

Ministry of Higher Education and Scientific Research

University of Babylon

Collage Of Information Technology

Information security department

**Intrusion Detection System using** **Decision Tree**

**A Graduate Project Submitted to the department of Information Security of the College of Information Technology, University of Babylon, in Partial Fulfillment of the Requirements for the bachelor’s degree in the Information Security of Information Technology.**

***Abbas Ghasan Noury Jreou***

**Supervised by:  
 *Prof. Dr. Eman Salih Alshamery***

**2023**



**Dedication**

This study is wholeheartedly dedicated to my beloved parents and my aunt, who have been my source of inspiration. They gave me the strength when I thought of giving up, and continually provided moral, emotional, and financial support. I wish them all the love and happiness in the world.

**Acknowledgments**

I would like to Express My Gratitude for Everyone Who Helped me during the Graduation Project Starting with Endless Thanks for My Supervisor Prof. Dr. Eman Salih Alshamery , Who does Keep any Effort in encouraging us to do a Great Job, providing me with Valuable Information and Advices to be Better Each Time.

Thanks for the Continuous Support and Kind Communication Which Had a Great Effect on me.

Thanks are extended to University of Babylon, College of information technology and Department of Information Security and the whole Doctors that helped me during my study in all stages.

**Supervisor Certification**

I certify that the Project entitled “Intrusion Detection system using Decision Tree” was prepared under my supervision at the Department of information Security / information technology college / University of Babylon , by the student Abbas Ghasan Noury as partial fulfillment of the requirements of the bachelor's degree in information Security.

Signature:

Name: Prof. Dr. Eman Salih Alshamery

Date: / / 2023

**The Head of Department Certification**

In view of the available recommendation, I forward the Project entitled “Intrusion Detection system using Decision Tree “ for debate by the examination committee.

Signature:

Name: Dr. Alharith A. Abdullah (Prof.)

Date: / / 2023

# **Abstract**

With the continuous increase of cyber dangers of all kinds and their rapid spread through networks and the Internet, the need arose to find effective solutions, including intrusion detection systems, but due to the emergence of new and diverse threats over time, we needed to increase the efficiency of these systems, which were previously dependent on the values of Statistical only in making decisions.

In this project, we will employ machine learning in the field of detecting intruders .Specifically,the decision tree (j48 classifier) is applied with (entropy)and (gini index) purity measures in our work, as this algorithm is fast And high accuracy. Also, we used labeled dataset about the network traffic called (CICIDS 2017), which contains 56,662 record and 77 features in training the system. When the system was put to the test, it showed high accuracy rates in finding its decisions as follows:

Entropy(67% train, 33% test) : 99.4%

Entropy(50% train, 50% test) : 99.2%

Gini index (67% train, 33% test) : 99.2%

Gini index (50% train, 50% test) : 99.3%

Also, the number of features has been reduced and specified by the system to (60), and the length of the tree limited to 10 levels only.

This allowed us to use this system to protect networks and increase their security, and it is also one of the systems that can be developed in the future.

Table Of Content

[**Abstract** VI](#_Toc135135197)

[**Chapter One: Introduction** 1](#_Toc135135198)

[**1.1** **Introduction** 2](#_Toc135135199)

[**1.2** **Problem Description** 3](#_Toc135135200)

[**1.3** **The Aim of project** 3](#_Toc135135201)

[**1.4** **Related Work** 3](#_Toc135135202)

[**1.5** **Outline** 4](#_Toc135135203)

[**Chapter Two: Theoretical Background** 5](#_Toc135135204)

[**2.1 Introduction To IDS** 6](#_Toc135135205)

[**2.2 History Of IDS** 6](#_Toc135135206)

[**2.3 Machine Learning** 7](#_Toc135135207)

[**2.3.1 Supervised Learning with Classification** 8](#_Toc135135208)

[**2.3.2 Unsupervised Learning with Classification** 8](#_Toc135135209)

[**2.3.3 Decision Tree Algorithm** 8](#_Toc135135210)

[**2.4 IDS with Machine Learning** 9](#_Toc135135211)

[**2.4.1 J48 (C4.5 Decision Tree) Classifier** 10](#_Toc135135212)

[**2.4.2 CICIDS 2017 Dataset** 10](#_Toc135135213)

[**2.5 Evaluation Methods** 11](#_Toc135135214)

[**Chapter Three: Practical** 12](#_Toc135135215)

[**3.1 System Architecture** 13](#_Toc135135216)

[**3.2 Dataset** 14](#_Toc135135217)

[**3.3 Building Step** 14](#_Toc135135218)

[**3.4 Simple Decision Tree Constructure** 15](#_Toc135135219)

[**3.5 J48 Algorithm** 16](#_Toc135135220)

[**Chapter Four: Implementation and results** 17](#_Toc135135221)

[**4.1 Introduction** 18](#_Toc135135222)

[**4.2 Interfaces** 18](#_Toc135135223)

[**Chapter five: Conclusion & Future work** 21](#_Toc135135224)

[**5.1 Conclusion** 22](#_Toc135135225)

[**5.2 Future Work** 22](#_Toc135135226)

[**References** 23](#_Toc135135227)

# **Chapter One: Introduction**

## **Introduction**

With wide population of networks around the world the risk become more serias where it effect not only peaple now , but also effect fenincials infstracture and big companies and also involved in medical health care information and more , so the need is arised for new approch to protect our network .

The Intrusion Detection System took big place in that protection cause now ids used in varity places and varity types of networks and framworks . The evilution of IDS make it now be able to prevent attacks now only detect it, which make it nesecery in every network needs protection .

In past time IDS was depend on his work on static values to take decision on pakets and there types , after evolutaion of attack and the new approches it takes , the need of new ways is arised.

The machine learning is envolved in IDS working mechanism , it can help to classify deferent types of attack , it is fast and accurate. Now devloper use Machine leanning algorithm & methods like dession tree , knn , deep learning... etc .

In this project we use **decision tree** and **J48** algorithm with the IDS , also we used offline dataset created by (university of new brunswick) called (CICIDS2017\_Sample) which containing information of network traffic contain 7 types of classification (Normal , Dos , Bot , Bruteforce , portscan , Webattack, filtration).

## **Problem Description**

With the increasing frequency and sophistication of cyberattacks, organizations need reliable and efficient intrusion detection systems to safeguard their networks against various threats, such as denial-of-service attacks, portscan attacks, web attacks, botnet, and unauthorized access attempts. The J48 algorithm, due to its simplicity and effectiveness, can be a suitable tool for building such systems. However, creating an IDS using J48 involves various challenges, such as selecting appropriate datasets, feature engineering, training and tuning the algorithm, handling evolving and unknown attacks, scalability, and performance. Addressing these challenges is crucial to develop a robust and effective intrusion detection system that can detect and prevent potential threats to the network.

## **The Aim of project**

This project aims to use machine learnnig in classifey traffic packets in classes depending on trees made by the algorithm (J48 Classifier), this trees havve been built by selecting features depends on how much they effect on the decision after some opration mentioned later .

The project will classifiey the packet into classes depends on their features and impurity masures (entropy and gini index)and other data related to it.

## **Related Work**

MeeraGandhi, G., Appavoo, K., & Srivasta, S. (2010) [12] evaluate the performance of a set of classifier algorithms of rules (JRIP, Decision Tree, PART, and OneR) and trees (J48, Random Forest, REPTree, NBTree). Based on the evaluation results, best algorithms for each attack category are chosen and two classifier algorithm selection models are proposed.

Sahu, S., & Mehtre, B. M. (2015, August) [13] used a new labelled network dataset, called Kyoto 2006+ dataset. In Kyoto 2006+ data set, every instant is labelled as normal (no attack), attack (known attack) and unknown attack. We use Decision Tree (J48) algorithm to classify the network packet that can be used for NIDS. For training and testing we used 134665 network instances. The generated rules works with 97.2% correctness for detecting the connection i.e., no attack, known attack or unknown attack.

Alkasassbeh, M., & Almseidin, M. (2018) [14] This research work illustrates how the Knowledge Discovery and Data Mining (or Knowledge Discovery in Databases) KDD dataset is very handy for testing and evaluating different Machine Learning Techniques. It mainly focuses on the KDD preprocess part in order to prepare a decent and fair experimental data set. The J48, MLP, and Bayes Network classifiers have been chosen for this study. It has been proven that the J48 classifier has achieved the highest accuracy rate for detecting and classifying all KDD dataset attacks, which are of type DOS, R2L, U2R, and PROBE.

## **Outline**

Chapter Two is going to discuss the theory part. IDS , Machine learning , Decision Tree

Chapter Three is about the practical part of the project and going to discuss the algorithms and what type of dataset we used and the design of system .

Chapter Four demonstrates the implementations and how to get the results.

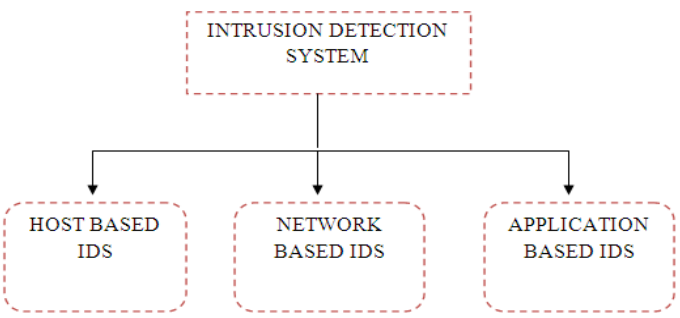
Chapter Five is the conclusion and the future work for the project.

# **Chapter Two: Theoretical Background**

## **2.1 Introduction To IDS**

An Intrusion Detection System is an application used for monitoring the network and protecting it from the intruder. With the rapid progress on the internet-based technology new application areas for computer network have emerged. In instances, the fields like business, finance, industry, security and healthcare sectors the LAN and WAN applications have progressed. All of these application areas made the network an attractive target for the abuse and a big vulnerability for the community [1]. As the internet emerging into the society, new stuffs like viruses and worms are imported. The malignant so, the users use different techniques like cracking of password, detecting unencrypted text are used to cause vulnerabilities to the system. Hence, security is needed for the users to secure their system from the intruders. IDS are used in network related activities, medical applications, credit card frauds, Insurance agency [2].

## **2.2 History Of IDS**

Intrusion detection concept was introduced in early 1980’s after the evolution of internet with surveillance end monitoring the threat [3]. There was a sudden rise in reputation and incorporation in security infrastructure. Since then, several events in IDS technology have advanced intrusion detection to its current state [4]. Then the detection appeared, and audit data and its importance led to terrific improvements in the subsystems of every operating system [4]. IDS and Host Based Intrusion Detection System (HIDS) were first defined. In 1983, SRI International and Dorothy Denning began working on a government project that launched a new effort into intrusion detection system development [3]. Around 1990s the revenues are generated, and intrusion detection market has been raised. Real secure is an intrusion detection network developed by ISS. After a year, Cisco recognized the priority for network intrusion detection and purchased the Wheel Group for attaining the security solutions [3]. The government actions like Federal Intrusion Detection Networks (FID Net) were designed under Presidential Decision Directive 63 is also adding impulse to the IDS [3].

**Figure1. IDS Types**

## **2.3 Machine Learning**

a branch of artificial intelligence that allows machines to evolve behaviors based on empirical data using algorithms and focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine learning is an application of AI that enables systems to learn and improve from experience without being explicitly programmed. Machine learning focuses on developing computer programs that can access data and use it to learn for themselves. Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

* Supervised ML
* Unsupervised ML
* Semi-Supervised ML
* Reinforcement Learning

### **2.3.1 Supervised Learning with Classification**

In this type of learning, the machine is provided with a given set of inputs with their desired outputs. The machine needs to study those given sets of inputs and outputs and find a general function that maps inputs to desired outputs. Supervised learning is inferring a function from a given set of data (inputs with their respective outputs). The training data consists of a set of examples with which the computer is trained. Each example, a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm supervises the training data and produces a general rule (function), which can be used for mapping new inputs [5]. Supervised learning comes in two different types:

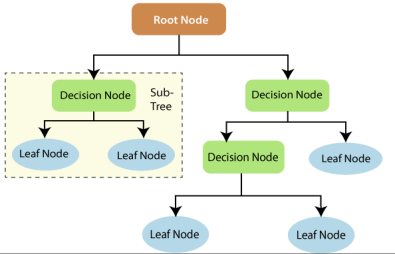
* Regression: The target output is a real number or a whole vector of real numbers.
* Classification: The target output is a class label like in the simplest case choosing between positive and negative. [6]

### **2.3.2 Unsupervised Learning with Classification**

Unsupervised learning studies how systems can learn to represent input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with **SUPERVISED LEARNING** or **REINFORCEMENT LEARNING**, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output. [6]

### **2.3.3 Decision Tree Algorithm**

One of the widely used techniques in machine learning is systems that create classifiers [7]. In machine learning, classification algorithms are capable of handling a vast volume of information. It can be used to make assumptions regarding categorical class names, to classify knowledge on the basis of training sets and class labels, and to classify newly obtainable data [8]. Classification algorithms in machine learning contain several algorithms, and in this work. Figure2. illustrate a structure of decision tree algorithm,



**Figure2. Decision Tree Structure**

## **2.4 IDS with Machine Learning**

The idea of applying machine learning techniques for intrusion detection is to automatically build the model based on the training data set. This data set contains a collection of data instances each of which can be described using a set of attributes (features) and the associated labels. The attributes can be of different types such as categorical or continuous. The nature of attributes determines the applicability of anomaly detection techniques. For example, distance-based methods are initially built to work with continuous features and usually do not provide satisfactory results on categorical attributes. The labels associated with data instances are usually in the form of binary values i.e., normal and anomalous. In contrast, some researchers have employed different types of attacks such as DoS, Brute Force, Port Scan and Probe rather than the anomalous label. This way learning techniques can provide more information about the types of anomalies. However, experimental results show that current learning techniques are not precise enough to recognize the type of anomalies. Since labeling is often done manually by human experts, obtaining an accurate labeled data set which is representative of all types of behaviors is quite expensive. As a result, based on the availability of the labels, there are three operating modes defined for anomaly detection techniques: Supervised Learning, Unsupervised Learning, semi supervised Learning.

### **2.4.1 J48 (C4.5 Decision Tree) Classifier**

The J48 algorithm is the most popular tree classifier. It was developed by Quinlan [10]. It is an ID3 algorithm extension which uses a predictive ML model. The J48 algorithm uses the improved tree pruning technique to reduce the number of classification errors. In addition, the J48 algorithm adopts a dividing-and-conquer greedy approach for inducing recursively the decision trees that contain the features of the dataset for performing an additional classification. The J48 classifier algorithm is divided into datasets based on the attribute values of data to distinguish the probable prediction. The J48 classification algorithm will build its decision tree based on the theoretical attribute values of the present training data. Furthermore, in the J48 algorithm, each feature calculates the gain value separately. The estimation process proceeds until the process of prediction gets completed. A suitable feature is defined as the function that offers much information about the data instances. Several studies aimed at exploring the impact of using the J48 algorithm to improve the accuracy of IDSs [11].

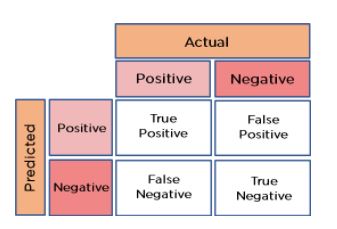
### **2.4.2 CICIDS 2017 Dataset**

Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) are the most powerful defense tools against sophisticated and ever-growing network attacks. Due to the lack of reliable test and validation datasets, anomaly-based intrusion detection approaches suffer from consistent and accurate performance development. The attacks included Brute Force attack, Heartbleed/ Denial-of-service (DoS), Web Attack, filtration, Bot, PortScan.

## **2.5 Evaluation Methods**

For assessing the efficiency level of the proposed model, the following metrics employ several features. These metrics are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [9].

The confusion matrix—as displayed in Figure3 calculates True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).



**Figure3. Confusion Matrix**

Accuracy is represented in a percentage. It refers to the degree to which the instances are predicted correctly. It is calculated as follows:

Accuracy=**TP+TN / TP+TN+FP+FN**

# **Chapter Three: Practical**

## **3.1 System Architecture**

An IDS module that used Machine Learning algorithm (DT – J48) and Dataset containing information about a Network traffic (CICIDS2017\_Sample.csv) to calculate Accuracy & Cross Validation mean with different splits size or different criterion.

This system needs a manual pre-process data go throw multi step, at first change the value from inf to 0, second step is deleted any row that containing missing values, third step converting multi categorical classification to multi numerical classification, forth step change the name of decision column to Label.

After the pre-process is complete, we can now upload the dataset to the module and train it to be able to create decisions based on training data, the value of decisions used to calculate the accuracy with original data.

Also, it can test individual samples in other datasets to make predictions based on what model is learned from training process.

CICIDS2017

IDS DATASET

56,662 Sample

77 features

7 classes

Data Splitting

(Holds Out)

Missing Value Processing

Apply Entropy

And

Gini Index

J48 Classifier

Makes Decisions

Multi classes

Embedded

Features

Selection

Evaluate

## **3.2 Dataset**

The dataset used in this project is (CICIDS2017\_sample), this dataset created by university of New Brunswick and collected by traffic of network for one week, it is containing 56,662 record and 77 features with 7 classification of the pakets that been collected . it needed a pre-process to fix missing value or inf values.

## **3.3 Building Step**

The system consists of two main functions, Evaluation function and Test function.

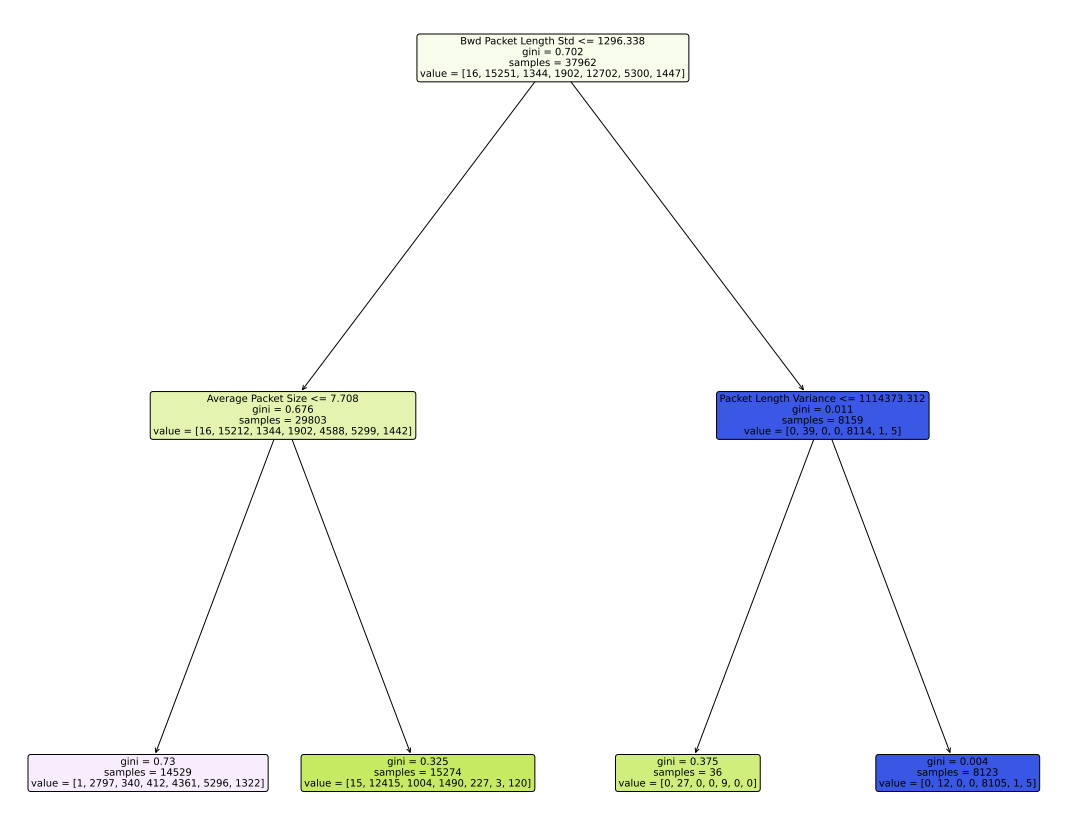
**Evaluation function:**

1. This function has five parameters which are: path of dataset, criterion type, training size, testing size, print option.
2. The function split dataset sample into train and test splits based on given value.
3. The function builds a classifier based on (j48) decision tree algorithm and uses criterion type based on given choice of user.
4. The function start fitting the values in the classifier to start training process.
5. When finish training is finished, the function starts the prediction process with the test split that is mentioned earlier.
6. The values that come from prediction in compared with actual value that is saved in dataset to calculate the accuracy.
7. The function then calculates the mean value of cross validation of different portions.
8. The result of accuracy and cross validation is returned to the user.

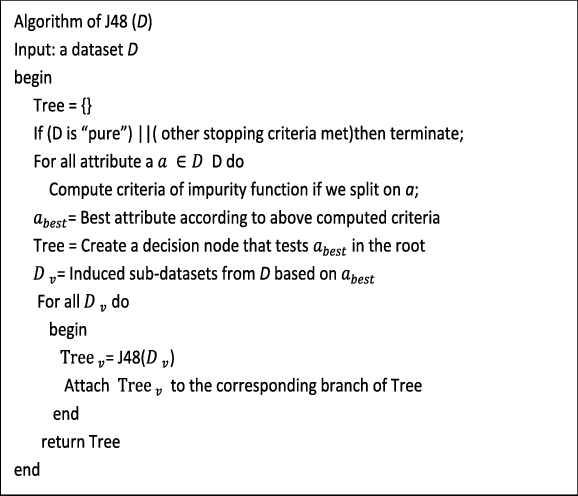
**Test function:**

1. This function has three parameters which are: path to train dataset, criterion type, path to test dataset.
2. The function builds a classifier based on (j48) decision tree algorithm and uses criterion type based on given choice of user.
3. The function start fitting the values of training dataset in the classifier to start training process.
4. When finish training is finished, the function reads the test dataset that was chosen by the user earlier.
5. The function starts the prediction process with test values to get the result.
6. The prediction that comes from the algorithm is returned to the user.

## **3.4 Simple Decision Tree Constructure**



## **3.5 J48 Algorithm**

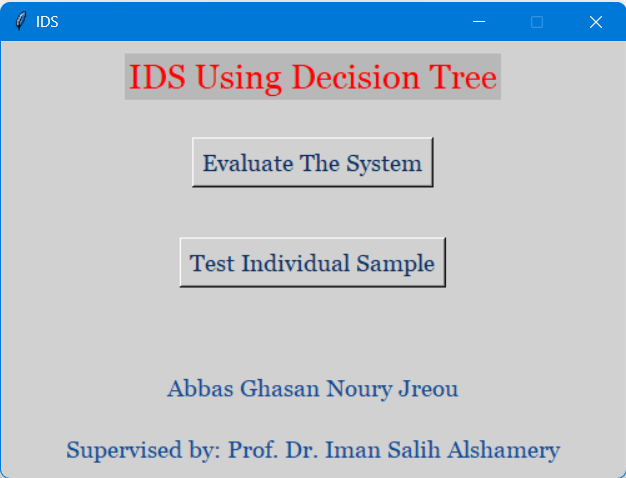


# **Chapter Four: Implementation and results**

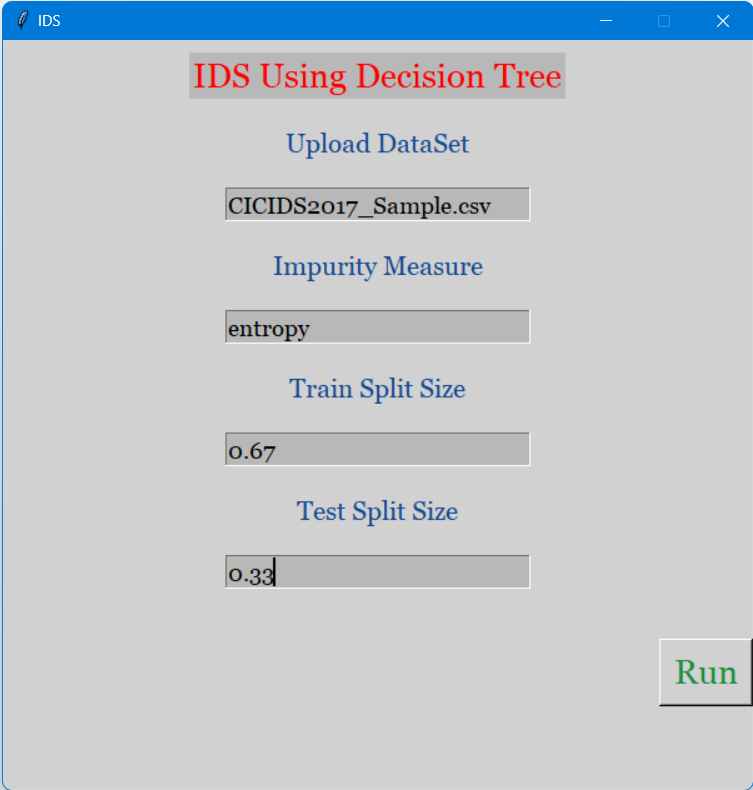
## **4.1 Introduction**

This project has four main Forms that complete the interface of the project, it written with python using tkinter library which provide widget to the user.

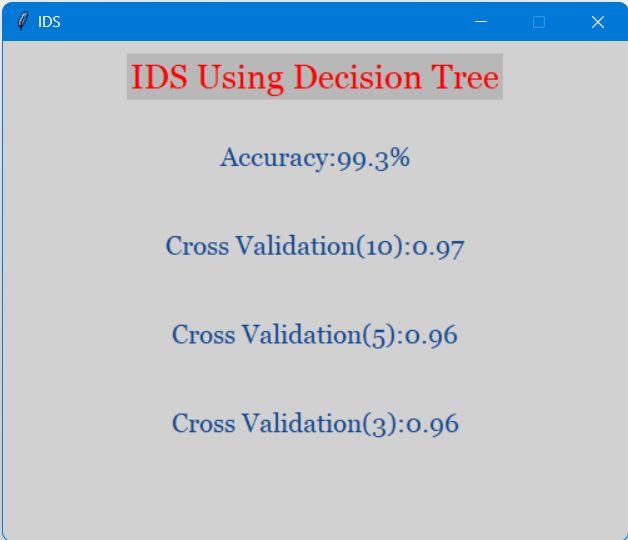
## **4.2 Interfaces**



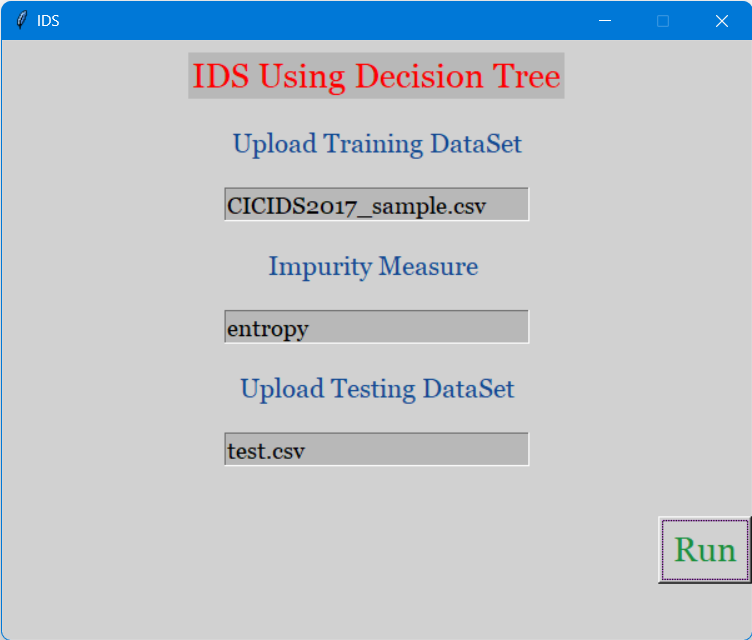
First page that consider as welcome page and give choice to the user if he/she want to Evaluate the system or test Individual Sample



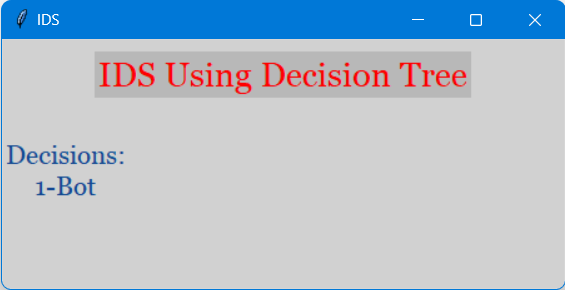
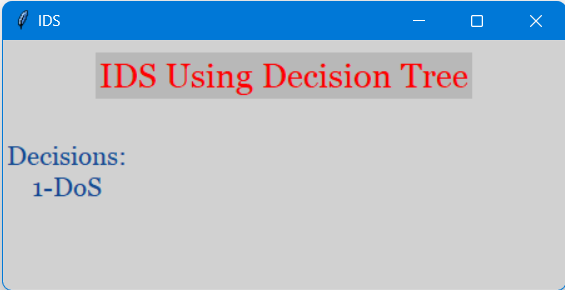
Evaluation page when user specified the parameter to the Evaluation function to start process and go to the results page.



The result returned after evaluation is complete represented in result Form.



Individual testing page to test sample from external dataset and give decisions.

Decisions Based on learning dataset and algorithm about the testing dataset given by the user in previous page.

# **Chapter five: Conclusion & Future work**

## **5.1 Conclusion**

1. Decision tree has proven its efficiency in finding the true prediction.
2. Using entropy as impurity measure gives better results than Gini Index.
3. The higher the number of Samples, the higher the efficiency of the system, but lower speed.
4. Imbedded Feature selection in the classifier reduces the time of training.

## **5.2 Future Work**

1. Implementing the system Realtime.
2. Optimize the J48 classifier using genetic algorithm in feature selection.
3. Applying the classifier on another types of attacks

# **References**

1. PeymanKabiri and Ali A. Ghorbani - “Research on Intrusion Detection and Response Survey”- International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005
2. Christopher Low – “Understanding Wireless attacks &detection “-GIAC Security Essentials Certification (GSEC) Practical Assignment 13 April 2005 -SANS Institute InfoSec Reading Room
3. Paul Innella- “The Evolution of Intrusion Detection Systems “-Tetrad Digital Integrity, LLC. International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, February 2015
4. Asmaa Shaker Ashoor, Prof. Sharad Gore – “Importance of Intrusion Detection System”-International Journal of Scientific & Engineering Research, Volume 2, Issue 1, January-2011.
5. R. Sathya, Annamma Abraham, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification”, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013.
6. Thomas G. Dietterich, “Machine-Learning Research”, AI Magazine Volume 18 Number 4 (1997).
7. R. Kumar and R. Verma, “Classification algorithms for data mining: A survey,” International Journal of Innovations in Engineering and Technology (IJIET), vol. 1, no. 2, pp. 7–14, 2012.
8. S. S. Nikam, “A comparative study of classification techniques in data mining algorithms,” Oriental journal of computer science & technology, vol. 8, no. 1, pp. 13–19, 2015.
9. Smadi, S.; Aslam, N.; Zhang, L. Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. Decis. Support Syst. 2018, 107, 88–102.
10. Quinlan, J.R. C4. 5: Programs for Machine Learning; Elsevier: Amsterdam, The Netherlands, 2014.
11. Aljawarneh, S.; Yassein, M.B.; Aljundi, M. An enhanced J48 classification algorithm for the anomaly intrusion detection systems. Clust. Comput. 2019, 22, 10549–10565.
12. MeeraGandhi, G., Appavoo, K., & Srivasta, S. (2010). Effective network intrusion detection using classifiers decision trees and decision rules. Int. J. Advanced network and application, Vol2.
13. Sahu, S., & Mehtre, B. M. (2015, August). Network intrusion detection system using J48 Decision Tree. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2023-2026). IEEE.
14. Alkasassbeh, M., & Almseidin, M. (2018). Machine learning methods for network intrusion detection. arXiv preprint arXiv:1809.02610.

الخلاصة

**مع التزايد المستمر للأخطار السيبرانية بكافة انواعها وانتشارها بسرعة خلال الشبكات و الانترنت ظهرت الحاجة الى ايجاد حلول فعالة و كان منها انظمة اكتشاف المتطفلين، لكن بسبب ظهور تهديدات جديدة و متنوعة مع مرور الزمن احتجنا الى زيادة كفاءة هذه الانظمة التي كانت في السابق معتمدة على قيم احصائية فقط في اتخاذ قراراتها.**

**في هذا المشروع سنقوم بتوظيف تعلم الآلة في مجال اكتشاف المتطفلين و لأن الهدف الاساسي من هذا المشروع هو زيادة أمنية الشبكات الرقمية، و خصوصاً قمنا باستعمال شجرة القرار (j48 classifier) مع مقياس نقاوة (entropy) و (Gini Index) في عملنا حيث تعد هذه الخوارزمية ذات سرعة و دقة عالية و ايضاً تم الاستعانة ببيانات خاصة بحركة شبكة خلال ايام الاسبوع تسمى (CICIDS 2017) و التي تحتوي على 56662 من القيود و 77 من الfeatures في تدريب النظام و عند وضع النظام تحت الاختبار اظهر في ايجاد قراراته نسب دقة عالية حسب الآتي :**

**Entropy (67% train, 33% test): 99.4%**

**Entropy (50% train, 50% test): 99.2%**

**Gini index (67% train, 33% test): 99.2%**

**Gini index (50% train, 50% test): 99.3%**

**وكذلك تم تقليص عدد الـfeatures و تحديدها من قبل النظام الى (60) مع تحديد طول عقد الشجرة ب10 مستويات فقط**

**مما اتاح لنا هذا الامر توظيف هذا النظام في حماية الشبكات وزيادة امنيتها وكذلك يعد من الانظمة القابلة للتطور السريع في المستقبل**



وزارة التعليم العالي والبحث العلمي

جامعة بابل

كلية تكنولوجيا المعلومات

قسم امنية المعلومات

**نظام اكتشاف المتطفلين باستعمال شجرات القرار**

**مشروع التخرج و المقدم لقسم أمنية المعلومات في كلية تكنولوجيا المعلومات في جامعة بابل و هو احد المتطلبات للحصول على درجة البكالوريوس في أمنية المعلومات في تكنولوجيا المعلومات**

***عباس غسان نوري جريو***

**المشرف:   
ا. د. ايمان صالح الشمري**

٢٠٢٣