Unauthorized Access Point Detection using Machine Learning Algorithms

[1]Sudhiksha.V, [2]Sufiyaan Ali Khan, [3] Suhaas.M
[4]Sumanth.G, [5]Sumedh Acharya.CH, [6] Sumedh.V,
[7] Prof R. Karthik
[123456]Student, [7] Assistant Professor

[1] 2111cs020572@mallareddyuniversity.ac.in, [2] 2111cs020573@mallareddyuniversity.ac.in,
[3]2111cs020574@mallareddyuniversity.ac.in, [4] 2111cs020575@mallareddyuniversity.ac.in,
[5]2111cs020576@mallareddyuniversity.ac.in, [6] 2111cs020577@mallareddyuniversity.ac.in
Department of Artificial Intelligence & Machine Learning
MallaReddy University, Hyderabad.

Abstract:

The increasing cybersecurity threats in the digital landscape, characterized by the persistent attempts of malicious entities to acquire valuable data, underscore the importance of implementing robust defenses to counter unauthorized access. Safeguarding against the relentless pursuit of valuable data by malicious actors is of paramount significance in the digital realm. A key element in this defense strategy involves the use of Unauthorized Access Point Detection Systems (UADPS), which play a crucial role in achieving a balance between high detection accuracy and minimizing false alarms. This document offers a thorough examination of UADPS, presenting a taxonomy of machine learning approaches, such as Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Multilayer Perceptron (MLP), to significantly improve detection capabilities. Through a meticulous evaluation, the paper reveals the nuanced strengths and weaknesses of recent UADPS implementations, highlighting the transformative impact of machine learning algorithms in bolstering information security. The incorporation of methodologies like SVM and KNN proves to be instrumental in addressing cybersecurity challenges. By scrutinizing datasets, performance metrics, and application scenarios, the paper contributes valuable insights to the ongoing conversation on enhancing UADPS accuracy, fostering innovation in cybersecurity practices, and advancing overall network security.

1. INTRODUCTION

In the realm of cybersecurity, the imperative for robust detection systems to identify unauthorized access points has grown substantially. This project is dedicated to the creation of a detection system for unauthorized access points using machine learning methodologies. The foundational work involves an in-depth analysis of the KDD99 dataset, a widely recognized benchmark dataset within the field of unauthorized access point detection.

Executing the project involves leveraging Python alongside essential data processing libraries such as NumPy and Pandas to thoroughly explore the KDD99 dataset. This dataset encapsulates network traffic data, with each record representing a network connection characterized by various features, including protocol type, service, and numerous statistical attributes. Rigorous steps are taken to load and cleanse the dataset to ensure data quality.

Commencing with an in-depth understanding of the dataset through exploratory data analysis, the project aims to discern the distribution of different classes of network traffic, focusing specifically on identifying normal activities and potential malicious incursions. The dataset contains a diverse array of outcomes, ranging from 'normal' to various types of unauthorized access point attempts, including 'neptune,' 'back,' and 'teardrop.'

The exploratory analysis extends to meticulous data cleaning, involving checks for null and duplicate values. Subsequently, the project delves into encoding categorical variables and analyzing the statistical distribution of various features. Visualizations are employed to comprehend the class distribution and glean insights into the prevalence of different types of network activities.

To optimize the dataset for machine learning models, additional preprocessing steps include handling categorical variables and scrutinizing numerical features. Emphasis is placed on the encoding of categorical variables to prepare the data adequately for training and testing machine learning models.

In the course of the exploratory process, attention is devoted to understanding the statistical characteristics of features such as duration, protocol type, service, and flag. This understanding is pivotal for selecting suitable machine learning algorithms and fine-tuning model performance.

This introduction establishes the groundwork for a comprehensive exploration of unauthorized access point detection using machine learning. Subsequent sections of the project will entail the development and training of machine learning models on the preprocessed dataset, with the ultimate aim of crafting an effective and precise Unauthorized Access Point Detection System.landscape.

2. Information Protection

Information protection is like building a fortress around sensitive data, ensuring it's safe from unwanted guests, changes, or harm. It's not just a one-time thing but a whole set of strategies, technologies, and rules to guard data from creation to disposal. Think of it as encoding secrets with encryption, setting up guards to decide who gets inside, keeping a watchful eye for trouble, and having clear rules for how data should be treated.

At its heart, information protection is about keeping things hush-hush (confidentiality), making sure things stay as they should (integrity), and making sure you can get to your stuff when you need it (availability). In today's digital world, where everything's connected, protecting information is key to staying safe, keeping trust intact, following the rules, and keeping what's valuable out of harm's way. It's not just about fancy tech; it's a mix of smart tech, teaching people, checking things often, and doing things the right way to make sure sensitive info stays safe.

Imagine it like a multi-layered shield – you figure out how important the info is, lock it up with codes, decide who gets the keys, keep an eye out for trouble, follow the rules of the land, and make sure everyone knows how to keep things safe. In the end, information protection is a big deal in today's world, where data is like treasure, and you want to make sure it's always in good hands.

3. LITERATURE REVIEW

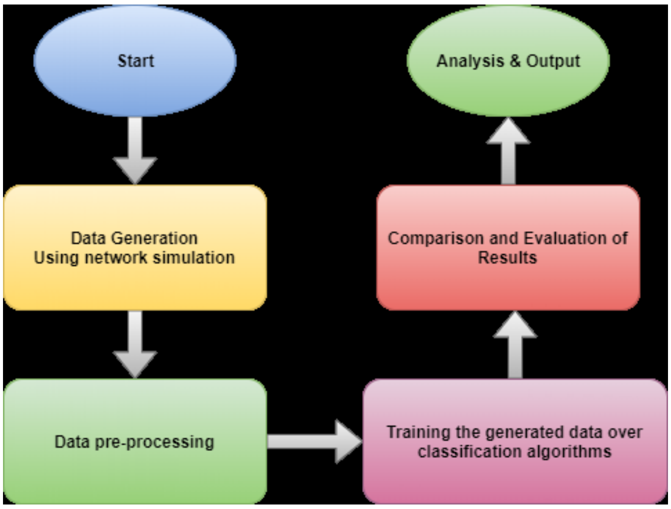
To thoroughly explore the detection of intruders in wireless networks using algorithms, we'll dive deep into the issue. We'll gather information from various sources, including practical evaluations and insights from respected researchers. Checking out what experts have already discovered in empirical literature will help us pinpoint the problems others have encountered in similar studies or in areas closely related to ours.

By delving into existing literature, we can uncover drawbacks and areas that need improvement, providing valuable guidance for our own research. This chapter aims to discuss crucial aspects of our study based on what other authors have presented in their empirical research. Additionally, we'll shed light on gaps in the literature to identify where our study can make a meaningful contribution.

Moving forward, we'll explore theories and models related to the application of algorithms for detecting unauthorized access by intruders in wireless networks. This chapter will also present a conceptual framework, offering a detailed examination of both dependent and independent variables crucial to our study. In simpler terms, we're setting the stage for our research by building on what others have found, identifying gaps, and establishing the theoretical groundwork for detecting intruders in wireless networks.

4. METHODOLOGY

The methodology comprises of five processes:
Load Data, Data Analysis, Data Preprocessing, Model Building and Prediction.


Start
Analysis & Output
Data Generation Using network simulation
Comparison and Evaluation of Results
Data pre-processing
Training the generated data over classification algorithms

Load Data and Data Analysis:

To construct a machine learning model, the essential first step involves obtaining a dataset—specific data formatted for the intended problem. Whether targeting business analytics or medical diagnostics, each dataset varies in content. Typically stored in a CSV file for code compatibility, datasets serve as the building blocks for machine learning models, influencing their accuracy and effectiveness. This process involves tailoring raw information to fit the specific intricacies of the problem at hand, emphasizing the pivotal role of quality datasets in training and enhancing machine learning models.

Data Preprocessing:

Getting data ready for a machine learning model is what we call data preprocessing, a key initial step in creating such models. In real-world machine learning projects, we often don't start with neat and organized data. It's like receiving a messy puzzle that needs sorting before we can put it together. Every time we work with data, it's a must to clean it up and structure it properly. That's where data preprocessing comes in – it's the task that takes care of cleaning and organizing the data so that it's all set for the machine learning magic to happen.

Model Building:

Building a model involves a series of steps to ensure its effectiveness. Initially, identify the type of training dataset needed and gather labeled data for it. Then, split the dataset into training, testing, and validation subsets. Understand the input features of the training dataset thoroughly to enable accurate predictions. Choose an appropriate algorithm—like support vector machine or decision tree—for the model. Implement the chosen algorithm on the training dataset. In some cases, validation sets, a subset of training data, are necessary for controlling parameters. This step-by-step process ensures a well-informed and robust model ready for making accurate predictions

4.1 Data Generation

The real time of network access point is not available publicly. Therefore, we have a generated a synthetic data using network simulation. The generated access point data will be mainly based on the connection type and connection type will has been classified into 3 categories 3G, 4G and Wired data. For each connection type we will generate a separate data and perform a classification over it. The generated data will have 2 features response time of network and the length of the data packet. These both the features will be used to perform the predictive analysis.

4.2 Training the Models

Now that we've prepared the synthetically generated data through pre-processing, the next step is training it using various classification models for a thorough comparison. In this analysis, we're employing three classification algorithms: SVM (Support Vector Machine), KNN (K-nearest neighbor), and the Ant Colony Optimization algorithm. Both SVM and KNN fall under supervised machine learning, providing guidance based on labeled data. On the other hand, the Ant Colony Optimization technique follows a genetic algorithm approach. The diverse data for different network types undergoes training with these models, paving the way for a comprehensive predictive examination of their performance.

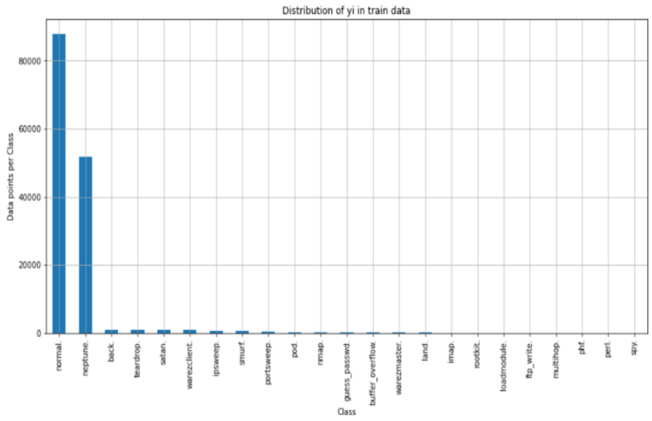
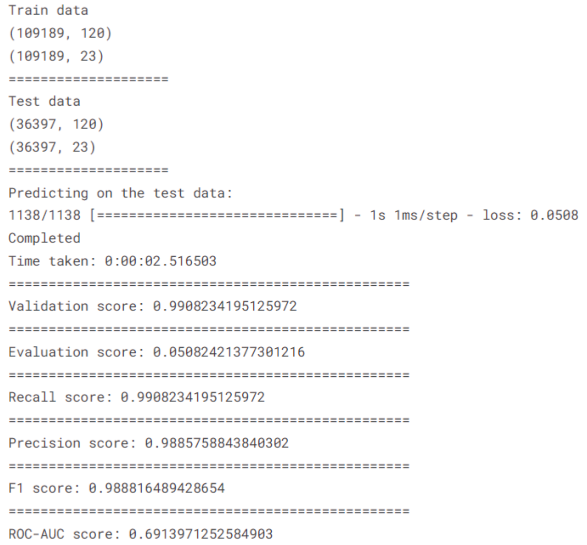


Fig 4.2 Data points per class

4.3 Evaluation metrics

There are two different approaches will be used for evaluation of each model. First the accuracy of each model will be calculated. later, the precision score will be calculated, in order to check for false positive values. This experiment will be performed over 1000 and 10000 different number of samples. the different subsets of dataset are used in order
to find out the impact on increasing the training data. The Evaluation has been carried
out by performing iterative operation with algorithm over the same data-set, in order to
_ne-tune the parameters.



```
Train data
(109189, 120)
(109189, 23)
====================
Test data
(36397, 120)
(36397, 23)
====================
Predicting on the test data:
1138/1138 [==============================] - 1s 1ms/step - loss: 0.0508
Completed
Time taken: 0:00:02.516503
==================================================
Validation score: 0.9908234195125972
==================================================
Evaluation score: 0.05082421377301216
==================================================
Recall score: 0.9908234195125972
==================================================
Precision score: 0.9885758843840302
==================================================
F1 score: 0.988816489428654
==================================================
ROC-AUC score: 0.6913971252584903
```

Fig 4.3 Evaluation metrics

4.4 Experimental Results

The provided Python code focuses on the analysis of the KDD99 dataset for Unauthorised Access Point Detection. Initially, the dataset is loaded into a Pandas DataFrame, and a brief exploration is performed by displaying a sample. Subsequently, a comprehensive dataset analysis is conducted, revealing its size, the number of features, and the distribution of network activity labels, with a notable dominance of the 'normal' class. Data cleaning operations follow, including checking for null values, removing duplicates, and handling missing data. The cleaned dataset is saved as a pickle file, although this operation is currently commented out. Exploratory Data Analysis (EDA) ensues, with visualizations depicting class distribution and numerical insights into output labels. Categorical features undergo encoding, and their value distributions are analyzed. Numeric features, including 'duration,' 'src_bytes,' and 'dst_bytes,' are explored for a better understanding of their characteristics. The analysis extends to categorical features such as 'protocol_type,' 'service,' and 'flag.' Summary statistics are generated, presenting unique values and their percentages for each feature. A bar chart is plotted to visualize the distribution of the 'same_srv_rate' feature. The code concludes by emphasizing the importance of this analysis for building an effective Unauthorised Access Point Detection system, hinting at potential next steps like feature scaling, model training, and evaluation.

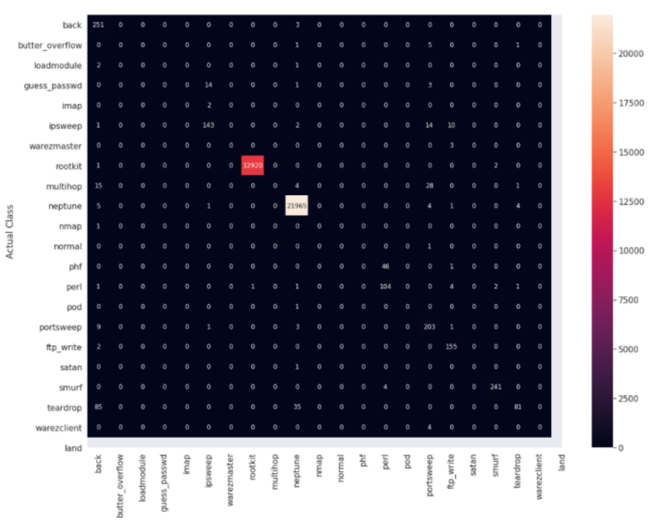


Fig 4.4 Confusion matrix

5. CONCLUSION

In contemporary times, network-based attacks pose a pervasive threat, leading to service disruptions and the compromise or manipulation of critical user information. This dissertation addresses the pressing issue of unauthorized access points within wireless

LAN networks, which have the potential to disrupt services and illicitly access vital user data, all without the need for specialized technical expertise. Detecting and mitigating these threats swiftly is crucial to avert potential disasters arising from their exploitation.

The proposed methodology focuses on the identification and classification of unauthorized access points, often initiated by malicious actors aiming to pilfer legitimate user credentials. These rogue points redirect the traffic of legitimate users through the attacker's system, facilitating the interception of sessions and sensitive data.

To achieve this, the project scrutinizes four classification algorithms—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP)—for their efficacy in detecting unauthorized access points. However, it acknowledges that practical considerations, such as scalability and seamless real-time integration into existing network infrastructures, are imperative for the successful implementation of countermeasures against evolving cyber threats.

5.2 References

[1] Wailinn Oo. (3 years ago). Retrieved from [https://www.kaggle.com/code/wailinnoo/intrusion-detection-system-using-kdd99-dataset]-2002.

[2]Bakkashetti Srinivas, Bhavana Puri, Yelaganamoni Vamshikrishna, B. Triveni. (June 2022). "Unauthorized Access Point Detection Using Machine Learning Algorithms for Information Protection." International Research Journal of Modernization in Engineering Technology and Science, Volume 04, Issue 06. e-ISSN: 2582-5208.

[3] Singh, S. (2020, January 12). Building an Intrusion Detection System using KDD Cup'99 Dataset. Analytics Vidhya. Retrieved from https://medium.com/analytics-vidhya/building-an-intrusion-detection-model-using-kdd-cup99-dataset-fb4cba4189ed.

[4]Juwale, A. M. (Year). Analysis and Detection of Unauthorized Access Points Using Various Machine Learning Algorithms. MSc Internship, Cyber Security, National College of Ireland. Student ID: X19129866. School of Computing.