

Identifying inappropriate access points using machine learning algorithms RandomForest and KNN

MSc Research Project
Cyber Security

Tushar Sanjaykumar Vaidya
Student ID: X20254083

School of Computing
National College of Ireland

Supervisor: Prof. Imran Khan

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name: Tushar Sanjaykumar Vaidya

 X20254083
Student ID:
Programme: MSc Cyber Security **Year:** 2022 - 2023
 MSc Research Project
Module:
 Imran Khan
Supervisor:
Submission Due Date: 01-02-23
Project Title: Identifying inappropriate access points using machine learning
 algorithms RandomForest and KNN
 6699
Word Count: **Page Count:**22.....

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: Tushar Sanjaykumar Vaidya

 01/02/23
Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
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.	<input type="checkbox"/>

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):	

Identifying inappropriate access points using machine learning algorithms RandomForest and KNN

Tushar Sanjaykumar Vaidya

X20254083

MSc in Cybersecurity

National College of Ireland

Abstract

A wireless network is being used by a wide range of users in today's world from large organizations to schools to government offices adding to this WiFi hotspots are available at many public places which are used by thousands of people. The wireless network of an organization may facilitate the transfer of highly important and sensitive information and even normal person using a wireless network may transfer their sensitive information through the network. The transfer of large amounts of data has attracted the attention of people who might be looking to steal or leak the sensitive information that is being transferred through a wireless network and these people are able to do this by configuring access points that are inappropriate. People may unknowingly connect to these inappropriate access points and may result in the attacker stealing valuable information which may lead to huge financial and personal losses. So, it is important to identify the presence of unauthorized access points in a wireless network and a system is proposed here to do so. The system proposed here will use machine learning classifiers for identifying the presence of unauthorized access points in a wireless network. The system will be developed based on the Aegean WiFi Intrusion Dataset (AWID) intrusion detection dataset. The data will be pre-processed and the important features from the data will be selected using the Analysis of variance (ANOVA) technique. These features will be used for training the machine learning classifiers K-nearest neighbor (KNN), Random Forest, Support Vector Machine (SVM), Genetic Algorithm (GA). The accuracy and precision of the machine learning classifiers will be found out for evaluating the effectiveness of the classifiers and the classifier with best performance will be used for creating desktop application that is able to identify the networks with inappropriate access points based on the network features provided as input to it. It found as the result of the approach that the machine learning classifiers are able to effectively identify wireless networks with inappropriate access points.

Keywords: WiFi hotspots, leak the sensitive information, Aegean WiFi, Analysis of variance technique, Random Forest algorithm, Accuracy.

1 Introduction

These days, wireless networks are commonly used. They are favored. compared to wired networks because they provide mobility, flexibility, and quick infrastructural growth in the telecommunications sector. But these wireless networks are more prone to attacks than wired networks. An access point in a wireless network is a device that creates a local area network

and allows other devices like computers to connect to the wireless network (Liu, Barber and DiGrande, 2009). The access point device is able to connect to wireless networks by plugging into a hub or a switch using a wire. This access point can be misused as it is also accessible by people with malicious intentions and such a person can obtain information of any kind sent through a wireless network. The access points of companies, schools etc can be considered authorized access points and these authorized access points can be used by the attackers for creating unauthorized or fake access points through which these attackers can obtain data transferred through the network. Two kinds of access points can be created using different equipment. The first type contains a wireless router connected to the main network of an organization. This access point can be considered the authorized or real access point (Figure (1)) . The second type of access point can be created by connecting to the real access point. These access points can be setup by using a device like a laptop having two wireless cards where one is connected to the real access point and the other acts as an access point (Figure (2)) (Tahaseen, 2019). These unauthorized access points can be configured in areas or networks which are tend to be accessed by a large number of people like schools, organizations and public hotspots. The users will unknowingly connect to the unauthorized access points and will be at risk of losing important or sensitive information that they are transferring through the wireless network.



Figure 1: Device connected to an authorized access point (Tahaseen, 2019)

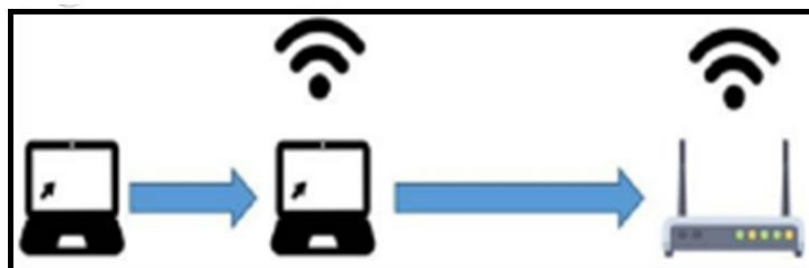


Figure 2: Device connected to an unauthorized access point (Tahaseen, 2019)

These access points or Wifi networks usually contains a number of levels for securing the data being transferred across the network. These security levels consists of Equivalent Privacy (WEP) (Ali Alsahlany, 2014), Media Access Control (MAC) filter (Sebastian Nixon and Yibrah Haile , 2017) and Wi-Fi Protected Access (WPA I & WPA II) (Vanhoef and Piessens, 2017). But these levels are unable to protect the wireless networks from the malicious activities occurring through unauthorized access points in the network (Alsahlany, Almusawy, and Alhassan, 2018). These unauthorized or inappropriate access points are not only responsible for the leak of important information but also for the degradation in the performance of the network (Calhoun et al., 2008). The data flowing through the networks

may also not have any added security like encryption. By configuring unauthorized access points the attacker is able to perform attacks like De-authentication, coffee latte and evil twin etc.

The existence of unauthorized access points will lead to different kinds of attacks on a wireless network resulting in huge data and financial loss as the victims of the attack can range from normal users to large organizations. Since the security layers in the wireless networks are unable to prevent these attacks a mechanism for detecting the unauthorized access points in a wireless network has to be devised. Such a system is proposed to be developed here. The machine learning algorithms have been used to perform different kinds of intrusion detection in networks and these algorithms have been highly successful detecting intrusions in networks over the years (Othman et al., 2018 ; Abdallah, Eleisah and Otoom, 2022). So machine learning techniques will be used for creating a system that is able to identify if an access point in a wireless network is authorized or not. The system will be developed in the form a desktop application that is able to identify the unauthorized access points in a network. The random forest , KNN , SVM and genetic algorithms will be used here and the machine learning algorithm with the best performance in identifying the unauthorized access points have to be found out . The best machine learning model will be the one that is used in the desktop application.

The report mainly consists of the ‘Introduction’ which gives an idea about unauthorized access points and the negative effects of such access points and on revealing the negative effects of the unauthorized access points the need for a system to identify these access points is also justified in the ‘Introduction’. Followed by the ‘Introduction’ section the report contains the ‘Literature review’ section where the effectiveness of the machine learning methods like KNN, SVM , random forest and genetic algorithm in detecting intrusions in wireless networks are studied in a critical manner. Followed by the ‘Literature review’ the ‘Methodology’ section comes up , this section defines the methods used here and the way in which data was collected in this approach and also the nature of the data. The ‘Methodology’ section is followed by the ‘Design specification and implementation ‘ section , this section describes the details about the final stage of implementation of the unauthorized access point detection system. The ‘Design specification and implementation’ section will be followed by the ‘Evaluation and discussion’ section , in this section the results obtained from the system developed will be evaluated. This section will also contain the main findings revealed from the unauthorized access point detection system developed here. The final section of the report is the ‘Conclusion and future enhancements’ section. This section contains round up of the details of the unauthorized access point detection system developed here. The section also contains the details about the advancements that can be made to the system in the future for improving the performance of the system.

1.1 Research questions

- Will the machine learning models , KNN, random forest, SVM and genetic algorithm, be effective in identifying unauthorized access points in wireless networks?
- How will the performance of the machine learning models in identifying unauthorized access points in wireless networks be evaluated ?

1.2 Aims of the research

- Create or identify a dataset that contains the data required for identifying unauthorized access points in wireless networks.
- Create and train the machine learning algorithms using the data in the dataset.
- Evaluate the performances of the machine learning models in identifying unauthorized access points in wireless networks and find out the model with the best performance.
- Create a desktop application that is able to identify the unauthorized access points in wireless networks.

2 Literature review

The existing systems that perform identification of unauthorized access points in wireless networks have been studied. The findings from the literature associated with the existing systems will help in understanding the results of the system developed here and the performances of the unauthorized access point detection systems will provide a context to the results obtained from the system developed here.

2.1 Network intrusion detection using machine learning

The detection of unauthorized or inappropriate access points in wireless is similar to the detection of intrusions in networks and the intrusion detection systems have been seen studied in several existing approaches (Khan et al., 2020). The Spark-Chi-SVM model was used for intrusion detection in (Othman et al., 2018). The features from the data were selected using the ChiSqSelector and the SVM classifier based on Apache Spark Big Data platform was used for detecting intrusions. The KDD99 dataset was used for training and testing the model. It is revealed in this approach that the Spark-Chi-SVM model is able to detect intrusions effectively. The main limitation of this approach is that it not able to detect the type of intrusion that has occurred in a network and is only able to detect if an intrusion has occurred in a network or not. It can be seen in this approach that the SVM is able to detect intrusions in networks. The machine learning model works based on features selected from the data so feature selection can be considered as an important step for training the machine learning models.

Rough set theory (RST) and SVM was used for detecting network intrusions in (Das et al., 2010). The data is preprocessed and here feature selection is performed using RST. The selected features are used for training the SVM so that a model could be created for identifying normal networks and networks in which intrusions have occurred .The performance of the model based on two feature selection techniques is observed and it is shown in this approach that the performance of the intrusion detection model is affected by the feature selection technique used . It can be seen from this approach that the data pre-processing and feature selection are important steps in training a machine learning model. The effectiveness of the SVM in detecting intrusions in networks is again displayed in this approach as it achieves an accuracy of 98.7%. This approach also detects if a network has intrusions or is normal and it does not specifically identify the type of intrusion but this can also be seen in a way that this is how intrusion detection approaches detect intrusions. Accuracy is used in this approach as a way to evaluate the performance of the model.

Machine learning techniques like Random Forest , J48 , Naïve Bayes and SVM were used for detecting intrusions in (Almutairi, Alhazmi and Munshi, 2022). This approach uses the NSL-KDD dataset for training the dataset. So a dataset containing network features can be considered as a requirement for training a machine learning model for intrusion detection. The performances of the machine learning models are compared using metrics like accuracy and precision. Here the intrusions are detected in both a multiclass, identifying the type of intrusion in a network, and binary , identifying if an intrusion has occurred in a network or not, manner. The random forest classifier achieves the best performance in this approach in detecting intrusions as it achieves an accuracy of 98.77% and a precision of 98.8% based on the NSL-KDD dataset . It can be seen that accuracy and precision are also two valid metrics that can be used for comparing the performances of machine learning algorithms.

The performance of the random forest classifier in detecting network intrusions is again studied in (Farnaaz and Jabbar, 2016). The approach uses the NSL-KDD dataset and the performs feature selection and pre-processing before training the random forest model. The accuracy is used as a metric for comparing the performance of the random forest with other machine learning models like J48 tree. It is seen that the random forest achieves an accuracy of 99.67% in identifying different kinds of intrusions in a network. But a main drawback that can be seen in this approach is that model proposed in the network, the random forest, has a lower performance than the J48 tree and the FSS-Symmetric Uncertainty is applied to the random forest for improving it's performance and it's performance was observed to improve. Even before applying the FSS-Symmetric Uncertainty the random forest exhibited a solid performance and was only slightly lower than the J48 and the FSS-Symmetric Uncertainty was used as a means to introduce a new method for improving the performance of machine learning classifiers.

The supervised machine learning classifiers KNN and Naïve bayes are used for intrusion detection in (T and Badugu, 2021). The necessity of a dataset containing network features was seen through all the literature studied till now the NSL-KDD dataset was used in a couple of approaches studied here and it can be again confirmed from this approach that the network features are required for creating an intrusion detection model as the dataset used in this approach is the CIDDS-001 dataset and this dataset also contains the network features and labels . The approach performs pre-processing but does not perform feature selection , this means that the feature selection is not mandatory for intrusion detection but the absence of feature selection may negatively affect the performances of the models used in this approach. The KNN was observed to achieve an accuracy of 92.3% which was way better than the naïve bayes. The labels used in the CIDDS-001 dataset are 'normal' , 'attacker' , 'suspicious' , 'victim' and 'unknown' . The labels like 'unknown' and 'suspicious' do not specify if the network has an intrusion or not and training the models with data having these kinds of labels cannot be considered a drawback of the system that may make the results of the intrusion detection in this approach less significant.

The KNN and decision tree was used for performing intrusion detection in (Ashwini Pathak and Sakshi Pathak, 2020). This approach performs feature selection using the ANOVA technique on the NSL-KDD dataset. The feature selection may have helped in improving the performance of the machine learning classifiers. This approach reveals that the KNN exhibits a lower accuracy than the decision tree classifier but the precision achieved by the KNN is greater than the decision tree. So it can be implied that precision has to be also considered as a relevant metric for comparing the performances of machine learning models. The feature selection technique ANOVA is observed to perform feature selection effectively and may have played a part in the performance of the models as both the KNN and decision tree were

able to high accuracy and precision values and overall exhibits a good performance in detecting intrusions.

GA is used for the detection of harmful connections in a network in (Suhaimi et al., 2019). This approaches the KDD cup 99 dataset . The GA consists of a fitness function , crossover and mutation and the generation of new chromosomes. It was found in this approach that the GA can be used for predicting network intrusions. But the performance in the GA in this approach is not measured using metrics like accuracy and precision and it was concluded that GA was effective by studying the chromosomes generated by the GA algorithm. But as the performance of the GA algorithm can't be measured with metrics like accuracy usually used for evaluating the performance of machine learning models the performance of the GA algorithm cannot be compared directly with other machine learning models so it will be difficult to determine if the GA algorithm is better in intrusion detection than machine learning models.

The GA is used for detecting intrusions in (Sazzadul Hoque, 2012). The performance of the network intrusion model is measured here using the metrics like accuracy and it was revealed in this approach that the network intrusion model is able to achieve a good performance in detecting intrusions. The GA was found to be highly effective in detecting intrusions in (Hashemi, Muda and Yassin, 2013). The detection rate of attack is used as the parameter for evaluating the performance of the model. The GA was again successfully used for detecting network intrusions in (Chandrakar et al., 2014). This approach also did not use metrics likes accuracy for evaluating the performance of the GA algorithm.

A Wireless Local Area Network Intrusion Detection System (WIDS) was proposed to discover attackers in wireless networks (Alotaibi and Elleithy, 2016). The Random Forests , Extra Trees , Bagging and a custom majority voting technique was used in this approach for detection. The best performance was exhibited by the Bagging classifier with an accuracy of 96.32%. The classifiers are trained using the AWID dataset ,the dataset is large and observed to be highly effective in training machine learning models for intrusion detection. The main limitation of the approach is that the data in the dataset is WEP specific and another limitation is that the possibility of an attacker using novel methods for bypassing detection is not considered here.

2.2 Unauthorized access point detection

The unauthorized access points or rouge access points(RAP) were identified using a novel method in (Wu et al., 2018). The RSS-based practical rogue access point detection (PRAPD) was used in this approach. The approach is able to achieve a good performance in detecting the RAPs . The performance of the RAP is evaluated in this approach using metrics like detection rates which cannot be directly compared to the performances of machine learning algorithms . Several approaches that performed RAP detection without using machine learning algorithms (Han et al., 2011; Yang, Song and Gu, 2012; Nakhila et al., 2018). All of these approaches identify RAP without using machine learning techniques but all of these approaches also use features of the network for identifying the RAPs.

Identifying unauthorized access points has been the focus of the approach here and this was exactly seen in (Ravula Arun Kumar et al, 2021). The KNN, SVM and decision tree classifiers where used in this approach . The approach used the RTT (Round Trip Time) dataset . The machine learning classifiers were trained and it was observed that the decision

tree exhibited the best performance with an accuracy of 99.99%. The KNN and SVM was also found to be very effective as the decision tree had only a slightly better accuracy in comparison with the two models. But the RTT dataset used in the approach contained only a small sample of data and due to this the results produced in the approach cannot be considered conclusive.

The RTT dataset and machine learning algorithms were again used for identifying unauthorized access points in (Srinivas et al., 2022). The SVM, Multilayer Perceptron (MLP) , KNN and Decision Tree classifiers were used here and it was found that Decision Tree achieved the best accuracy in classification with a value of 96.56%. The data was read in the form of a CSV file and preprocessed . This approach also has the limitation of using a dataset that has a relatively small amount of samples.

2.3 Summary

Research paper	Approach	Dataset	Algorithms	Limitations	Upsides
Othman et al., 2018	Spark-Chi-SVM model was used for intrusion detection.	KDD99 dataset	SVM	Type intrusions are not detected.	The use of the Spark-Chi-SVM model makes the detection effective.
Das et al., 2010	RST and SVM was used for detecting network intrusions	KDD99 CUP dataset.	SVM	Type intrusions are not detected.	Achieves very high accuracy during detection.
Almutairi, Alhazmi and Munshi, 2022	Machine learning for intrusion detection	NSL-KDD	Random Forest , J48 , Naïve Bayes and SVM	The NSL-KDD dataset has several problems.	Very high accuracy and precision values are achieved.
Farnaaz and Jabbar, 2016	Machine learning for intrusion detection	NSL-KDD	random forest classifier	the random forest, has a lower performance.	The FSS-Symmetric Uncertainty is applied for improving the performance of the RF.
T and Badugu, 2021	Machine learning for intrusion detection	NSL-KDD and CIDDS-001	KNN and Naïve bayes	labels like ‘unknown’ and ‘suspicious’ in the CIDDS-001 dataset do not	Two different data sets were used.

				specify if the network has an intrusion or not.	
Ashwini Pathak and Sakshi Pathak, 2020	Machine learning for intrusion detection	NSL-KDD	KNN and decision tree	NSL-KDD dataset has several problems.	The feature selection is observed to improve the accuracy .
Suhaimi et al., 2019	GA is used for the detection of harmful connections in a network	KDD Cup 99	GA	Performance of the GA algorithm can't be measured with metrics like accuracy.	GA was observed to be very effective.
Sazzadul Hoque, 2012	GA is used for detecting intrusions	KDD99 dataset	GA	Performance of the GA algorithm can't be measured with metrics like accuracy.	GA was observed to be very effective.
Hashemi, Muda and Yassin, 2013	GA is used for detecting intrusions	KDD Cup 99 dataset	GA	Performance of the GA algorithm can't be measured with metrics like accuracy.	GA was observed to be very effective.
Chandrakar et al., 2014	GA is used for detecting intrusions	Audit dataset	GA	Performance of the GA algorithm can't be measured with metrics like accuracy.	GA was observed to be very effective.
Alotaibi and Elleithy, 2016	WIDS) was proposed to discover attackers in wireless networks.	AWID dataset	Random Forests, Extra Trees , Bagging	data in the dataset is WEP specific	The dataset contains a large amount of data.
Wu et al., 2018	unauthorized access points or rouge access points(RAP) detection.	RSS vectors	RSS-based practical rouge access point detection.	RAP is evaluated in this approach using metrics like detection rates and cannot be	Good performance in detection.

				compared to machine learning algorithms.	
Han et al., 2011	unauthorized access points or rouge access points(RAP) detection.	Real-time data	client-centric approach that uses the round-trip time between the DNS server and the user.	cannot be directly compared to machine learning algorithms.	Good performance in detection of RAP's.
Yang, Song and Gu, 2012	unauthorized access points or rouge access points (RAP) detection.	Real-time data	user-side evil twin detection method.	cannot be directly compared to machine learning algorithms.	Good performance in detection of RAP's.
Nakhila et al., 2018	unauthorized access points or rouge access points (RAP) detection.	Real-time data	real-time client-side detection method.	cannot be directly compared to machine learning algorithms.	Good performance in detection of RAP's.
Ravula Arun Kumar et al, 2021	unauthorized access point identification.	RTT dataset	KNN, SVM and decision tree classifiers.	Small amount of samples in the dataset.	High accuracy in detection.
Srinivas et al., 2022	unauthorized access point identification.	RTT dataset	SVM, Multilayer Perceptron (MLP), KNN and Decision Tree classifiers	Small number of samples in the dataset.	Effective performance in detection.

Table (1): Research summary

The main focus of the research is to find out the effectiveness of machine learning models in identifying inappropriate or unauthorized access points in wireless networks. But the literature that corresponds to existing approaches that use machine learning algorithms for identifying unauthorized access points were limited. The working of the machine learning algorithms in detecting network intrusions are similar to the working of the machine learning algorithms in identifying unauthorized access points. It is clear from the literature that both the intrusion detection and access point detection requires a dataset containing different features of networks. It can be seen that the machine learning classifiers , like KNN, SVM , Random forest and GA are highly effective in detecting intrusions and unauthorized access points. The pre-processing and feature selection are revealed to be important steps before

training a machine learning model. Feature selection methods like ChiSqSelector , RST and ANOVA are found to be highly effective.

Only two approaches were found to directly use machine learning classifiers for identifying inappropriate or unauthorized access points but the main limitation of those two approaches was that both the approaches used the RTT dataset which contained a relatively small sample size. So in the approach proposed here the inappropriate or unauthorized access points will be identified by using machine learning algorithms with ANOVA as the feature selection method and the dataset used in this approach will contain a large amount of data related to network features of wireless networks and that dataset will be the AWID intrusion detection dataset.

3 Research Methodology

3.1 Overall working

A system for identifying inappropriate or unauthorized access points in a wireless network using machine learning classifiers is being developed here. The machine learning classifiers KNN,SVM,GA and random forest will be trained using the data in the AWID intrusion detection dataset. During training the data in the AWID intrusion detection dataset is read and pre-processed after this the important features from the data will be selected using the ANOVA method. Using these features, the KNN,SVM,GA and random forest classifiers will be trained. I would like to suggest a new Identifying inappropriate access points using feature selection methods and algorithms as shown in the below image.

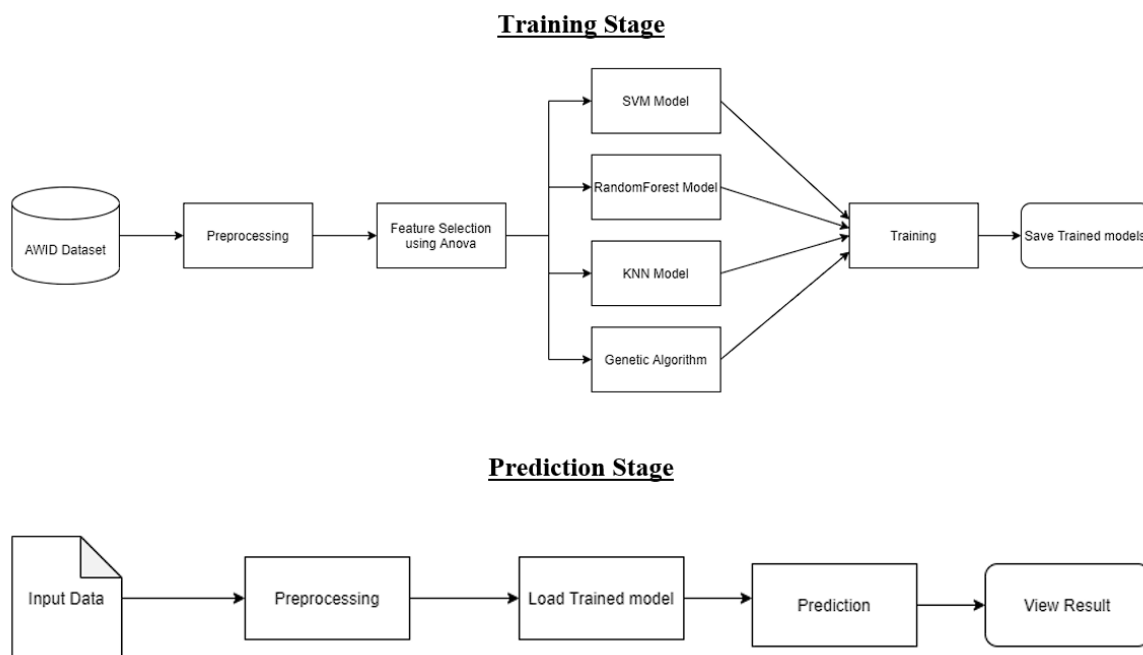


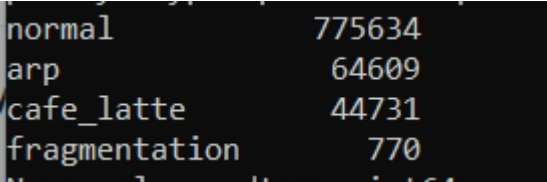
Figure 3: Proposed flow

3.2 Data collection

The data that will be used here is obtained from the AWID intrusion detection dataset (Marlowe, 2022). The dataset consists of different features of the wireless network as data. The data will be in a .CSV file. Each row of the .CSV file corresponds to the features of a network or data. The columns of the file will contain the values that corresponds to the different features associated with networks and one of the columns will contain the labels or classes associated with networks. The column that represents the label will contain the data as text and it will be the values of different types of attacks caused by unauthorized access points. These labels are 'amok', 'arp', 'Deauthentication', 'authentication_request', 'beacon', 'caffe_latte', 'fragmentation', 'normal', 'probe_response' and 'evil_twin'. Here, the label 'normal' represents a network in which no attack has taken place and all the other labels represent the networks in which an attack corresponding to the label has taken place. This dataset is selected as it contains a large number of data samples.

3.3 Pre-processing

The pre-processing is done so that all the unwanted data present in the dataset are removed. The pre-processing will help in improving the performance of the machine learning models and it will also help in reducing the time taken for training the machine learning models. During pre-processing the columns and rows containing the 'null' values will be discarded. After removing all the columns and rows containing 'null' values. The data that correspond to only four labels will remain in the dataset, which are 'normal', 'arp', 'caffe_latte' and 'fragmentation' and all the data corresponding to the other labels are removed.



normal	775634
arp	64609
cafe_latte	44731
fragmentation	770

Figure 4: pre-processing data

The machine learning models use the labels for learning and the labels must contain numerical values as machine learning models are able to perform better using numerical values just like machines. So all the string or text values of the labels corresponding to the data must be replaced with a numerical value. Here, a binary classification will be performed by the machine learning models and for this all the data corresponding to the labels that are attacks will be assigned a particular numerical value and all the data corresponding to the label 'normal' will be represented a different numerical value. The values of all the labels that represent attacks present in the form of text, 'arp', 'caffe_latte' and 'fragmentation' will be replaced by the numerical value 1 and the data with label 'normal' will be replaced with the numerical value 0.

3.4 Feature selection

After pre-processing the important features will be selected from the remaining data. The feature selection will help in reducing time while training and also will help the machine learning models in achieving a more accurate result. The ANOVA technique is used here and from the entire features in the dataset and the most important 14 features will be selected from the data. These features will now be used for training the machine learning classifiers.

3.5 Model creation

The important features from the data will be used for training the machine learning models. As seen from the literature the machine learning is highly effective detecting intrusions in networks. The data will be split into a training and testing test . 20% of the data will be used for testing the performance or evaluating the machine learning classifiers and 80% of the data will be used for training the machine learning classifiers. The data will be scaled before training the machine learning models to make sure that all the values of the features are in specific ranges and not in random ranges as that may affect the performance of the machine learning models.

```
Training set
(708595, 14)
(708595,)

Testing set
(177149, 14)
(177149,)
```

Figure 5: Training and testing evaluate output

3.5.1 KNN classifier

The KNN is a supervised machine learning classifier and it performs classification based on the number of nearest neighbors . This number is provided as a parameter to the classifier. The KNN works by determining the data points which are nearby based on the parameter value provided . On finding the data points that are nearest the algorithm uses majority voting for finding out the class or label the appears the greatest number of times. This classifier is loaded and trained using the training data and saved. The number of nearest neighbors will be set as 7 here.

3.5.2 Random forest classifier

Random forest is a group of classification trees created from randomly chosen samples from the training data. Several tree structures are created in the random forest and the each tree will perform a prediction the final prediction by the random forest is obtained by aggregating the predictions of each trees. The random forest is trained using the data in the training set and here the number of trees to created is set as 100.

3.5.3 SVM classifier

The SVM performs linear classification. The SVM works by creating hyperplanes that sperate different types of data samples. The same kind of data will be in the same side of the plane and the it performs classification by linear separation of the data. The kernel parameter determines the way in which the SVM classifies the data and here the value of the kernel parameter is set as linear and then trained using the training data.

3.5.4 Genetic Algorithm

The GA works performs searching and works by mimicking natural evolution. Here the genetic algorithm is applied to a Logistic regression (LR) classifier . The GA performs mutation , selection , inheritance and crossover and since here it is applied the LR classifier it is trained using the training data.

3.6 Evaluation

The performance of the machine learning models will be evaluated by measuring the performance metrics accuracy and precision. The values of both accuracy and precision these models will be compared for finding the model with the best performance. The performance metrics will be found by testing the trained machine learning models with the testing data.

4 Design specification and implementation

4.1 Design specification

The system that detects the unauthorized access points in a wireless was created on:

- A PC with 8GB RAM and with an i7 processor
- Python is the programming language.
- Tkinter is the Python library is used for creating designing the desktop application

4.2 Implementation

The pre-processing is performed using the methods present in the ‘pandas’ library in Python. The feature selection is performed by using the ‘SelectKBest()’ method imported from the ‘sklearn’ library in Python. The ANOVA technique is implemented by assigning the argument ‘score_func’ of the ‘SelectKBest()’ method with the value ‘f_classif’ which represents the ANOVA method and is imported from the ‘sklearn’ library in Python. Then the machine learning classifiers are loaded. The random forest classifier is loaded first by importing the method ‘RandomForestClassifier()’ from the ‘sklearn’ library in Python. Then the KNN classifier is loaded by importing the method ‘KNeighborsClassifier()’ from the ‘sklearn’ library in Python, the GA algorithm is applied to the logistic regression classifier which is defined by importing the ‘LogisticRegression()’ method from the ‘sklearn’ library in Python, the fitness function ,crossover and mutation and the generation of new chromosomes are also performed. Finally the SVM classifier is loaded using the ‘SVC()’ method imported from the ‘sklearn’ library in Python. All the machine learning classifiers are trained using the ‘fit()’ method. The ‘predict()’ method is used for testing all the trained machine learning classifiers.

A system will be implemented in the form of a desktop application. The desktop application will consist of a login interface and on logging the interface where the unauthorized access points can be found out will be displayed. The interface will consist of input fields where the different features associated with a wireless network can be given as input. A total 14 inputs or values of features can be provided. After providing the features and clicking a button ‘Predict’ in the interface.

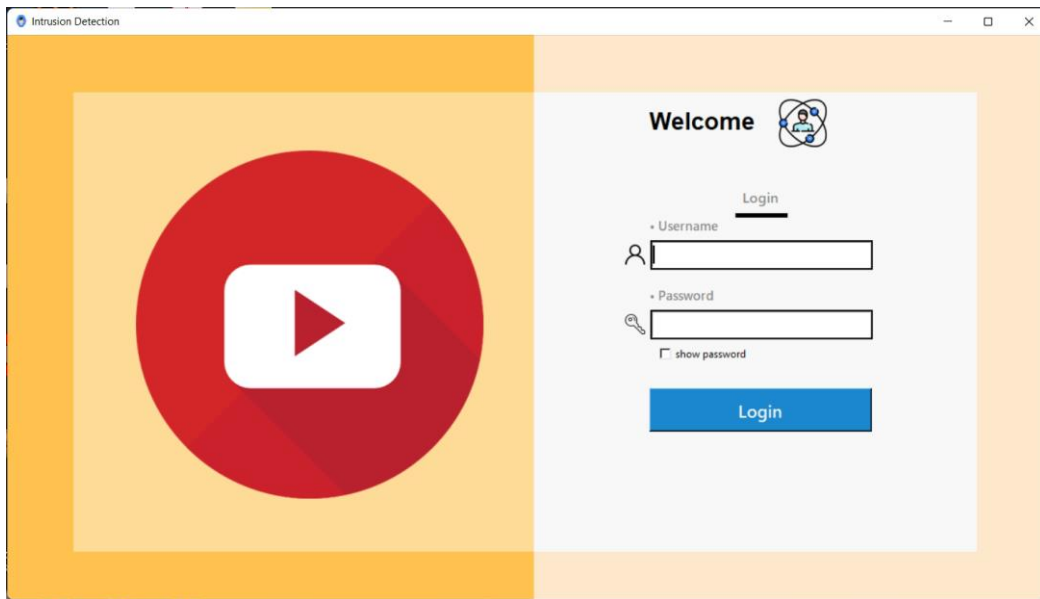


Figure 6: Desktop application

Once we login with admin credential then after we can be able to fill the input. The system will use the inputs and provide these inputs to the trained machine learning classifier which is loaded here. The machine learning classifier receives the features as inputs and predicts if the network features provided as input correspond to a network in which an unauthorized access point attack has occurred. The output will be displayed as the text 'Normal' if no unauthorized access point attack has occurred in a network whose features is given as the input and the text displayed will be 'Intrusion Detected' if the system identifies that an unauthorized access point attack has occurred in a network whose features is given as the input.

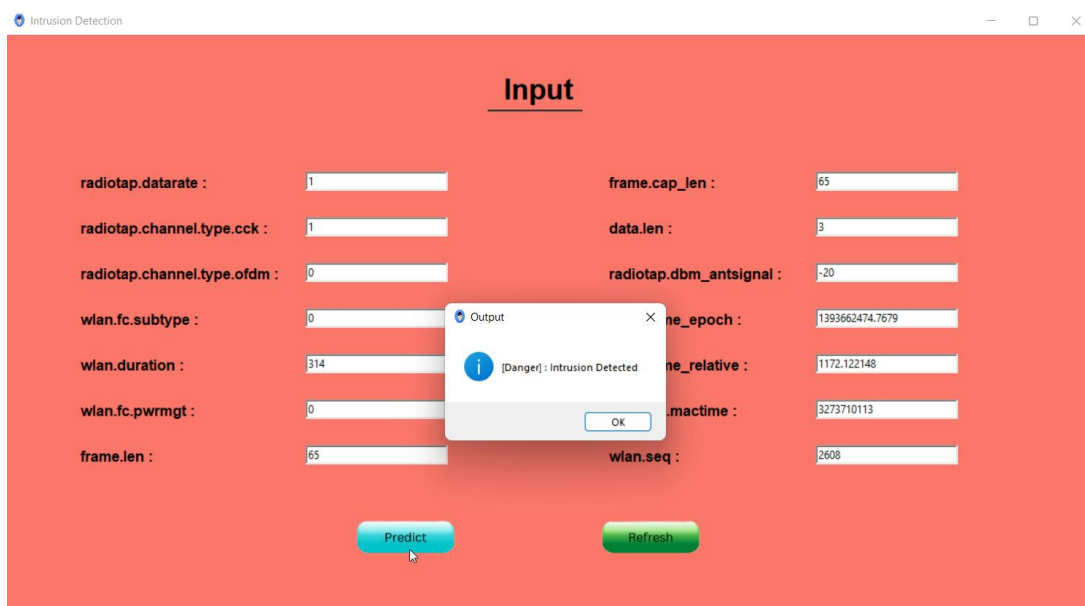


Figure 7: Identifying inappropriate access points

5 Evaluation and discussion

5.1 Evaluation and results

The trained and saved machine learning models are tested using the data in the testing set for finding out the performance of the machine learning classifiers in identifying the presence of unauthorized access points in a wireless network. The accuracy and precision achieved by the machine learning classifiers will be used as the metrics for evaluating performances of the machine learning classifiers.

```
Accuracy score for RF is :100.0%
Precision score for RF is :100.0%
Accuracy score for KNN is :99.7%
Precision score for KNN is :97.3%
Accuracy score for Genetic Algorithm is :99.5%
Precision score for Genetic Algorithm is :96.39999999999999%
Accuracy score for SVM is :99.5%
Precision score for SVM is :96.5%
```

Figure 8: Trained model output

5.1.1 Accuracy

The accuracy can be defined as the ratio of the number of unauthorized access points attacks correctly identified by a machine learning classifier to the total number of predictions made by the machine learning classifier and the accuracy achieved by 4 machine learning classifiers can be seen in figure (9).

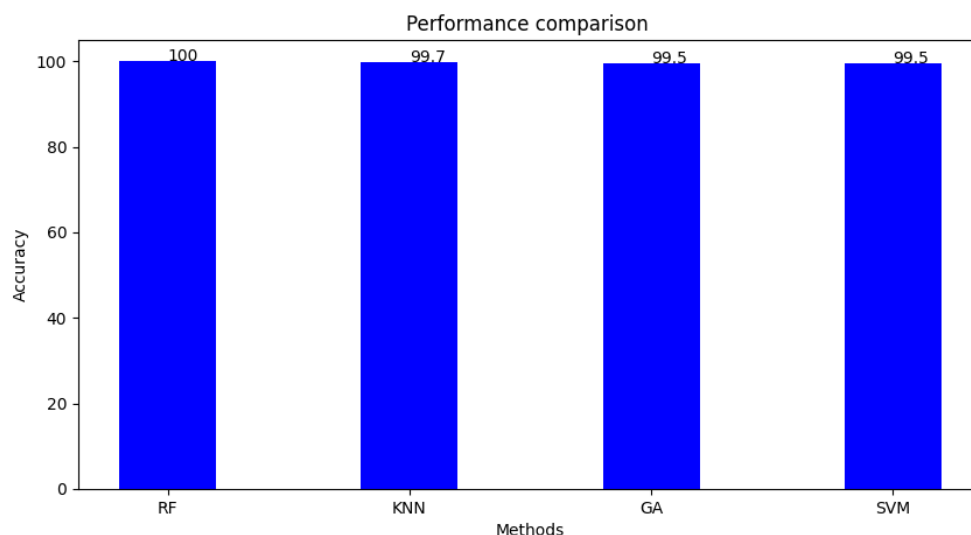


Figure 9: The comparison of accuracies achieved by the machine learning classifiers

The accuracy will be represented in percentage and it can be seen that the best value of accuracy is achieved by the random forest algorithm with a value of 100% (Figure(9)).

5.1.2 Precision

The comparison of the values of precision achieved by the machine learning classifiers will be displayed in figure (10).

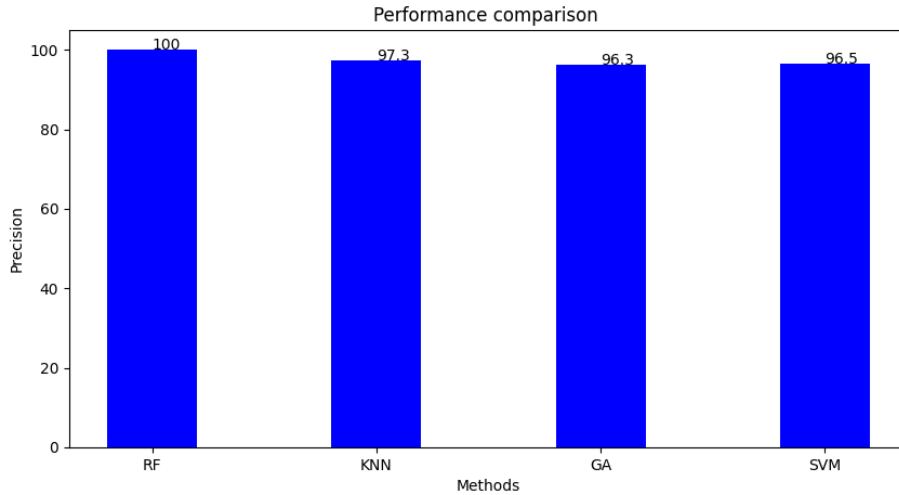


Figure 10: The comparison of precision achieved by the machine learning classifiers

Precision will also be represented in percentage and it can be seen that the best precision is achieved by the random forest model with a value of 100% (Figure (10)).

5.2 Discussion

The machine learning classifiers are trained using the features from the pre-processed data in the dataset. The accuracy and precision values of the machine learning classifiers are found out and it is seen that the random forest classifier showcases the best performance with both its precision and accuracy value reaching 100%. It was seen from the study of the existing literature that machine learning classifiers are highly effective in identifying unauthorized access points in wireless networks this observation is supported by the results of the study performed here as it was seen that all the machine learning classifiers achieved good performances. It was also found in the study of the literature that pre-processing and feature selection will result in the machine learning models producing a good performance and as pre-processing and feature selection are also used here along with good performances from machine learning models, so the result of the study here supports the observations made when studying the literature. A dataset having a relatively small sample was seen as the limitations of two of the existing approaches that performed the unauthorized access point detection using machine learning and here a dataset having a large number of data samples are used and it can be seen that the random forest model is able to achieve an accuracy greater than the accuracy achieved in the existing approaches which implies that the use of a large dataset may have resulted in the performance of the machine learning model being better.

System	Accuracy
(Ravula Arun Kumar et al, 2021)	99.99%

(Srinivas et al., 2022)	96.56%
System developed here	100%

Table 1: The accuracies achieved by the existing systems and the system developed here

The system developed here is able to achieve better accuracy than the existing systems that have been studied here (Table (1)).

5.2.1 Critical analysis

The system is able to achieve a very good performance in detecting unauthorized access points in wireless networks, but it has its own limitations. One main limitation of the system developed here is that it is not able to detect the type of unauthorized access point attack occurring in a wireless network as the system developed here only is able to identify if an unauthorized access point attack has occurred on a wireless network or not and the type of attack is not identified. Accuracy and Precision is maybe different if we change the dataset because it depends on the dataset. Another limitation is the size of the dataset used because initially the dataset was quite large and had 8 labels corresponding to different attacks but the some of the labels and a part of the data were removed during pre-processing reducing the size of the dataset considerably and as the size of the data is reduced the significance of the results may also be reduced.

6 Conclusion and future enhancements

The main research questions posed at the beginning of the research have been answered , the first research question is answered as it clearly seen that the KNN ,SVM, Random forest and GA are really effective in identifying the unauthorized access point attack in a wireless network with the random forest being the most effective and the second research question is also answered as the performances of the machine learning models are evaluated by comparing the performance metrics' accuracy and precision and it can be seen that the random forest achieves the highest values for accuracy and precision with both values being 100%. The data from the AWID intrusion detection dataset is used here. The data is pre-processed and the important feature from the data is extracted using the ANOVA technique and these features are used for training the machine learning classifiers. The machine learning classifier with is,best performance is , here the random forest , is successfully used for creating a desktop application that is able to predict the unauthorized access point attack in a wireless network.

Enhancements can be made to the system that it is able to detect different types of unauthorized access point attacks. The detection is now performed using machine learning models in the future deep learning models can be used for performing the detection of inappropriate access points in wireless networks.

7 Video Presentation

- Presentation - https://youtu.be/118_hiazSTc
- Demo - https://youtu.be/YF_DJaKTSYs

References

- Abdallah, E.E., Eleisah, W. and Otoom, A.F. (2022). Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey. *Procedia Computer Science*, 201, pp.205–212. doi: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(3), pp.193–206. doi:10.12913/22998624/149934.
- Alotaibi, B. and Elleithy, K. (2016). A majority voting technique for Wireless Intrusion Detection Systems. 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT). doi:10.1109/lisat.2016.7494133.
- Alsahlany, Ali & Almusawy, Alhassan & Alfatlawy, Zainalabdin. (2018). Risk analysis of a fake access point attack against Wi-Fi network. *International Journal of Scientific and Engineering Research*. 9. 322-326.
- Alsahlany, Ali. (2014). Experimental Analysis of WLAN Security Weakness By Cracking 64 & 128 bit WEP Key. *The islamic college university journal*. 9. 165-176.
- Ashwini Pathak and Sakshi Pathak (2020). Study on Decision Tree and KNN Algorithm for Intrusion Detection System. *International Journal of Engineering Research and*, V9(05). doi:10.17577/ijertv9is050303.
- Awad, F., Al-Refai, M. and Al-Qerem, A. (2017). Rogue access point localization using particle swarm optimization. 2017 8th International Conference on Information and Communication Systems (ICICS). doi:10.1109/iacs.2017.7921985.
- Calhoun, P.R., Friday, R.J., Jr, R.B.O., Galloway, B., Frascone, D.A., Dietrich, P.F. and Jain, S.K. (2008). Discovery of rogue access point location in wireless network environments. [online] Available at: <https://patents.google.com/patent/US8089974B2/en> [Accessed 17 Nov. 2022].
- Chandrakar, O., Singh, R., Lal, B. and Barik (2014). Application of Genetic Algorithm in Intrusion Detection System. [online] Available at: <https://core.ac.uk/download/pdf/234676822.pdf> [Accessed 27 Nov. 2021].
- Das, V., Pathak, V., Sharma, S., Sreevathsan, Srikanth, MVVNS. and Gireesh Kumar, T. (2010). Network Intrusion Detection System Based On Machine Learning Algorithms. *International Journal of Computer Science and Information Technology*, 2(6), pp.138–151. doi:10.5121/ijcsit.2010.2613.
- Dr. Sebastian Nixon , Yibrah Haile .(20127). " 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 (IJETTCS)* , Volume 6, Issue 3, May - June 2017 , pp. 293-300 , ISSN 2278-6856.
- Farnaaz, N. and Jabbar, M.A. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer Science*, [online] 89, pp.213–217. doi:10.1016/j.procs.2016.06.047.

- Han, H., Sheng, B., Tan, C.C., Li, Q. and Lu, S. (2011). A Timing-Based Scheme for Rogue AP Detection. *IEEE Transactions on Parallel and Distributed Systems*, 22(11), pp.1912–1925. doi:10.1109/tpds.2011.125.
- Hashemi, V.M., Muda, Z. and Yassin, W. (2013). Improving Intrusion Detection Using Genetic Algorithm. *Information Technology Journal*, 12(11), pp.2167–2173. doi:10.3923/itj.2013.2167.2173.
- Khan, K., Mehmood, A., Khan, S., Khan, M.A., Iqbal, Z. and Mashwani, W.K. (2020). A survey on intrusion detection and prevention in wireless ad-hoc networks. *Journal of Systems Architecture*, 105, p.101701. doi:10.1016/j.sysarc.2019.101701.
- Liu, D., Barber, B. and DiGrande, L. (2009). Introduction to Networking. Cisco CCNA/CCENT Exam 640-802, 640-822, 640-816 Preparation Kit, pp.1–46. doi:10.1016/b978-1-59749-306-2.00005-1.
- Marlowe, B. (2022). CIS 660 - Final Project. [online] GitHub. Available at: <https://github.com/Bee-Mar/AWID-Intrusion-Detection> [Accessed 18 Nov. 2022].
- Nakhila, O., Amjad, M.F., Dondyk, E. and Zou, C. (2018). Gateway independent user-side wi-fi Evil Twin Attack detection using virtual wireless clients. *Computers & Security*, 74, pp.41–54. doi:10.1016/j.cose.2017.12.009.
- Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T. and Al-Hashida, A.Y. (2018). Intrusion detection model using machine learning algorithm on Big Data environment. *Journal of Big Data*, 5(1). doi:10.1186/s40537-018-0145-4.
- Ravula Arun Kumar, Pasula Sreya, A.Sai Vamshi, S.Anoop Kumar. (2021). Detection of Illegitimate Access Point Using Machine Learning. *International Journal Of Engineering And Computer Science* .(10)7. pp.25359-25361 .doi : 10.18535/ijecs/v10i7.4595
- Sazzadul Hoque, M. (2012). An Implementation of Intrusion Detection System Using Genetic Algorithm. *International Journal of Network Security & Its Applications*, 4(2), pp.109–120. doi:10.5121/ijnsa.2012.4208.
- Srinivas, B., Puri, B., Vamshikrishna, Y., Triveni, B. and Reddy, T. (n.d.). UNAUTHORIZED ACCESS POINT DETECTION USING MACHINE LEARNING ALGORITHMS FOR INFORMATION PROTECTION. [online] *International Research Journal of Modernization in Engineering Technology and Science*, Peer-Reviewed, Open Access, pp.2582–5208. Available at: https://www.irjmets.com/uploadedfiles/paper/issue_6_june_2022/27396/final/fin_irjmets1656693956.pdf [Accessed 18 Nov. 2022].
- Suhaimi, H., Suliman, S.I., Musirin, I., Harun, A.F. and Mohamad, R. (2019). Network intrusion detection system by using genetic algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(3), p.1593. doi: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*, [online] 12(7), pp.8226–8235. Available at: <https://www.tojqi.net/index.php/journal/article/view/5151/3648> [Accessed 18 Nov. 2022].

Tahaseen, U. (2019). Survey on Unauthorized Access Point Detection using Machine Learning Algorithms. [online] International Journal of Engineering Science and Computing. Available at: [https://ijesc.org/upload/6e7285c783e179e62099b800036f6b27.Survey%20on%20Unauthorized%20Access%20Point%20Detection%20using%20Machine%20Learning%20Algorithms%20\(1\).pdf](https://ijesc.org/upload/6e7285c783e179e62099b800036f6b27.Survey%20on%20Unauthorized%20Access%20Point%20Detection%20using%20Machine%20Learning%20Algorithms%20(1).pdf) [Accessed 17 Nov. 2022].

Vanhoef, M. and Piessens, F. (2017). Key Reinstallation Attacks. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17. [online] doi:10.1145/3133956.3134027.

Wu, W., Gu, X., Dong, K., Shi, X. and Yang, M. (2018). PRAPD: A novel received signal strength-based approach for practical rogue access point detection. International Journal of Distributed Sensor Networks, 14(8), p.155014771879583. doi:10.1177/1550147718795838.

Yang, C., Song, Y. and Gu, G. (2012). Active User-Side Evil Twin Access Point Detection Using Statistical Techniques. IEEE Transactions on Information Forensics and Security, 7(5), pp.1638–1651. doi:10.1109/tifs.2012.2207383.