National College of Ireland

# Enhancing Network Security by Detecting Rogue Access Points using Ensemble Machine Learning Algorithms

MSc Research Project
M.Sc Cybersecurity

Haroon Ali Mohamed Ibrahim Maraicar
Student ID: 22186549

School of Computing
National College of Ireland

Supervisor: Prof. Imran Khan

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | | | |
|---|---|---|---|
| **Student Name:** | Haroon Ali Mohamed Ibrahim Maraicar | | |
| **Student ID:** | 22186549 | | |
| **Programme:** | Msc Cybersecurity | **Year:** | 2023-2024 |
| **Module:** | Msc Research Project (Practicum) | | |
| **Supervisor:** | Prof. Imran Khan | | |
| **Submission Due Date:** | 12-08-2024 | | |
| **Project Title:** | Enhancing Network Security by Detecting Rogue Access Points using Ensemble Machine Learning Algorithms | | |
| **Word Count:** | 8271 | **Page Count:** | 20 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Haroon Ali

**Date:** 11-08-24

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Enhancing Network Security by Detecting Rogue Access Points using Ensemble Machine Learning Algorithms

Haroon Ali Mohamed Ibrahim Maraicar
22186549

**Abstract**

Unauthorised access points (APs) in wireless networks pose substantial threat to security, potentially resulting in data breaches and unauthorised access. Conventional security measures frequently fail to successfully identify these breaches, requiring the use of more sophisticated methods. considering the increasing reliance on wireless networks, especially in contexts where security is crucial, it is essential to create strong detection systems to identify unauthorised access points and protect the integrity of the network. This study utilised ensemble machine learning models, specifically Random Forest, Gradient Boosting, and AdaBoost, along with ANOVA feature selection, to identify rogue Access Points (APs) using the AWID dataset. The application of ensemble approaches, enhanced through the utilisation of Grid Search and cross-validation, greatly enhanced the accuracy of detection. The results indicated that the ensemble models surpassed the traditional models, with Gradient Boosting obtaining the highest level of accuracy. The SMOTE method was utilised to tackle the issue of data imbalance, resulting in improved model performance. However, the evaluation of certain metrics encountered difficulties due to constraints in computational resources. The results of this study add to the existing body of knowledge on network security by showing that ensemble learning approaches are more effective than traditional techniques such as KNN and SVM in identifying rogue access points (APs). The created model provides a reliable tool for network administrators, potentially reducing the likelihood of data breaches. Future research should prioritise the integration of threat intelligence to improve detection capabilities and investigate the system's capacity to recognise certain types of attacks.

**Keywords- AWID dataset, Machine learning classifiers, ANOVA, Access Points**

# 1 Introduction

Wireless networks are integral to the internet. Almost every household now has a wireless connection. Wireless networks are preferred due to their advantages over wired networks, such as mobility and flexibility. Moreover, it is also used in the telecom industry for quick infrastructure growth. With advancements comes disadvantages too. These wireless networks are vulnerable to cyber attacks than the wired connections. Wireless network contains access points. These access points are devices which create the local area network. Through this device other computers and wireless devices connect to the internet(Liu, Barber and DiGrande, 2009). The access point can be connected to wireless networks by connecting into

a hub or switch. Access points from businesses, educational institutions, and similar organisations can be authorised access points. Two types of access points can be established utilising distinct equipment. The first form consists of a wireless router that is linked to the primary network of an organisation. This access point might be regarded as the authorised or legitimate access point (Figure 1). An alternative access point can be established by establishing a connection with the authentic access point(Vaidya, 2023). To set up these access points, a laptop with two wireless cards can be used. One card is connected to the actual access point, while the other card functions as an access point itself .These unauthorised access points can be set up in regions or networks that are commonly visited by many individuals, such as schools, organisations, and public hotspots. Users may inadvertently connect to unauthorised access points, putting them at risk of compromising crucial or sensitive information being transmitted across the wireless network. Attackers can use this access point and receive any type of information provided through a wireless network. Attackers can use these authorised points to create unreliable or unauthorised access points that will allow them to obtain data that is transferred across the network.

**Figure 1: Authorized access point connection**

**Figure 2: Unauthorized access point connection**

These access points or Wi-Fi networks typically incorporate multiple layers of security to protect the data being transmitted over the network. The security levels include Equivalent Privacy (WEP) (Alsahlany, 2014), Media Access Control (MAC) filter (Nixon and Haile, 2017), and Wi-Fi Protected Access (WPA I & WPA II) (Vanhoef and Piessens, 2017). However, these levels are ineffective in safeguarding wireless networks from harmful activities that arise from unauthorised access points within the network (Alsahlany, Alfatlawy and Almusawy, 2019) .Unauthorised or improper access points are responsible for both the leakage of crucial information and the decline in network performance. The data transmitted across the networks may lack additional security measures such as encryption. The attacker can carry out several attacks by setting up unauthorised access points.
During the initial phase of the initiative in October 2019, a comprehensive search of over 100 structures on the Microsoft campus revealed over 1,000 unauthorised access points (APs). Microsoft uses machine learning and advanced techniques to detect rogue APs in their corporate network(Gantenbein, 2023). Approximately 20% of organisations are believed to own unauthorised access points (APs) within their networks, hence exposing the network to various targeted cyber-attacks(Zheng *et al.*, 2014).

Unauthorised access points can result in many types of attacks on a wireless network, leading to significant loss of data and finances for victims, which can include both individual users and major organisations. Due to the inability of security layers in wireless networks to prevent such attacks, it is necessary to develop a technique for detecting unauthorised access

points in a wireless network. Machine learning algorithms have been effectively utilised for many types of intrusion detection in networks, consistently demonstrating high success rates in identifying intrusions throughout the years(Othman *et al.*, 2018; Abdallah, Eleisah and Otoom, 2022). Machine learning techniques will be employed to develop a system capable of determining the authorization status of an access point in a wireless network. The Random Forest, Gradient Boosting classifier, and Adaboost algorithms will be utilised to identify the unauthorised access points. The machine learning approach with the highest performance in this task will be determined.

## 1.1 Research Questions

- How do ensemble machine learning models contribute to more robust intrusion detection systems compared to traditional methods like KNN and SVM in wireless network security?
- How do these models differ in terms of accuracy, precision, computational efficiency, and recall rates for monitoring and protecting wireless network infrastructure?

## 1.2 Objectives

- To preprocess and utilize the AWID dataset for model training.
- To develop an ensemble machine learning model for detecting rogue access points.
- To evaluate the model's effectiveness.
- To enhance wireless network security using the developed model.

## 1.3 Contribution

- **Literature Contribution:**
    - Demonstrates the effectiveness of ensemble models in cybersecurity applications.
    - Provides new insights and comparisons with traditional machine learning algorithms.
    - Expands the use of the AWID dataset in the context of wireless network security research.
    - Bridges the gap between theoretical research and practical implementation in network security.
- **Practical Contribution:**
    - Enhances the detection accuracy of rogue access points, improving overall network security.
    - Offers a scalable and adaptable solution for various wireless network environments.
    - Provides a framework for continuous monitoring and updating, ensuring long-term reliability.
    - Reduces the risk of unauthorized access and data breaches, protecting sensitive information.
    - Supports network administrators with a robust tool for maintaining secure wireless infrastructure.

## 1.4 Structure

The report primarily comprises a 'Introduction' section that provides an overview of unauthorised access points and highlights the detrimental consequences associated with them. By explaining the negative effects of unauthorised access points, the 'Introduction' also establishes the necessity for a system to identify and mitigate such access points. The report includes a 'Literature review' part that examines the usefulness of ensemble machine learning algorithms such As Gradient Boost, Adaboost, and Random Forest in detecting intrusions in wireless networks. After the 'Literature review', the following part is the 'Methodology'. This section outlines the specific methods employed and the approach taken for data collection, as well as the characteristics of the data. The section titled 'Design specification and implementation' follows the 'Methodology' section. This section provides detailed information about the last stage of implementing the unauthorised access point detection system. The section titled 'Evaluation and discussion' will follow the 'Design specification and implementation' section. In this section, the produced outcomes from the developed system will be assessed. This part will also present the primary discoveries uncovered by the unauthorised access point detection system built in this study. The last part of the report is the 'Conclusion and future work' section. This section provides a comprehensive summary of the unauthorised access point detection system that was created in this study. The section also includes information on potential enhancements that can be implemented in the future to enhance the system's performance.

# 2 Related Work

Previous research has investigated the current technologies that detect unauthorised access points in wireless networks. The insights gained from the existing literature on related systems will help in comprehending the outcomes of the system developed in this study. Additionally, the performance of unauthorised access point detection systems will serve as a framework for interpreting the results produced from the developed system.

## 2.1 Utilizing machine learning for network intrusion detection

The identification of unauthorized or inappropriate access points in wireless networks is similar to the detection of intrusions in networks. Intrusion detection systems have been extensively examined in several existing methodologies (Khan *et al.*, 2020).The intrusion detection in (Othman *et al.*, 2018) utilised the Spark-Chi-SVM model. The ChiSqSelector was utilised to choose characteristics from the data, while the SVM classifier, implemented on the Apache Spark Big Data platform, was employed for intrusion detection. The KDD99 dataset was utilised to train and evaluate the model. This approach demonstrates the efficacy of the Spark-Chi-SVM model in effectively detecting intrusions. The primary constraint of this approach is its inability to identify the specific type of intrusion that took place within a network. It just has the capability of determining whether an intrusion has taken place or not within a network. This approach demonstrates the capability of SVM to identify network intrusions. Machine learning models operate by utilising selected characteristics from the data, making feature selection a crucial component in training these models.

## 2.2 Improving intrusion detection using Rough Set Theory and SVM

The authors of the study (Vipin *et al.*, 2010) employed rough set theory (RST) and support vector machines (SVM) to identify network intrusions. The data undergoes preprocessing, followed by feature selection using RST. The chosen features are utilised to train the Support Vector Machine (SVM) to develop a model that can distinguish between normal networks and networks that have experienced intrusions. The performance of the intrusion detection

model is evaluated using two feature selection strategies. It is demonstrated that the choice of feature selection strategy significantly impacts the model's performance. This approach highlights the significance of data pre-processing and feature selection as crucial stages in training a machine learning model.
The efficacy of the Support Vector Machine (SVM) in identifying network intrusions is once again demonstrated in this study, with an accuracy rate of 98.7%. This approach similarly determines whether a network has been compromised or is operating normally, without explicitly identifying the precise type of compromise. However, it may be assumed that this is how intrusion detection approaches typically identify intrusions.
The evaluation of the model's performance in this approach relies on the use of accuracy.

## 2.3 Performance Comparison of Random Forest, J48, Naïve Bayes, and SVM

The detection of intrusions in (Almutairi, Alhazmi and Munshi, 2022) was accomplished using machine learning algorithms such as Random Forest, J48, Naïve Bayes, and SVM. This approach utilises the NSL-KDD dataset to train the dataset. A dataset that includes network features is necessary for building a machine learning model for intrusion detection.
The machine learning models are evaluated by comparing their performances using metrics such as accuracy and precision. Here, intrusions are identified using both a multi-class approach, which identifies the type of intrusion in a network, and a binary approach, which determines whether an intrusion has occurred in a network or not. The random forest classifier has superior performance in detecting intrusions in this strategy, with an accuracy of 98.77% and a precision of 98.8% based on the NSL-KDD dataset. Accuracy and precision are both relevant criteria for measuring the performances of machine learning algorithms.

## 2.4 Comparison of Random Forest and J48

The study carried out by (Farnaaz and Jabbar, 2016) reevaluates the effectiveness of the random forest classifier in identifying network intrusions. The approach uses the NSL-KDD dataset and conducts feature selection and pre-processing before to training the random forest model. The accuracy serves as a statistic for evaluating and comparing the performance of the random forest model with other machine learning models, such as the J48 tree. The random forest algorithm demonstrates a remarkable accuracy rate of 99.67% in accurately detecting various types of intrusions within a network. However, a significant limitation of this strategy is that the proposed model in the network, the random forest, exhibits lesser performance compared to the J48 tree. To enhance its performance, the FSS-Symmetric Uncertainty is applied to the random forest, resulting in observed improvements.
Prior to implementing the FSS-Symmetric Uncertainty, the random forest shown a strong performance that was only marginally inferior to the J48. The FSS-Symmetric Uncertainty was employed as a technique to present a novel approach for enhancing the performance of machine learning classifiers.

## 2.5 Comparative Study of KNN and Naïve Bayes Using the CIDDS-001 Dataset

The study conducted by (T and Badugu, 2021) use the supervised machine learning classifiers KNN and Naïve Bayes for the purpose of intrusion detection. The importance of having a dataset that includes network features has been consistently emphasised in the literature. The NSL-KDD dataset has been utilised in several approaches examined in this study, further confirming the necessity of network features for developing an intrusion

detection model. In this approach, the CIDDS-001 dataset was used, which also includes both network features and labels. The approach includes pre-processing but does not include feature selection. This implies that feature selection is not required for intrusion detection, but the lack of feature selection may have a detrimental impact on the performance of the models employed in this approach. The K-nearest neighbours (KNN) algorithm demonstrated a superior accuracy of 92.3%, surpassing that of the naïve Bayes classifier. The CIDDS-001 dataset employs the following labels: 'normal', 'attacker', 'suspect', 'victim', and 'unknown'. The labels 'unknown' and 'suspect' do not indicate whether the network has experienced an intrusion or not. Training the models with data that includes these labels should not be seen as a disadvantage of the system, since it does not diminish the significance of the intrusion detection results in this approach.

## 2.6 Ensemble Machine Learning for IoT Security

(Abbas *et al.*, 2021) presents a sophisticated intrusion detection system (IDS) specifically developed to improve security in Internet of Things (IoT) settings. The authors address the growing security risks caused by the fast growth of IoT networks by utilising ensemble machine learning techniques, notably Logistic Regression, Naïve Bayes, and Decision Tree classifiers, which are merged through a voting process. The model is trained and assessed using the CICIDS2017 dataset, which consists of genuine network traffic and diverse attack categories. The utilisation of the ensemble approach greatly enhances the accuracy of detection and minimises the occurrence of false alarms, resulting in accuracy rates reaching as high as 99.68%. This study enhances the current pool of knowledge by demonstrating the effectiveness of ensemble learning in Intrusion Detection Systems (IDS) and offers a feasible and adaptable method for addressing real-world security issues in the Internet of Things (IoT).

## 2.7 Effectiveness of ANOVA Feature Selection Using KNN and Decision Tree

In the study conducted by (Pathak and Pathak, 2020), the KNN and decision tree algorithms were employed for the purpose of intrusion detection. This method utilises the ANOVA technique to conduct feature selection on the NSL-KDD dataset. The utilisation of feature selection likely contributed to the enhancement of the performance of the machine learning classifiers. Upon analysis, it is evident that the KNN algorithm demonstrates a lesser level of accuracy compared to the decision tree classifier. However, the precision attained by the KNN algorithm surpasses that of the choice tree. Therefore, it can be inferred that accuracy must also be considered as a significant measure when comparing the performances of machine learning models. The ANOVA feature selection technique is found to be effective in selecting features and likely contributed to the high accuracy and precision values achieved by both the KNN and decision tree models. Overall, these models performed well in identifying intrusions.

## 2.8 Genetic Algorithm Using the KDD Cup 99 Dataset

Genetic Algorithm (GA) is used for identifying rogue connections within a network, as stated by (Suhaimi *et al.*, 2019). This relates to the utilisation of the KDD cup 99 dataset. The Genetic Algorithm (GA) comprises a fitness function, crossover, mutation, and the creation of additional chromosomes. This research demonstrated that the Genetic Algorithm (GA) may be utilised for accurately predicting network breaches. However, the evaluation of the

GA's success in this technique does not rely on measurements such as accuracy and precision. Instead, the effectiveness of the GA was determined by analysing the chromosomes produced by the GA algorithm. However, since the performance of the GA algorithm cannot be evaluated using metrics such as accuracy, which are typically used to assess the performance of machine learning models, it is not possible to directly compare the performance of the GA algorithm with other machine learning models. Therefore, it is challenging to determine whether the GA algorithm is superior to machine learning models in intrusion detection.

## 2.9 Genetic Algorithm for Network Intrusion Detection: A Multi-Study Review

The Genetic Algorithm (GA) is used for the purpose of identifying and detecting attack, as mentioned in the work of (Hoque, Mukit and Bikas, 2012). The effectiveness of the network intrusion model is assessed in this study utilising measures such as accuracy. The findings indicate that the model demonstrates a commendable performance in identifying intrusions. The effectiveness of the Genetic Algorithm (GA) in identifying intrusions was demonstrated in a study conducted by (Hashemi, Muda and Yassin, 2013). The attack detection rate serves as the metric for assessing the model's performance. The genetic algorithm (GA) was effectively employed for the identification of network intrusions in the study conducted by(Chandrakar *et al.*, 2014). This approach also did not utilise indicators such as accuracy to assess the success of the GA algorithm.

## 2.10 Intrusion Detection System for WLANs Using Ensemble Learning Techniques

An Intrusion Detection System (IDS) for Wireless Local Area Networks (WLANs) was put forward to identify unauthorised individuals attempting to get access to wireless networks (Alotaibi and Elleithy, 2015).This approach utilised Random Forests, Extra Trees, Bagging, and a customised majority voting technique for detection. The Bagging classifier demonstrated the highest performance, achieving an accuracy of 96.32%. The classifiers are trained using the AWID dataset, which is a sizable dataset known for its efficacy in training machine learning models for intrusion detection. The primary constraint of this methodology is that the dataset only contains data related to WEP, and another drawback is that it does not consider the potential for attackers to employ new methods to evade detection.

## 2.11 Non-Machine Learning Approaches for Rogue Access Point Identification

The identification of unauthorised access points or rogue access points (RAP) was accomplished through the utilisation of a unique methodology as described in the study conducted by Wu et al. in 2018. This solution utilised the RSS-based practical rogue access point detection (PRAPD). The technique has excellent performance in detecting the RAPs. In this approach, the evaluation of the RAP is conducted using measures such as detection rates. It is important to note that these metrics cannot be directly compared to the performances of machine learning algorithms. Several methodologies that conducted RAP detection without employing machine learning techniques include the studies by Han et al. (2011), Yang, Song, and Gu (2012), and Nakhila et al. (2018). All of these methods detect RAP (Rogue Access Points) without employing machine learning techniques. However, all of these methods rely on network properties to identify the RAPs.

## 2.12 Performance Comparison of KNN, SVM, and Decision Tree in Identifying Unauthorized Access Points with RTT Dataset

The primary objective of this technique has been to identify unauthorised access points, as demonstrated in the study conducted by (Kumar *et al.*, 2021). This technique utilised the KNN, SVM, and decision tree classifiers. The methodology employed the Round-Trip Time (RTT) dataset. After training the machine learning classifiers, it was determined that the decision tree had the highest performance, achieving an accuracy of 99.99%. The KNN and SVM models were likewise observed to be highly successful, with the decision tree model only slightly surpassing them in terms of accuracy. However, the approach's conclusions cannot be regarded conclusive because the RTT dataset used in the study was limited to a tiny sample of data. In the study conducted by (Srinivas *et al.*, 2022), the RTT dataset and machine learning techniques were once again employed to detect unauthorised access points. In this study, the SVM, Multilayer Perceptron (MLP), KNN, and Decision Tree classifiers were employed. The results revealed that the Decision Tree classifier attained the highest accuracy in classification, with a value of 96.56%. The data was imported as a CSV file and performed preprocessing. This approach is further limited using a dataset that contains a very small number of samples.

| Study | Algorithms/Techniques Used | Dataset | Key Findings | Limitations | Relevance to this Research |
|---|---|---|---|---|---|
| Khan et al. (2020), Othman et al. (2018) | Spark-Chi-SVM | KDD99 | Effective in detecting intrusions | Cannot identify specific types of intrusions | Highlights the potential of SVM – based approach |
| Vipin et al. (2010) | Rough Set Theory (RST), SVM | KDD99 CUP | High accuracy (98.7%) in identifying intrusions | Does not identify precise type of compromise | Shows high accuracy obtained by SVM |
| Almutairi, Alhazmi, and Munshi (2022) | Random Forest, J48, Naïve Bayes, SVM | NSL-KDD | Random Forest: accuracy 98.77%, precision 98.8% | Several problems in the dataset | Random Forest shows good performance |
| Farnaaz and Jabbar (2016) | Random Forest, J48 | NSL-KDD | Random Forest: accuracy 99.67% | Limited to dataset; lesser performance compared to J48 | Study suggests exploring ensemble methods |
| T and Badugu (2021) | KNN, Naïve Bayes | CIDDS-001 | KNN: accuracy 92.3% | No feature selection | Highlights need for feature selection to improve KNN accuracy |
| Abbas et al. (2021) | Logistic Regression, Naïve Bayes, Decision | CICIDS2017 | High accuracy | high computational | Gives insights into |

| | | | | | |
|---|---|---|---|---|---|
| | Tree, SGDClassifier, Random Forest, Linear SVM | | (up to 99.68%) | power and resource requirements | tradeoffs between computational power and accuracy |
| Pathak and Pathak (2020) | KNN, Decision Tree | NSL-KDD | Decision Tree higher accuracy; KNN higher precision | Several problems in the dataset | Robustness of Decision trees under varied conditions |
| Suhaimi et al. (2019) | Genetic Algorithm (GA) | KDD Cup 99 | Effective in predicting breaches | Performance not evaluated using accuracy/precision | Non-traditional methods demonstration |
| Hoque, Mukit, and Bikas (2012); Hashemi, Muda, and Yassin (2013); Chandrakar et al. (2014) | Genetic Algorithm (GA) | Various | Effective in identifying intrusions | Does not use accuracy for evaluation | Evaluation beyond traditional metrics |
| Alotaibi and Elleithy (2015) | Random Forests, Extra Trees, Bagging | AWID | Bagging: accuracy 96.32% | Dataset only contains WEP data | Ensemble methods introduction |
| Wu et al. (2018); Han et al. (2011); Yang, Song, and Gu (2012); Nakhila et al. (2018) | RSS-based PRAPD, Network properties | Real time data | Effective in detecting RAPs | Metrics not comparable to machine learning algorithms | Real time detection which can enhance practical application |
| Kumar et al. (2021); Srinivas et al. (2022) | KNN, SVM, Decision Tree, Multilayer Perceptron (MLP) | RTT | Kumar: Decision Tree accuracy 99.99%; Srinivas: Decision Tree accuracy 96.56% | Limited dataset size | Shows effectiveness of Decision trees on small datasets. |

**Table 1: Literature review: summary**

The study aims at assessing the new existing machine learning models of ensemble, on the assessment of the unauthorised access points in wireless networks. there is lack of written work describing the application of machine learning algorithms for this particular use.

Therefore, it can be understood that the procedures used by the programmes of the machine learning algorithm to detect intrusions to the network bear similarity to procedures used in detecting points of unauthorised access. Each of them requires some network description that should comprise various aspects of the network environment. KNN classifier, SVM classifier, Random Forest classifier and SDG classifier together proved the effectiveness of the proposed classifiers in detecting intrusion and unauthorized access points. Before applying machine learning, one must carry out feature engineering and selection for the model. Several strategies that have been deemed helpful include ChiSqSelector, RST, and ANOVA are illustrated to be very helpful in this regard.

Only two methods have been identified where machine learning classifiers are used to detect intrusions. Yes, both methods employed the RTT dataset, and it must be stated that this data set has relatively small sample size. On the other hand, the suggested method will detect unauthorised access points by conducting ensemble machine learning methods including Random Forest, SVM, Gradient Boosting, and AdaBoost concerning the ANOVA method will be used for feature selection. This approach will use the AWID intrusion detection dataset of a large amount of data concerning the network characteristics of wireless networks to ensure the reliable and accurate identification. As the study shows, when using several models in parallel, using efficient feature selection, one can achieve a very high level of accuracy, and hence a significant improvement in the detection of APs that have been compromised.

# 3 Research Methodology

## 3.1 Overall working

Here, I am developing a system that utilises machine learning classifiers to identify and label any unauthorised access points in a wireless network. The machine learning classifiers KNN, SVM, random forest, gradient boosting, and Adaboost, will be trained using the data from the AWID intrusion detection dataset. The AWID intrusion detection dataset is read and pre-processed during training. Following that, the ANOVA approach is used to identify the significant features from the data. The KNN, SVM, Random Forest, Gradient Boosting, and Adaboost classifiers will be trained using these features. The following figure illustrates the proposed approach of this research.

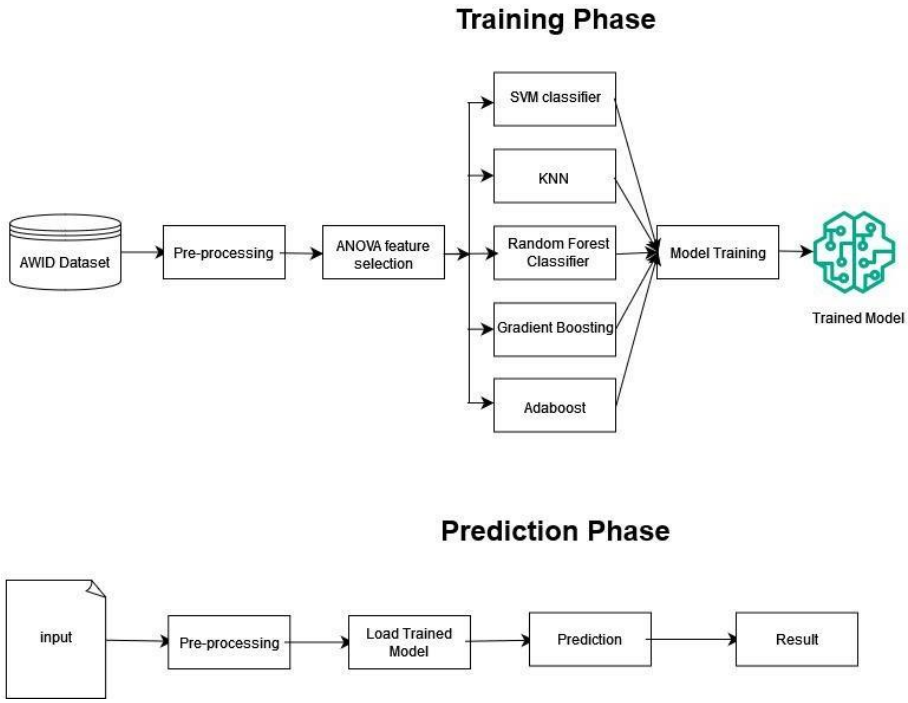


**Figure 3: Proposed Approach**

## 3.2 Data Collection

The data utilised in this research was collected from the AWID intrusion detection dataset. The dataset comprises several attributes of the wireless network.
The information will be stored in a .CSV file. Each row in the .CSV file represents the characteristics of a network or data. The file's columns will contain values corresponding to various attributes related to networks, and one of the columns will contain the labels or classes linked with the networks. The column representing the label will contain textual data, including the values of various types of attacks resulting from unauthorised access points. The labels include ‘Flooding’, ‘Impersonation’, ‘Injection’, ‘Normal’. In this context, the label ‘Normal’ denotes a network that has not experienced any attack, whereas all the other labels indicate networks that have experienced an attack corresponding to the given label. This dataset is chosen due to its substantial volume of data samples.

## 3.3 Data Pre-processing

The pre-processing phase involves the elimination of any unnecessary data within the dataset. Pre-processing enhances the performance of machine learning models and reduces training time. During the pre-processing stage, any columns and rows that have 'null' values more than 50 percent was removed.

Machine learning models rely on numerical labels for training, as they are more effective in improving the performance of the models, like how machines operate. All the string or text values of the labels associated with the data must be replaced with a numerical value. In this case, machine learning models will conduct binary classification. All data associated with attack labels will be assigned a certain numerical value, while data associated with the 'normal' label will be represented by a different numerical value. This is known as One Hot Encoding(OHE). The textual labels ‘Flooding’, ‘Impersonation’, ‘Injection’ will be

substituted with the numerical value 1, whereas the label 'Normal' will be substituted with the numerical value 0. This is how machines process string data to train the model. In this scenario, the labels are assigned values 1 and 0.
Pre-processing also involved normalizing the features of the dataset to make sure the data is consistent and reliable.

## 3.4 Feature Engineering

Following the pre-processing stage, the significant features will be chosen from the remaining data. Feature selection improves training efficiency and enhances the accuracy of machine learning models. The ANOVA approach is utilised to identify the most relevant attributes from the full dataset because ANOVA has shown efficient results in the literature. The 14 most relevant features were selected which had a huge effect on model training. The machine learning classifiers will now utilise these features for training.

## 3.5 Model Development

The key characteristics extracted from the data will be utilised to train the machine learning models. Based on the studies, machine learning is a highly effective method for identifying intrusions in networks. The data will be divided into a training set and a testing set. 20% of the data will be allotted for testing the performance and evaluating the machine learning classifiers, while the remaining 80% will be allocated for training the machine learning classifiers. Prior to training the machine learning models, the data will undergo scaling to ensure that all feature values are within predetermined ranges, as opposed to random ranges, which could potentially impact the performance of the machine learning models. The steps are important for enhancing the reliability of intrusion detection in detecting rogue access points. The models went through cross-validation to prevent overfitting and ensure reliable performance. The hyperparameters were optimised using grid search to fit the models to the specific attributes of network traffic data.
To optimise hyperparameters, techniques such as grid search and random search systematically examine various combinations of parameters, whereas Bayesian optimisation provides a more efficient and probabilistic approach. Model performance validation often includes cross-validation, which divides the data into many folds to assure accurate findings, or holdout validation, which is faster but might be affected by how the data is divided. Performance evaluation in machine learning involves assessing key metrics such as accuracy, precision, recall, and ROC-AUC. To prevent overfitting and ensure the model's ability to generalise to new data, regularisation approaches, early stopping, and data augmentation are used.

### 3.5.1 KNN Classifier

The KNN algorithm is a supervised machine learning classifier that categorises data points by their closeness to a predetermined number of closest neighbours (K), which is given as an input to the classifier. The algorithm discovers the nearest data points by calculating their distance using a certain metric, such as Euclidean distance. It then uses majority voting to select the most frequent class or label among these neighbouring points. In the event of a tie during voting, other approaches, such as opting for the class of the nearest neighbour, might be employed to resolve it. In this case, the value of K is determined to be 7, which is selected based on cross-validation outcomes to achieve a balance between model bias and variance,

thereby providing the best possible performance. The classifier is trained using the training data and then stored for future utilisation.

### 3.5.2 SVM Classifier

The Support Vector Machine (SVM) is a robust technique used in linear classification. It accomplishes this by creating hyperplanes that effectively distinguish diverse groups of data samples. Instances that share the same class are located on one side of the hyperplane, allowing for categorisation based on linear separation. The kernel parameter is essential as it determines the approach used by the SVM to perform the classification task. In this scenario, a linear kernel is selected, which makes the classifier well-suited for data that can be separated linearly. However, Support Vector Machines (SVMs) have the capability to utilise alternative kernel functions, such as polynomial or radial basis function (RBF) kernels, in order to effectively handle intricate and non-linear data distributions. Moreover, hyperparameters such as the regularisation parameter (C) have a substantial impact on balancing the trade-off between minimising error on the training data and preserving a margin for class separation. In this scenario, the kernel parameter is configured as linear, and the Support Vector Machine (SVM) is trained using the training data. Hyperparameters such as C may be adjusted to enhance performance.

### 3.5.3 Random Forest Classifier

A random forest is a technique in ensemble learning that aggregates the predictions of many different decision trees. Each decision tree is created using randomly picked samples from the training data. The technique, referred to as bagging (Bootstrap Aggregating), improves the accuracy of the model and mitigates overfitting by ensuring that each tree is constructed using a slightly different subset of data and features. The random forest combines the predictions of all the individual trees to generate its final projections, usually by using majority voting for classification problems or averaging for regression tasks. In this scenario, the random forest algorithm is trained using the data from the training set, resulting in the formation of 100 decision trees. The selection of 100 trees was determined by parameter tuning and validation procedures to achieve a harmonious trade-off between computational efficiency and model performance.

### 3.5.4 Gradient Boosting Classifier

Gradient Boosting is an ensemble learning technique that builds a series of decision trees in a sequential fashion. Each tree is trained to rectify the mistakes caused by the previous trees. This strategy gradually enhances the overall model performance by specifically targeting the remaining faults from prior rounds. The Gradient Boosting classifier operates by iteratively optimising a loss function, where each subsequent tree minimises this loss, hence improving prediction accuracy. For this scenario, the boosting algorithm uses 100 stages, which corresponds to the number of trees incorporated into the ensemble. Additionally, a learning rate of 0.1 is specified to regulate the impact of each tree on the final model. The parameters of boosting stages and learning rate are of utmost importance. Excessive stages or a high learning rate might cause overfitting, whilst insufficient stages or a low learning rate can lead to underfitting. The values were selected through the process of hyperparameter tuning and validation to achieve a compromise between model complexity and performance. The ultimate forecast is formed by aggregating the contributions of all individual trees to construct the final model.

### 3.5.5 AdaBoost Classifier

AdaBoost, also known as Adaptive Boosting, is a technique in ensemble learning that combines multiple weak classifiers to create a more powerful and precise classifier. The method applies weights to each training sample and adjusts these weights after training each weak classifier. More precisely, it amplifies the importance of incorrectly classified data, guaranteeing that future classifiers prioritise the more difficult situations. This iterative procedure improves the overall model's capacity to rectify faults. In this scenario, a decision tree with a maximum depth of 1, referred to as a "decision stump," is utilised as the fundamental estimator, and the model incorporates 50 boosting stages. The selection of 50 stages was determined by a meticulous tuning procedure, which involved finding a balance between attaining a high level of accuracy and preventing overfitting. The AdaBoost classifier's final prediction is generated by combining the predictions of all weak classifiers using a weighted majority vote. The influence of each classifier on the conclusion is defined by its performance, with better-performing classifiers having a bigger effect. The utilisation of a weighted strategy guarantees that the model consistently enhances its accuracy with the addition of more steps.

### 3.5.6 Evaluation of Models

The machine learning models will be evaluated using important metrics such as accuracy, precision, and recall. Accuracy is a metric that evaluates the overall correctness of a model by determining the ratio of correct predictions (both positive and negative) to the total number of predictions. Precision measures the model's capacity to accurately identify positive instances, which is computed by dividing the number of true positives by the sum of true positives and false positives. Recall, however, quantifies the model's ability to correctly identify any relevant examples. It is calculated by dividing the number of true positives by the total of true positives and false negatives. The performance of different models will be evaluated by comparing these measures to identify the one with the highest performance. The evaluation will be conducted by assessing the model's capacity to achieve a balance between accuracy, precision, and recall. Emphasis will be placed on reducing the occurrence of false positives and false negatives, as these are crucial for preserving network security. The ultimate selection of the optimal model will consider not only the highest individual metric values but also the overall efficacy in identifying unauthorised access locations while minimising errors.

# 4 Design Specification

## 4.1 System Architecture

The system architecture comprises several levels, beginning with the ingestion of data from network traffic records from the AWID dataset, followed by preprocessing and feature extraction to highlight significant network features. The processed data is subsequently fed into the ensemble learning models which comprises of Random Forest, Gradient Boosting, and AdaBoost classifiers. The final stage include the identification and notification mechanism, which quickly notifies administrators of possible unauthorised access points in real-time.

## 4.2 Requirements

The platform utilised for building the system for identifying rogue access points was as follows:

- The **hardware** required for this task is a Windows PC equipped with 24GB of RAM and an Intel i5 CPU. This choice was made due to its optimal combination of processing power and memory capacity. However, it is important to note that there may be restrictions when dealing with exceptionally big datasets or real-time processing demands.
- **Software**: The system was developed using Python as the programming language, renowned for its wide range of libraries and frameworks that are well-suited for machine learning applications. The coding was performed using Jupyter Notebook, accessed through Anaconda Navigator, as well as Google Colab. These platforms offer user-friendly environments for interactive development and collaboration. However, it is important to note that there may be limitations in terms of computational resources and execution time, particularly when using Google Colab's free tier.

# 5 Implementation

## 5.1 Process

The preprocessing phase commences by importing the AWID dataset into a Jupyter Notebook utilising the Pandas library. The dataset is then read from a CSV file and stored as a DataFrame, facilitating easy manipulation and analysis. Missing values are managed by either removing rows and columns containing null values or replacing them with median values, ensuring the dataset maintains its robustness. In order to ensure that no individual feature has a disproportionate impact on the model, it is necessary to scale all features to a given range, thus standardising the data. The ANOVA technique is employed for feature selection, utilising SelectKBest with the f_classif score function to identify the most significant features for the model.

Afterwards, multiple machine learning models such as Random Forest, Gradient Boosting, AdaBoost, SVM, and KNN are established and trained using the preprocessed dataset. The models are optimised using Grid Search with cross-validation to determine the optimal combination of hyperparameters. SMOTE, which stands for Synthetic Minority Over-sampling Technique, is utilised to correct class imbalance problems in a dataset by creating synthetic instances for the minority class. Ultimately, the models are assessed and their performance is graphically represented using Matplotlib, offering valuable insights into their efficacy in identifying unauthorised access points.

## 5.2 Challenges

- Hyperparameter optimisation is accomplished through the use of Grid Search with cross-validation. This method systematically investigates a range of hyperparameters in order to determine the optimal combination for each model, hence ensuring the highest level of performance.
- To tackle the issue of data imbalance, i utilised the oversampling technique SMOTE (Synthetic Minority Over-sampling Technique). This method generates synthetic

samples for the minority class, guaranteeing that the models were trained on a dataset that had equal representation of each class.

- Training the ensemble models on the big dataset posed difficulties in Google Colab due to the platform's inability to handle the substantial computational workload, resulting in frequent disruptions of the session. To address this issue, the training process was transferred from Anaconda Navigator to Jupyter Notebook. This transition allowed for a more reliable environment, enabling the consistent saving of outputs. As a result, interruptions would no longer result in the loss of progress.
- The computational resources were limited and the model's precision, recall could not be evaluated even though the model training was running for 38 hours. The dataset was also broken down into smaller subsets to around 10000 rows to train the model faster. But the precision and recall could still not be printed for all the ensemble models due to limited computational resource.

# 6 Evaluation

## 6.1 Results

The trained and saved machine learning models are evaluated using the data in the testing set to determine the effectiveness of the machine learning classifiers in detecting unauthorised access points in a wireless network. The machine learning classifiers' accuracy, precision and recall will be used as metrics for measuring their performance.

```
KNN Accuracy: 0.89, Precision: 0.79, Recall: 0.89

Random Forest Accuracy: 0.9359661857471383

Gradient Boosting Accuracy: 0.9381707928106171
AdaBoost Accuracy: 0.9121600267225098
```

**Figure 4: Trained Model Output**

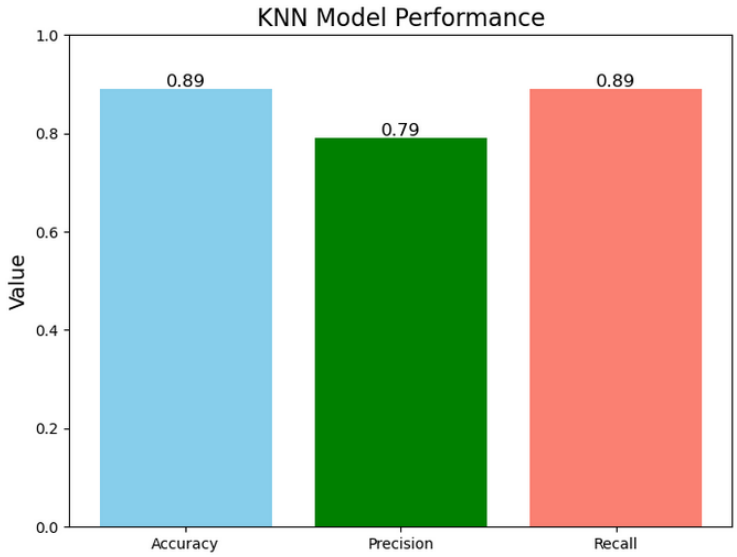


**Figure 5: KNN model performance**

## 6.2 Accuracy

The accuracy is the proportion of unauthorised access point attacks correctly identified by a machine learning classifier, divided by the total number of predictions made by the classifier. The accuracy achieved by five machine learning classifiers is shown in the figure.

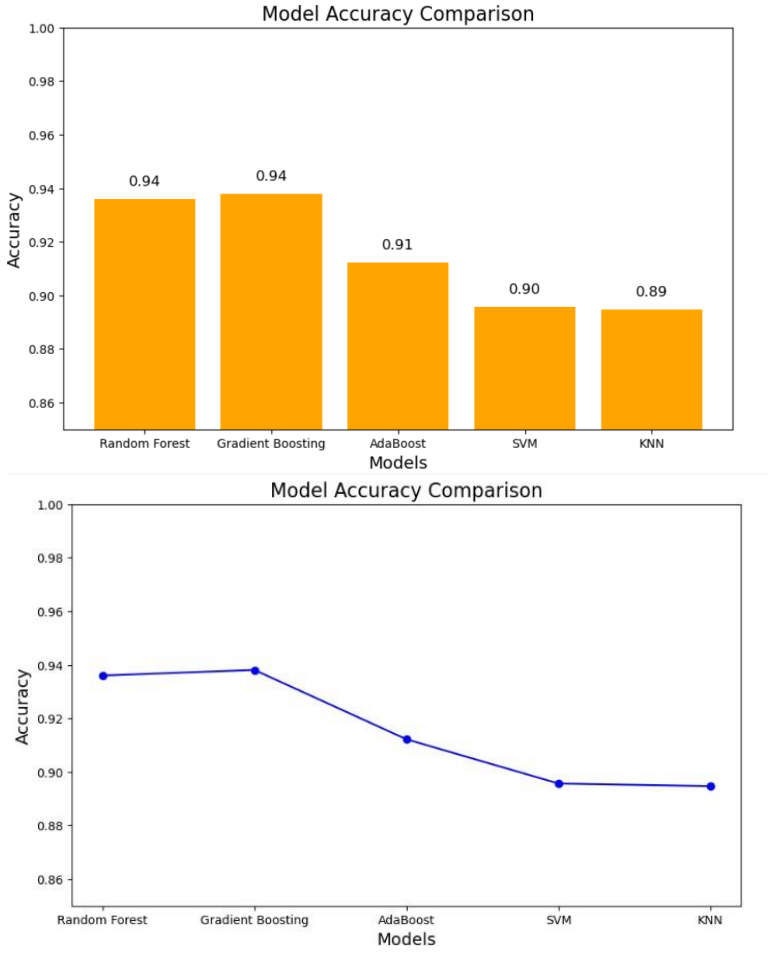


**Figure 6: Accuracy of models**

## 6.3 Precision

Precision refers to the degree of accuracy shown by machine learning models in successfully detecting unauthorised access points in wireless networks. Precision is a metric that quantifies the accuracy of a model in identifying unauthorised access points by measuring the proportion of correctly detected unauthorised access points out of all the access points recognised as unauthorised. This statistic is essential for reducing the occurrence of false alarms, ensuring that when an alert is triggered for a rogue access point, it is extremely probable to be a genuine positive.

## 6.4 Recall

Recall, also referred to as sensitivity or true positive rate, is a performance indicator utilised in the classification process to assess the model's ability to accurately detect all relevant events within a dataset. the term "recall" refers to the measure of the model's ability to correctly detect the proportion of true rogue access points. Recall is vital as it signifies the efficacy of machine learning models in detecting all instances of unauthorised access points.

A high recall value indicates that the model can identify most of the unauthorised access points, therefore minimising the chances of undetected security breaches.

## 6.5 Discussion

The model accuracy were all above 89 percent with Gradient Boosting showing a high performance of 93.8 percent and KNN and SVM having accuracies 89.46 percent and 89.56 percent respectively. The precision and recall could not be obtained for the ensemble models due to the resource limitation. Random sampling was also used to train the models faster but still results could not be obtained with ensemble models. Based on the Accuracy score, ensemble models have shown higher accuracy when compared to traditional models. If there was a higher computational power and more time, the results for all of these models could be obtained. The ensemble model developed shows better results for detecting rogue access points.

## 6.6 Critical Analysis

The Research objective was achieved demonstrating that Ensemble models better detect the Rogue AP's and make the network secure. The Preprocessing of AWID dataset was way complicated than expected because of the number of missing values and constant values. A dataset of smaller data values should have been chosen to make this work easier. Resources to train the model were very limited. So the model training took days to complete and the heat emissions in the laptop was surreal. Future implications of this model suggest using this in Operational Technology sectors to identify the Rogue AP's as soon as possible with no damage to the network infrastructure. Also integrating this model with threat intelligence can yield a 100 percent detection rate.

# 7 Conclusion and Future Work

The research was aimed at finding whether the ensemble machine learning models are better in detecting rogue access points than the traditional models using metrics like Accuracy, Precision and recall. The research found that ensemble models are better in detecting rogue access points with Gradient Boosting classifier showing the highest accuracy. The other metrics like recall and precision could not be obtained due to the high computational power required. Without the feature selection the model training was taking a lot of time even when doing random sampling of only 10000 rows. After selecting the most relevant 14 features using ANOVA, the model was trained quicker. The findings of this rogue access points is crucial for network security especially in today's age where almost all connections are now wireless. With a better detection algorithm, the networks can be safer from attackers thereby improving cybersecurity in general.

Improvements can be made to this system to detect what type of attack is taking place in a network. The system in this research can only detect if the network has an attack or not and cannot detect what type of attack. Also, in the future threat intelligence can be integrated to the ensemble models to have a 100 percent failproof system. By incorporating threat intelligence into ensemble machine learning classifiers, the capacity to identify unauthorised access points can be greatly improved. This is achieved by utilising current and pertinent information, which enhances the accuracy, adaptability, and overall efficacy of the model. This technique provides a more dynamic and informed defence against developing threats in network environments.

# References


Abbas, A. *et al.* (2021) 'A New Ensemble-Based Intrusion Detection System for Internet of Things', *ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING*, 47. Available at: https://doi.org/10.1007/s13369-021-06086-5.

Abdallah, E., Eleisah, W. and Otoom, A. (2022) 'Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey', *Procedia Computer Science*, 201, pp. 205–212. Available at: https://doi.org/10.1016/j.procs.2022.03.029.

Almutairi, Y., Alhazmi, B. and Munshi, A. (2022) 'Network Intrusion Detection Using Machine Learning Techniques', *Advances in Science and Technology Research Journal*, 16, pp. 193–206. Available at: https://doi.org/10.12913/22998624/149934.

Alotaibi, B. and Elleithy, K. (2015) 'An empirical fingerprint framework to detect Rogue Access Points', in *2015 Long Island Systems, Applications and Technology*. *2015 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, Farmingdale, NY, USA: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/LISAT.2015.7160206.

Alsahlany, A. (2014) 'Experimental Analysis of WLAN Security Weakness By Cracking 64 & 128 bit WEP Key', *The islamic college university journal*, 9, pp. 165–176.

Alsahlany, A., Alfatlawy, Z. and Almusawy, A. (2019) 'Experimental Evaluation of Different Penetration Security Levels in Wireless Local Area Network', *Journal of Communications*, 13. Available at: https://doi.org/10.12720/jcm.13.12.723-729.

Chandrakar, O. *et al.* (2014) 'Application of Genetic Algorithm in Intrusion Detection System'.

Farnaaz, N. and Jabbar, M.A. (2016) 'Random Forest Modeling for Network Intrusion Detection System', *Procedia Computer Science*, 89, pp. 213–217. Available at: https://doi.org/10.1016/j.procs.2016.06.047.

Gantenbein, D. (2023) 'Finding and remediating rogue access points on the Microsoft corporate network', *Inside Track Blog*, 11 August. Available at: https://www.microsoft.com/insidetrack/blog/finding-rogue-access-points-on-the-microsoft-corporate-network/ (Accessed: 14 July 2024).

Hashemi, V.M., Muda, Z. and Yassin, W. (2013) 'Improving Intrusion Detection Using Genetic Algorithm', *Information Technology Journal*, 12(11), pp. 2167–2173. Available at: https://doi.org/10.3923/itj.2013.2167.2173.

Hoque, M.S., Mukit, M.A. and Bikas, M.A.N. (2012) 'An Implementation of Intrusion Detection System Using Genetic Algorithm', *International Journal of Network Security & Its Applications*, 4(2), pp. 109–120. Available at: https://doi.org/10.5121/ijnsa.2012.4208.

Khan, A. *et al.* (2020) 'A Survey of the Recent Architectures of Deep Convolutional Neural Networks', *Artificial Intelligence Review*, 53(8), pp. 5455–5516. Available at: https://doi.org/10.1007/s10462-020-09825-6.

Liu, D., Barber, B. and DiGrande, L. (2009) *Cisco CCNA/CCENT exam 640-802, 640-822, 640-816 preparation kit*.

Nixon, J.S. and Haile, Y. (2017) 'Analyzing Vulnerabilities on WLAN Security Protocols and Enhance its Security by using Pseudo Random MAC Address', *International Journal of Emerging Trends & Technology in Computer Science*, 6, pp. 293–300.

Othman, S.M. *et al.* (2018) 'Intrusion detection model using machine learning algorithm on Big Data environment', *Journal of Big Data*, 5(1), p. 34. Available at: https://doi.org/10.1186/s40537-018-0145-4.

Pathak, A. and Pathak, S. (2020) 'Study on Decision Tree and KNN Algorithm for Intrusion Detection System',

*International Journal of Engineering Research and*, V9. Available at: https://doi.org/10.17577/IJERTV9IS050303.

Srinivas, B. *et al.* (2022) 'UNAUTHORIZED ACCESS POINT DETECTION USING MACHINELEARNING ALGORITHMS FOR INFORMATION PROTECTION', *Open Access*, 04(06).

Suhaimi, H. *et al.* (2019) 'Network intrusion detection system by using genetic algorithm', *Indonesian Journal of Electrical Engineering and Computer Science*, 16, p. 1593. Available at: https://doi.org/10.11591/ijeecs.v16.i3.pp1593-1599.

T, R.D. and Badugu, D.S. (2021) 'Network Intrusion Detection System Using KNN and Naive Bayes Classifiers', *Turkish Online Journal of Qualitative Inquiry*, 12(7), pp. 8226–8235.

Vaidya, T.S. (2023) 'Identifying inappropriate access points using machine learning algorithms RandomForest and KNN'.

Vanhoef, M. and Piessens, F. (2017) 'Key Reinstallation Attacks: Forcing Nonce Reuse in WPA2', in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. *CCS '17: 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas Texas USA: ACM, pp. 1313–1328. Available at: https://doi.org/10.1145/3133956.3134027.

Vipin, D. *et al.* (2010) 'Network Intrusion Detection System Based On Machine Learning Algorithms', *International Journal of Computer Science & Information Technology*, 2. Available at: https://doi.org/10.5121/ijcsit.2010.2613.

Zheng, X. *et al.* (2014) 'Accurate rogue access point localization leveraging fine-grained channel information', in *2014 IEEE Conference on Communications and Network Security*. *2014 IEEE Conference on Communications and Network Security (CNS)*, San Francisco, CA, USA: IEEE, pp. 211–219. Available at: https://doi.org/10.1109/CNS.2014.6997488.