T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF

ARTIFICIAL INETLLIGENCE HEAD OF THE DEPARTMENT

ENHACING INDUSTRIAL SYSTEM RELIABILITY THROUGH PREDICTIVE ANALYSIS AND INTRUSION DETECTION USING MACHINE LEARNING

MASTER'S THESIS

HOUDA SAMIRY

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF

ARTIFICIAL INTELLIGENCE HEAD OF THE DEPARTMENT

ENHACING INDUSTRIAL SYSTEM RELIABILITY THROUGH PREDICTIVE ANALYSIS AND INTRUSION DETECTION USING MACHINE LEARNING

MASTER'S THESIS

THESIS ADVISOR

DR. MUSTAFA OZBAYRAK

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program Name:	ARTIFICIAL INTELLIGENCE
Student's Name and Surname:	Houda samiry
Name Of The Thesis:	Enhancing Industrial System Reliability Through Predictive Analysis and Intrusion Detection Using Machine Learning
Thesis Defense Date:	07/06/2024

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

Assoc. Prof. Yücel Batu SALMAN
Institute Director

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

	Title/Name	**Institution**	**Signature**
Thesis Advisor's	Dr.Mustafa Ozbayrak	Bahcesehir University	
Member's	Dr.Ismail Duru	Bahcesehir University	
Member's	Dr.Nizamettin Aydin	Istanbul Technical University	

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

Name, Last Name: Houda Samiry

Signature:

ABSTRACT

ENHACING INDUSTRIAL SYSTEM RELIABILITY THROUGH PREDICTIVE ANALYSIS AND INTRUSION DETECTION USING MACHINE LEARNING

Houda Samiry

Master's Program in Artificial Intelligence

Supervisor: Dr.Mustafa Ozbayrak

Mai 2024,48 pages

This research focuses on the application of the concept of predictive analysis and IDS through improved machine learning to promote the effectiveness and safety of industrial systems. To predict the failures of equipment and to detect the cyber threats, several models such as Random Forest and Logistic Regression were used based on the CICIDS 2017 dataset. The study revealed high levels of accuracy and resonance especially with the random forest model that performed well in the two prediction tasks. Major conclusions included the significance of data preprocessing, feature selection and the general utilitarian values of consolidating all the mentioned technologies. The study unearthed issues of data quality, model size considerations, and real-time processing and provided insights for actuation along with ideal policy adherence. Some of the future research areas of interest involve the enhancement of data fusion techniques, exploring the possibilities in edge computing, and machine learning algorithms with learning capabilities. The result of this combined approach is not only reduced downtime and operational expenses associated with maintenance, but also enhanced security – a critical requirement of today's industrial landscape. Therefore, this study highlights the significance of the constant improvement and further research in the field of predictive maintenance and IDS to provide reliable and protected industrial processes.

ÖZ

MAKİNE EĞİTİMİNİ KULLANARAK ÖNGÖRÜCÜ ANALİZ VE SİZİN GİDERME TESPİTİ YOLUYLA ENDÜSTRİYEL SİSTEM GÜVENİLİRLİĞİNİN ARTIRILMASI

Houda Samiry

Yapay Zeka Yüksek Lisans Program

Tez Danışmanı: Dr.Mustafa Ozbayrak

Mayıs 2024, 48 sayfa

Bu araştırma, endüstriyel sistemlerin etkinliğini ve güvenliğini artırmak için gelişmiş makine öğrenimi yoluyla tahmine dayalı analiz ve IDS kavramının uygulanmasına odaklanmaktadır. Ekipman arızalarını tahmin etmek ve siber tehditleri tespit etmek amacıyla CICIDS 2017 veri setini temel alan Rastgele Orman ve Lojistik Regresyon gibi çeşitli modeller kullanıldı. Çalışma, özellikle iki tahmin görevinde iyi performans gösteren rastgele orman modeliyle yüksek düzeyde doğruluk ve rezonans ortaya çıkardı. Başlıca sonuçlar, veri ön işlemenin önemini, özellik seçimini ve bahsedilen tüm teknolojileri birleştirmenin genel faydacı değerlerini içeriyordu. Çalışma, veri kalitesi, model boyutuyla ilgili hususlar ve gerçek zamanlı işlemeyle ilgili sorunları ortaya çıkardı ve ideal politika bağlılığının yanı sıra harekete geçirmeye yönelik öngörüler sağladı. Gelecekteki ilgi duyulan araştırma alanlarından bazıları, veri birleştirme tekniklerinin geliştirilmesini, uç bilişimdeki olanakların araştırılmasını ve öğrenme yeteneklerine sahip makine öğrenimi algoritmalarını içermektedir. Bu birleşik yaklaşımın sonucu, yalnızca bakımla ilgili arıza sürelerinin ve operasyonel masrafların azalması değil, aynı zamanda günümüzün endüstriyel ortamının kritik bir gereksinimi olan gelişmiş güvenliktir. Bu nedenle bu çalışma, güvenilir ve korumalı endüstriyel süreçler sağlamak için kestirimci bakım ve IDS alanında sürekli iyileştirmenin ve daha fazla araştırmanın önemini vurgulamaktadır.

To my supervisor, whose guidance and support made this research possible, and to my parents, sister, and husband, whose unwavering encouragement and love have been my pillars of strength throughout this journey.

ACKNOWLEDGEMENTS

I am deeply indebted to my supervisor, Dr.Mustafa OZBAYRAK, for his invaluable guidance and unwavering support throughout this research endeavor. His mentorship has been instrumental in shaping the trajectory of this study, and I am grateful for his expertise and encouragement.

To my dear parents, Aziza and Mohammed, and my beloved sister, Zineb, I extend my heartfelt gratitude for their unconditional love, unwavering support, and endless sacrifices. Their belief in me has been my greatest motivation, and I am forever grateful for their unwavering presence in my life.

Lastly, to my beloved husband, Rachid, your unwavering faith in me and your unwavering support have been my greatest motivation. Your love and encouragementhave given me the courage to overcome challenges and strive for greatness.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

IDS: Intrusion Detection Systems

IoT: Internet of Things

IIoT: Industrial Internet of Things

CSV: Comma-Separated Values

DoS: Denial of Service

DDoS: Distributed Denial of Service

FTP: File Transfer Protocol

SSH: Secure Shell

XSS: Cross-Site Scripting

SQL: Structured Query Language

KDE: Kernel Density Estimate

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

CBM: Condition-Based Maintenance

AUC: Area Under the Curve

ROC: Receiver Operating Characteristic

PCA: Principal Component Analysis

RFE: Recursive Feature Elimination

MAE: Mean Absolute Error

RMSE: Root Mean Square Error

SHAP: SHapley Additive exPlanations

Chapter 1

Introduction

1.1 Introduction

In today's industrial environment, system stability and protection has emerged as a crucial consideration. The growth in systems' complexity and integration within industries coupled with the introduction of IoT and other technologies require new ways for continuity of operations and protection from threats. The application of machine learning in proactive analysis and detection of intrusion has been affirmed to be critical in tackling these challenges. This chapter provides background information of the study, development of the research problem, objectives of the study and the research questions that guide the study in enhancing reliability in industrial systems using predictive analysis and intrusion detection using machine learning.

1.2 Background of the Research

Reliability has emerged as the most critical factor in most of the contemporary industrial settings. Due to advancing technologies and the rapid increase in the Internet of Things, industrial systems are evolving and integrating closer. Thus, the continuity of operations is pertinent (Shukla, Samia, & Davis, 2022). Malfunctions put not only billions of dollars at stake but also significant dangers associated with safety and the environment. This makes the reliability of these systems of supreme importance (Shukla et al., 2022). With machine learning, predictive analysis and intrusion detection have emerged as necessities to solve issues in such systems due to its ability to predict failure and identify security breaches before they escalate into major catastrophes.

Predictive analysis in industrial systems uses past data and complex formulas to indicate where likely failure may occur and where maintenance is required. Machine learning helps the creation of predictive models in big datasets where future system behavior is determined (Sghir, Adadi, & Lahmer, 2022). This is because such an approach helps organizations maintain their maintenance up to date so that the probability of system failure is reduced alongside its associated risks. Similarly,

predictive analytics is used where processes are ongoing, for instance in manufacturing plants and electricity supply, and transport facilities (Sghir et al., 2022).

Alongside with predictive analysis, intrusion detection systems are the key to protecting industrial systems from threats. As the industry processes get digitized, the risk factor is higher and the mentioned systems are vulnerable to cyber threats all the time. Machine-learning-based IDS can analyze the network traffic for anomaly detection and potential security threats in real-time (Rojek et al., 2023). These are the systems that, because of learning based on past incidents and adaptation to the new emerging threats, become the most muscular guard against cyber intrusions. Thus, it makes the IDS integrated with machine learning more tolerant in differentiation between benign and malicious activities, therefore enhancing more accuracy and reliability toward threat detection.

This merging of predictive analysis with intrusion detection created by machine learning provides a complete approach to increasing the reliability of industrial systems. This way, organizations can be prepared to predict and prevent any kind of system failure and secure their operations against cyber threats. Hence, a combination of these two strategies shall be adapted toward a higher level of operational resilience that will enable industries to remain in continuous production and service provision. Integration of both technologies generally portrays a trend towards intelligent, data-oriented solutions in industrial management and ushers in more resilient, more secure industrial systems.

1.3 Statement of the Problem

Ensuring operational reliability and safety in the industrial applications is a tasking; it is even hard for the present sophisticated and integrated systems. Maintenance and security challenges have always been treated via traditional and conventional approaches that tend to breakdown these challenges into manageable small bits and as such are more often than not reactive (Neupane et al., 2022). Some of the maintenance activities include regular scheduled or reactive maintenance that leads to high operational costs and more downtimes. The third reason that make industrial operations vulnerable to cyber threats is that traditional security measures are often outdated and unable to adapt to emerging risks (Neupane et al., 2022). This

now leads us to the fact where the high relevance of advanced proactive solutions comes into play for better reliability and security.

The problem is the difficulty of identifying system failures that may occur in the future. This predictive gap leads to unpredictable downtime something that is very damaging as it is very costly and disruptive. This is detrimental to all industries as every business depends on a consistent flow of production since disruptions have ripple effects on production timelines, supplies, and final output. Besides, the data carrying capacity of monitoring and maintenance systems for large industrial facilities is more than traditional methods. Without the ability to analyze that data effectively, valuable information is lost that can therefore present problems before they arise.

Against this background, the current research answers the imperative need for an increased level of industrial system reliability and security through predictive analysis and machine learning-driven intrusion detection. Machine learning could be utilized to process big datasets for accurate and timely equipment failure prediction to enable early detection of cyber threats. These technologies promise to reshape the approaches of industries to maintenance and security, from reactive to proactive. Such a proactive position not only reduces downtime and operational costs but also uplifts the level of security for infrastructure protection against evolving cyber threats. Integrating predictive analysis and advanced intrusion detection systems marks a strong stride toward fostering resilience and security in industrial operations.

1.4 Purpose of the Study

The main objective of this study is to investigate and present how predictive analysis could be effectively integrated with machine learning-based IDS for the improvement of dependability and security in industrial systems. This is due to the very fact that industries are increasingly using complex, interlinked systems whose failure cannot only be predicted but also fended off at all costs. Therefore, the study will be oriented toward comprehensive learning on how machine learning technologies can be instrumental in ensuring adequate prediction of system failures and alerting security breaches to enhance the continuity of operation and reliability of the state.

Given that, the project will focus on the development and application of machine learning models for predictive analysis using different ML models. These models, through the analysis of historical industrial system data, can reveal patterns and abnormalities that might point toward failures. The idea is to make a robust predictive analysis strategy, which would bring about minimum downtime.

The study will also seek to understand the application of machine learning algorithms within intrusion detection systems. Therefore, the developed advanced IDS should be able to identify all forms of cyber threats and respond according to the threat in real time. The fact that most traditional IDS do not recognize new or advanced attacks has, in a critical manner, left holes in industrial security. Enabled with the help of machine learning techniques, it aims to develop an adaptive and intelligent IDS system that learns from earlier incidents, recognizes emerging threats, and distinguishes between benign and malignant activities. In this regard, the present study is steered toward fortifying industrial systems against cyber-attacks, which will guarantee the integrity and safety of critical operations.

Ultimately, the study establishes a link between predictive analysis and cybersecurity in industrial systems by machine learning. It aims to base contributions for developing more resilient industrial infrastructures since it reveals the potential of these technologies in increasing system reliability and security. The resulting findings should inform best practices and guide future innovations in industrial system management, thereby setting the path for smarter, more secured, and efficient industrial operations.

1.5 Research Questions

1. How can the integration of predictive analysis and machine learning-based intrusion detection systems enhance the reliability and security of industrial systems?
2. What are the most effective machine learning algorithms and techniques for predictive analysis and intrusion detection in the context of industrial systems, and how do they compare in terms of accuracy, efficiency, and scalability?

Chapter 2

Literature Review

2.1 Introduction

This chapter provides an overview the theoretical frameworks and existing literature regarding predictive analysis and IDS. In this chapter, the author delves into the development of IDS, the various kinds of IDS, and the crucial importance of integrating machine learning into IDS. Moreover, it analyzes the integration of predictive analysis with IDS in terms of advantages, use cases, issues, and recommendations. Lastly, some of the current trends and emerging technologies are highlighted to help in outlining gaps in knowledge that would be valuable for future research in this important area.

2.2 Theoretical Framework

System reliability in an industrial application is the ability of a system to perform its mandatory functions under certain conditions, which are indicated during a given period (Breznická, Kohutiar, Krbaťa, Eckert, & Mikuš, 2023). It is essential because a reliable system ensures unwavering production, reduced downtime, and decreased maintenance costs. High reliability and maximum availability of equipment in industrial operations are essential for the continuity of safety and high levels of efficiency and productivity (Breznická et al., 2023). Systems failure may lead to substantial financial losses, unsafety, or cause environmental hazards. Given this, high levels of system reliability are considered paramount in any industrial organization. Of importance, different predictive and proactive failure-preventing maintenance strategies have been developed and put into practice.

Predictive analytics is indeed based on a proactive method that uses historical and current data to determine a future, even in the event of system breakdown. It comprises statistical techniques and machine learning algorithms to observe trends in data patterns that possibly show the way to some problems (Achouch et al., 2022). Some key elements or concepts of predictive analysis include data collection, preprocessing, and model selection and evaluation (Achouch et al., 2022). Data collection is the process of collecting required information from existing data sources,

such as sensors and log files. Data preprocessing includes both cleaning and transformation in making data appropriate for analysis (Shukla et al., 2022). Model selection consists of selecting appropriate algorithms for analyzing the data, while evaluation involves assessing the performance of the model using such metrics as accuracy, precision, and recall (Yacouby & Axman, 2020). Predictive analysis facilities assist an organization in carrying out the maintenance at the most appropriate time to minimize cases of random failures and therefore enhancing the reliability of the systems (Rojek, Jasiulewicz-Kaczmarek, Piechowski, & Mikołajewski, 2023).

Intrusion Detection Systems is a cybersecurity tool to identify unauthorized access or suspicious activities in hosts or the network. IDS can be categorized into two significant groups: NIDS: Network based IDS and HIDS: Host based IDS (Lansky et al., 2021). Network-based IDS aims at observing network activities in a bid to apprehend packets that may be behaving uncharacteristically (Lewandowska, 2024). However, the host-based IDS looks at the behavior and state of various devices at the same time (Martins et al., 2022). Classification of IDS according to methods of detection can also be signature-based and anomaly-based (Lansky et al., 2021). This type identifies the detected threats by analyzing the incoming data against the patterns of the attacks. On the other hand, anomaly-based IDS defines every behavior that has the potential of emerging with new forms of threats not known before (Martins et al., 2022). An effective IDS is so vital in guarding industrial systems from any form of cyber-attack that may affect system availability and integrity.

Machine learning is a subfield of artificial intelligence where algorithms are trained to anticipate outcomes and make decisions without being coded (Dhall et al., 2019). In this regard, machine learning can contribute to algorithms that sort through millions of entries, analyze them, and deduce potential future occurrences, including equipment breakdowns (Rojek et al., 2023). Some of the most frequently used machine learning methods in predictive analysis are regression analysis, decision trees, and neural networks. These algorithms can analyze historical data by understanding the components responsible for system failures and predicting when they are likely to happen (Dhall et al., 2019). This, in turn, enables organizations to execute their maintenance activities in good time so that there is an improvement in system reliability.

Machine learning is crucial in identifying and responding to cyber threats within intrusion detection systems. Traditional IDSes rely on predefined signatures and rules, in which recognition remains limited to new sophisticated attacks (Thapliyal & Thapliyal, 2024). An IDS based on machine learning can detect new types of attacks because it uses the training of historical data and keeps updating its database until the threats emerge (Thapliyal & Thapliyal, 2024). The main approaches are supervised learning, in which algorithms are trained using labeled datasets of normal and malignant activities; the other is unsupervised learning—where algorithms identify patterns based on data sets not being labeled. Recently, deep learning, a subset of machine learning, has also gained increasing popularity in IDS by harnessing neural networks to extract features from raw data and improve detection performance automatically (Thapliyal & Thapliyal, 2024). By equipping IDS with machine learning capabilities, industrial system security can be made more robust and adaptive.

2.3 Predictive Analysis in Industrial Systems

Predictive analysis in industrial systems makes a way to leap the classical ideas of maintenance to data-driven approaches with proactivity. Traditionally, this field has been dominated either by reactive maintenance strategies or scheduled strategies, where equipment is either repaired after the breakdown or replaced at regular intervals, irrespective of the condition it is in (Echkina, Levichev, & Sushko, 2024). Often, these result in too much downtime or unwarranted maintenance activities, both expensive and inefficient. Predictive analysis using machine-learning algorithms and data analytics can predict equipment failure long before it occurs, thus optimizing maintenance schedules and reducing operational disruptions (Echkina et al., 2024). According to Çınar et al. (2020), due to the accurate prediction of equipment degradation and potential failures, the application of predictive analysis strategies can result in enormous cost savings and system reliability improvement.

The core of predictive analysis lies in handling and analyzing massive volumes of historical and real-time data to indicate tendencies toward future problems. Machine learning algorithms use regression models, decision trees, neural networks, etc., in developing predictive models that can learn from historical data and predict future events with a high degree of accuracy (Çınar et al., 2020). For instance, Sang, Xu, and

de Vrieze (2021) applied machine learning techniques to manufacturing data to predict machine tool failures. They achieved savings of a significant percentage of the cost associated with unplanned downtimes and maintenance actions. The overall process includes four distinct steps, starting with the data collection and preprocessing to model training and validation (Sang et al., 2021). Proper data preprocessing involving cleaning, normalization, and feature extraction is paramount in determining model accuracy and reliability in prediction models.

One advantage of predictive analytics is the ability to gain real-time insights into the health of industrial systems. Predictive models are enabled to pick early signs of wear and tear in equipment or deviations from normal operating conditions that would eventually lead to a failure, being designed to monitor performance of equipment continuously and to analyze the flows of data from sensors and other IoT devices (Aljohani, 2023). This real-time capability is an important feature when operational continuity is critical, for example, in the manufacturing, energy, and transportation sectors (Aljohani, 2023). For example, the energy sector has employed predictive analytics in monitoring and maintaining wind turbines; this has led to a general improvement in performance and cost reduction about maintenance (Udo & Muhammad, 2021). According to Fox et al. (2022), failure prediction makes it possible for proactive action ahead of failure and, therefore avoids expensive downtimes while prolonging equipment life.

In addition, predictive analytics supports the shift toward condition-based maintenance (CBM), a strategy in which maintenance is performed based on the actual condition of equipment and not per a set timetable (Teixeira, Lopes, & Braga, 2020). This will not only optimize maintenance efforts but also enhance the overall efficiency and reliability of industrial systems. Alaswad and Xiang (2017) also authored a comprehensive review of the same area, which remarked that predictive analytics, as leveraged by CBM, greatly enhanced the effectiveness of maintenance programs by ensuring resources were committed according to actual need and not upon schedules that are more or less arbitrary. This has been facilitated by the improvements in sensor technology and the increase in ample data availability.

However, despite the advantages, predictive analysis also brings diverse challenges to implementation in industrial systems. One of the significant challenges involves high-quality, comprehensive data for practical training of predictive models (Attaran & Attaran, 2019). Incomplete or noisy data highly reduces prediction accuracy, leading to false positives or negatives. Secondly, incorporating predictive analytics with industrial systems already in place requires a fair amount of investment in technology and skills development that most organizations cannot afford (Aljohani, 2023). According Kumar and Garg (2018), although predictive analysis promises considerable benefits, the adoption thereof is essentially encumbered by technological and organizational challenges. However, the advancement of technology and the increased realization of the value of predictive analysis for organizations are steadily beginning to address these challenges.

2.4 Machine Learning Techniques for Predictive Analysis

Machine learning techniques have transformed predictive analysis in industrial systems by predicting equipment failures and maintenance needs much more accurately and timely. One such core machine learning method applied to predictive analytics is regression analysis, the most vital technique used to predict a continuous outcome variable based on one or more predictor variables (Sghir et al., 2022). The popularity of linear regression models is much attributed to their simplicity and interpretability. They have proved applicable in many industrial cases to predict the equipment's life and failure time. However, the complex non-linear relationships within industrial datasets have been a challenge for linear models (Sghir et al., 2022). Hence more complicated techniques apply, for instance, polynomial regression and support vector regression.

Other significant groups of machine learning algorithms used in predictive analysis are classification algorithms. These prediction algorithms work for categorical predictions, for example, whether a piece of equipment will fail in some specified period (Alnuaimi & Albaldawi, 2024). The technique that has been popular due to its robustness and successful application with large amounts of data whose dimensionality is high is that of decision trees and random forests (Alnuaimi & Albaldawi, 2024). Decision trees can be thought of as constructed based on the rationale to partition the

data recursively into subsets about the best predictors so that at the end of it, a model is produced for making decisions in the shape of a tree (Blockeel, Devos, Frénay, Nanfack, & Nijssen, 2023). An ensemble learning method, random forests combine multiple decision trees to enhance the predictive capability of a model and minimize overfitting (Blockeel et al., 2023).

Among neural networks, there is a lot of buzz around deep learning models, as they can model complex, non-linear relationships well. Deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), automatically extract features from raw data and represent the dependencies over time, respectively (Alzubaidi et al., 2021). The inclusion of CNNs in the approach for the processing and analyzing of sensor data related to rotating machinery seems to have been successful in many other works related to the predictive analysis of rotating machinery systems (Alzubaidi et al., 2021). RNNs, by their nature of handling sequence data, are particularly powerful in predicting time series problems, such as remaining useful life in equipment prediction (Alzubaidi et al., 2021). These neural network architectures exhibit superior performance in capturing complex patterns and dependencies with industrial data.

Feature engineering which is selecting, modifying, and creating new features from raw data is a crucial process in increasing machine learning performance in predictive analytics (Verdonck, Baesens, Óskarsdóttir, & vanden Broucke, 2021). The basic idea is to have domain knowledge to carry out practical feature engineering and conduct iterative experiments. Data dimensionality reduction is commonly done by techniques such as principal component analysis (Verdonck et al., 2021). It helps in simplifying models and reducing computational load; therefore, it is easier to interpret and visualize high-dimensional data. Feature selection methods, such as recursive feature elimination, assist in identifying relevant features that help in improving performance and preventing overfitting of a model (Verdonck et al., 2021).

Performance evaluation of machine learning models in predictive analysis is very important so that the reliability of these models can be gained for the industrial applications. Key evaluation metrics are accuracy, precision, recall, F1 score for classification models, and MAE and RMSE for regression models (Abdullah & Said,

2023). Cross-validation techniques include k-fold cross-validation to assess model generalization and avoid overfitting by dividing data into training and testing sets multiple times (Abdullah & Said, 2023). Model interpretability is also crucial, especially in the industry domain, where insights into 'how decisions are made' based on model predictions prove valuable. Techniques such as SHAP—SHapley Additive exPlanations—support feature importance and behavior of models, which support confidence and acceptance of machine learning models (Rozemberczki et al., 2022).

2.5 Intrusion Detection Systems

Intrusion detection systems have come a long way since their inception because these modern times, with the increase in sophistication of threats, demanded a much better system (Khraisat, Gondal, Vamplew, & Kamruzzaman, 2019). The idea of IDS first originated from one James P. Anderson back in the early 1980s and involved monitoring system logs for unauthorized access patterns. In his 1980 paper "Computer Security Threat Monitoring and Surveillance," he defined IDSs as a program that is set aside, which closely observes a log to see whether an unauthorized user is trying to get in or not (Neupane et al., 2022). The further the growth of the internet regarding networked systems, the more the requirement for real-time monitoring and automatic detection mechanisms, which had to be manifested with the enhancement in the IDS technology (Khraisat et al., 2019). The predecessor IDS relied heavily on predefined rules and signatures to detect known threats. While that was very effective against the then-present attack vectors, it generally faltered when it came to novel, sophisticated attacks (Neupane et al., 2022). This development has, over time, evolved into an impressive integration of statistical methods, machine learning, and artificial intelligence towards developing IDS with advanced capabilities for the detection of a broader range of threats while still adhering to improved accuracy and efficiency in the detection process.

Based on these working principles, intrusion detection systems come as network-based IDS and host-based IDS. NIDS intercepts and analyzes network traffic for signs of malicious activities. These operate strategically in the network, such as on a router or a switch, to capture and analyze data packets on the fly (Trisolino, 2023). One class of IDS is well-suited for detecting attacks across the network, such as denial-

of-service (DoS) attacks, port scanning, and malware propagation. While NIDS are installed at the choke points within the network to inspect all traffic, HIDS are deployed on an individual host or device and control mainly the activities that take place within a single computer system, including accessing files, creation of processes, and logins (Al Lail, Garcia, & Olivo, 2023). HIDS makes it possible to detect particularly insider threats and attacks that bypass network-level defenses (Al Lail, Garcia, & Olivo, 2023). All this shows that everyone has its strengths therefore each can be effectively used in combination with the other for the total security coverage.

A large body of research has investigated intrusion detection using the CICIDS-2017 dataset, which is also the subject of our investigation. For example, Stiawan et al. (2020) tested multiple supervised machine learning algorithms on a modified CICIDS-2017 dataset and achieved a high accuracy of 98.87% by picking important characteristics. Similarly, investigations by Faker et al. in 2019, Vinayakumar et al. in 2019, Yang et al. in 2019, and Zhang et al. in 2019 found high accuracy rates using this dataset.

Another work of Al Lail et al. (2023) compared four algorithms: RF, LSVM, GNB, and LG using the CICIDS 2017 dataset. The proposed Random Forest (RF) model is the most accurate one based on the macro-average F1-score of 97% and outperforms other models in terms of the prediction of different types of attacks. On the other hand, Gaussian Naive Bayes (GNB) Classification frequently classified attacks as normal which is an indication of its inefficiency. The PRC plot of RF was nearly close to the ideal diagonal line signifying that it had high precision as well as recall on most of the classes. These results suggest that the RF model is a powerful tool for intrusion detection and highlights its advantage over other models.

The work of Al Lail et al. (2023) gives the general information regarding machine learning in intrusion detection and discusses the results of Random Forest algorithm. From the confusion matrices and the classification reports, the evaluation was extensive, and it gave everyone the understanding of the limitations as well as the advantages of each model. The precision-recall curve of the RF model is nearly perfect in practice, allowing it to identify various forms of attack. On the other hand, poor performance of GNB implies that it is not well suited to face diversified or multiple

types of threats. This comparison shows the significance of using the right models in machine learning for designing reliable IDSs.

In the other hand Almomani et al. performed a comparative assessment of many Machine Learning (ML) classifiers for Network Intrusion Detection Systems (NIDS) at the 2021 International Conference on Information Technology (ICIT). Finding the best classifier to identify and categorize network breaches in the face of growing security threats was the study's main goal. Classifiers like k-Nearest Neighbors, Decision Tree, Adaptive Boosting, Random Forest, Multilayer Perceptron, Gradient Boosting, Logistic Regression, Multinomial Naive Bayesian, Gaussian Naive Bayesian, Bernoulli Naive Bayesian, and Gradient Boosting were evaluated using the UNSW- NB15 dataset. F-measure metrics were used to ensure accuracy and precision. With an accuracy of 87%, precision of 98%, and F-measure of 84%, RandomForest emerged as the best classifier in the trial,but this work did not use balancing which is important to ensure the reliability of the results.

Also, another study was made In order to improve intrusion detection system performance, a variety of machine learning classifiers were assessed in this work utilizing the KDD intrusion dataset. The random forest classifier displayed the highest average accuracy rate, while the decision table classifier achieved the lowest false negative rate. There were 2% other attacks, 19% regular packets, and 79% DOS attacks in the dataset. Using 148,753 cases, training models were constructed, and 60,000 randomly selected instances were used for testing. The results showed that no single algorithm was able to effectively handle every kind of attack; the random forest classifier, with its 93.77% accuracy rate, showed the highest level of performance. Effective intrusion detection requires taking into account not just true positive and average accuracy rates, but also false negative and false positive rates (Almseidin et al., 2017).

2.6 Integration of Predictive Analysis and IDS

The concept of integrating predictive analysis and IDS originated from boosting both the reliability and security of industrial systems using the combination of predictive analysis and intrusion detection systems (Abdelmoumin, Rawat, & Rahman, 2021). Predictive analysis aims to predict equipment failures to schedule proactive maintenance activities through data-driven techniques. The use of historical and real-time data in predicting potential issues reduces downtime and extends machinery life (Abdelmoumin et al., 2021). On the contrary, IDS is designed to perform the network traffic and system behavior monitoring required to detect and respond to security threats. Integration of the two domains would focus on a holistic system that predicted and prevented mechanical failures while ensuring safety against cyber threats to realize continuous and secure operations.

When predictive analysis and IDS are combined, it leads to many benefits regarding improving resilience. Both technologies, when combined, can guarantee a double-layer defense mechanism wherein predictive analysis ensures the operational integrity of physical assets while IDS protects digital infrastructure (Çınar et al., 2020). This synergy allows the detection of early mechanical anomalies and cyber intrusions, reducing the risks of catastrophic failure and security breaches (Achouch et al., 2022). In addition, the integration would enable better allocation of resources because one can optimize the efforts put into maintenance with set maintenance schedules and cybersecurity measures (Çınar et al., 2020).

However, despite the many advantages, predictive analysis coupled with IDS has to overcome some technical and operational obstacles during integration. One of the major problems is the challenge of managing data; the two systems produce vast amounts of data, which must be derived and analyzed in real time. As a result, data quality and consistency must be ascertained since any mistake would lead to a wrong prediction or an obscure property of the target being missed (Zhu et al., 2020). Also, the issues of interoperability for various systems and technologies whose interface in most cases would demand overwhelming customization and configuration for them to inter-operate seamlessly. Besides, these systems require substantial monetary and

technical investment in the integration process. There is also risk in the system because it may become too complex to manage or troubleshoot (Gupta & Abdelsalam, 2020).

2.7 Research Gaps

The CICIDS 2017 dataset has been extensively used for enhancing intrusion detection systems (IDS) by applying machine learning algorithms such as decision trees, support vector machines, and neural networks to improve detection accuracy and efficiency. However, there is a significant gap in integrating predictive analysis with IDS, particularly for industrial systems. Most research has focused solely on detection rather than predicting system failures that could be preemptively addressed.

Moreover, robust algorithms like Random Forest (RF) and Logistic Regression have not been extensively utilized for this dual purpose with the CICIDS 2017 dataset. These algorithms are known for handling large datasets and providing clear probabilistic classifications, making them suitable for both detection and prediction tasks. This study aims to fill this gap by leveraging these algorithms to develop a predictive analysis framework integrated with IDS, thereby enhancing both detection capabilities and predictive maintenance.

Another critical gap is the imbalance in the dataset, which many studies fail to address, leading to biased model performance and reduced generalizability. Balancing the dataset is crucial for developing robust machine learning models that perform accurately across various types of network traffic and threat scenarios.

Another critical research gap is in the real-time processing and decision-making section. However, further improvements are required to create dynamic predictive analysis and IDS integration with robust real-time operational environments. This is a challenging task toward presenting robust real-time performance of the system that can also keep up with the low latency in both preventing equipment failures and quickly responding to emerging cyber threats. Moreover, research on the human and organizational factors of such integrated systems has been scant. The training of personnel, change management, and ensuring the acceptance of those advanced technologies within an industrial set-up are vital areas that require further exploration.

Chapter 3

Research Methodology

3.1 Introduction

This chapter highlights the process by which this research was conducted to improve reliability of industrial systems through predictive analysis and intrusion detection using machine learning. It explains data acquisition, initial data preparation, data exploration, variable selection, model selection, and model assessment. Thus, leveraging CICIDS 2017 dataset the study employs machine learning approaches to anticipate system weaknesses and identify cyber threats. This methodological approach is meant to be used as the basis for creating efficient PdM and IDS solutions that enhance the security and dependability of industrial processes.

3.2 Data Collection and Preprocessing

3.2.1 Data Collection

The data which was used in this study is the CICIDS 2017 Dataset accessed from Kaggle which consist of sensor data for a period of one week. When joined, the obtained DataFrame contains 2,830,743 rows and 79 columns after the concatenation operation. This dataset is of paramount important in the field of network intrusion detection because of its size and difference of cyber threats. It includes various network traffic situation and attack types making it a rather valuable set of data for training and testing IDS.

The target labels are of different types, which classify different activities in the network between normal traffic and different types of cyber threats. Among the most used tags, “BENIGN” refers to normal network activity, which is not dangerous for the network, and it accounts for 2,096,484 of all cases. This label helps in benchmarking against the possible other malicious activities on the same platform. On the other hand, tags including ‘DoS Hulk’ (172,849) and ‘DDoS’ (128,016) refer to Denial of Service (DoS) typically used to flood a network service with traffic to make it unavailable. The term “PortScan” was recorded 90,819 times and refers to

reconnaissance attempts, which involve probing network ports for weaknesses in anticipation of an assault.

In this context, vulnerabilities refer to recognizing risks or threats that exhibit such patterns or behaviors. For instance, activities labeled as “FTP-Patator,” or “SSH-Patator,” and different types of web attacks, including “Brute Force,” “XSS,” and “SQL Injection” with relative frequency point to attempts to discover vulnerabilities in different protocol and web applications. Sneak attacks and exceptional cases such as “Heartbleed” (11 cases) represent a particular type of attempt to penetrate organizational network protection measures. In this dataset, to define what a vulnerability is, it is necessary to determine which patterns or intentional actions are deviations from benign activities that might threaten the integrity and confidentiality of the network. Recognizing these complex patterns and behaviours within networks is essential for designing and deploying suitable intrusion detection systems. Thus, based on the CICIDS 2017 Dataset, the main idea of this research is to analyze modern machine learning approaches to improve the effectiveness of the recognition and prevention of cyber threats. Due to its large number of attack scenarios, the models that are trained are not only tested against the known sets of threats, but they also can survive the new and constantly changing approaches to cyber-attacks.

3.2.2 Data Cleaning

In order to clean the raw data and transform it into suitable format for analysis, a specific function was coded. Firstly, the function was used to load and merge several CSV files from Google Drive, creating a huge DataFrame with a shape of (2830743, 79). The cleaning process started with the identification and elimination of any cases in which the same data points were repeated in the dataset. Thus, using pandas’ drop_duplicates function, the rows of the dataset were sorted and all duplicates removed so that the final dataset had a shape of (2522362, 79).

After that, tackling of missing values was done on the entire dataset. In the case of numerical columns such as float64 or int64 containing the important features for analysis, missing values were replaced with the median of the feature column. This ensures that numerical features remain statistically sound and do not contain features that are overwhelmingly skewed by outliers. On the other hand, in categorical or

object-type columns, missing values were replaced with the mode which is the most frequent value in the selected column. These steps were significant in ensuring that the dataset was complete and intact which would not be an issue in subsequent analysis.

The quality of data determines the effectiveness and accuracy of any machine learning models that are used in it. In this regard, the comprehensive cleaning procedure conducted on the dataset sought to improve the necessary predictive models for a network intrusion detection system. After handling the first issue of having duplicate rows and the second issue of having some rows with missing values, the cleaned dataset to be used for training/evaluating machine learning algorithms is (2522362, 79). This reduces possibilities of bias resulting from such issues as incomplete or wrong data hence enhancing the credibility of the insights that may be inferred from such network intrusion patterns.

3.3 Exploratory Data Analysis

The initial exploration of the dataset begins by displaying the first few rows and the summary statistics for numerical features. This step serves the purpose to gain first insights into the structure of the data i.e., its range, data type, and the presence of outliers. For instance, the dataset contains 79 features which are both numerical and categorical features, with names like ‘Flow Duration’, ‘Total Fwd Packets’, and ‘Label’. Mean, median, mode, range, variability and measures of skewness provide the information about the location, spread, and the shape of the data distributions. The descriptive statistics including the mean, median, standard deviation and range of each numerical variables are checked to ensure that the data does not have extreme outliers or extreme skewness that may have a negative impact on any eventual analysis and model performance. Understanding these characteristics is essential when choosing proper preprocessing procedures: normalization, scaling, or imputation of missing values.

Switching to the correlation analysis, we begin with the analysis of correlation coefficients between numerical features to identify the relationships and possible multicollinearity. Looking at the correlation matrix, one can get information about linear dependence between variables, which coefficient can be between -1 and 1. The higher the value near to 1 there is a strong positive relationship, and if the value is near

to -1 there is strong negative relationship. For instance, parameters such as ‘Total Fwd Packets’ and ‘Total Backward Packets’ are almost perfectly correlated at 0. 99 which implies that the two parameters are virtually identical. Such correlations aid in feature selection and reduction by identifying the important variables to include in the model, without the inclusion of redundant features that would only cause multicollinearity. Furthermore, this analysis shows the variables with small correlation coefficients or no correlation at all, suggesting those variables may be trimmed or transformed in order to improve the model.

The categorical data analysis aims to determine the distribution of classes belonging to the ‘Label’ column which presents different types of network attacks. This analysis is significant to establish the level of class imbalance and the proportion of each type of attacks. For example, the majority of the entries are labelled with ‘BENIGN’ which amounts to 2,096,484 while attack categories such as ‘Heartbleed’ are rarely seen with a total of 11. Graphing these distributions through bar chart and pie chart will make it easy to note such skewed distributions that will need to be handled using oversampling, undersampling or even setting class weights. Such techniques are helpful in training models that are not inclined towards the majority class, and therefore enhance detection performance for the minority classes. The categorical analysis thus provides the foundation upon which a solid and sustainable model for network intrusion detection could be formulated.

3.4 Feature Selection and Engineering

Feature selection is a fundamental factor in improving the efficiency and interpretability of models in the area of machine learning for network intrusion detection. Choosing the right features enables the model to work on portions of the data that contain the most relevant information and, thus, enhance prediction while avoiding overspecialization. Regarding feature selection for this project, the first step involved was to construct a list of all the possible features that one could think of and available in the dataset. These features included packet statistics, flow characteristics, flag counts and all the elements vital for establishing suspicious behaviors of a network indicative of intrusion.

The approach to selecting features was based on first identifying a large pool of possible features which are prerequisites for intrusion detection as identified by experts in network protocols and common intrusion type. Next, features were selected according to their ability and significance in classification between normal and attack scenarios. Thus, in this evaluation, traditional statistical tools such as correlation analysis were used to eliminate multicollinearity and make the selected features independent. Only those features which showed high degree of discriminative capability between the classes were selected for use, other features which were deemed to be unnecessary or irrelevant were hence eliminated.

Feature engineering also had a significant function in injecting new variables or transformed attributes into the dataset to provide better representation of the patterns within the network traffic. For example, derived features: mean packet size, forward/_backward packets, and moments above 2 of packet size provided additional information about the networks. Use of domain knowledge was instrumental in this step because it helped in selection and transformation of features into separate features defining the nature of network traffic that is associated with different forms of attacks. Such domain knowledge allowed for the engineering of features that are easily interpretable and understandable, making it easier to detect and identify intrusion to computer networks.

The set of features extracted in this manner was useful for constructing the intrusion detection network model and to achieve the best performance by using only the discriminant and informative features of the given data set. Thus, though accelerating the steps of the anomaly identification process, this approach was helpful in defining the specifics of network operation in order to counteract threat activity and introduce safeguards where necessary.

3.5 Dataset Balancing

The final process on feature engineering involved dealing with class imbalance where the dataset was balanced to ensure that machine learning models are trained on all the classes. In this study, the dataset's class of BENIGN comprised the majority, while the classes representing different forms of attacks were in the minority. To strike a balance, a function that randomly selects from the majority class with the size of the combined

minority classes was developed. This reduces the bias towards the majority class making the model more effective in detecting and distinguishing between network intrusions. The balanced dataset obtained reduces chances of bias and offers fair training for the intrusion detection system.

3.6 Model Selection and Evaluation

In choosing models for IDS, a strategic approach involves to consider the complexity a relationship in the data and the simplicity and computational capacity needed for the detection of intrusions. For this project, two different models were chosen: Random Forest and Logistic Regression, followed by the evaluation of their performances due to the nature of network intrusion detection.

1. Random Forest Classifier: This ensemble method builds up several decision trees at the time of training and presents the class which repeats most often or the mean of the individual trees if it is a regression tree (Waskle, Parashar, & Singh, 2020). Random Forest is well suited for high dimensional problems with various interactions with other variables and are less likely to over fit than single decision trees. It is especially useful when the mapping between input features and outcome is non-linear or when there exist complex interactions between the features (Waskle et al., 2020).
2. Logistic Regression Classifier: Logistic regression seeks to estimate the probability of an event that is binary in nature using a logistic function. This is a very basic yet strong model, easy to understand, hence good as default model especially in binary classification problems such as intrusion detection (Al-Haija, Saleh, & Alnabhan, 2021). The fundamental assumption of logistic regression is that there is a linear relationship between the log odds of the outcome and the predictor variables, meaning that the decision boundaries between classes are fairly straightforward and/or linear (Al-Haija, Saleh, & Alnabhan, 2021).

The data set was randomly split into training and testing data sets with the ratio of 80:20 to check the ability of each model to perform on unseen data. Before training the data, the features were normalized through StandardScaler. Normalization standardizes the data into the range of zero mean and unit standard deviation to make

the features equally effective for the model fitting process and not allow features with large scales to dominate the training.

Each model's performance was comprehensively evaluated using key metrics:

1. Accuracy: The ratio of the number of correctly classified instances to the total number of instances (Vujovic, 2021).
2. Classification Report: Precision, recall, F1-score, and support are described for each class, allowing the reader to see the model performance for each class of network intrusions in detail (Vujovic, 2021).
3. Confusion Matrix: Also known as an error matrix, confusion matrix is a graphical representation of the model performance that indicates the number of true positives, true negatives, false positives and false negatives (Vujovic, 2021). The confusion matrix gives information about the kind of mistakes the model is making and is useful in judging the reliability of the model.

For each of the described models these evaluation metrics were computed and plotted in order to provide an accurate comparison with each other in terms of the ability to detect network intrusions. The confusion matrix visualization for each model enabled the determination of how the models were capable of classifying various classes of network traffics.

Thus, the proposed criterion for selecting the models for real world network intrusion detection also evaluates their generalization ability, interpretability, as well as computational efficiency by sorting them based on their accuracy.

3.7 Predictive Analysis

Predictive analysis when used in the context of network intrusion detection entails an attempt to forecast the possibility of different activities on a network being indicative of some form of vulnerability. For this study, two models were employed: Random Forest and Logistic Regression are two such models. The Random Forest model with 100 estimators was used for predicting how likely a point belongs to the “VULNERABILITY” class, and the probability distribution of the points was visualized. The histogram followed by KDE of the predicted probabilities presented

here demonstrated the distribution of the frequency considering the model's confidence level on vulnerability predictions. This visualization proves useful in gaining insights on the model especially on when it predicts high probabilities for the vulnerabilities.

In the same way, the training and usage of the Logistic Regression model was for the prediction of probabilities of the vulnerabilities. The plot of the output probabilities of the logistic model was also created, which had a different pattern as that of Random Forest model. Thus, by comparing these plots, it is possible to estimate the peculiarities of the prediction of each model. The histograms and KDE plots provide information regarding the models, especially their capability of discriminating normal from possibly risky behaviors. This type of comparative analysis helps in choosing the model that is most suitable for practical application in intrusion detection systems, which in addition to the ability to accurately determine the presence of vulnerabilities also provides an understanding of the probability of a given prediction.

Chapter 4

Findings and Discussion

4.1 Introduction

This chapter focuses on the presentation of the findings and discussion on the use of predictive Analysis and IDS with the use of machine learning. The data analysis was based on the CICIDS 2017 dataset which contains various network traffic scenarios and cybersecurity threats. Methods applied were data cleaning and data exploration, feature extraction, and classification using Random Forest and Logistic Regression to predict and identify risks.

4.2 Data Analysis Results

4.2.1 Descriptive Statistics

The first step in analyzing the dataset is to familiarize oneself with its general characteristics. The first several rows display multiple network characteristics including ‘Destination Port,’ ‘Flow Duration,’ and ‘Total Fwd Packets,’ amongst others. Based on this brief overview of the data, one can observe that there are numerous attributes relevant to the analysis of network traffic. For example, the ‘Flow Duration’ values range from very short to long times; the ‘Total Fwd Packets’ and ‘Total Backward Packets’ show different packet numbers associated with different flows in the network. Such variations in feature values are crucial in training reliable machine learning models that differentiate between normal and malicious behaviors.

The summary statistics give out the measures of dispersion and centrality with reference to the numerical characters. For instance, the mean of ‘Flow Duration’ is 16,581,320 with the standard deviation of 35,224,260 to show a high variability in network flow durations. Likewise, ‘Total Fwd Packets’ and ‘Total Length of Fwd Packets’ are highly volatile and have the mean of 10.28 and 611.58, respectively. The fact that minimum and maximum values of such features as ‘Flow Duration’ and ‘Total Length of Bwd Packets’ are spread across orders of magnitude speaks for the presence of outliers. These fluctuations imply that proper data preprocessing must be done to

enhance the accuracy of the model. Such extremes have to be detected and addressed in creating accurate prediction models for intrusion detection.

4.2.2 Correlation Analysis

Using the correlation matrix, essential conclusions were made about the dependencies between numerical characteristics. Interestingly, there is a very strong positive correlation (0.999) between 'Total Fwd Packets' and 'Total Backward Packets,' which demonstrates that these features escalate together, which might point towards bidirectionality of network flows. It also reveals moderate relationships with other attributes such as 'Total Length of Fwd Packets' (0.0635) and 'Idle Mean' (0.764) suggesting that there could be some patterns in network duration concerning packet length and idle time. These correlations are important for feature selection and for engineering since features that are highly correlated may be redundant and can be candidates for removal or transformation.

Categorical data analysis was focused on the distribution of the 'Label' column, which describes network activities. Most of the cases included in the dataset are 'BENIGN' traffic accounting to 2,096,484. But it also encompasses different kinds of attack: 'DoS Hulk' (172,849 cases) and 'DDoS' (128,016 cases), which are important for training and testing the IDS. There are attacks with a low occurrence such as 'Heartbleed' which appeared 11 times challenging the generalization of the model due to class imbalance. To counter this, methods like oversampling or class weighting must be applied to enhance the model's focus on less frequent but impactful threats. This detailed overview of the dataset provides a good starting point for further work on the predictive analysis and evaluation of the model.

4.3 Feature Importance

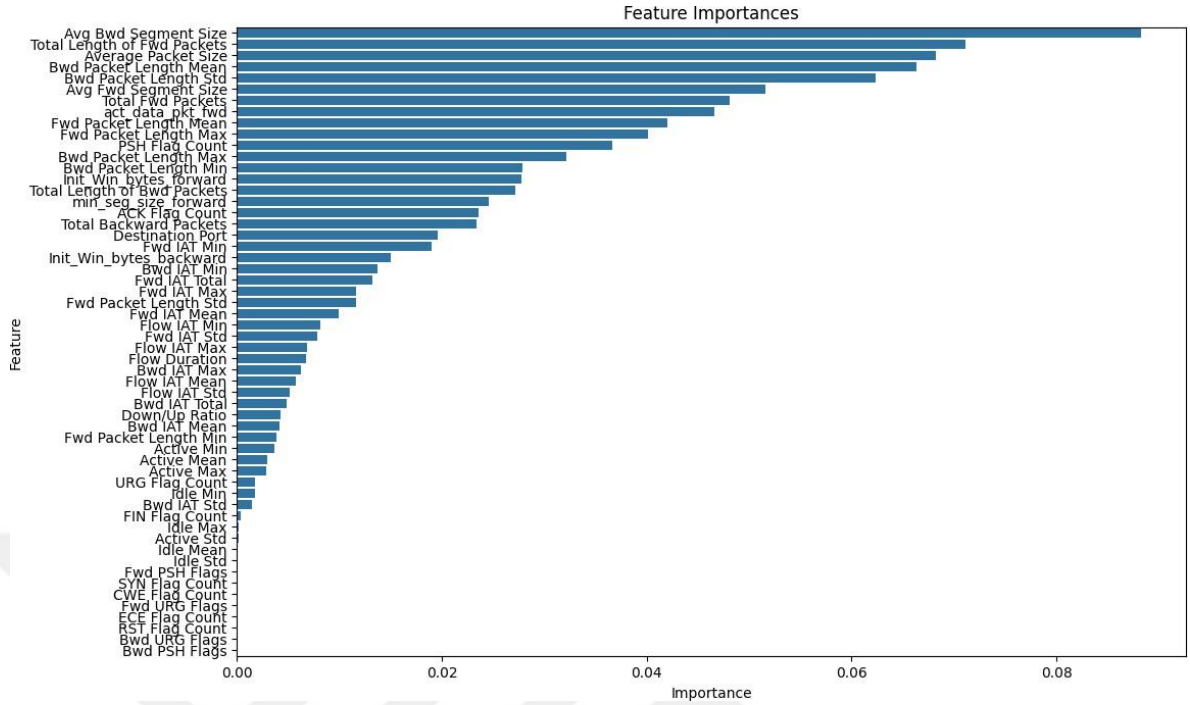

Figure 1. Feature Importances

The feature importance plot for the Random Forest classifier is very helpful in understanding which of the attributes are the most predictive for network intrusions. Of which, the very highest-ranked feature, 'Avg Bwd Segment Size,' shows the importance of the average size of backward segments in indicating possible malicious activity. Some of the other most important features are 'Total Length of Fwd Packets' and 'Average Packet Size,' which illustrate the importance of packet features in classifying normal and malicious traffic. 'Bwd Packet Length Mean' and 'Bwd Packet Length Std' are two features that focus, to a greater extent on backward packet length metrics. This provides a deep understanding of feature importance, which allows more focused feature engineering and selection, ensuring the most informative attributes are brought forward in predictive models to increase model accuracy and efficiency in network intrusion detection.

4.4 Model Performance

4.4.1 Random Forest Classifier

To reduce processing time and resources, the classifier was trained using 600,000 rows of the cleaned data. The developed model has been evaluated on differentiating between benign network traffic instances and various types of vulnerabilities. Comprehensive performance metrics reveal that the Random Forest model performed with an accuracy of 99.96%, proving its effectiveness in distinguishing benign from malicious activities within the network traffic.

```
Comprehensive Report for Random Forest Classifier:
--------------------------------------------------
Accuracy: 0.9996

Classification Report:
              precision    recall  f1-score   support

      BENIGN       1.00      1.00      1.00     85186
VULNERABILITY      1.00      1.00      1.00     34814

    accuracy                           1.00    120000
   macro avg       1.00      1.00      1.00    120000
weighted avg       1.00      1.00      1.00    120000
```

Figure 2. Comprehensive Report for Random Forest Classifier

As the classification report shows, a Random Forest classifier returns nearly perfect precision, recall, and F1-scores for the classes 'BENIGN' and 'VULNERABILITY.' Both precision and recall, 1.00, are very high for the class 'BENIGN,' showing it is the model that has correctly classified almost all instances of benign traffic without wrongly classifying those as vulnerabilities. This is likewise for 'VULNERABILITY' in which precision and recall are 1.00, respectively, indicating that the model identified real positive cases for vulnerabilities within a minimized false-positive ratio. These measures give a collective F1-score for both classes as 1.00, demonstrating that the model is equal and harmonious regarding its precision and recall.

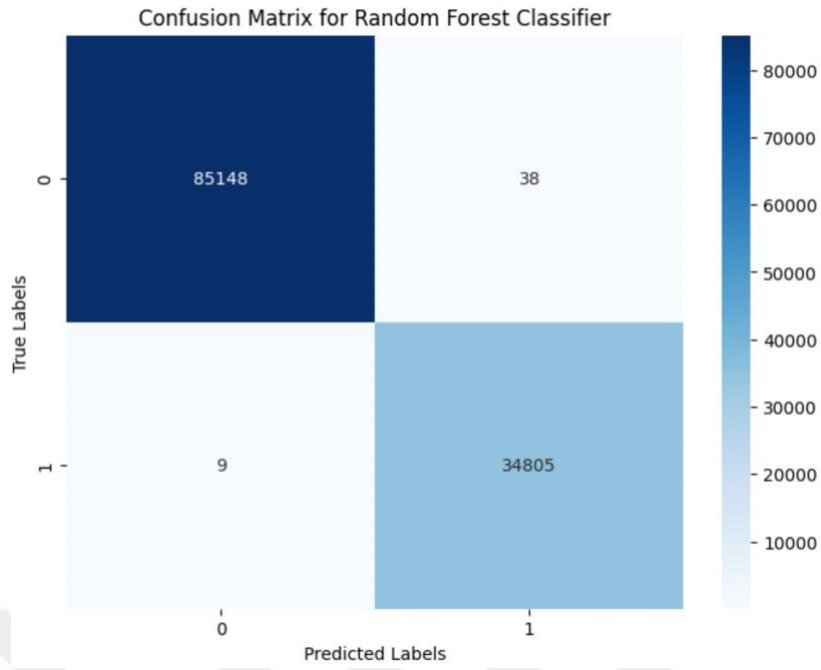

Figure 3. Confusion Matrix for RFC

The confusion matrix provides a visualization with a way to report model performance. This shows that out of 85,186 benign traffic instances, there were only 38 cases of wrong classification as vulnerability. In the case of 34,814 instances of vulnerabilities, only nine have been misclassified into benign traffic. The low number of misclassifications only assures the robustness and reliability of the model for network intrusion detection. The high accuracy and precision of the measurements, integrated with the in-depth confusion matrix, further assert the strength of the Random Forest classifier as a powerful tool for use in network intrusion detection tasks requiring very accurate and reliable predictions.

4.4.2 Logistic Regression Classifier

The Logistic Regression classifier was also trained over 600,000 cleaned datasets using a selection based on the processing time. Performance was evaluated by classifying network traffic as benign or indicative of vulnerabilities. The accuracy obtained for this model was 97.75%, which was pretty high but slightly lower than that of the Random Forest Classifier.

```
Comprehensive Report for Logistic Regression Classifier:
--------------------------------------------------
Accuracy: 0.9775

Classification Report:
              precision    recall  f1-score   support

      BENIGN       0.99      0.98      0.98     85186
VULNERABILITY      0.95      0.98      0.96     34814

    accuracy                           0.98    120000
   macro avg       0.97      0.98      0.97    120000
weighted avg       0.98      0.98      0.98    120000
```

Figure 4. Comprehensive Report for Logistic Regression Classifier

The report gave a detailed insight into the precision and recall of the Logistic Regression classifier concerning the 'BENIGN' and 'VULNERABILITY' classes. For the class 'BENIGN,' we achieved a precision of 0.99, recall of 0.98, and thus f1-score of 0.98. That said, the model is good at identifying benign traffic very well, with very few false positives. The class 'VULNERABILITY' had a precision of 0.95 and a recall of 0.98, giving an F1 score of 0.96. This shows the model's efficiency in recognizing actual vulnerabilities without letting the false negative rate go high. The general accuracy and the high values in precision and recall indicate robust performance across both classes.

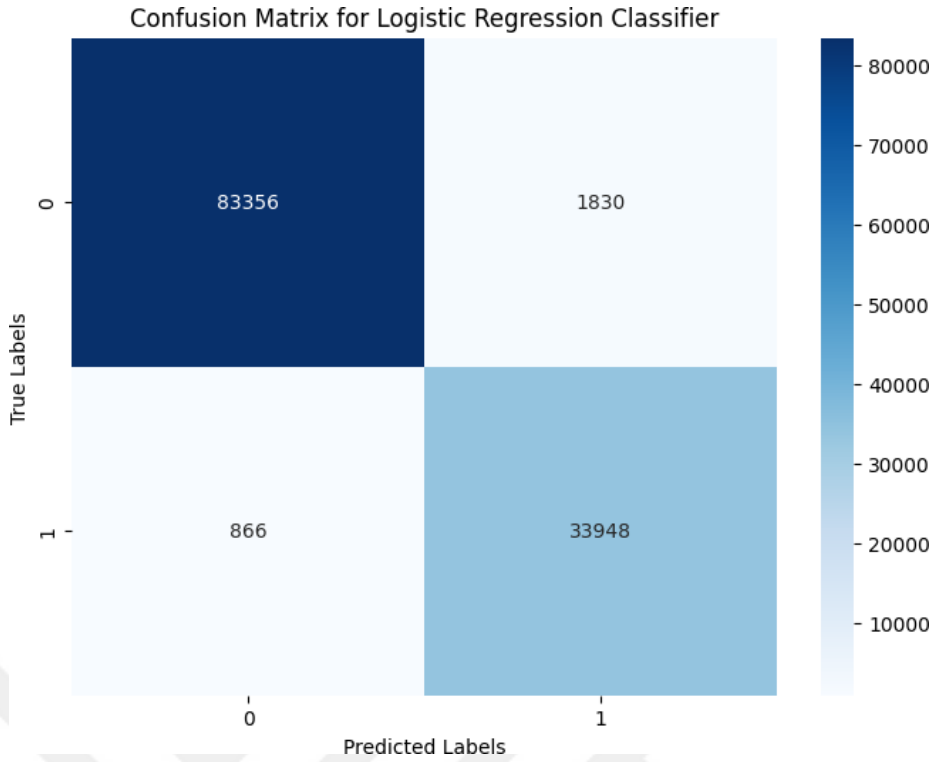

Figure 5. Confusion Matrix for LRC

The confusion matrix of the Logistic Regression classifier shows the number of correct predictions versus incorrect predictions. It misclassified 1,830 benign cases out of 85,186 and 866 vulnerability cases out of 34,814. However, in most of the predictions, it rendered proper identification of network intrusions, which proves it to be a sound model. The model strength in detecting threats is underlined by the lower misclassification rate in the 'VULNERABILITY' class compared to that of the 'BENIGN' class. These results bring out that although Logistic Regression is highly effective, still more optimization and other additional techniques, such as feature engineering or ensemble methods, may be required to bring about the performance levels of more complex models like Random Forest in network intrusion detection tasks.

4.5 Predictive Analysis and Probability Predictions

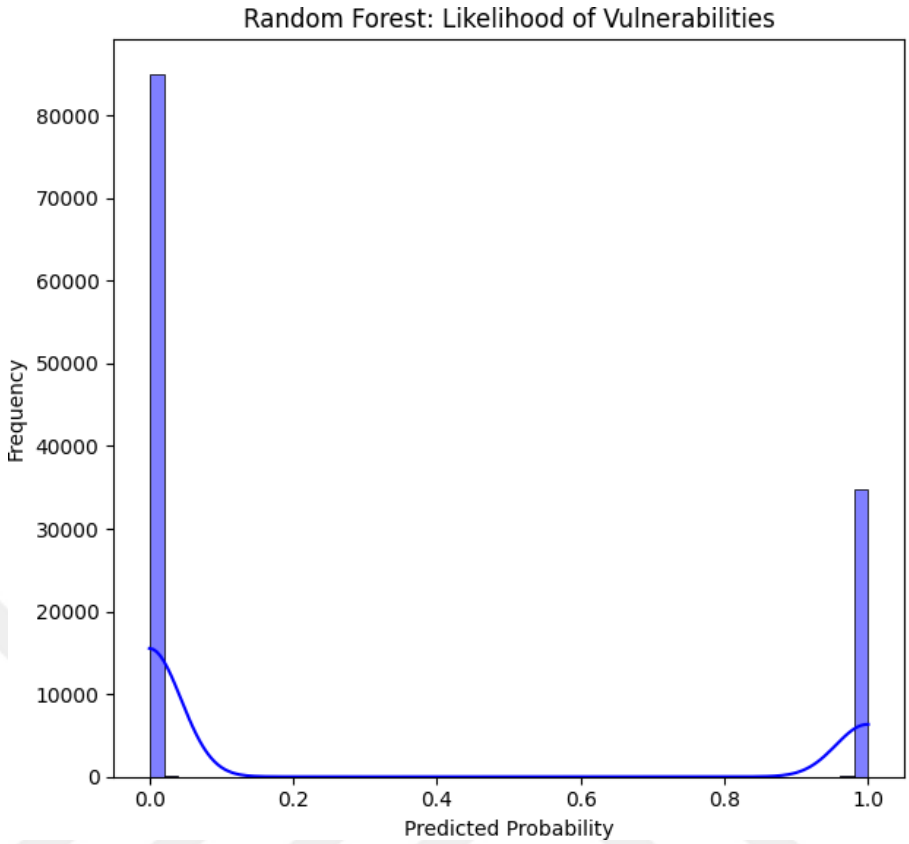

Figure 6. Random Forest: Likelihood of Vulnerabilities

The probability predictions made by the Random Forest model for network vulnerabilities were displayed in a histogram and a KDE plot. The resulting plot was of the bimodal nature, proving that most of the given predictions referred to the probabilities of 0 and 1. This clearly shows that the Random Forest model was quite certain in the predictions made without much confusion of the instance as either benign or vulnerabilities. The spikes at both ends of the probability range indicate that the model has the ability to clearly differentiate between the two classes of data thereby reducing the decision uncertainty.

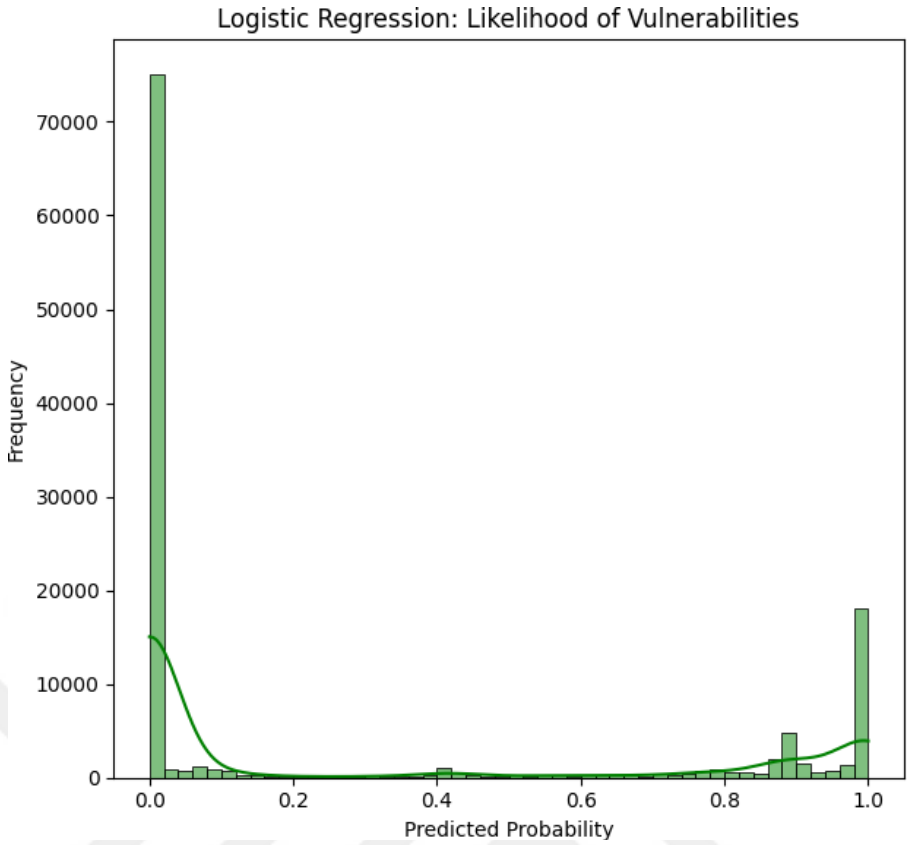

Figure 7. Logistic Regression: Likelihood of Vulnerabilities

Similarly, the probability predictions of Logistic Regression model were represented through histogram and KDE graph. The probabilities were also normally distributed, although extended outside the range of 0-1 as in the case of the Random Forest model above. This implies that whereas the Logistic Regression model was also very confident with its classifications, it was slightly less so as compared to the Random Forest model. The number of instances with an intermediate probability of belonging to class one was higher, meaning that the Logistic Regression model moved away from a decisive outcome more frequently compared to the Random Forest model, likely because of the linear approach of the Logistic Regression model in comparison to the decision boundaries of the Random Forest model.

Both models performed well in terms of accuracy when estimating the risk of network threats. However, the Random Forest model seems to be more confident with its predictions as can be seen from the higher intensity of the peaks at 0 and 1 in the probability density plot. This makes it possible to infer that the Random Forest model

will be more appropriate for situations that demand higher accuracy of classification. The other model used was the Logistic Regression model, albeit less accurate it offered a more diverse range of probability estimates. The following comparative analysis will focus on the advantages and disadvantages of each model and can be useful for choosing the appropriate model when considering the requirement of network intrusion detection tasks.

4.6 Integration of Predictive Analysis and IDS

The use of predictive analysis coupled with intrusion detection systems has been found to be very effective as postulated in this study. To be specific, Random Forest and Logistic Regression models have been shown to be effective in predicting both conditions that may lead to failures in equipment as well as cyber threats. Random Forest classifier with more accuracy and confidence in its predictions emphasizes the aspect of the method's ability to deal with nonlinear interactions in data and its advisability for use in real-time processes in industries. This is made possible through the integration of mechanical health and cybersecurity, to ensure that operations continue apace without interruptions due to mechanical breakdowns or security vulnerabilities.

The advantages of integrating a predictive analysis and IDS include the following in relation to operational efficiency. Firstly, the prediction analysis by using the machine learning enables one to find out the status of an equipment and when it is best to bring it for service. It is an action that is taken with an aim of minimizing the occurrence of unanticipated outages, decrease the costs of maintaining important equipment and enhancing its longevity. Likewise, the IDS component increases the general security of industrial systems by detecting and responding to cyber threats when they arise. The integrated approach helps to maintain the industrial environment both productive and secure and solve the two main problems: availability of the system and security from the cyber threats at the same time.

4.7 Challenges and Limitations

One of the most significant challenges was the effort involved in cleaning and preprocessing datasets to achieve high quality. The CICIDS 2017 dataset, as large as

it may be, also had missing values, duplicate entries, and noisy data, something that needed to be worked on carefully. Addressing these challenges required defining functions to detect and remove duplicate records, and fill in missing values using statistical means, along with ensuring uniformity of data across features. These preprocessing steps were necessary but at the same time tedious and rather cumbersome and again raised the question of data quality assurance. Low quality data malfunctions the system and makes the machine learning algorithms ineffective on prediction of intrusions in the network or the failure of some equipments.

The models chosen for the current analysis, including Random Forest and Logistic Regression, showed satisfactory performance yet had some drawbacks as well. Another problem that was found to be significant was scalability and computational performance due to the amount of data. Training models on millions of rows was resource-intensive and time-consuming and had to be downsized to accommodate day-to-day applications. Such a reduction, as beneficial as it was, might have removed some useful data patterns which might affect the robustness of the model. Additionally, using these models in real-time systems increases the level of challenge. Handling real-time data involves models that should be capable of analyzing data real-time, a condition that would prove hard for Random Forest to achieve due to its resource-consuming nature.

Real-time data processing and the ability to learn continually are more challenges worth mentioning. In dynamic industrial settings, there is always a continuous generation of data and, therefore, a need to develop models that can minimize a requirement of retraining from scratch with every new set of data. This ongoing education is essential to ensuring that the model remains valid and updated as and when needed. However, there are technical challenges that arise from the use of continuous learning mechanisms include how to manage computational load as well as how to ensure that the models are updating nicely on the go without necessarily compromising on their performance. The pressure to integrate updates on a regular basis while trying to preserve rather high levels of prediction accuracy underlines the necessity to create more efficient and adaptive approaches based on machine learning for both integrated PM and IDS applications.

4.8 Best Practices and Recommendations

Real-time integration of predictive analysis and IDS into an effective and efficient predictive maintenance process needs a good planning and design. Some of the best practices include designing for modularity so that the different components of the system, such as the predictive maintenance and IDS modules can be developed and deployed as separate units. This modularity allows for simpler updates and/or maintenance of each of the system components. It is also important to establish effective data feeds, from sensor scans of industrial processes and network traffic to the analytical models. Data cleaning, normalization and feature extraction should be done automatically so that there is always high quality data being fed into the models. When it comes to selecting models to use, it is possible to use the Random Forests because they are resistant to overfitting and can be applied in most datasets or situations due to their versatility, while on the other hand, choosing models like Logistic Regression for interpretability and computational efficiency because they do not require much computational power.

The industrial environment should be taken into consideration when deploying models due to the prevailing needs of the industries. For example, although Random Forest models are highly precise, they could be computationally demanding; therefore, it is possible to use them for batch processing and off-line analysis. Logistic Regression models, being less computationally intensive as compared to other models, can be used in the case of real-time monitoring when a quick decision has to be made. The usage of containerization such as Docker can help in the deployment of these models in different industrial systems at scale. It is also recommended to monitor the performance of the model constantly by evaluating the accuracy, precision, recall, and F1-score to make sure they always are efficient. Incorporating strategies for learning with incremental batches enables models to stay updated with new data trends without necessarily being retrained, enhancing their utility and reliability.

Chapter 5

Conclusion and Recommendations

5.1 Introduction

This chapter provides a summary of major findings, and implications. It provides insights on how to apply the concept of predictive analysis and IDS in industrial environments. Further, it explains the difficulties that were experienced during the study and provides recommendations for further study that can enhance the literature. The purpose is to give the reader a clear view of the opportunities and challenges of using these technologies and propose approaches to increase the dependability and protect the industrial systems using new machine learning solutions.

5.2 Summary of Findings

5.2.1 Key Findings from Data Analysis

Table 1. Key Findings from Data Analysis

Aspect	Findings
Descriptive Statistics	- Dataset included 2,522,362 rows and 79 columns after cleaning. - Key features include 'Flow Duration,' 'Total Fwd Packets,' and 'Total Length of Fwd Packets.'
Correlation Analysis	- High positive correlation (0.999) between 'Total Fwd Packets' and 'Total Backward Packets.' - Moderate correlations of 'Flow Duration' with several features like 'Total Length of Fwd Packets' (0.0635) and 'Idle Mean' (0.764).
Feature Importance	- 'Avg Bwd Segment Size,' 'Total Length of Fwd Packets,' and 'Average Packet Size' identified as top features.

	- These features significantly contribute to model performance.

5.2.2 Model Performance

Table 2. Model Performance Summary

Model	Accuracy	Precision (BENIGN)	Recall (BENIGN)	F1-Score (BENIGN)	Precision (VULNERABILITY)	Recall (VULNERABILITY)	F1-Score (VULNERABILITY)
Random Forest	99.96 %	1.00	1.00	1.00	1.00	1.00	1.00
Logistic Regression	97.75 %	0.99	0.98	0.98	0.95	0.98	0.96

5.2.3 Confusion Matrix

Table 3. Confusion Matrix Summary

Model	True Positives	True Negatives	False Positives	False Negatives
Random Forest	34,805	85,148	38	9
Logistic Regression	33,948	83,356	1,830	866

5.2.4 Probability Predictions

Table 4. Probability Predictions

Model	Insights from Probability Distributions
Random Forest	- Highly confident predictions with probabilities clustering around 0 and 1. - Indicates clear distinction between benign and malicious activities.
Logistic Regression	- Slightly more spread in probability predictions compared to Random Forest. - Shows more intermediate probabilities, suggesting more conservative and varied confidence levels in classifications.

5.3 Implications of the Findings

Practical implications of the findings of this study are important and valuable for industrial systems. Due to the high accuracy and confidence that has been observed from the Random Forest model, the suggested model is suitable for real-time predictive maintenance and intrusion detection. This model, therefore, aids in the reduction of time taken for equipment failure, low costs on maintenance, and the general improvement of productivity in organizations. The combination of PA and IDS provides a multiprotect approach that can attend to simultaneous mechanical and cyber issues with equipment. By applying such models, better maintenance plans, resource management, and cybersecurity measures can be achieved, thus enhancing the overall reliability of industrial facilities.

From a theoretical standpoint, this research enhances the existing literature covering the application of machine learning in predictive analysis and cybersecurity domains. The study shows that using predictive analysis integrated with IDS is suitable

in handling the multifaceted problems of Industrial systems maintenance and protection. It underscores the relevance of feature selection and data preprocessing in improving the model's performance. Comparing the results of Random Forest and Logistic Regression models gives insights into the advantages and drawbacks of different machine learning strategies that can be used to decide which approach will be finest for a set of operations. Moreover, the study highlights the importance of learning and reaction to events in real time in machine learning models, which opens the doors to future developments in adaptive and scalable predictive maintenance and IDS systems. Such theoretical groundwork can be useful for future studies and applications of the machine learning technologies and their implementation into the enhancement of the reliability and security of industrial systems.

5.4 Recommendations

For effective integration of the predictive analysis and IDS, the following recommendations should be made: Start with pilot projects to establish and validate the models in specific scenarios before large-scale application. This makes it possible to assess prospective challenges ahead of time and address them. The use of complex machine learning models such as Random Forest for PM and Logistic Regression for real-time IDS may be optimal in terms of both risk and efficiency. Ensuring availability of high-quality data that feeds from the industrial sensor and different network logs to the analytical models requires reliable data pipelines. By automating data preprocessing step including cleaning, normalization, and feature extraction, data quality will always be maintained. Moreover, such technologies as Docker can help to deploy the application at scale across the different systems. Secondary mechanisms should be put in place constantly checking the models and updating them with the changing data trends without having to train them all over again.

Companies need to coordinate their predictive maintenance and IDS plans with their overall business objectives and security plans. This is done by incorporating them into the general structure of risk management for optimum security and productivity. Policies for data management needs to be established to ensure that data used by predictive models is accurate and safe from external and internal threats. This involves establishing standards in data accessibility, storage and transfer and meeting legal

requirements. Training and development interventions are essential in ensuring that staff is prepared to work with these integrated systems to the best of their abilities. Skills maintenance can be achieved through staff seminars and training that makes the employees aware of the new technologies and methodologies. Also, the culture of encouraging more innovation and improvement will help identify other technologies earlier and adopt them accordingly. Thus, stressing the use of predictive maintenance and IDS integration at all levels of the company organizational structure, businesses improve their operational performance and security status and achieve lasting success.

5.5 Future Research Directions

More research should concentrate on applying predictive maintenance and intrusions detection system optimization in practice. One important and prospective direction is the enhancement of data fusion processes capable of merging various types of information including sensor data, operational logs, and network traffic. Better integration of data will enhance the understanding of system health and security, and thus of more accurate prognosis and discovery. Moreover, optimization of edge computing and federated learning can enhance real time data analysis and the continuous learning process. Edge computing minimizes latency and bandwidth utilization through processing of data near the sources while federated learning makes it possible to train models across decentralized devices and data.

Further advancements in the field of machine learning are also necessary for enhancing the functionalities of predictive maintenance and IDS. It is recommended that further studies focus on the development of the adaptive algorithms that would help optimize the process depending on the current conditions of data input and outputs. Compared to the traditional models, these adaptive models can give more powerful and versatile solutions that are immune to performance decay with changes in conditions. Furthermore, research into more recent techniques in machine learning and artificial intelligence like reinforcement learning and improved neural network structures can provide increased prediction power and speed. There is also a need for more research in developing more interpretable models so that the decision supporting systems produced would be easily comprehensible to industrial personnel. By focusing on these research directions in the future, it will be possible to establish smarter and

more effective industrial systems, as well as apply predictive maintenance and cyber-security on a larger scale.

5.6 Conclusion

This study provides some of the most critical insights about integrating predictive analysis and IDS using advanced machine learning techniques. The models have indicated that Random Forest and Logistic Regression perform well in making predictions regarding equipment failure and cyber threat detection. These results underline that, to attain high-accuracy models, one must seek appropriate data preprocessing and feature selection. The comparative analysis between Random Forest and Logistic Regression would furnish the framework for model selection considering particular operational needs and, therefore, the necessity to intertwine robust predictive maintenance with real-time IDS capabilities.

Real-time integration of the developed predictive Analysis approach within the IDS solution notably enhances the management of such industrial systems. A two-layer approach guarantees continuous operation without interruption and with the optimal level of protection against potential equipment breakdowns or cyber threats. Such integration in real-life scenarios creates space for reduction time, optimization of maintenance and schedules, and enhanced cybersecurity, leading to better industrial processes. These benefits collectively show precisely why these systems need to be adopted in the contemporary industrial setting: one that is challenging and unpredictable.

Further development and research in this area is needed to serve current problems and pursue future possibilities better. Subsequent advances in machine learning, data fusion and real-time processing will improve the prospects for predictive maintenance and IDS. Adopting an adaptable and enlarging model will ensure that the mentioned systems are effective in dynamic industrial environments. Thus, focusing on current research will result into creation of intelligent, efficient, and safe industrial systems. This shall further be beneficial to the broader adoption and evolution of predictive maintenance and cybersecurity. Therefore, to secure and improve the operational reliability of the industrial systems across the world, it is imperative to adapt to these changes in technology with time.

References

Abdelmoumin, G., Rawat, D. B., & Rahman, A. (2021). On the Performance of Machine Learning Models for Anomaly-Based Intelligent Intrusion Detection Systems for the Internet of Things. *IEEE Internet of Things Journal*, *9*(6), 1–1. https://doi.org/10.1109/jiot.2021.3103829

Abdullah, M., & Said, S. (2023). Performance Evaluation of Machine Learning Regression Models for Rainfall Prediction. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-3258529/v1

Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On Predictive Maintenance in Industry 4.0: Overview, Models, and Challenges. *Applied Sciences*, *12*(16), 8081. https://doi.org/10.3390/app12168081

Al Lail, M., Garcia, A., & Olivo, S. (2023). Machine Learning for Network Intrusion Detection—A Comparative Study. *Future Internet*, *15*(7), 243. https://doi.org/10.3390/fi15070243

Al-Haija, Q. A., Saleh, E., & Alnabhan, M. (2021). Detecting Port Scan Attacks Using Logistic Regression. *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*. https://doi.org/10.1109/isaect53699.2021.9668562

Alaswad, S., & Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability Engineering & System Safety*, *157*, 54–63. https://doi.org/10.1016/j.ress.2016.08.009

Aljohani, A. (2023). Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility. *Sustainability*, *15*(20), 15088. mdpi. Retrieved from mdpi.

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *Bio Web of Conferences/BIO Web of Conferences*, *97*, 00133–00133. https://doi.org/10.1051/bioconf/20249700133

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., … Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00444-8

Attaran, M., & Attaran, S. (2019). Opportunities and Challenges of Implementing Predictive Analytics for Competitive Advantage. *Advances in Business Strategy and Competitive Advantage*, *9*(2), 64–90. https://doi.org/10.4018/978-1-5225-5718-0.ch004

Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1124553

Breznická, A., Kohutiar, M., Krbaťa, M., Eckert, M., & Mikuš, P. (2023). Reliability Analysis during the Life Cycle of a Technical System and the Monitoring of Reliability Properties. *Systems*, *11*(12), 556. https://doi.org/10.3390/systems11120556

Çınar, Z. M., Nuhu, A. A., Zeeshan, Q., Korhan, O., Asmael, M., & Safaei, B. (2020). Machine Learning in Predictive Maintenance towards Sustainable

Smart Manufacturing in Industry 4.0. *Sustainability*, *12*(19), 8211. https://doi.org/10.3390/su12198211

Dhall, D., Kaur, R., & Juneja, M. (2019). Machine Learning: A Review of the Algorithms and Its Applications. *Lecture Notes in Electrical Engineering*, *597*, 47–63. https://doi.org/10.1007/978-3-030-29407-6_5

Echkina, E., Levichev, A., & Sushko, A. (2024). Predictive analysis in industry. *Journal of Physics: Conference Series*, *2701*(1), 012109–012109. https://doi.org/10.1088/1742-6596/2701/1/012109

Fox, H., Pillai, A. C., Friedrich, D., Collu, M., Dawood, T., & Johanning, L. (2022). A Review of Predictive and Prescriptive Offshore Wind Farm Operation and Maintenance. *Energies*, *15*(2), 504. https://doi.org/10.3390/en15020504

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, *2*(1), 1–22. https://doi.org/10.1186/s42400-019-0038-7

Kumar, V., & Garg, M. L. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, *182*(1), 31–37. Researchgate. https://doi.org/10.5120/ijca2018917434

Lansky, J., Ali, S., Mohammadi, M., Majeed, M. K., Karim, S. H. T., Rashidi, S., … Rahmani, A. M. (2021). Deep Learning-Based Intrusion Detection Systems: A Systematic Review. *IEEE Access*, *9*, 101574–101599. https://doi.org/10.1109/access.2021.3097247

Lewandowska, N. (2024). Intrusion Detection Systems: Categories, attack detection and response. *IEEE Access*. https://doi.org/10.20944/preprints202402.0008.v1

Martins, I., Resende, J. S., Sousa, P. R., Silva, S., Antunes, L., & Gama, J. (2022). Host-based IDS: A review and open issues of an anomaly detection system in IoT. *Future Generation Computer Systems*, *133*, 95–113. https://doi.org/10.1016/j.future.2022.03.001

Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. *IEEE Access*, *10*, 112392–112415. https://doi.org/10.1109/ACCESS.2022.3216617

Rojek, I., Jasiulewicz-Kaczmarek, M., Piechowski, M., & Mikołajewski, D. (2023). An Artificial Intelligence Approach for Improving Maintenance to Supervise Machine Failures and Support Their Repair. *Applied Sciences*, *13*(8), 4971. https://doi.org/10.3390/app13084971

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). Predictive Coding: Towards a Future of Deep Learning beyond Backpropagation? *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2022/774

Sang, G. M., Xu, L., & de Vrieze, P. (2021). A Predictive Maintenance Model for Flexible Manufacturing in the Context of Industry 4.0. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.663466

Sghir, N., Adadi, A., & Lahmer, M. (2022). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, *28*. https://doi.org/10.1007/s10639-022-11536-0

Shukla, K., Samia , N.-M., & Davis, S. (2022). A heuristic approach on predictive maintenance techniques: Limitations and scope. *Advances in Mechanical*

Engineering, *14*(6), 168781322211010-168781322211010. https://doi.org/10.1177/16878132221101009

Teixeira, H. N., Lopes, I., & Braga, A. C. (2020). Condition-based maintenance implementation: a literature review. *Procedia Manufacturing*, *51*, 228–235. https://doi.org/10.1016/j.promfg.2020.10.033

Thapliyal, V., & Thapliyal, P. (2024). Machine Learning for Cybersecurity: Threat Detection, Prevention, and Response. *Darpan International Research Analysis*, *12*(1), 1–7. https://doi.org/10.36676/dira.v12.i1.01

Trisolino, A. (2023, October 27). Analysis of Security Configuration for IDS/IPS. Retrieved from webthesis.biblio.polito.it website: https://webthesis.biblio.polito.it/29003/

Udo, W., & Muhammad, Y. (2021). Data-Driven Predictive Maintenance of Wind Turbine Based on SCADA Data. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2021.3132684

Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. *Machine Learning*. https://doi.org/10.1007/s10994-021-06042-2

Vujovic, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/ijacsa.2021.0120670

Waskle, S., Parashar, L., & Singh, U. (2020). Intrusion Detection System Using PCA with Random Forest Approach. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. https://doi.org/10.1109/icesc48915.2020.9155656

Yacouby, R., & Axman, D. (2020, November 1). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. https://doi.org/10.18653/v1/2020.eval4nlp-1.9