzing and correlating events in real time. In general, SIEM technologies are rule-based, meaning that all the unusual or

nonstandard traffic goes through filtering (Joshi et al., 2025). It means that any security-relevant data may be ignored by SIEM solutions. While SIEM technology provides certain automation and helps analyze data, its efficiency still relies on the use of predefined rules and human interaction. With the increase of data volume, it becomes unreasonable to employ humans as the primary resource for data processing, as argued by Mahboubi et al. (2024).

Intrusion Detection Systems (IDS) have been developed as an important complement to SIEM systems to enable the detection of both known and unknown attack patterns in network traffic. However, traditional signature-based IDS (SIDS) are limited to detecting attacks that are already documented in signature databases, and thus lack the capacity to identify new, as yet unknown threats such as zero-day attacks (Khraisat, Gondal, Vamplew, & Kamruzzaman, 2019).

1.1.1 Pattern Recognition And Anomaly Detection

According to Ortiz-Garcés et al. (2024), one of the major tools used in cybersecurity is pattern recognition. Pattern recognition is helpful in identifying anomalies that do not follow the patterns of usual behaviors in network traffic. After defining the normal patterns of activities within a network, any anomaly could be detected, which may possibly pose a threat to the network.

Traditional approaches to attacks based on signatures suffer from the limitation that they can detect only those attacks that are known before since they depend on threat patterns observed previously. The process of pattern recognition and anomaly detection helps in identifying any new type of attack through dynamic analysis of traffic flows and activities. According to Ortiz-Garcés et al. (2024), dynamic analysis is required because cyberattacks constantly evolve using innovative techniques in order to bypass conventional techniques used for detecting these cyberattacks.

It has been observed that the rise in security threats has resulted in the deployment of machine learning and artificial intelligence in order to improve traffic analysis and intrusion detection (Han et al., 2023). Intrusion detection systems based on networks employ machine learning algorithms in their functioning by training them with past data in order to differentiate

between normal and abnormal network traffic. One other method of detecting intrusions is through anomaly detection.

Standard datasets are used for the purpose of developing and testing intrusion detectors. One such dataset is CICIDS2017 (Sharafaldin, Lashkari & Ghorbani, 2018), which consists of simulated network traffic data including both normal traffic and 14 different cyberattacks. The traffic has been converted to bidirectional flows using the CICFlowMeter tool, which generates 79 statistical functions per flow. This provides a comprehensive basis for training and validating intrusion detection models.

CICIDS2017 thus constitutes a stable and representative basis for developing and evaluating machine learning models in network security, which is of great importance in dealing with today's increasing cyber threats.

1.2 PROBLEM STATEMENT AND RESEARCH QUESTIONS

Despite the exponential increase in data volumes in corporate networks, driven by technologies such as AI, IoT and cloud storage, a large part of the collected network traffic remains insufficiently analysed. In many organizations, network traffic is stored in platforms such as Azure Data Lake and is still analyzed manually by threat hunting and cybersecurity specialists.

The manual analysis of large amounts of network data is both resource-intensive, timeintensive and difficult to scale. It is imperative that firms have the ability to invest in skilled security experts to examine traffic flows and logs all the time in order to detect any intrusion attempts or other anomalies. As a result, it becomes costly for firms to operate and difficult to scale their operations as the amount of data increases. Furthermore, the use of manual techniques increases the likelihood of attacks being detected too late or even not at all dues to human error. In some situations, firms are not able to allocate the necessary resources to be able to monitor traffic flows continually. This is why the use of AI in threat hunting becomes crucial because, according to Mahboubi et al. (2024), automated systems can handle much more data than humans.

Against this background, a number of key research questions arise:

- To what extent can automated machine learning models contribute to the detection of threats in raw network traffic?
- How is the accuracy of an AI-based model affected by class imbalance in the CICIDS2017 dataset when detecting different types of attacks?

1.3 PURPOSE

The aim of this research paper is to analyze how machine learning and artificial intelligence can be used for the automation of the analysis process of raw data network traffic for the purpose of detecting anomalies and patterns of cyber attacks. It will also be discussed how machine learning and artificial intelligence can help in the analysis process, which is traditionally done manually, for detecting patterns in network security.

CHAPTER-2

LITERATURE REVIEW AND THEORETICAL BACKGROUND

2.1 RAW DATA IN NETWORK SECURITY

2.1.1 What Is Raw Data and its importance?

Raw data is defined as information collected directly from instruments, sensors, or other sources before any extensive processing has taken place. According to Borgman (2015), this data often forms the basis of scientific analysis and can include, for example, network logs, traffic patterns, and other digital signals.

Therefore, especially in security-critical applications such as intrusion detection, it is often necessary to analyze information in its most original form. Preserving different levels of processing of data is seen as an important aspect of research infrastructure, as it allows different types of analysis to be carried out while maintaining transparency and reproducibility (Borgman, 2015, pp. 48–52).

2.1.2 Log Files and Network Traffic As Security Sources

In current IT systems, a huge amount of network raw data is produced daily, namely system logs, traffic flows, authentication requests, firewall actions, and applications logs. All together, these sources of information give us a comprehensive view of what happens within the network. Network raw traffic can contain not only regular but also suspicious actions and traffic flows that can be useful in cybersecurity. According to Avila, Khoury, Khoury, and Petrillo (2021), analyzing security logs is important for detecting leaks and attacks because they may contain anomalies and leaks of information.

2.1.3 The Need For Analysis Of Unfiltered Security Systems

Although modern IT systems produce a huge amount of network data, their potential is not fully used. According to Avila et al. (2021), security systems such as SIEM are mainly interested in finding out events that correspond to some previously known rules or signatures.

This creates a blind spot for unknown, unusual, or slowly developing attacks, which therefore risk remaining undetected.

To complement these systems, manual analysis in the form of so-called threat hunting is sometimes used. However, this method is resource-intensive and requires both time, expertise and continuous staffing, making it difficult to apply on a larger scale. According to Avila et al., few organizations can effectively analyze all of the existing traffic systematically because there is not enough capacity. Therefore, new approaches should be sought to solve this problem.

The authors point out that it is necessary to develop automated ways to conduct log analysis, especially those that are related to machine learning and artificial intelligence. These tools can be used not only to detect anomalies in the network traffic but also to constantly improve their performance and differentiate between normal and deviant behavior by applying dynamic algorithms instead of static signatures.

2.2 THREAT HUNTING AND MANUAL ANALYSIS LIMITATIONS

Threat hunting is the process of proactively searching for signs of any threat using analysis of the data from network and computer systems, where the objective is to find patterns or anomalies that could be indicative of an attack.

In contrast to typical security systems that respond to alerts generated by them, threat hunting relies on hypothesis-driven approach, whereby security experts manually analyze indicators that cannot be detected by signature-based systems. According to Belal and Sundaram (2022), the aim is to detect such cyberattacks that can evade automatic protection through the simulation of normal network behavior.

Even though threat hunting allows flexible application and discovery of new threats, it requires access to expert knowledge and resources. Threat hunting, in particular, is mostly performed in a targeted manner and not continuously. It is clear that they currently do not allocate resources for manual analysis of the extensive data generated by XDR platforms and other log sources. At the same time, they see clear potential for improvement in using machine learning to strengthen the ability to detect threats.

As Belal and Sundaram (2022) note, traditional detection systems were not meant to be used for processing large amounts of data. The problem is that current networks generate huge amounts of information that cannot be effectively processed manually, thus making it very likely that some possible dangers will remain unnoticed.

Furthermore, this challenge is confirmed by security experts, who state that only selected parts of the raw data are manually reviewed, and that some data is purged or archived after 30–90 days depending on policy.

2.3 PROACTIVE CYBERSECURITY THROUGH AI

The sophistication of the attacks has been a change in the cyber security strategy from reactive to predictive. The emergence of new technologies has helped this transformation take place. The use of large amounts of data on the development of cyber attacks, for instance, is possible thanks to the improvements in artificial intelligence and machine learning techniques; for example, Belal and Sundaram (2022) developed an AI-based method for detecting cyber attacks through data processing.

One of the key concepts involved in this changeover is automation. While rules-based methods can detect anomalies in network traffic, ML algorithms can also be used to analyze all kinds of anomalies that cannot be handled by rules-based techniques, as stated by Shaukat et al. (2020). The advantage of such algorithms is that they have the ability to constantly learn about new threats that might arise.

Moreover, Tang, Alazab and Luo (2017) demonstrate that proactive analysis, combined with AI, can not only be used to detect existing attacks, but also to predict possible attacks through the recognition of trends in historic and contextual data. Predictive systems then give an organisation the ability to take action proactively – cutting response time and raising the accuracy of advanced attack detection.

67

From one perspective, this example demonstrates what can be achieved through the application of artificial intelligence and machine learning from a technical standpoint; from another perspective, this is a theoretical foundation of cybersecurity.

2.4 MACHINE LEARNING - FUNDAMENTALS AND METHODS

Machine learning refers to an academic discipline that involves more than one discipline, and the objective of this discipline is the creation of computer programs that can identify patterns and draw conclusions from such patterns irrespective of whether there are any instructions from the programmer beforehand. As stated by Li et al. (2024), machine learning requires the use of statistical and probabilistic techniques to identify patterns and predict them. However, it was not until the 1980s that Herbert A. Simon, the Nobel Laureate and an excellent researcher in artificial intelligence and decision-making theory, defined learning as the process of improving one's performance on the basis of experience.

49

2.4.1 Supervised Learning

Supervised learning can be seen as the earliest kind of machine learning, wherein the algorithm is trained by using an input-output (or labeled) dataset. Some of the most prevalent forms of supervised learning include regression and classification (both for continuous and discrete values respectively), making supervised learning extremely useful for intrusion detection, diagnostics, and finance (Li, Lin, and Zeng, 2024).

2.4.2 Unsupervised Learning

Unsupervised learning comes into play in cases where the data does not have any preexisting labels. In this case, the algorithm tries to find some form of pattern in the data by using clustering or dimensionality reduction techniques, among others. The main application areas of unsupervised learning are those where there is no clear understanding of what is normal behavior and what is abnormal (Li, Lin, and Zeng, 2024).

2.4.3 Semi-Supervised Learning

Semi-supervised learning involves both labeled and unlabeled data points during training, and it proves useful when labeling becomes expensive. As stated by Zhou et al. (2023), semisupervised learning methods have proven to be useful in improving the performance of models, particularly in anomaly detection where there are very few attacks.

2.5 ANOMALY DETECTION AND PATTERN RECOGNITION

Pattern recognition is an essential part of anomaly detection, which is crucial in achieving cybersecurity. This is because most security threats, including data intrusion and exfiltration, exhibit statistical patterns different from those of normal network traffic. Pattern recognition helps in detecting such anomalies (Zamanzadeh Darban et al., 2024).

Moreover, the authors claim that, unlike signature-based approaches, which depend on the existing knowledge about the patterns of attacks, anomaly detection allows for the detection of previously unknown attacks. It is because by building models based on the data from normal traffic, it is possible to spot potential abnormalities before they become actual problems.

Also, the authors state that the decision regarding the type of anomaly detection method to use largely depends on the availability of labeled data. In cases when there is no labeled data available, which is common in the field of network security, unsupervised techniques like clustering or autoencoders are used, while supervised methods are employed in cases when there is labeled data available for both normal and abnormal traffic. Besides, traditional statistical methods are also considered, which demonstrates the versatility of the approach.

2.6 AZURE MACHINE LEARNING AND AUTOML

Various prior works have emphasized the necessity of deploying machine learning in an automated manner due to the need for processing large volumes of complex and unstructured datasets (Tang, Alazab & Luo, 2017; Belal & Sundaram, 2022). In the case of cybersecurity, the importance of cloud computing platforms, such as Azure Machine Learning (Azure ML), has grown because they allow developers to create ML models while providing various types

of automation, including AutoML. Thus, combining such tools allows applying machine learning in practice in the face of specific requirements related to data storage capacity and analysis speed.

It is worth noting that previous works have proven that Azure ML can be effectively used for classification purposes, time-series analysis, and real-time monitoring for safety-critical applications (Fakhraee, Balakreshnan & Masanz, 2023). Thanks to MLOps, which is readily available in Azure ML, it becomes possible to control the entire lifecycle of the model through tracking the data pre-processing, training, deployment, monitoring, and re-training phases.

24

2.6.1 Technical Overview: Azure Machine Learning

36

Azure Machine Learning is a cloud-based framework offered by Microsoft which allows developers to create, train, deploy and manage their machine learning models. The tool is able to use both the workflow based on programming code as well as graphical interface (GUI), as well as provide support for various existing open-source platforms such as PyTorch, TensorFlow, and scikit-learn (Microsoft, 2023).

One of the major strengths of Azure ML is the tight connection between various services provided by Azure, thus allowing for the creation of consistent and scalable infrastructure for machine learning. Some examples of those include: Azure ML studio for management of experiments, models and pipeline; Azure Key Vault for secure authentication keys management; Application Insights for logging and performance tracking; Azure Container Registry for Docker images management; and Azure Storage for structured management of data required for model training and validation (Microsoft, 2023). Therefore, thanks to all these capabilities, Azure ML offers a comprehensive environment for machine learning model lifecycle management.

3

2.6.2 AutoML: Automated Model Development In Azure

The concept of Automated Machine Learning (AutoML) in Azure Machine Learning entails an automation of various stages involved in the development of ML models, including selecting the right algorithm, carrying out the feature engineering process, hyperparameter

tuning, and data splitting into training, testing, and validating datasets. This is achieved by creating multiple pipelines simultaneously in order to evaluate several configurations, which are then compared based on their ability to produce the best performance according to the predefined criterion. The selected model that shows superior performance becomes the one to be applied as a standard for use (Microsoft, 2025).

Many scholars prove the effectiveness of using AutoML for building effective models requiring minimal human involvement. This method has been found quite effective in different sectors, including cybersecurity, where it is utilized to develop ML models forecasting incidents, network classification, and intrusions (Google Developers, 2024). The AutoML technique may also be applied in classification, regression, forecasting, computer vision, and NLP, among others. Some special types of models include BERT in NLP and Auto-ARIMA in time series analysis.

In addition to the AutoML service, Azure ML supports other development environments tailored to specific needs of different user groups. Azure ML Designer is a no-code environment that allows building ML models without programming skills. The platform provides a graphical environment that allows building a workflow from data import through the model development stage to its training and deployment as an endpoint, with capabilities for advanced data preprocessing and training on CPU/GPU cluster environments. The Designer makes it possible for users to have a straightforward way of creating models by not writing any code and then publishing them to REST API endpoints (Fakhraee, Balakreshnan & Masanz, 2023).

In more complex scenarios, a code-first method can also be applied. The latter means having control over the whole process of model development in terms of code through Python SDK or Azure CLI. With the help of a code-first method, users have full access to their environment, settings, and resources. This method is suitable to apply in cases where it is necessary to work with the applications where CI/CD and flexibility matter.

2.7 CICIDS2017 – A STANDARDIZED DATASET

CICIDS2017 Data Set is an experimental data set, which is used when developing models using machine learning to detect intrusions. The said data set has been described as modern

and extremely valuable in this field of study by Zhao et al., 2023. The data set has 2,827,876 flows that are categorized into those classified as being normal (BENIGN) at 80.3% and fourteen different types of attacks that account for the other 19.7%.

Unlike traditional log files, CICIDS2017 is based on network sessions collected in a controlled test environment. Using CICFlowMeter, NetFlow-like attributes were extracted from raw PCAP traffic, resulting in 79 features per data point. Each data record represents a bidirectional flow between two parties and is annotated as benign or belonging to a specific attack category. The above arrangement allows this data set to be ideal for supervised learning algorithms since it already comes with labeled data and realistic network traffic. The B-Profile mechanism was utilized to create realistic network patterns through simulating behavior from twenty-five user profiles (Sharafaldin, Lashkari & Ghorbani, 2018). They interact using protocols like HTTP, HTTPS, FTP, SSH, and email. The collection of data took place over five days ranging from Monday, July 3 to Friday, July 7, 2017, where only benign traffic was created on the first day, while an attack took place on all other days (Zhao, Hu, & Liu, 2023).

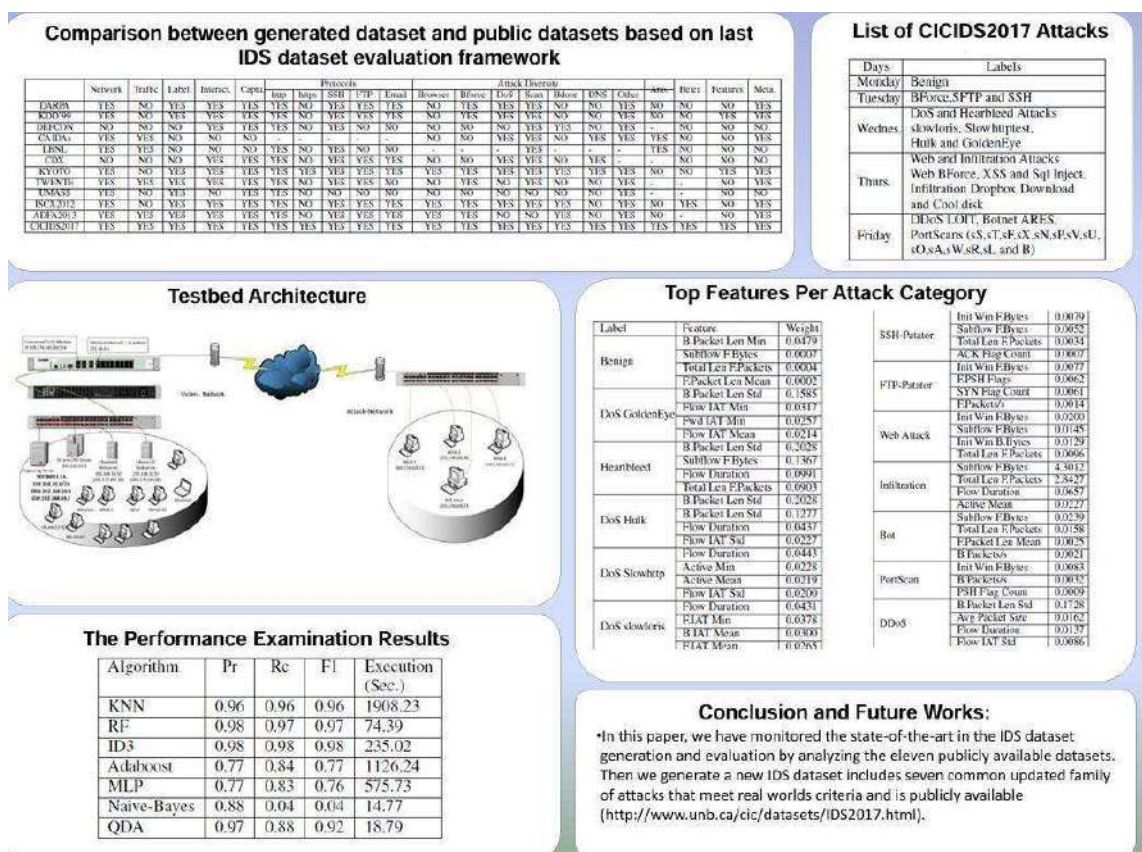


Figure 1: Overview of dataset compared w.r.t protocol diversity, traffic types and metadata

2.8 PERFORMANCE METRICS FOR MODEL EVALUATION

Uddin in his article titled "Addressing Accuracy Paradox Using Enhanced Weighted Performance Metric in Machine Learning" talks about the problem of using only the accuracy measure when evaluating machine learning models especially when there is class imbalance within the dataset. It should be noted that the accuracy measure cannot indicate the true worth of the model because it ignores the balance ratio between the classes. This situation is referred to as the accuracy paradox, whereby a model may attain high accuracy scores through prediction of the majority class. In most cases, this approach leaves out the less represented class in favour of the dominant class. For instance, in an intruder detection system, attacks will remain unnoticed if the majority class dominates.

Most machine learning algorithms assume balanced data sets, meaning they will focus on the majority class, thereby ignoring the minority class, which, in most cases, is intrusions/attacks. This problem is referred to as class imbalance, whereby the machine learning model can be highly accurate yet fail to detect important attacks (Vargas, Aranda, Costa, Pereira, & Barbosa, 2023). Therefore, a set of complementary metrics is used that capture different aspects of the model's strengths and weaknesses. Below, the most commonly used performance metrics in Azure AutoML and their relevance in the context of this study are presented.

2.8.1 Classification Outcomes: TP, TN, FP, And FN

In order to interpret the metrics of performance correctly, one needs to know about the four fundamental outcomes in the context of classification models, which are discussed by Grigorev (2021).

- **True Positive (TP):** The output that the model classifies as being a true positive (e.g. an attack), is indeed so.
- **True Negative (TN):** The model predicts a negative outcome (e.g. benign traffic) and this is correct, i.e. normal traffic that is correctly classified.

- False Positive (FP): The model classifies benign traffic as an attack, a false alarm in other words.
- False Negative (FN): A real attack is classified as benign, the threat has been completely missed.

In other words, for the purposes of cybersecurity, the right classification should not have any false positives (FP) or false negatives (FN). The four criteria are the basic ones used when evaluating any classification measure discussed below.

2.8.2 Accuracy

Accuracy describes the rate at which classification is done correctly relative to the total number of cases. This measure ranges from 0 to 1, with a higher accuracy approaching one (Microsoft, 2025). To give an example, if there are 100 predictions made by a classifier and 90 of them are correct (correspond to the label of the data point), then the accuracy equals 0.90 or 90%.

Although the metric is easy to interpret, it can give a false impression of the model's performance in scenarios with unbalanced data. In intrusion detection, where normal traffic often constitutes a clear majority, a model can achieve high accuracy by simply classifying all traffic as normal despite completely failing to identify attacks. Accuracy is calculated using the following formula: (Uddin, 2019).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.8.3 Precision

Precision measures a model's ability to avoid falsely classifying negative examples as positive. The range is [1,0] and the closer to 1, the more secure the model is (Microsoft 2025). The formula below is the general calculation for precision (scikit, 2025).

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.8.4 Recall

Recall measures a model's ability to find all true positive examples. It shows how many of the actual positive cases the model managed to find. The interval here is also between [1,0] where the closer to 1 the model hits, the better it is at not missing positive examples (Microsoft, 2025). The computation goes thus (scikit, 2025):

$$\text{Recall} = \frac{TP}{TP + FN}$$

The higher the recall, the more appropriate it is when there is a need for no threat to be missed out, despite there being lots of false alarms. This is the situation in cybersecurity since each threat that remains undetected may lead to dire consequences. Such a model will hence be beneficial since it reduces the amount of false negatives – attacks that are missed by the model. This explanation is justified by the analysis of the different approaches used in evaluating classification models provided in Chapter 4 in the book Machine Learning Bookcamp. “Recall is the fraction of correctly classified positive examples among all positive examples...The higher the recall, the fewer false negatives we have.” (Grigorev 2021, Chap.: Evaluation Metrics For Classification, Paragraph: Precision and Recall).

The book uses a customer churn prediction problem as an example, but the principle is general: high recall means that the model is better at identifying the positive, important observations, which in security-critical environments corresponds to real threats.

16

8

2.8.5 F1 score

4

F1 score is the harmonic average of Precision and Recall, giving weight to both of them, which becomes very important in imbalanced data where both False Positives and False Negatives are harmful. F1 score doesn't consider True Negative and can be used when

detecting threats is more critical than reducing False Positive (Microsoft, 2025). Formula for the calculation of the F1 score is given below:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.8.6 AUC – Area Under the ROC Curve

AUC stands for “area under the curve,” showing a model’s ability to separate one class from another at any threshold. The area under the ROC graph is taken in this case, in which the dependency of True Positive Rate on False Positive Rate is plotted, see Figure 2, and falls in the interval [0,1]. When the value approaches 1, we have perfect separation; whereas, the value around 0.5 shows that the model performs even worse than random guessing.

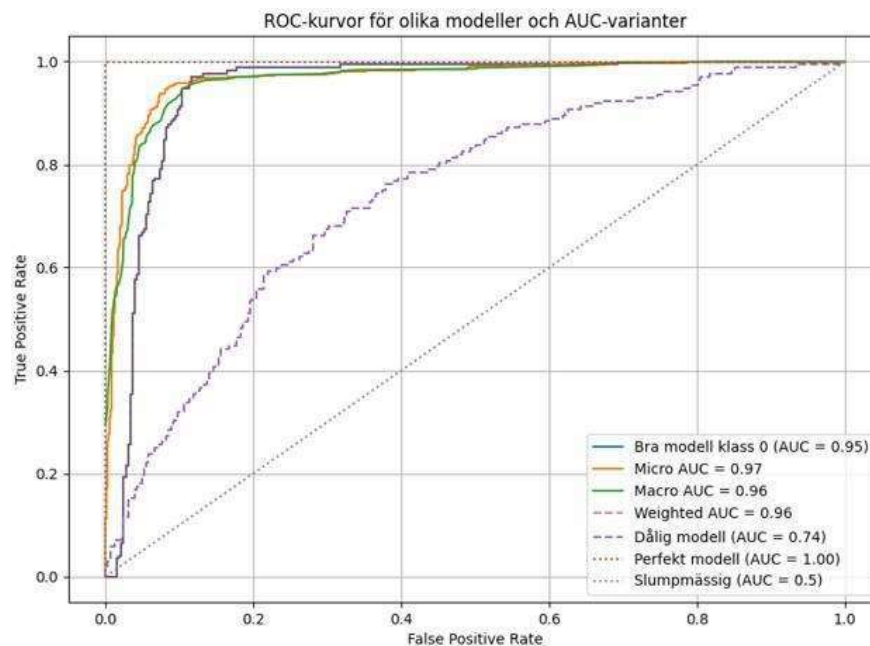


Figure 2: ROC curves for different models and AUC variants. A model with a high AUC (closer to 1.0) has good separation ability between classes, while a model close to 0.5 performs like chance

2.8.7 Confusion Matrix

The confusion matrix offers an in-depth view of the results of classification of the model. It indicates how many instances belonging to the actual class (the y-axis) have been classified

under another class (the x-axis) by the model. The ideal situation for such a model is where most observations lie on the diagonal line indicating that the prediction and actual class match (Microsoft, 2025). Figures 3 and 4 demonstrate how such a matrix appears in both cases.

The term "confusion" in relation to the confusion matrix comes from the model's tendency to confuse classes (Microsoft, 2025). This could be exemplified in the case of a intrusion detection model where the true positive would indicate that an attack has been correctly detected and the false negative means that a real intrusion was left undetected.

The confusion matrix is therefore a central tool in the evaluation of models in safety-critical applications.

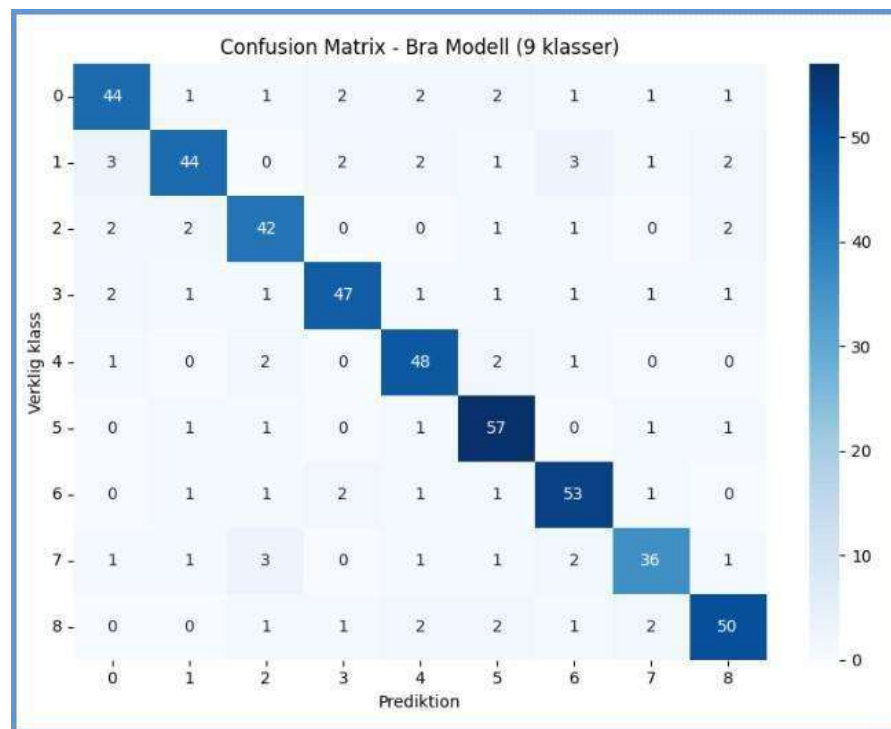


Figure 3: Confusion matrix for a well-trained model. Most observations lie along the diagonal, indicating good accuracy

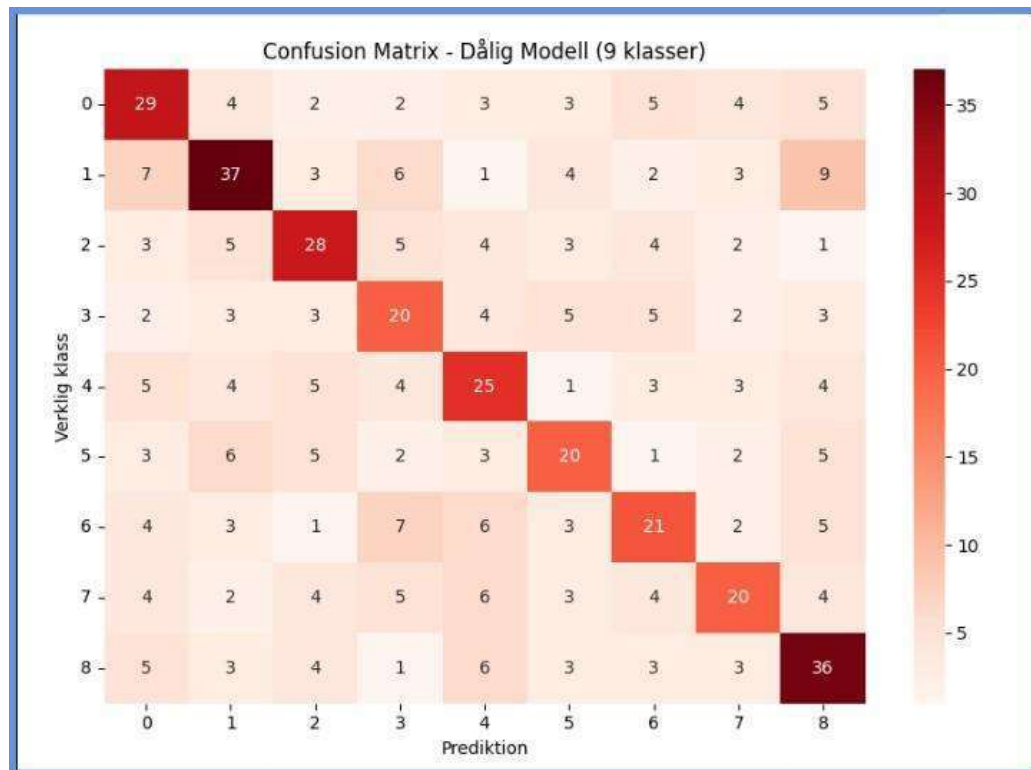


Figure 4: Confusion matrix for a weak model. Many misclassifications occur, which is visible in the offdiagonal distribution

2.8.8 Macro, Micro And Weighted Calculations

For multi-class problems or unbalanced data, three common strategies are used to summarize results (Microsoft, 2025), see a summary in Table 1.

Table 1: Macro, Micro and Weighted calculations

Metric	Description	Application/Context
Macro	Calculates the metric for each class and averages them	When each class, regardless of size, should affect the result equally, especially for unbalanced data
Micro	Treats each individual prediction (in all classes) as a single large binary problem. Adds together all TP, FP, FN, TN.	When the interest is in the overall performance.

Weighted	Calculates the metric for each class, but weights each class according to how many true examples it has.	When classes occur in different frequencies.
----------	--	--

Azure AutoML automatically calculates several of these combinations, for example macro F1, micro precision, or weighted recall, depending on the selected evaluation strategy.

2.8.9 Choosing A Primary Metric In Azure AutoMLml

When configuring an AutoML job in Azure, a primary metric is chosen for which the model should be optimized. The choice should be adapted to the purpose of the analysis. Table 2 is inspired by Microsoft Learn's documentation on automated training in Azure Machine Learning, where different primary evaluation metrics are linked to use cases (Microsoft, 2025).

Table 2: Primary values in AutoML

Metrics	Application/Examples
Accuracy	Image classification, balanced datasets, sentiment analysis
AUC-weighted	Fraud detection, image classification, anomaly and spam detection, anomaly detection
Average precision scoreweighted	Text classification, especially when interpreting sentiment in texts
Norm-macro-recall	Identification of unusual events, e.g. customer churn in unbalanced datasets
Precision-score-weighted	Text classification, e.g. topic analysis or identification of subjectivity in content

In this study, recall, F1-score and AUC are mainly used as metrics, since the goal is to maximize the detection of real threats (minimize false negatives) without overloading the system with false positives.

2.9 CLASSIFICATION

Classification is a form of supervised learning where the goal is to predict a discrete output variable based on input data represented as feature vectors. In this method, a model is trained using labeled data, where each data point consists of a pair (x_i, y_i) where x_i is a feature vector (input) and y_i is a category (output). The next step of the algorithm is training to map an observation into one of the specified classes in the output space. In contrast to regression where the target value is continuous, the classification model will produce an outcome that is comprised of a finite number of classes. Examples of classification applications include but are not limited to email spam filtering, image detection, and various network threat detection (Li, Lin, & Zeng, 2024).

2.9.1 Multiclass Classification

Multiclass classification is considered one area of machine learning where the classification algorithm aims to categorize more than two classes within a data set. Unlike in the binary classification problem where the output of the classifier must belong to only two categories, for example, "attack" or "benign," the multiclass classifier attempts to classify the input into at least three categories. Multi-class classification plays a vital role in cybersecurity, especially during the real-time detection of IoT attacks. In this respect, multi-class classification helps to differentiate among various classes of malicious data, such as DDoS attacks, port scan, and unknown attacks, and normal data. One of the most popular techniques that can be utilized here is One-vs-Rest (OVR) classification because of its ability to create multiple binary classifiers. Each binary classifier corresponds to each class versus the remaining classes.

2.9.2 Classification Algorithms In Auto ML

The algorithms included in AutoML's classification models are presented in Table 3 (Microsoft, 2025).

Table 3: Classification models in AutoML

AI-Algorithm
Logistic Regression
LightGBM
Gradient Boosting
Decision Tree
K-Nearest Neighbour
Linear SVC
Support Vector Classification (SVC)
Random Forest
Extremely Randomized Trees
XGBoost
Naïve Bayes
Stochastic Gradient Descent (SGD)

There are other algorithms besides the ones mentioned above, such as the meta-algorithm VotingEnsemble in AutoML, which is not a machine learning model but rather an aggregate model consisting of several base models to enhance prediction accuracy. Ensemble models are created during the last phase of the autoML process, and Azure uses Caruana's ensemble selection algorithm to pick the base models for the ensemble. The algorithm creates the initial ensemble from the best five models up to five models whose score is within 5% of the best score.

More models will then be added one by one, and each time, an improvement in accuracy is considered. VotingEnsemble is thus iteratively optimized based on which combinations give the best results (Microsoft, 2025).

2.9.2.1 Caruana Ensemble Selection Algorithm

2 To improve the performance of machine learning models, the Ensemble Selection method uses a library of several different models, trained with varying algorithms (such as SVM, KNN, ANN, decision trees, etc.) and different hyperparameter settings. The goal is to create a diversified set of models where some perform well on different metrics such as accuracy, F1 score or ROC-AUC. Through a so-called forward stepwise selection strategy, the models are selected that at each step improve the overall performance of the ensemble model according to a specific metric, such as accuracy or ROC area. The selection is done iteratively by successively adding the models that best complement the already existing ensemble. Initially, the process often starts with the five best-performing models to minimize the risk of poor starting performance. Each new model added to the ensemble must increase its overall performance to be included (Caruana, R., Niculescu-Mizil, A., Crew, G. & Ksikes, A., 2004).

To reduce the risk of overfitting, the method has been improved by three techniques: (1) selection with replacement, which allows weighting of particularly strong models, (2) initialization with a selection of the best-performing models before starting the greedy selection, and (3) bagging of the ensemble selection, where several sub-ensemble are created from random subsets of models and their average is used as the final result. Studies show that ensemble selection often outperforms both the individually bestperforming models and other ensemble methods such as bagging, boosting, and Bayesian model aggregation (Caruana et al., 2004).

13 2.10 DATA SCALING AND PREPROCESSING IN MACHINE LEARNING

In machine learning, data preprocessing is a central part of model development, especially when variables have different scales. One technique that could be used in such a case would be standardization, which involves scaling of the variables in such a way that their mean is zero and standard deviation is one. Such a practice increases the efficiency of scale-sensitive algorithms like SVM, KNN, and gradient descent algorithm. The Azure Machine Learning framework uses StandardScalerWrapper for achieving this purpose (Microsoft, 2025). Apart from standardization, there may be other methods applied based on data. Data can be scaled to the range [-1,1] by using MaxAbsScaler. SparseNormalizer is mainly used when the data

is sparse, such as in cases with many zero values, e.g. in log data and normalizes each row so that its length is 1. The application of TruncatedSVDWrapper, which retains the most important features from high-dimensional matrices without centering the input data, may be considered an attempt to reduce the number of features used in analysis to enhance the model performance and interpretation. Particularly, this may prove beneficial for applications involving a huge amount of input data (Ahsan et al., 2022).

2.11 PREVIOUS RESEARCH ON AI IN NETWORK SECURITY

It has been well proven that artificial intelligence (AI) and machine learning (ML) play crucial roles in ensuring network security and implementing various intrusion detection approaches, anomaly-based detection, and traffic flow pattern recognition. With the growing number of threats posed to modern networks, there has been a growing need for efficient approaches aimed at ensuring security by leveraging the use of AI and ML (Ahsan et al., 2022; Zuech, Khoshgoftaar & Wald, 2015).

2.11.1 Use Of CICIDS2017 In Previous Studies

CICIDS2017 is one of the most established and frequently used benchmark datasets in intrusion detection research. CICIDS2017 dataset stands out with its ability to provide realistic network traffic within a controlled environment, which consists of pre-classified data as either normal or malicious. The diversity of traffic data and large amount of data can be considered useful in testing both supervised and unsupervised machine learning models (Zhao, Hu, & Liu, 2023.; Khraisat et al., 2019).

In addition to the diverse data presented in CICIDS2017, there is also a considerable number of statistical properties related to the network, namely, 79 per flow, that allows using CICIDS2017 in testing machine learning models on auto ML services like Azure ML without spending too much effort on feature engineering (Sharafaldin, Lashkari, & Ghorbani, 2018). As reported by Zhao, Hu, & Liu (2023.), various deep learning models were tested on CICIDS2017, and the LSTM models demonstrated high efficiency in testing. These

networks showed to be especially effective for working with sequential data where timing and order are important aspects, e.g., for detecting escalation of attacks.

Firstly, according to Ahsan et al. (2022), CICIDS2017 can be regarded as an essential element in the testing process aimed at verifying the effectiveness of developed artificial intelligence systems for network security. Nevertheless, according to Khraisat et al. (2019), the issue of the existing skew between types of attacks becomes apparent. The problem may hinder the model's detection of infrequent types of attacks, so certain measures should be employed, including oversampling, using weighted losses, or creating synthetic data.

Secondly, it should be noted that some researchers have pointed out specific limitations of CICIDS2017 due to the fact that the data in question was gathered in laboratory conditions. The fact that attacks were performed in a relatively short period of time and their characteristics were predetermined means that models that rely on CICIDS2017 may experience certain difficulties while detecting prolonged or more complex attacks in real life (Zhao, Hu, and Liu, 2023).

2.11.2 Models That Have Shown High Accuracy

As reported in multiple studies, there is a high precision and recall in using different ML approaches for classifying the network traffic on the basis of CICIDS2017 dataset. As indicated by Ahsan et al. (2022), classic ML techniques like Random Forest and XGBoost prove their superiority in detecting attacks with distinct signature, i.e., DDoS, PortScan and Brute Force, where in some cases the accuracy rate exceeds 95%.

According to Belal & Sundaram (2022), one of the strengths of the ensemble methods is their good generalization capacity. Additionally, it has been claimed that hybrid approaches such as those utilizing the CNN-LSTM approach are particularly effective in recognizing spatial and temporal patterns, which may prove to be very beneficial for attacks with several phases.

Although there are many advantages of modern methods like CNN-LSTM models, they are quite complex to implement due to their high computational demands and need for specialized skills. Thus, simpler approaches such as Decision Tree or Logistic Regression

remain relevant because of their ability to offer faster deployment and easier implementation, even if this means sacrificing predictive performance (Khraisat et al., 2019).

Consequently, previous research has proven that there is not one solution that may be referred to as the best model for intrusion detection as it varies based on many factors, including the type of threat, resources at hand, the requirement for model interpretability, and others. With this said, several authors recommend taking an automated approach and employing such technologies as Azure AutoML which offers a systematic approach in selecting the optimal model based on set criteria (Microsoft, 2025; Ahsan et al., 2022).

2.12 PROBLEMS RELATED TO USING ARTIFICIAL INTELLIGENCE IN THREAT DETECTION

Although artificial intelligence can improve the efficiency of threat detection significantly, the technology itself poses a number of challenges. Some of those challenges include the problem of data imbalance, lack of model interpretability, and poor generalization capacity of models.

2.12.1 Bias In Training Data

AI-based detection models depend on training on datasets that represent the variability of cases that happen in actual networks. Should one group of attacks or traffic be present in abundance in the training set, whereas other groups are highly scarce, there is a possibility of building an imbalanced predictive capacity in the model. This would lead to detecting frequent attacks with high precision, yet leaving uncommon, although critical attacks unnoticed (Khraisat et al., 2019; Zhao, Hu, and Liu, 2023).

One example of such a problem is related to popular CICIDS2017 dataset that contains some attack classes extremely rarely, such as Heartbleed and Infiltration attacks. These attacks significantly affect the recall rate for the specific attacks in the model. The problem is described by Belal and Sundaram (2022) and includes techniques like class balancing, synthetic minority oversampling technique, and ensemble learning models.

2.12.2 Lack Of Transparency And Interpretability

Another central problem when applying AI in cybersecurity is the limited interpretability of the models. More specifically, deep learning models in general have been associated with the so-called “black box” issue, making it difficult to understand the reasoning behind a certain classification choice. The presence of such issues becomes a serious challenge in security-relevant contexts when decisions have to be justified (Ahsan et al., 2022). With that in mind, a new idea called explainable artificial intelligence (XAI) has become increasingly popular among scientists in recent times.

That is, in line with systematic studies done by Mohale and Obagbuwa (2025), SHAP and LIME can act as explanation methods to help in understanding how models operate in IDSs. While the simplicity of some models such as decision trees makes it easy to understand them, their accuracy was found to be relatively low compared to other models. In effect, the problem raised by Mohale and Obagbuwa (2025) is the absence of standard tools to use XAI.

2.12.3 Generalizability To Real-World Environments

Another issue that arises with the use of such machine learning models involves the lack of their generalization to real-world conditions since real-world networks exhibit great diversity, constant change, and include legitimate anomalies, as well as novel and unknown types of cyberattacks (Khraisat et al., 2019; Avila et al., 2021). Thus, algorithms that have proven themselves in testing environments cannot necessarily deliver similar results in practice.

Zuech, Khoshgoftaar and Wald (2015) emphasize that this lack of transferability is a major obstacle to the practical application of AI in organizations' security work. To counteract this, it is proposed to continuously retrain the models with current traffic data. But this entails the use of substantial resources such as computational power, expertise, and data management skills, which may not be available to all companies.

2.13 SUSTAINABLE DIGITALIZATION AND CLOUD INFRASTRUCTURE



14

With the advent of digital services, there is a greater need for computational power. However, this entails the greater demand for energy consumption in the information technology industry. According to Buyya, Ilager and Arroba (2023), data centers, the backbone of any cloud computing infrastructure, will contribute to a growing number of electricity usage. It has been predicted that by 2025, data centers worldwide will utilize 20 percent of the total electricity used in the world and produce 5.5 percent of carbon dioxide emissions.

However, cloud-based services are considered to have the potential to reduce the IT sector's climate impact compared to traditional, locally operated systems. Through resource sharing and large-scale optimization in cloud providers' data centres, opportunities for increased energy efficiency and better integration of renewable energy are created (Yenugula, Sahoo & Goswami, 2024). These solutions also support circular economy principles through, for example, hardware reuse and more efficient workload distribution.

To enable more sustainable digitalisation, organisations and system developers should consider environmental factors in their technical choices. Loeser, Erekan and Zarnekow (2017) argue that sustainability principles should be integrated already in the design phase of IT services and not only considered as a secondary requirement. This includes, among other things, energy use, carbon footprint, material recycling and the ability to manage waste in a sustainable way.

CHAPTER-3

METHODOLOGY

2 This methodological design was employed in order to perform network traffic analysis, detect malicious actions in the data, and assess the efficiency of various machine learning algorithms in the cyberattacks detection process. The research methodology consists of several stages that include data collection, preprocessing, solving class imbalance problem, modeling and evaluation of results.

45 38 The reason for the selection of the CICIDS2017 dataset as the base for this research is that it offers real network traffic data that contain various kinds of cyberattacks alongside the benign activities. The dataset was created at the Canadian Institute for Cybersecurity. This dataset incorporates modern cyberattacks such as Distributed Denial of Service (DDoS) Attacks, Brute Force Attack, Port Scan, Web Attack, Botnet Traffic, DoS Attack. The dataset comprises network flows extracted from the network traffic by the CICFlowMeter with such metrics as flow duration, packet length, byte/s, packet/s and others. Since the complete dataset includes millions of records and involves considerable computations, roughly 20% of the data set was selected using stratified sampling technique in order to maintain original class distribution and decrease computational complexity at the same time.

In preparation of the data set, several preprocessing steps have been applied prior to model training. Duplicate values, missing data and invalid values have been eliminated to increase quality of the data and stabilize training process. Infinitely high values obtained during calculation of network flows were handled properly. Classification structure of the target variable was represented by two classification scenarios. Binary classification involved grouping all kinds of attacks to one ATTACK class and comparison with the BENIGN traffic. Multiclass classification included maintaining separate attacks for evaluating model's ability to classify different attacks.

2 Data normalization and feature scaling was done to normalize variation within numerical features and improve performance of machine learning algorithms. Due to imbalance nature of the CICIDS2017 dataset in which majority is made up of benign traffic that exceeds

several minor attacks, RandomOverSampler was applied in order to make the dataset balanced. Over-sampling involved duplicating the minority classes and varying the samples slightly to avoid biased towards majority classes.

6 The resulting dataset was subsequently split into training and testing datasets at the rate of 80 percent vs. 20 percent. The stratified train-test splitting was applied in order to guarantee that there was consistent representation among both datasets. While the former served as training data for machine learning models, the latter provided means of evaluating models' performance on unseen data.

In order to have something to compare AI-driven techniques against, a rule-based approach to intrusion detection was applied to simulate traditional SIEM and signature-based detectors. As such, the baseline detector determined whether the particular network activity is suspicious based on threshold-based rules which included extremely high numbers of packets transmitted, flows with suspicious duration, and unusual byte transmission. Thus, the rule-based system was introduced as a baseline system for comparative purposes.

11 Supervised ML algorithms applied in the course of the project included logistic regression, random forest, XGBoost and lightGBM. Logistic regression was used as a baseline technique in terms of linear classification problems. In turn, the latter two were utilized due to their outstanding ability to work with multi-dimensional and nonlinear traffic.

6 The resulting dataset was subsequently split into training and testing datasets at the rate of 80 percent vs. 20 percent. The stratified train-test splitting was applied in order to guarantee that there was consistent representation among both datasets. While the former served as training data for machine learning models, the latter provided means of evaluating models' performance on unseen data.

In order to have something to compare AI-driven techniques against, a rule-based approach to intrusion detection was applied to simulate traditional SIEM and signature-based detectors. As such, the baseline detector determined whether the particular network activity is suspicious based on threshold-based rules which included extremely high numbers of packets transmitted, flows with suspicious duration, and unusual byte transmission. Thus, the rule-based system was introduced as a baseline system for comparative purposes.

11 Supervised ML algorithms applied in the course of the project included logistic regression, random forest, XGBoost and lightGBM. Logistic regression was used as a baseline technique in terms of linear classification problems. In turn, the latter two were utilized due to their outstanding ability to work with multi-dimensional and nonlinear traffic.

5 Additionally, comparative experiments were conducted for assessing the effects of applying various imbalance management approaches on the performance of the models. There was studied how models trained through oversampling outperform other models without any balancing methods in terms of better detecting attacks. Lastly, there was discussed the results obtained concerning the effectiveness assessment of machine learning-based intrusion detection systems.

3.1 DATASET EXPLANATION

2 The data source used in the paper was the CICIDS2017 dataset created by the Canadian Institute for Cybersecurity (CIC). The reason behind selecting this particular dataset is its ability to provide a comprehensive picture of the current traffic network scenario including both normal and attack activity. CICIDS2017 dataset is widely used in intrusion detection research area for it simulates real-life situations including attacks. Dataset link: <https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset>.

1 20 The data is collected using network flow records created through traffic capture analysis by means of the CICFlowMeter application. Every network flow is characterized by 79 statistical attributes that characterize the network traffic. Such attributes can range from flow duration, total number of packets in the forward direction, total number of packets in the reverse direction, flow-bytes-per-second, flow-packets-per-second, packet size statistics, inter-arrival times, flag counts, header sizes, active time, idle time, and many others. Attributes provide comprehensive information regarding the communication process and can be helpful in distinguishing between normal and malicious traffic flows.

19 CICIDS2017 not only comprises normal network traffic but also includes numerous types of cyberattacks, namely Distributed Denial of Service (DDoS) attack, Denial of Service

19 (DoS) attack, Brute Force attack, Port Scanning attack, Web Attack, Botnet attack, Heartbleed attack, and Infiltration attack. Thus, the presence of a variety of cyberattack types allows evaluating the effectiveness of both binary and multi-class intrusion detection systems. The binary classification involves distinguishing between benign and malicious flows whereas multi-class classification distinguishes an attack type within network traffic.

51 The initial data set includes millions of traffic samples gathered for several days with simulated network operations. Because of the very large number of data and processing difficulties, about 20% of the total amount of data was sampled by the stratified method. The stratification of the data helped preserve the original ratio of benign and attacked classes while simplifying computations.

The data set is characterized by an evident imbalance between benign traffic and some attack classes on one hand, and uncommon attacks like Heartbleed or Infiltration on the other hand. Imbalance makes it difficult to train models since machine learning algorithms may get biased toward majority classes and ignore minority classes of attacks. Thus, balance procedures should be performed to enhance results.

17 The CICIDS2017 database is an appropriate benchmark for the evaluation of intrusion detection systems based on the realistic generation of traffic and the inclusion of various attack scenarios, allowing for an effective detection of not only well-known but unknown attacks using machine learning techniques.

3.2 PREPROCESSING

Data preprocessing is an essential step when working with intrusion detection based on machine learning as raw data might contain inconsistencies, repetitions, missing fields, and noise affecting the performance of the algorithms. For that reason, the following preprocessing steps have been conducted to prepare the CICIDS2017 data set for training.

Firstly, the data set has been loaded into the analysis environment and checked for missing fields, repetitions, and incorrect numeric fields. Then, all duplicate records were excluded to avoid a possible bias in training algorithms and enhance their generalizability. Finally, missing and undefined fields were detected and treated appropriately.

The data set had some infinities due to network flow calculations especially on packet rates and byte transmission rates. The infinities were converted to missing values which were then dropped to ensure stable model training during computation. This was necessary to make sure that there were only valid numeric values of the features in our data.

From this cleaned data set, two classification labels were made for the target. In the binary class labels, all attacks fall under ATTACK class while normal is BENIGN class. On the other hand, in the multiclass class labels, attacks have their own labels as it helps in identifying the type of attack. Categorical target labels were then encoded using label encoding.

Feature scaling and normalization were used to decrease variance between numerical features and help the models converge. Considering that some of the numerical features of the dataset are characterized by very different values, scaling would allow not letting one of the features affect the learning process too much.

Next, class imbalance in the dataset was considered. Some attack classes include far fewer examples than benign data and the majority of attacks. Therefore, oversampling technique, RandomOverSampler, was used on the training set in order to duplicate examples from the minority classes and balance them out.

Feature scaling and normalization were used to decrease variance between numerical features and help the models converge. Considering that some of the numerical features of the dataset are characterized by very different values, scaling would allow not letting one of the features affect the learning process too much.

Next, class imbalance in the dataset was considered. Some attack classes include far fewer examples than benign data and the majority of attacks. Therefore, oversampling technique, RandomOverSampler, was used on the training set in order to duplicate examples from the minority classes and balance them out.

3.3 ALGORITHM WORKFLOW

The suggested intrusion detection system utilizes a supervised machine learning approach that can be used to automatically detect malicious behavior in network traffic using the CICIDS2017 dataset. This approach is based on a series of steps such as dataset acquisition,

pre-processing, imbalance handling, machine learning model training, and attack classification among others. Overall, this process was performed by utilizing machine learning libraries in Python as well as Microsoft Azure AutoML.

Dataset Acquisition (Step 1)

59

Initially, the task was to acquire the CICIDS2017 dataset, which is a collection of network traffic flows consisting of both normal traffic and attacks. Specifically, this dataset includes statistics on flows that were gathered using the CICFlowMeter tool.

The data was imported to the analysis workspace using data processing libraries such as Pandas and NumPy.

Step 2: Data Sampling

Since the original dataset consists of several millions of data entries, analyzing the entire data will need substantial computation capability. In an attempt to reduce computation cost but maintaining data class ratio at the same time, 20% of data was sampled using stratified sampling. Stratified sampling ensures equal proportions of benign traffic and attacks in the sampled data.

Step 3: Data Cleansing and Preprocessing

A number of data preprocessing methods have been applied before fitting the datasets into the model.

Data preprocessing included:

- Duplicate data elimination
- Missing value identification and imputation

- Infinite/invalid value elimination
- Data normalization
- Target encoding

However, infinite values that arose in the computation of the flows have been substituted and filtered out to avoid possible instability in the machine learning training phase. The target labels were further encoded using two classification methods:

- Binary classification (BENIGN vs ATTACK)
- Multiclass classification (each attack category)

Feature scaling and normalization methods were applied as well to guarantee numerical uniformity of the traffic data.

Step 4: Managing Class Imbalance Problem

CICIDS2017 suffers from extreme class imbalance since benign traffic data and some attack classes include significantly more instances than other attacks such as Heartbleed and Infiltration attacks.

To manage the problem of class imbalance, RandomOverSampler has been applied in the training process. Over-sampling involved duplicating instances from minority classes to ensure adequate representation among all classes. Over-sampling contributed to avoiding classifier bias against majority classes and improved detection of rare attacks.

Step 5: Train-Test Split

After data processing, the dataset was then subjected to a train-test split, with the 80/20 split ratio being applied.

- 80% of the data would be used for training the model
- The other 20% of the data would be used for testing and validating

Stratified train-test split technique would be used to ensure that the class balance would remain consistent throughout both data subsets.

Step 6: Rule-Based Intrusion Detection as a Baseline

Prior to training machine learning models, rule-based detection was performed to simulate the behavior of a traditional intrusion detection system (IDS).

A rule-based intrusion detector would identify unusual packets in terms of such threshold parameters as:

- Very high packet transmission rate
- Abnormal flow duration value
- Suspicious byte transfer rate
- Too many connection attempts

This baseline algorithm provided for comparison purposes for the performance evaluation of other models.

Step 7: ML Models Training

Supervised machine learning algorithms were trained based on the preprocessed data.



Logistic Regression

Logistic Regression was implemented as a baseline algorithm for linear classification of malicious traffic versus benign traffic.

14

Random Forest

Random Forest algorithm uses several decision trees to enhance accuracy of prediction and avoid overfitting.

XGBoost

XGBoost is a gradient boosting model that could learn complicated traffic patterns through iterative error reduction. The algorithm is famous for its high efficiency and good predictive power.

LightGBM

LightGBM was applied in both binary and multiclass intrusion detection problems where its efficient gradient boosting approach allowed fast model training without any loss in classification accuracy.

Step 8: Performance Metrics for Model Evaluation

- The trained models were analyzed based on various performance metrics as a way of evaluating the models' ability to perform intrusion detection.

25

- The performance metrics included:

Accuracy

Precision

Recall

F1-score

ROC-AUC

25

Confusion matrix

Precision-recall curve

33

Step 9: Feature Importance Analysis Using Ensemble Models

50

Feature importance analysis was performed using machine learning algorithms such as Random Forest and LightGBM.

The important traffic features were selected from those that influenced intrusion detection and they include:

1

- Flow bytes per second
- Flow packets per second
- Packet length features
- Flow duration
- Packet inter-arrival times

This made it possible to determine which features of traffic were most related to the attacks.

Step 10: Multi-Class Attack Classification

Besides binary attacks, several multiclass attacks were also tested to see whether specific categories of attacks could be detected.

The multi-class attacks include:

- DDoS
- DoS
- PortScan

- Brute Force
- Botnet
- Web attacks
- Infiltration attacks

The main idea of the multi-class classification was to evaluate how efficient the AI-based IDS could be in different situations.

Step 11: Performance Comparison

Finally, all machine learning algorithms were compared to the rule-based method for detection of cyber-attacks.

The performance comparison was based on the following criteria:

- Detection accuracy
- False positives
- Attack detection ability
- Class imbalance effects
- Computational complexity

It became evident that machine learning-based intrusion detection systems were able to detect attack patterns more accurately than the rule-based system.

CHAPTER-4

RESULTS

From the conducted experiment, one can conclude that AI-based intrusion detection systems perform very well when it comes to recognizing attacks present in network traffic. Several supervised machine learning models have been trained and tested on CICIDS2017 data, and then compared to a rule-based IDS used for solving similar cybersecurity issues. As seen from the received results, a substantial boost in detection precision and classification ability can be provided by an intelligent approach.

Both preprocessing and balancing have significantly impacted model results. Once the data had been cleaned, imbalanced, and prepared, it became evident that machine learning algorithms worked much better at recognizing both dominant and rare attack classes. It is especially worth noting that the application of RandomOverSampler helped increase recall rate of underrepresented attack types which previously lacked precision.

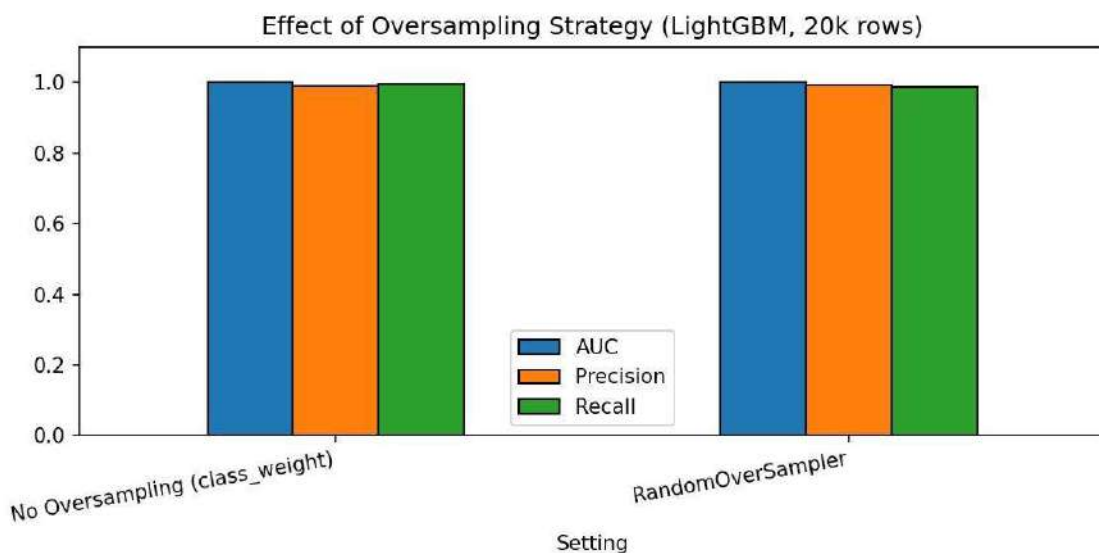


Figure 5: Effect of oversampling strategy

4

The baseline approach was characterized by higher numbers of false positives and false negatives than the machine learning approaches.

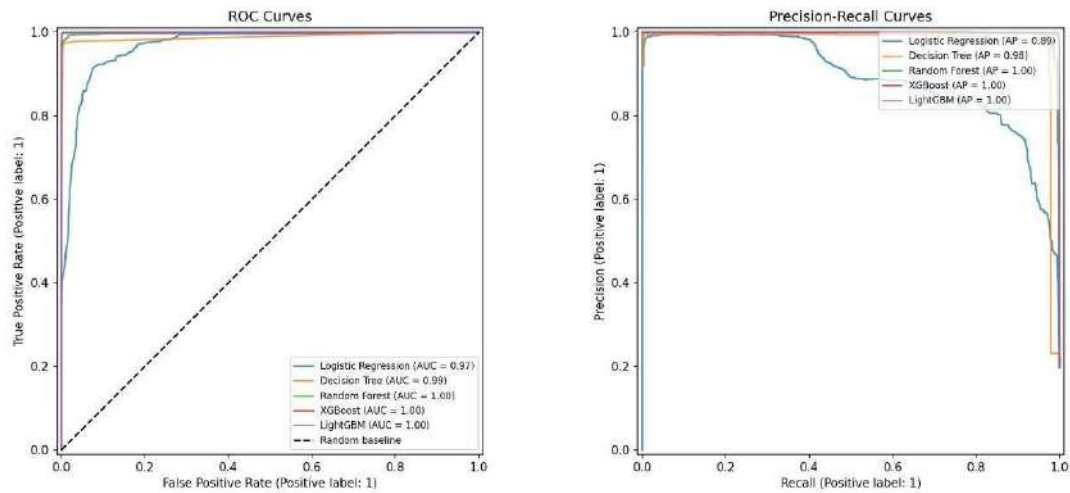


Figure 6: ROC curves and Precision-Recall curves

Logistic Regression performed poorly compared to other algorithms in achieving a relatively high AP score of 0.89. Therefore, it can be observed that ensemble models, including Random Forest, XGBoost, and LightGBM, perform better than Decision Tree and Logistic Regression.

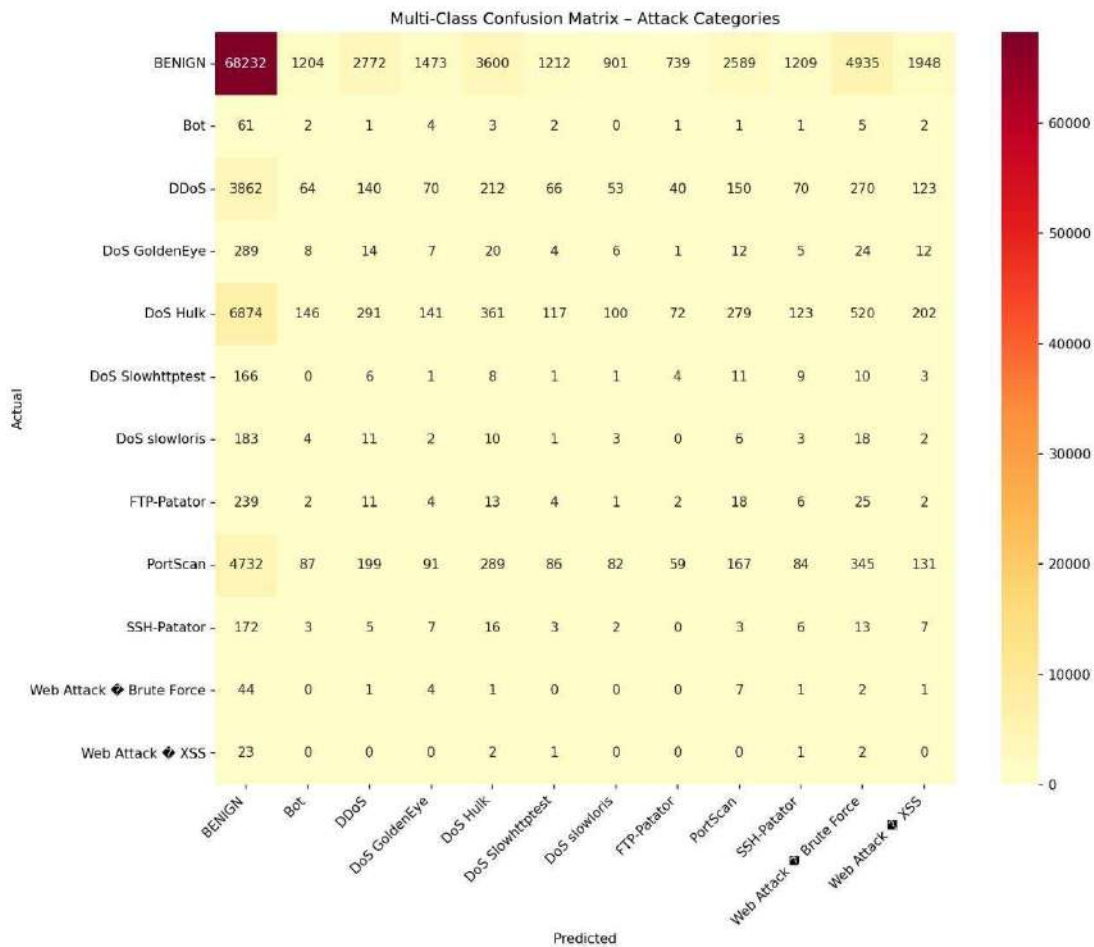


Figure 7: Multi-class confusion matrix

In the following diagram, the multi-class confusion matrix of attack classification is presented for the network intrusion detection system. Here, the rows refer to the true class labels of the attack category, while the columns denote the predicted labels provided by the classifier algorithm. The cells along the diagonals show the number of correctly classified samples, whereas the other cells exhibit misclassified samples. Darker shading signifies greater number of samples, as indicated by the color scale at the right. The classification model shows an excellent performance rate for the BENIGN class with 68,232 classified instances.

12 However, some samples categorized as BENIGN fall under the wrong attack categories such as DoS Hulk, PortScan, and Web Attacks – Brute Force. The DoS Hulk and PortScan attack types comprise many samples, while there are some instances of their confusion with the BENIGN category. Small attack categories such as Bot, DoS Slowhttptest, DoS slowloris,

9

SSH-Patator, and Web Attack – XSS have fewer samples and rare cases of being mislabeled.

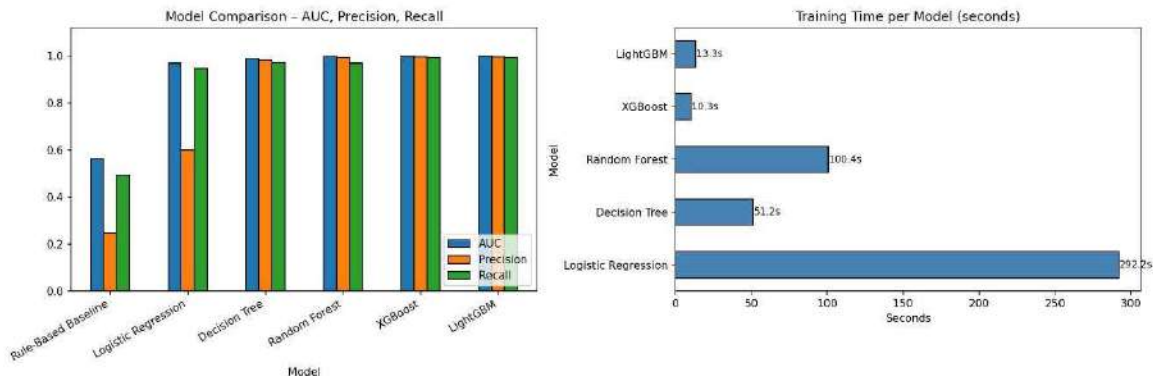


Figure 8: Model Comparison

This figure illustrates the comparison between various machine learning techniques using performance metrics and training time. The first subplot on the left side provides a comparison among the following six models through their Area Under the Curve (AUC), Precision, and Recall values: Rule-Based Baseline model, Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Among these models, the Rule-Based Baseline model shows the least performance, and Logistic Regression attains the best AUC value but a comparatively low precision value than other machine learning techniques. The Decision Tree, Random Forest, XGBoost, and LightGBM techniques consistently deliver outstanding performance in all three categories.

The second subplot shows the amount of time taken for training each algorithm, in seconds. The Logistic Regression model took the most training time, which was around 292.2 seconds, whereas Random Forest took 100.4 seconds and Decision Tree took 51.2 seconds. On the other hand, the training time for XGBoost was approximately 10.3 seconds and for LightGBM was around 13.3 seconds. It can be concluded that the gradient boosting algorithms are not only more accurate than others but also efficient.

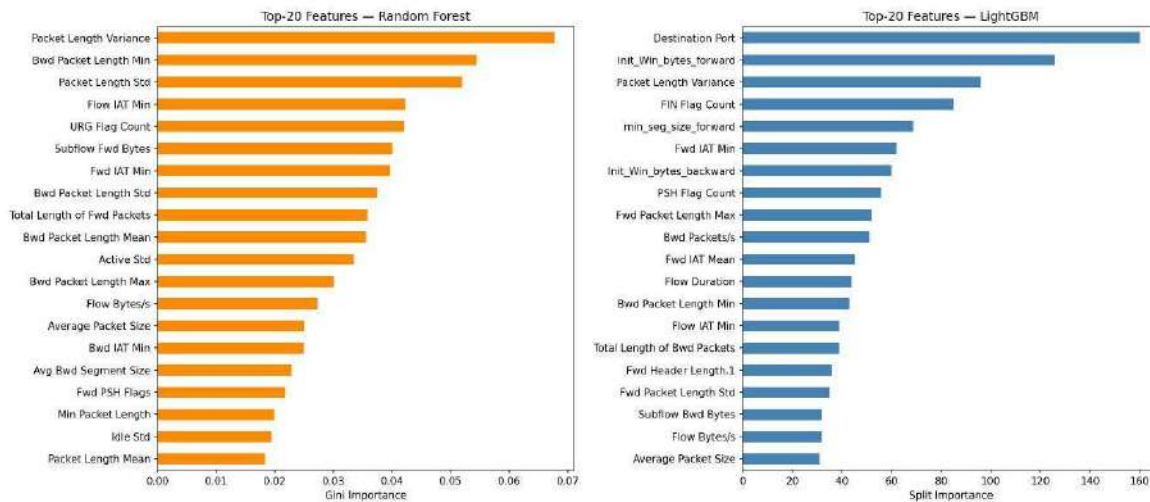


Figure 9: Features

The figure above shows the top 20 most significant features extracted from both the Random Forest and LightGBM algorithms for detecting network intrusions and classifying network attacks. Feature Importance analysis plays a critical role in understanding which network traffic features affect the prediction process the most.

The left graph shows the importance scores of all features calculated by using the Random Forest algorithm. As can be seen in the graph, Packet Length Variance turns out to be the most important feature among all, followed by Bwd Packet Length Min, Packet Length Std, and Flow IAT Min. Some other important features include some statistics-based features from traffic flows like packet lengths, inter-arrival times, and byte counts in packets.

The right subplot illustrates the importance of features as generated by the LightGBM classifier. Here, the dominant feature is Destination Port, followed by Init_Win_bytes_forward, Packet Length Variance, and FIN Flag Count. Other important features include various forward/backward packet metrics, flow duration, and other flagrelated metrics. As opposed to Random Forest, LightGBM pays closer attention to protocol and connection-related features.

Generally speaking, from the above illustration, it is evident that both models view packet length parameters and flow-related metrics as the main criteria for intrusion detection. At the same time, the differences in feature importances reveal the differences in the learning patterns generated by these two models.

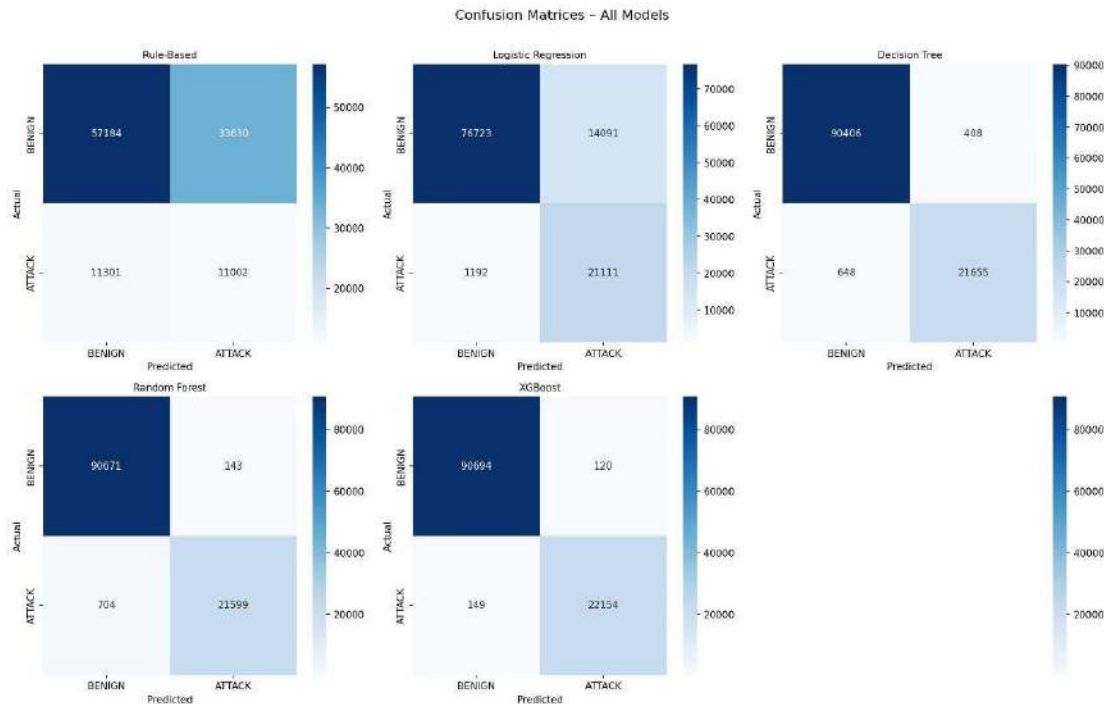


Figure 10: Confusion Matrices

Figure above shows the confusion matrixes of the five classification models namely Rulebased Baseline, **Logistic Regression, Decision Tree, Random Forest, and XGBoost** used for binary classification **intrusion detection**, where the network traffic can be either class BENIGN or ATTACK. The confusion matrix illustrates the connection between actual classes and predicted classes with the diagonals showing the correct classification and nondiagonals showing the wrong classification, where darker color intensity signifies a higher number of samples.

Rule-Based classifier model shows lower performance compared to other classifiers, with many benign samples being wrongly classified as attacks and also many attacks being classified as benign traffic.

The Decision Tree model achieves substantially higher accuracy, with very few misclassifications compared to the previous models. The algorithm Random Forest boosts the model by reducing the cases of misclassification to achieve very high accuracy in

separating benign and malicious traffic. Of all the models, the one that performs the best is the XGBoost with the least number of errors in classification.

Overall, the graph shows an increase in effectiveness in detecting intrusions from simpler to more complex models. Ensemble methods like Random Forest and XGBoost have improved classification, making them more reliable in detecting malicious activities within the network.

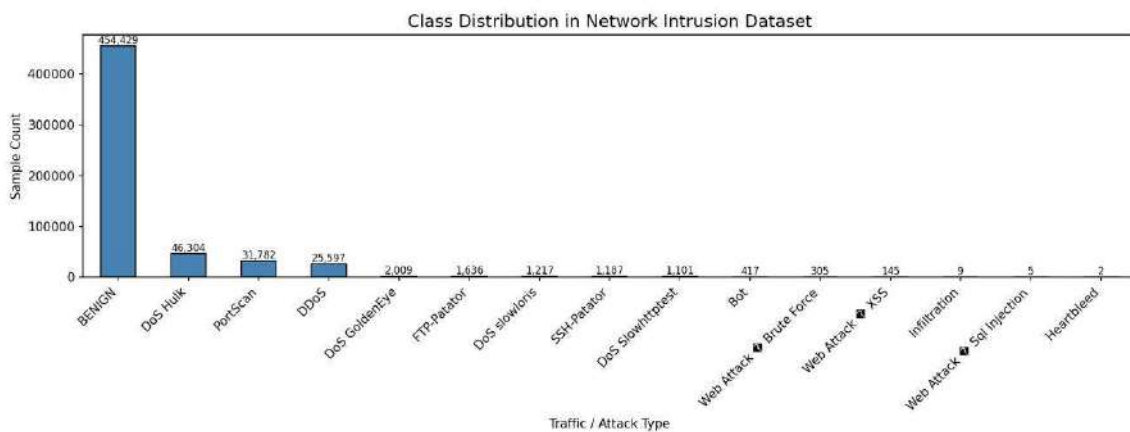


Figure 11: Class Distribution

As depicted in the figure below, there is an obvious imbalance between various classes in the Network Intrusion Dataset. The largest number of samples belongs to the BENIGN class (454,429), which is obviously the largest of all attack classes. On the side of the malicious traffic classes, DoS Hulk (46,304), PortScan (31,782) and DDoS (25,597) stand out as they include quite high numbers of data, while some classes such as Heartbleed (2), SQL Injection (5), and Infiltration (9) have very low data values.

The best results have been achieved through the application of ensemble methods such as Random Forest, XGBoost and LightGBM. All these models exhibited good capacity to learn non-linear relationships among the different network traffic features, and could classify normal and abnormal patterns accurately.

Logistic regression algorithm performed relatively well in classification but it did not produce better results compared to ensemble learning algorithms due to its inability to model complex patterns of network traffic.

Random Forest performed with high accuracy while showing robust resistance to overfitting due to its decision-tree-based ensemble architecture. It coped efficiently with the problem of processing numerous input features and performed stably in terms of different types of attacks.

XGBoost additionally improved intrusion detection system performance by continuously minimizing classification errors during the training process. It performed with high precision and recall rates, which proves its high capability of detecting attacks without false positives. LightGBM proved to be the most performant algorithm out of all considered approaches. Its gradient boosting framework ensured efficient and highly accurate training process while maintaining a relatively low amount of computation. LightGBM proved its high performance on multiclass attack classification task by performing well at classifying several types of attacks into distinct classes.

As it was shown by evaluation metrics, machine learning approaches were characterized by high accuracy, precision, recall and F1-score values. High values of precision imply relatively small number of false positive cases, whereas recall demonstrates the ability to detect attacks correctly. High F1-score indicates that precision and recall of the models have been properly balanced, which is crucial in this application.

The confusion matrix analysis further substantiated the efficiency of machine learning models by showing that most of the benign and malicious traffic samples were classified accurately while only a few attacks samples suffered from classification mistakes. The ensemble learning models showed substantially less classification mistakes than Logistic Regression and rule-based detection system.

The ROC-AUC analysis showed high capability of separation between benign and malicious traffic. The ensemble learning models generated the highest possible area-under-curve values indicating their ability to differentiate attack traffic on different decision thresholds. At the same time, precision recall graphs indicated that despite having problems of dealing with imbalanced classes, the models performed quite stably at that point.

Multiclass classification analysis showed that the developed models can recognize multiple categories of attack with good accuracy. The DDoS, DoS, PortScan and Brute Force attacks have been classified accurately with good prediction performance. However, the rare categories of attack like Heartbleed and Infiltration attack have shown relatively lower classification performance due to insufficient sample representation.

Feature importance assessment revealed several traffic characteristics that played an important role in intrusion detection. Traffic characteristics that included packet rate, flow duration time, packet size, and byte rate exhibited the most significant impact on attacks' classification. They provided meaningful insights in terms of determining whether traffic behavior is malicious or benign.

Comparative assessment of different models indicated that intrusion detection systems based on artificial intelligence principles proved to be superior to those based on rules in regard to their performance. Machine learning models were capable of detecting complex attacks that were not easily detected with threshold-based approaches.

CHAPTER-5

CONCLUSION

17 The main aim of this research was to test whether machine-learning intrusion detection systems would be able to detect cyber attacks in large network traffic effectively reducing reliance on rule-based security monitoring practices.

66 The results of this study show that machine-learning algorithms are capable of detecting intrusions with high accuracy, reliability, and preciseness. Such machine-learning algorithms as Logistic Regression, Random Forest, XGBoost, and LightGBM were successfully used to classify between normal and malicious traffic based on statistically derived network flows from the CICIDS2017 data set. Among different models, those employing ensemble learning technique, i.e., Random Forest, XGBoost, and LightGBM proved to outperform other algorithms as they allowed handling nonlinear relations and patterns within cyber-security data.

It should be mentioned that the conducted study stressed the relevance of data preprocessing and balancing when enhancing intrusions detection effectiveness. Thus, duplicate deletion, invalid values handling, feature scaling, and application of the RandomOverSampler tool helped increase the quality of input data, thus enabling the models to better detect minority classes of intrusions. It has been proven once again that imbalance handling is a prerequisite for obtaining reliable classification results for cybersecurity-related datasets.

Finally, comparing the results obtained by applying machine learning methods with those attained during implementation of the rule-based model proved superiority of AI-powered solutions. As compared to rule-based models which are capable of detecting only predefined attack patterns, machine learning models demonstrated the ability to detect unknown intrusion types. This fact confirms that machine learning-based IDS is more flexible and scalable than traditional systems.

It has been also proven that machine learning techniques can effectively classify a dataset consisting of five categories of cyber intrusions: DDoS, DoS, PortScan, Brute Force, and

Botnet attacks. Despite the difficulty in identifying several infrequent attack types due to insufficient training examples, the obtained results have demonstrated high performance of the implemented solution.

The analysis showed that the parameters which describe network traffic flows are important for detecting malicious activities. Thus, the significance of the statistical flow-based analysis was shown for detecting intrusions in computer networks. In general, the research proved that AI and machine learning are effective approaches that could be used to build efficient and scalable IDS systems.

61 In summary, the investigation confirms that machine learning-based and AI intrusion detection systems are an effective and efficient solution for modern computer network security. The introduction of AI technology into modern cybersecurity solutions is expected to increase efficiency in threat detection by reducing the number of false positives and negatives in real time. At the same time, the use of such approaches can help to decrease the workload of security specialists in cybersecurity companies.

54

CHAPTER-6

DISCUSSION

Due to the emergence of new technologies, such as digital technologies, cloud computing, IoT, and AI applications, there have been tremendous changes in the amount of complex traffic created by companies. The current cybersecurity systems that mostly depend on manual monitoring and signature-based detection become less efficient in dealing with cyber attacks. Thus, in order to find out whether the machine learning algorithms and AI could solve the issue of intrusion detection, this paper has investigated the applicability of these technologies in network security.

According to the results of the analysis, machine learning technology is able to detect intrusions by analyzing network traffic at scale and finding malicious patterns with a high degree of precision. When comparing machine learning and rules-based system, it can be claimed that the former one proved its effectiveness in detecting both known and unknown attacks. Consequently, machine learning technology offers a much better solution when dealing with cybersecurity.

31 Among the notable observations from the study is the superior performance of ensemble learning models such as Random Forest, XGBoost, and LightGBM. This is attributed to the fact that they can learn complex relationships between traffic features. As such, they were able to perform better in classification compared to Logistic Regression since they could capture nonlinearity among traffic features. In addition, the combination of weak learners by ensemble learning models made them more effective at intrusion detection. It should be noted that LightGBM had an advantage in terms of efficiency when working with large data sets.

Another important observation from the study concerns the preprocessing of traffic data and feature engineering in intruder detection systems. CICIDS2017 was observed to have duplicate records, missing values, infinitely valued fields, and an extremely skewed class distribution which made training difficult. However, removing these problems led to enhanced performance of the models. In addition, it was found that statistical characteristics

such as the rate of byte transfers, the packet transfer rate, and the duration of flows are crucial factors in detecting attacks based on traffic features analysis.

Furthermore, another important issue in discussing is the effect of class imbalance on machine learning algorithms. Given the presence of underrepresented attack classes in the CICIDS2017 dataset, the tendency of machine learning algorithms to predict majority classes appeared. It should be noted that using the RandomOverSampler algorithm helped to leverage the information of minority classes and reduce misclassified attacks.

Nevertheless, using an oversampling technique may be risky, as it may cause overfitting. Thus, choosing proper approaches to achieve balance is still critical.

The performance of the machine learning models was compared with the rule-based baseline approach to show that traditional ways of detecting hazardous materials had major drawbacks. Being based on threshold values and known attack signatures, rule-based solutions have a number of limitations. For example, they are unable to identify new attack types because there aren't any signatures to identify them. Machine learning methods can help spot patterns of network activity, meanwhile. This enables machine learning algorithms to deal so much better with new threats and zero-day attacks.

Despite these overall beneficial results, there may still be some limitations that are present. The experiments were carried out in CICIDS2017 dataset in a controlled lab environment. The real-world network environment may be more unpredictable than the simulation network environment, which can result in performance degradation when deploying ML solutions developed in a simulated environment with historical data as they encounter new threat types in the wild. There is a need for regular model updates to make good threat detection.

The next restriction is related with the description of the sophisticated machine learning models. Ensemble methods and boosting can be viewed as black box techniques that are not easily comprehensible in terms of how specific predictions are generated. This is an important aspect of enterprise cybersecurity, because security decision makers are asked to demonstrate security related decisions by explaining some of the alerts and events. This is why it is a necessity in practice to have explainability.

CHAPTER-7

FUTURE SCOPE

Increasing sophistication of cyberattacks together with the constantly growing amount of data processed in terms of network traffic provides an excellent ground for further progress in terms of AI-based intrusion detection systems. Even though the current research showed how machine learning techniques are able to identify malicious activities within network traffic, there are still certain areas of improvement.

An interesting direction to consider for future work will be the use of deep learning algorithms for intrusion detection purposes. Such algorithms include CNNs, RNNs, LSTM networks, as well as transformer architecture, which may help to achieve a higher level of complexity in terms of attack detection by taking into account both temporal dependencies and more complicated patterns in traffic. Moreover, such algorithms may help to detect very sophisticated traffic patterns.

Another important point to consider in future research would be real-time intrusion detection systems that integrate data from active network traffic monitoring platforms. Currently, the study has been performed on offline CICIDS2017 dataset in a controlled environment. In practice, any intrusion detection system has to process streaming traffic constantly and react immediately in case of detecting malicious activity. Machine learning algorithms may be seamlessly integrated with real-time cloud-based systems like Microsoft Azure.

A related issue that requires future consideration is the improvement of techniques used to handle class imbalances.

This makes it difficult to understand decisions made by algorithms used in intrusion detection systems. XAI can help improve trust in intrusion detection systems by providing explanations about the attacks detected by the system. Future research can also be done on the incorporation of intrusion detection systems with SIEM solutions and automated incident response. The combination of AI-driven threat detection together with automated response

can help organizations react rapidly when cyber attacks happen. Cybersecurity technologies that work autonomously will have a significant contribution to cybersecurity management.

Federated learning and distributed AI models can be incorporated into cybersecurity practices for collaborative threat detection. Federated learning refers to **the process of training machine learning models** using **data** from different parties without the exchange of sensitive network data. The incorporation of federated learning will be beneficial to large cybersecurity networks.

Future research can also consider the application of hybrid detection systems which will incorporate signature-based detection alongside anomaly-based machine learning models. Moreover, future research can analyze the performance of the intrusion detection models based on recent and more varied cybersecurity datasets. With time, new methods of attacks emerge, and hence training of the models on more recent datasets that capture the current threat landscape would become necessary.

Lastly, optimization for efficiency is another area for future research in this field. For instance, when deploying AI-based intrusion detection systems on large corporate networks, efficient use of resources becomes a priority since the system would have to handle a huge volume of network traffic with minimal latency and computational cost. Future studies could, therefore, develop lightweight machine learning algorithms for edge and cloud computing cybersecurity applications.