

ANOMALY DETECTION ON SELF-ORGANIZING NETWORKS: CONVENTIONAL VERSUS CONTEMPORARY MACHINELEARNING

Submitted in partial fulfillment of the
requirements for the award of
Bachelor of Technology degree in Information Technology

By

SAIHANEESH S (Reg.No – 39120098)

NIKHIL K (Reg.No - 39120075)



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

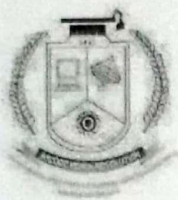
SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A” by NAAC

**JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119**

May - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —A|| grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **SAIHANEESH S(Reg.No - 39120098)** and **NIKHIL K(Reg.No -39120075)** who carried out the Project Phase-1 entitled "**ANOMALY DETECTION IN SELF-ORGANIZING NETWORKS: CONVENTIONAL VERSUS CONTEMPORARY MACHINE LEARNING**" under my supervision from January 2023 to May 2023.

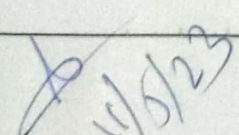

Internal Guide

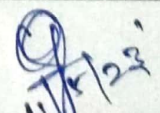
Dr. P.JEYANTHI, M.E., Ph.D


Head of the Department

Dr. R.SUBHASHINI, M.E., Ph.D.

Submitted for Viva voce Examination held on 11-05-23


Internal Examiner


External Examiner

DECLARATION

I, SAIHANEESH S (Reg.No- 39120098), hereby declare that the Project Phase-1 Report entitled "ANOMALY DETECTION IN SELF-ORGANIZING NETWORKS: CONVENTIONAL VERSUS CONTEMPORARY MACHINE LEARNING" done by me under the guidance of Dr. P.JEYANTHI, M.E., Ph.D is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Information Technology.

DATE: 11-05-23

PLACE: Chennai

S. Sai Haneesh
SAIHANEESH S

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr.R.SUBHASHINI, M.E., Ph.D.**, Head of the Department of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. P.JEYANTHI M.E.,Ph.D.**,for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-1 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

With the evolution in wireless communication, there are many security threats over the internet. The intrusion detection system (IDS) helps to find the attacks on the system and the intruders are detected. Previously various machine learning (ML) techniques are applied on the IDS and tried to improve the results on the detection of intruders and to increase the accuracy of the IDS. This paper has proposed an approach to develop efficient IDS by using the principal component analysis (PCA) and the random forest classification algorithm. Where the PCA will help to organize the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more efficiently in terms of accuracy as compared to other techniques like SVM, Naive Bayes, and Decision Tree. The results obtained by proposed method are having the values for performance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %.

Chapter No	TITLE		Page No.
	ABSTRACT		5
	LIST OF FIGURES		6
	LIST OF TABLES		7
1	INTRODUCTION		8
2	LITERATURE SURVEY		21
	2.1	A Proposed Wireless Intrusion Detection Prevention and Attack System	21
	2.2	Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm	21
	2.3	On the Selection of Decision Trees in Random Forests	22
	2.4	Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction	23
	2.5	The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection	24
3	REQUIREMENTS ANALYSIS		25
	3.1	System Requirements	25
	3.2	System Architecture	26
4	DESCRIPTION OF PROPOSED SYSTEM		27
	Module Description		27
	4.1	Data collection Module	27
	4.2	Data Preparation Module	29
	4.3	Analyze and Prediction Module	30
5	IMPLEMENTATION DETAILS		33
	5.1	ALGORITHMS	33
	5.2	TESTING	34
6	RESULT AND DISCUSSION		36
7	CONCLUSION		37
	REFERENCES		38

	APPENDIX	41
	A. SOURCE CODE	41
	B. SCREENSHOTS	44
	C. RESEARCH PAPER	59

FIG_NO	FIGURE NAME	Page No.
1.1	Traditional Programming	9
1.2	Machine Learning	9
1.3	Learning Phase	10
1.4	Inference from Model	11
1.5	Machine learning Algorithms	12
3.1	System Architecture	27

CHAPTER 1

INTRODUCTION

1.1 What is Machine Learning?

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights.

The breakthrough comes with the idea that a machine can singularly learn from the data(i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.

A typical machine learning task is to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

1.1.1 ***Machine Learning vs. Traditional Programming***

Traditional programming differs significantly from machine learning. In traditional programming, a programmer codes all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When

the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

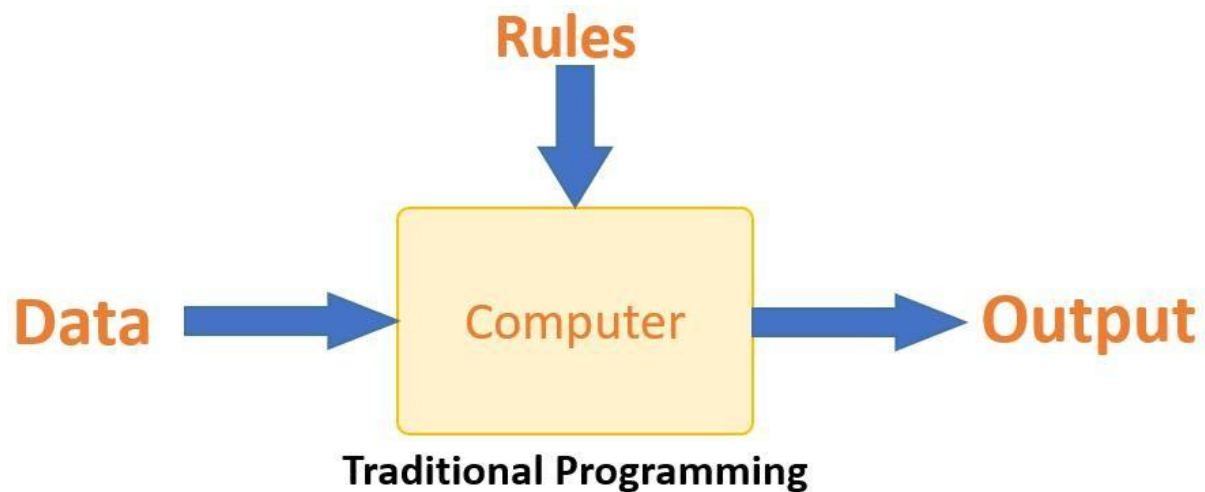


Fig.1.1 Traditional Programming

Machine learning is supposed to overcome this issue. The machine learns how the input and output data are correlated and it writes a rule. The programmers do not need to write new rules each time there is new data. The algorithms adapt in response to new data and experiences to improve efficacy over time.

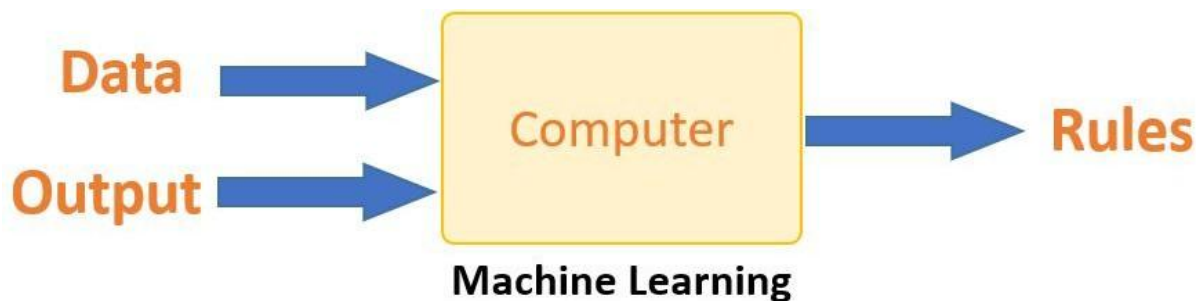


Fig.1.2 Machine Learning

1.1.2 How does Machine Learning Work?

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the **learning** and **inference**. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the **data**. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a **feature vector**. You can think of a feature vector as a subset of data that is used to tackle a problem.

The machine uses some fancy algorithms to simplify the reality and transform this discovery into a **model**. Therefore, the learning stage is used to describe the data and summarize it into a model.

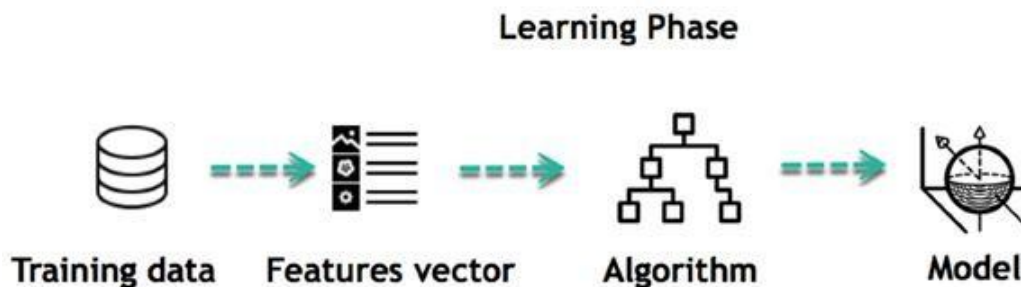


Fig.1.3 Learning Phase

For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model

1.1.3 Inferring

When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

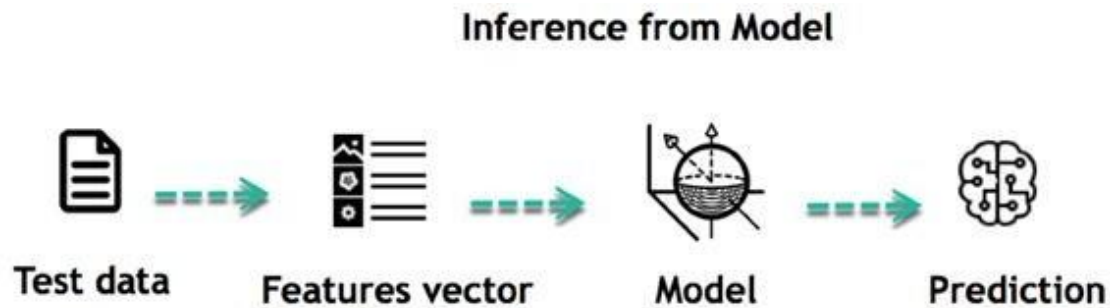


Fig.1.4 Inference from Model

The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

Machine Learning Algorithms and here they are Used?

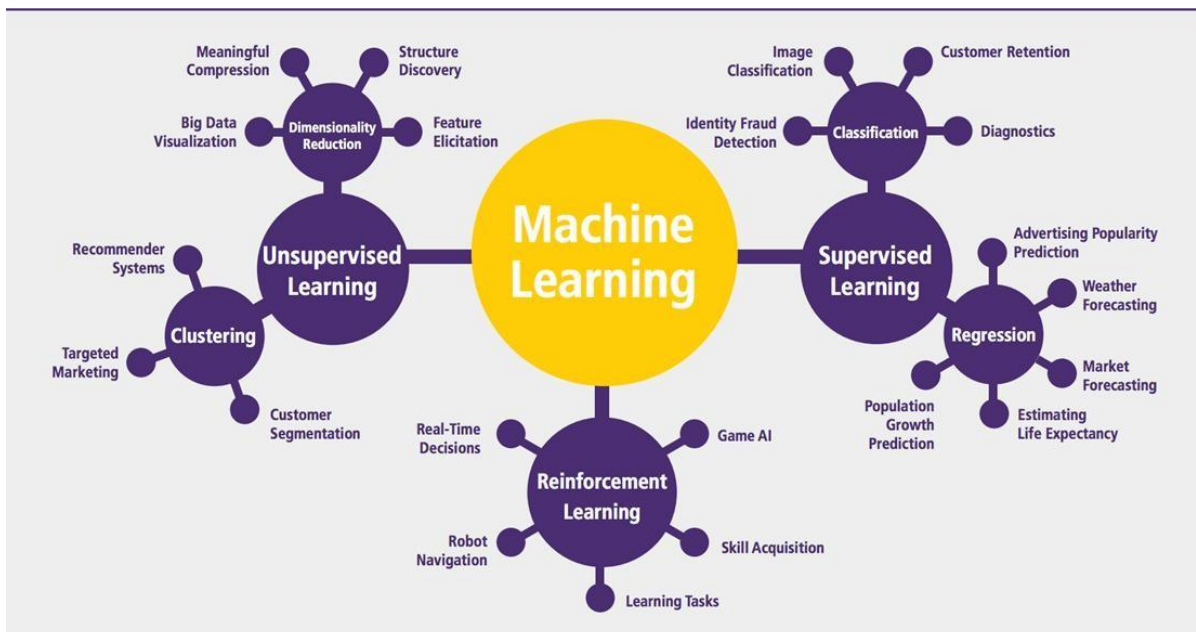


Fig.1.5 Machine learning Algorithms

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms

1.2 SUPERVISED LEARNING

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans. You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

- Classification task
- Regression task

1.3 CLASSIFICATION

Imagine you want to predict the gender of a customer for a commercial. You will start gathering data on the height, weight, job, salary, purchasing basket, etc. from your customer database. You know the gender of each of your customer, it can only be male or female. The objective of the classifier will be to assign a probability of being a male or a female (i.e., the label) based on the information (i.e., features you have collected). When the model learned how to recognize male or female, you can use new data to make a prediction. For instance, you just got new information from an unknown customer, and you want to know if it is a male or female. If the classifier predicts male = 70%, it means the algorithm is sure at 70% that this customer is a male, and 30% it is a female.

The label can be of two or more classes. The above Machine learning example has only two classes, but if a classifier needs to predict object, it has dozens of classes (e.g., glass, table, shoes, etc. each object represents a class)

1.4 REGRESSION

When the output is a continuous value, the task is a regression. For instance, a financial analyst may need to forecast the value of a stock based on a range of feature like equity, previous stock performances, macroeconomics index. The system will be trained to estimate the price of the stocks with the lowest possible error

1.4.1 UNSUPERVISED LEARNING

In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns)

You can use it when you do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you

Algorithm	Description	Type
K-means clustering	Puts data into some groups (k) that each contains data With similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters)	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommendation system	Help to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension Reduction

1.4.2 MACHINE LEARNING (ML) ALGORITHM:

There are plenty of machine learning algorithms. The choice of the algorithm is based on the objective.

In the Machine learning example below, the task is to predict the type of flower among the three varieties. The predictions are based on the length and the width of the petal. The picture depicts the results of ten different algorithms. The picture on the top left is the dataset. The data is classified into three categories: red, light blue and dark blue. There are some groupings. For instance, from the second image, everything in the

upper left belongs to the red category, in the middle part, there is a mixture of uncertainty and light blue while the bottom corresponds to the dark category. The other images show different algorithms and how they try to classified the data.

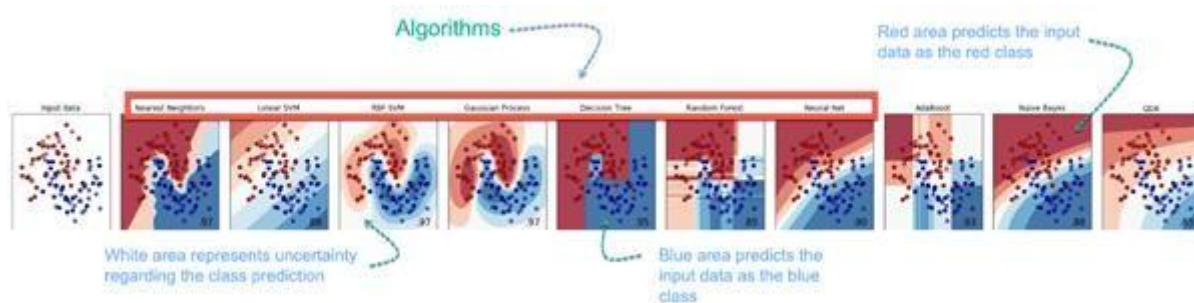


Fig.1.6 Algorithm

1.4.3 CHALLENGES AND LIMITATIONS OF MACHINE LEARNING

The primary challenge of machine learning is the lack of data or the diversity in the dataset. A machine cannot learn if there is no data available. Besides, a dataset with a lack of diversity gives the machine a hard time. A machine needs to have heterogeneity to learn meaningful insight. It is rare that an algorithm can extract information when there are no or few variations. It is recommended to have at least 20 observations per group to help the machine learn. This constraint leads to poor evaluation and prediction.

1.5 APPLICATION OF MACHINE LEARNING

- Machine learning, which assists humans with their day-to-day tasks, personally or commercially without having complete control of the output. Such machine learning is used in different ways such as Virtual Assistant, Data analysis, software solutions. The primary user is to reduce errors due to human bias.

1.5.1 Automation:

- Machine learning, which works entirely autonomously in any field without the need for any human intervention. For example, robots performing the essential process steps in manufacturing plants.

1.5.2 Finance Industry

- Machine learning is growing in popularity in the finance industry. Banks are mainly using ML to find patterns inside the data but also to prevent fraud.

1.5.3 Government organization³

- The government makes use of ML to manage public safety and utilities. Take the example of China with the massive face recognition. The government uses Artificial intelligence to prevent jaywalker.

1.5.4 Healthcare industry

- Healthcare was one of the first industry to use machine learning with image detection.

1.6 MARKETING

- Broad use of AI is done in marketing thanks to abundant access to data. Before the age of mass data, researchers develop advanced mathematical tools like Bayesian analysis to estimate the value of a customer. With the boom of data, marketing department relies on AI to optimize the customer relationship and marketing campaign.

1.6.1 Example of application of Machine Learning in Supply Chain

Machine learning gives terrific results for visual pattern recognition, opening up many potential applications in physical inspection and maintenance across the entire supply chain network.

Unsupervised learning can quickly search for comparable patterns in the diverse dataset. In turn, the machine can perform quality inspection throughout the logistics hub, shipment with damage and wear.

For instance, IBM's Watson platform can determine shipping container damage. Watson combines visual and systems-based data to track, report and make recommendations in real-time.

In past year stock manager relies extensively on the primary method to evaluate and forecast the inventory. When combining big data and machine learning, better forecasting techniques have been implemented (an improvement of 20 to 30 % over

traditional forecasting tools). In term of sales, it means an increase of 2 to 3 % due to the potential reduction in inventory costs.

1.6.2 Example of Machine Learning Google Car

For example, everybody knows the Google car. The car is full of lasers on the roof which are telling it where it is regarding the surrounding area. It has radar in the front, which is informing the car of the speed and motion of all the cars around it. It uses all of that data to figure out not only how to drive the car but also to figure out and predict what potential drivers around the car are going to do. What's impressive is that the car is processing almost a gigabyte a second of data.

1.6.3 Why is Machine Learning Important?

Machine learning is the best tool so far to analyze, understand and identify a pattern in the data. One of the main ideas behind machine learning is that the computer can be trained to automate tasks that would be exhaustive or impossible for a human being. The clear breach from the traditional analysis is that machine learning can take decisions with minimal human intervention.

Take the following example for this ML tutorial; a retail agent can estimate the price of a house based on his own experience and his knowledge of the market.

A machine can be trained to translate the knowledge of an expert into features. The features are all the characteristics of a house, neighborhood, economic environment, etc. that make the price difference. For the expert, it took him probably some years to master the art of estimate the price of a house. His expertise is getting better and better after each sale.

For the machine, it takes millions of data, (i.e., example) to master this art. At the very beginning of its learning, the machine makes a mistake, somehow like the junior salesman. Once the machine sees all the example, it got enough knowledge to make its estimation. At the same time, with incredible accuracy. The machine is also able to adjust its mistake accordingly.

Most of the big company have understood the value of machine learning and holding data. McKinsey have estimated that the value of analytics ranges from \$9.5 trillion

to \$15.4 trillion while \$5 to 7 trillion can be attributed to the most advanced AI techniques.

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

1.7 OVERVIEW

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In

cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. For example, to train a system for the task of digital character recognition, the MNIST dataset of handwritten digits has often been used.

1.7.1 Machine learning approaches

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

1.7.2 Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

1.7.3 Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

1.7.4 Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

Other approaches have been developed which don't fit neatly into this three-fold categorisation, and sometimes more than one is used by the same machine learning system. For example topic modeling, dimensionality reduction or meta learning.

As of 2020, deep learning has become the dominant approach for much ongoing work in the field of machine learning.

1.7.5 History and relationships to other fields

The term machine learning was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence. A representative book of the machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973. In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal.

Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E ." This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turing's proposal in his paper "Computing Machinery and Intelligence", in which the question "Can machines think?" is replaced with the question "Can machines do what we (as thinking entities) can do?".

Modern day machine learning has two objectives, one is to classify data based on models which have been developed, the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning in order to train it to classify the cancerous moles. Whereas, a machine learning algorithm for stock trading may inform the trader of future potential predictions.

CHAPTER 2

LITERATURE SURVEY

2.1 A Proposed Wireless Intrusion Detection Prevention and Attack System

2.1.1 AUTHORS:

1. JafarAbo Nada
2. Mohammad Rasmi Al-Mosa

This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. With the rapid deployment of wireless networks, the concept of network security has faced a lot of risks so it must provide security solutions. The classical methods of protecting networks from attacks are no longer adequate. For example, the intrusion detection system that works with wired networks has become useless with wireless networks. The Wireless technologies have opened a new field for network users. Because of its ease of use and setup, this technology has become popular and changing rapidly. However, the fear of the wireless world and the first threat is security. This is due to the nature of this network. With this increasing concern, it is necessary to start thinking about a security solution. This paper intends to propose a new wireless intrusion detection prevention and attack system to enhance the network security. Therefore, the paper will discuss the development of an intrusion detection system on wireless networks which is Wireless Intrusion Detection Prevention and Attack System “WIDPAS”. It is based on three main tasks: monitoring, analysis and defense. Through which it monitors denial of service attacks or false networks and then analyzes the attack and identifies the attacker and then protects the network users.

2.2 Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm

2.2.1 AUTHORS:

1. Kinam Park
2. Youngrok Song
3. Yun-Gyung Cheong

In this paper, we present the results of our experiments to evaluate the performance of detecting different types of attacks (e.g., IDS, Malware, and Shellcode). We analyze the recognition performance by applying the Random Forest algorithm to the various datasets that are constructed from the Kyoto 2006+ dataset, which is the latest network packet data collected for developing Intrusion Detection Systems. We conclude with discussions and future research projects.

2.3 On the Selection of Decision Trees in Random Forests

2.3.1 AUTHORS:

1. S. Bernard, L. Heutte
2. S. Adam

In this paper we present a study on the random forest (RF) family of ensemble methods. In a "classical" RF induction process a fixed number of randomized decision trees are inducted to form an ensemble. This kind of algorithm presents two main drawbacks : (i) the number of trees has to be fixed a priori (ii) the interpretability and analysis capacities offered by decision tree classifiers are lost due to the randomization principle. This kind of process in which trees are independently added to the ensemble, offers no guarantee that all those trees will cooperate effectively in the same committee. This statement rises two questions: are there any decision trees in a RF that provide the deterioration of ensemble performance? If so, is it possible to form a more accurate committee via removal of decision trees with poor performance? The

answer to these questions is tackled as a classifier selection problem. We thus show that better subsets of decision trees can be obtained even using a sub-optimal classifier selection method. This proves that “classical” RF induction process, for which randomized trees are arbitrary added to the ensemble, is not the best approach to produce accurate RF classifiers. We also show the interest in designing RF by adding trees in a more dependent way than it is traditionally done in “classical” RF induction algorithms.

2.4 Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction

2.4.1AUTHORS:

1. A. Tesfahun
2. D. Lalitha Bhaskari

Intrusion Detection Systems (IDS) have become crucial components in computer and network security. NSL-KDD intrusion detection dataset which is an enhanced version of KDDCUP'99 dataset was used as the experiment dataset in this paper. Because of inherent characteristics of intrusion detection, still there is huge imbalance between the classes in the NSL-KDD dataset, which makes harder to apply machine learning effectively in the area of intrusion detection. In dealing with class imbalance in this paper Synthetic Minority Over sampling Technique (SMOTE) is applied to the training dataset. A feature selection method based on Information Gain is presented and used to construct a reduced feature subset of NSL-KDD dataset. Random Forests are used as a classifier for the proposed intrusion detection framework. Empirical results show that Random Forests classifier with SMOTE and information gain based feature selection gives better performance in designing IDS that is efficient and effective for network intrusion detection.

2.5 The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection

2.5.1 AUTHORS:

1. Le, T.-T.-H., Kang,
2. H., & Kim, H.

A device or software appliance monitors a network or systems for malicious activity is an Intrusion Detection System (IDS). Conventional IDS does not detect elaborate cyber-attacks such as a low-rate DoS attack as well as unknown attacks. Machine Learning has attracted more and more interests in recent years to overcome these limitations. In this paper, we propose a novel method to improve intrusion detection accuracy of Gated Recurrent Unit (GRU) by embedding the proposed PCA-Scale with two options including PCA-Standardized and PCA-MinMax into the layer of GRU. Both optional methods explicitly enforce the learned object feature maps by affecting the direction of maximum variance with positive covariance. This approach can be applied to GRU model with negligible additional computation cost. We present experimental results on two real-world datasets such as KDD Cup 99 and NSL-KDD demonstrate that GRU model trained with PCA-Scaled method achieves remarkable performance improvements.

CHAPTER 3

REQUIREMENTS ANALYSIS

3.1 HARDWARE REQUIREMENTS:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 512 Mb.

3.2 SOFTWARE REQUIREMENTS:

- Operating system : Windows 7.
- Coding Language : Python
- Database : MYSQL

SYSTEM DESIGN

3.4 SYSTEM ARCHITECTURE:

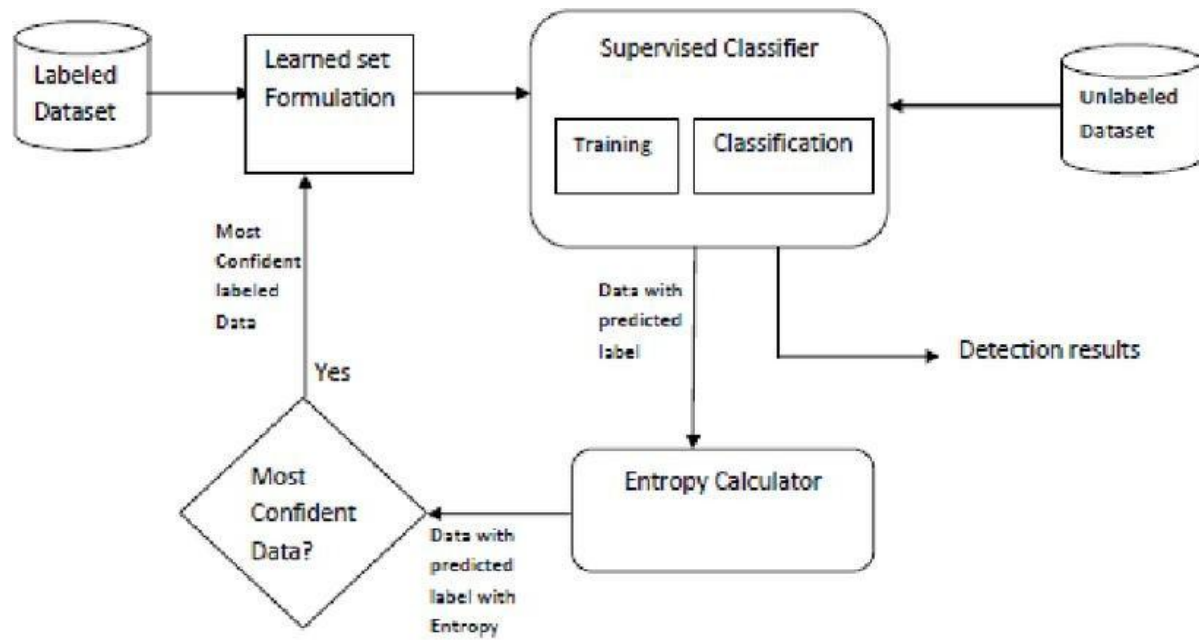


Fig.3.1 ARCHITECTURE DIAGRAM

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

4.1 DATA COLLECTION:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

The dataset used in this Intrusion Detection System dataset taken from kdd

Link: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

4.1.1 Dataset:

The dataset consists of 125974 individual data. There are 42 columns in the dataset, which are described below.

Feature name	Description	Type
Hot	number of “hot” indicators	continuous
Num_failed_logins	number of failed login attempts	continuous
Logged_in	1 if successfully logged in; 0 otherwise	discrete
Num_compromised	number of “compromised” conditions	continuous
Root_shell	1 if root shell is obtained; 0 otherwise	discrete
Su_attempted	1 if “su root” command attempted; 0 otherwise	discrete
Num_root	number of “root” accesses	continuous

Num_file_creations	number of file creation operations	continuous
Num_shells	number of shell prompts	continuous
Num_access_files	number of operations on access control files	continuous
Num_outbound_cmds	number of outbound commands in an ftp session	continuous
Is_hot_login	1 if the login belongs to the “hot” list; 0 otherwise	discrete
Is_guest_login	1 if the login is a “guest”login; 0 otherwise	discrete
Serror_rate	% of connections that have “SYN” errors	continuous
Serror_rate	% of connections that have “REJ” errors	continuous
Same_srv_rate	% of connections to the same service	continuous
Diff_srv_rate	% of connections to different services	continuous
Srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
Srv_serror_rate	% of connections that have “SYN” errors	continuous
Srv_rerror_rate	% of connections that have “REJ” errors	continuous
Srv_diff_host_rate	% of connections to different hosts	continuous

4.2 DATA PREPARATION:

we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

Split into training and evaluation sets

4.2.1 Model Selection

The principal component analysis is the technique that is used, especially for the reduction of the dimension of the given dataset. The principal component analysis is one of the most efficient and an accurate method for reducing the dimensions of data, and it provides the desired results. This method reduces the aspects of the given dataset into a desired number of attributes called principal components.

This method takes all the input as the dataset, which is having a high number of attributes so as the dimension of the dataset is very high. This method reduces the size of the dataset by taking the data points on the same axis. The data points are shifted on a single axis, and the principal components are carried out. The PCA can be performed

1. Take the dataset with all dimensions d .
2. Calculate the mean vector for each dimension d .
3. Calculate the covariance matrix for the whole dataset.
4. Calculate the eigen vectors ($e_1, e_2, e_3 \dots e_d$), and eigen values ($v_1, v_2, v_3, \dots v_d$).
5. Perform sorting of eigenvalue in decreasing order and select n eigenvector with the highest eigenvalues to get a matrix of $d \times n = M$.
6. By using this M form a new sample space.
7. The obtained sample spaces are the principal components.

Random Forest is one of the most powerful methods that is used in machine learning for classification problems. The random forest comes in the category of the supervised classification algorithm. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the classifier.

4.3 ANALYZE AND PREDICTION:

In the actual dataset, we chose only 9 features :

We got a accuracy of 99.1% on test set.

CHAPTER 5

IMPLEMENTATION DETAILS

5.1 ALGORITHMS

PCA (Principal Component Analysis) and Random Forest are two popular machine learning algorithms that can be used for anomaly detection in self-organizing networks. Here is a comparison of using these algorithms for conventional versus contemporary machine learning:

1. **Conventional Machine Learning:** In conventional machine learning, PCA and Random Forest can be used as follows:
 - **PCA:** PCA can be used to reduce the dimensionality of the input data by projecting it onto a lower-dimensional subspace. The resulting projection can then be used as input to a conventional machine learning algorithm, such as a support vector machine (SVM) or logistic regression, for anomaly detection.
 - **Random Forest:** Random Forest is an ensemble learning algorithm that can be used for both classification and regression problems. In the case of anomaly detection, it can be used as a classifier to detect anomalies based on the input data.
1. **Contemporary Machine Learning:** In contemporary machine learning, PCA and Random Forest can be used as follows:
 - **PCA:** In contemporary machine learning, PCA can be used as a preprocessing step to automatically learn the features from the input data. The resulting features can then be used as input to a contemporary machine learning algorithm, such as a deep neural network or autoencoder, for anomaly detection.
 - **Random Forest:** Random Forest can be used in contemporary machine learning as an ensemble method to combine the outputs of multiple contemporary machine learning models, such as deep neural networks or autoencoders, to detect anomalies.

The key difference between using PCA and Random Forest in conventional versus contemporary machine learning is in their usage. In conventional machine learning, they are used as standalone algorithms or in combination with other conventional machine learning algorithms. In contrast, in contemporary machine learning, they are used in combination with other contemporary machine learning algorithms, such as deep neural networks or autoencoders.

5.2 TESTING

Testing is a critical step in the development of anomaly detection systems for self-organizing networks. Here is a brief overview of the testing process for both conventional and contemporary machine learning algorithms:

1. Conventional Machine Learning Algorithms: The testing process for conventional machine learning algorithms involves the following steps:

- Test Data Preparation: Collecting test data from the self-organizing network and preparing it in a suitable format for analysis.
- Model Evaluation: Evaluating the performance of the trained model on the test data using suitable metrics such as accuracy, precision, recall, and F1 score.
- Performance Comparison: Comparing the performance of the trained model with other conventional machine learning algorithms to select the best model.
- Deployment: Deploying the selected model in the self-organizing network to monitor and detect anomalies in real-time.

1. Contemporary Machine Learning Algorithms: The testing process for contemporary machine learning algorithms involves the following steps:

- Test Data Preparation: Collecting test data from the self-organizing network and preparing it in a suitable format for analysis.
- Model Evaluation: Evaluating the performance of the trained model on the test data using suitable metrics such as accuracy, precision, recall, and F1 score.

- Performance Comparison: Comparing the performance of the trained model with other contemporary machine learning algorithms to select the best model.
- Deployment: Deploying the selected model in the self-organizing network to monitor and detect anomalies in real-time.

The key difference between the testing process for conventional and contemporary machine learning algorithms is in the performance comparison step. In conventional machine learning, different algorithms are compared based on their performance on the test data. In contrast, contemporary machine learning algorithms are compared based on their performance on validation data, and the model selection is done based on the validation performance.

CHAPTER 6

RESULT AND DISCUSSION

Conventional Machine Learning: In general, conventional machine learning algorithms, such as PCA and Random Forest, can achieve good accuracy in detecting anomalies in self-organizing networks when the data is well-defined and has a clear separation between normal and anomalous data. However, their accuracy may decrease when the data is complex and has subtle differences between normal and anomalous data. In such cases, the algorithms may require additional feature engineering or the use of more complex models.

Contemporary Machine Learning: Contemporary machine learning algorithms, such as deep neural networks and autoencoders, have shown promising results in detecting anomalies in self-organizing networks. These algorithms can automatically learn relevant features from the input data, which can capture complex relationships between different variables. Additionally, these algorithms can handle high-dimensional and complex data, which may not be possible with conventional machine learning algorithms. However, the accuracy of contemporary machine learning algorithms can be influenced by factors such as the choice of architecture, the number of training examples, and the quality of the training data.

The results obtained by proposed method are having the values for performance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %.

CHAPTER 7

CONCLUSION

As the involvement of the systems over the internet increasing rapidly, the security concerns have also seen. The proposed approach deals with the detection of intruders over the internet efficiently. The proposed algorithm has performed well as compared to the previously applied algorithms such as SVM, Naive Bayes, and Decision Tree. The detection rates and the false error rates can be improved at a great extent by the proposed approach. The dataset used here is the knowledge discovery dataset.

REFERENCES

1. JafarAbo Nada; Mohammad Rasmi Al-Mosa, 2018 International Arab Conference on Information Technology (ACIT), A Proposed Wireless Intrusion Detection Prevention and Attack System
2. Kinam Park; Youngrok Song; Yun-Gyung Cheong, 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigData Service), Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm
3. S. Bernard, L. Heutte and S. Adam "On the Selection of Decision Trees in Random Forests" Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009, 978-1-4244-3553-1/09/\$25.00 ©2009 IEEE
4. A. Tesfahun, D. Lalitha Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction" 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 978-0-4799-2235-2/13 \$26.00 © 2013 IEEE
5. Le, T.-T.-H., Kang, H., & Kim, H. (2019). The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection. 2019 International Conference on Platform Technology and Service (PlatCon). Doi:10.1109/platcon.2019.8668960
6. Anish Halimaa A, Dr K.Sundarakantham: Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) 978-1-5386-9439-8/19/\$31.00 ©2019 IEEE "MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM."
7. Mengmeng Ge, Xiping Fu, Naeem Syed, Zubair Baig, Gideon Teo, Antonio Robles- Kelly (2019). Deep Learning-Based Intrusion Detect ion for IoT Networks, 2019 IEEE

24th Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 256-265, Japan.

8. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, "An Investigation on Intrusion Detection System Using Machine Learning" 978-1-5386-9276-9/18/\$31.00 c2018IEEE.

9. Rohit Kumar Singh Gautam, Er. Amit Doegar; 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms."

10. Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahma, 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)"Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection."

11. L. Haripriya, M.A. Jabbar, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)" Role of Machine Learning in Intrusion Detection System: Review"

12. Nimmy Krishnan, A. Salim, 2018 International CET Conference on Control, Communication, and Computing (IC4) " Machine Learning-Based Intrusion Detection for Virtualized Infrastructures"

13. Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) "Feature extraction using Deep Learning for Intrusion Detection System."

14. Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, Rashmi Bhattad, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)"A Review of Machine Learning Methodologies for Network Intrusion Detection."

15. Iftikhar Ahmad , Mohammad Basher, Muhammad Javed Iqbal, Aneel Rahim, IEEE Access (Volume: 6) Page(s): 33789 – 33795 “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for IntrusionDetection.”

16. B. Riyaz, S. Ganapathy, 2018 International Conference on Recent Trends in Advanced Computing (ICRTAC)” An Intelligent Fuzzy Rule-based Feature Select ion forEffective Intrusion Detection.”

APPENDIX

A. SOURCE CODE

```
import numpy as np
import pandas as pd
from flask import Flask, request, jsonify, render_template, redirect, flash, send_file
from sklearn.preprocessing import MinMaxScaler
from werkzeug.utils import secure_filename
import pickle
from sklearn.decomposition import PCA
```

```
app = Flask(__name__) #Initialize the flask App
```

```
model = pickle.load(open('sk.pkl', 'rb'))
pca = pickle.load(open('kdd.pkl', 'rb'))
```

```
@app.route('/')
@app.route('/')
@app.route('/index')
def index():
    return render_template('index.html')
@app.route('/login')
def login():
    return render_template('login.html')
@app.route('/chart')
def chart():
    return render_template('chart.html')
@app.route('/abstract')
```

```

def abstract():
    return render_template('abstract.html')
@app.route('/performance')
def performance():
    return render_template('performance.html')
@app.route('/future')
def future():
    return render_template('future.html')
@app.route('/upload')
def upload():
    return render_template('upload.html')
@app.route('/preview',methods=["POST"])
def preview():
    if request.method == 'POST':
        dataset = request.files['datasetfile']
        df = pd.read_csv(dataset,encoding = 'unicode_escape')
        df.set_index('Id', inplace=True)
        return render_template("preview.html",df_view = df)

@app.route('/home')

def home():
    return render_template('test.html')

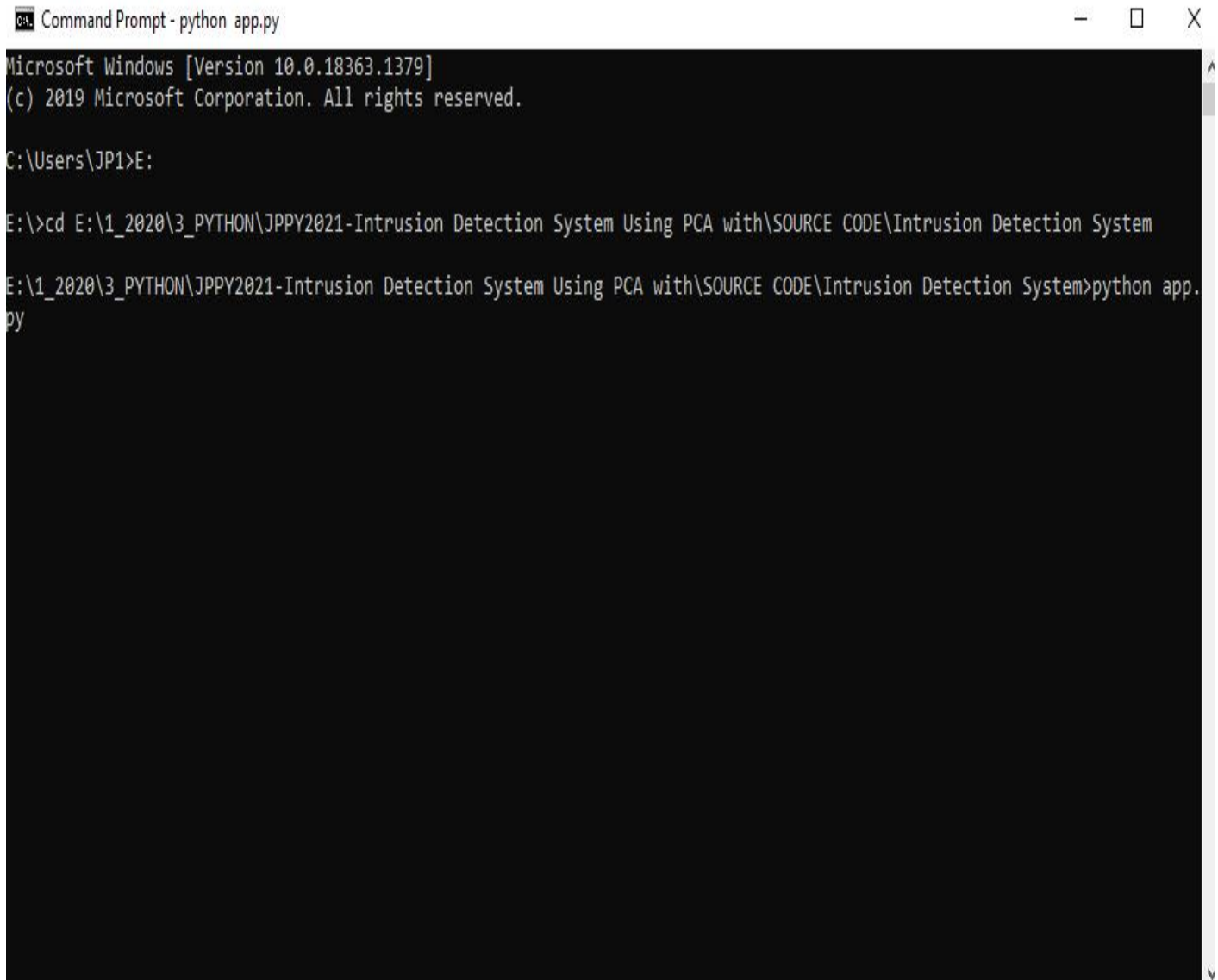
@app.route('/predict',methods=['POST'])
def predict():

```

```
features = [float(x) for x in request.form.values()]
final_features = [np.array(features)]
final_features = pca.transform(final_features)
prediction = model.predict(final_features)
output = prediction
print(output)
return render_template('test.html', prediction_text= output)
```

```
if __name__ == "__main__":
    app.run(debug=True)
```

B. SCREENSHOTS



```
Command Prompt - python app.py
Microsoft Windows [Version 10.0.18363.1379]
(c) 2019 Microsoft Corporation. All rights reserved.

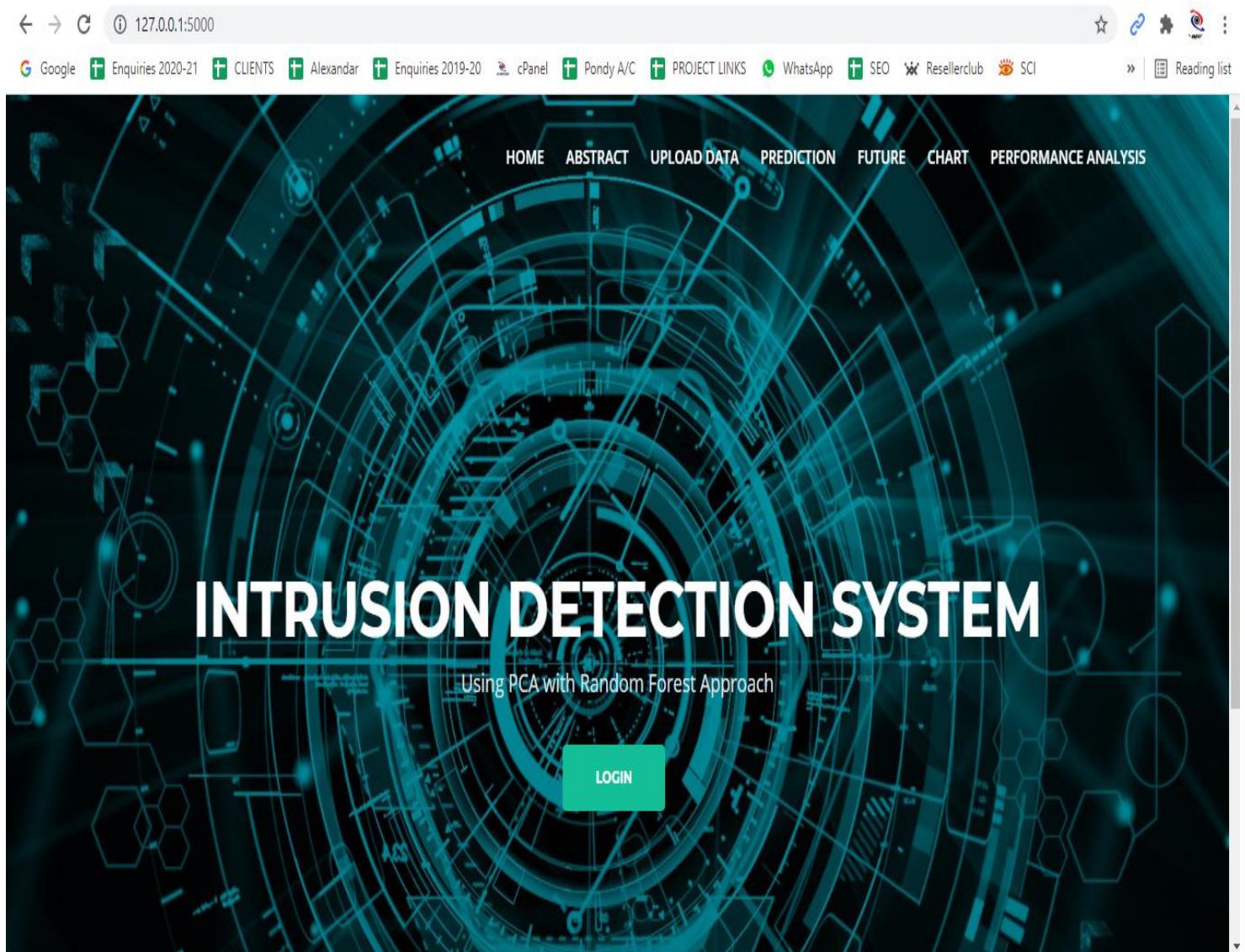
C:\Users\JP1>E:

E:\>cd E:\1_2020\3_PYTHON\JPPY2021-Intrusion Detection System Using PCA with\SOURCE CODE\Intrusion Detection System

E:\1_2020\3_PYTHON\JPPY2021-Intrusion Detection System Using PCA with\SOURCE CODE\Intrusion Detection System>python app.py
py
```

Select Command Prompt - python app.py

```
C:\Users\JP1\AppData\Roaming\Python\Python38\site-packages\sklearn\base.py:313: UserWarning: Trying to unpickle estimator
RandomForestClassifier from version 0.23.2 when using version 0.22.2. This might lead to breaking code or invalid results. Use at your own risk.
  warnings.warn(
C:\Users\JP1\AppData\Roaming\Python\Python38\site-packages\sklearn\base.py:313: UserWarning: Trying to unpickle estimator
PCA from version 0.23.2 when using version 0.22.2. This might lead to breaking code or invalid results. Use at your own risk.
  warnings.warn(
* Serving Flask app "app" (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Restarting with stat
C:\Users\JP1\AppData\Roaming\Python\Python38\site-packages\sklearn\base.py:313: UserWarning: Trying to unpickle estimator
DecisionTreeClassifier from version 0.23.2 when using version 0.22.2. This might lead to breaking code or invalid results. Use at your own risk.
  warnings.warn(
C:\Users\JP1\AppData\Roaming\Python\Python38\site-packages\sklearn\base.py:313: UserWarning: Trying to unpickle estimator
RandomForestClassifier from version 0.23.2 when using version 0.22.2. This might lead to breaking code or invalid results. Use at your own risk.
  warnings.warn(
C:\Users\JP1\AppData\Roaming\Python\Python38\site-packages\sklearn\base.py:313: UserWarning: Trying to unpickle estimator
PCA from version 0.23.2 when using version 0.22.2. This might lead to breaking code or invalid results. Use at your own risk.
  warnings.warn(
* Debugger is active!
* Debugger PIN: 224-664-831
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

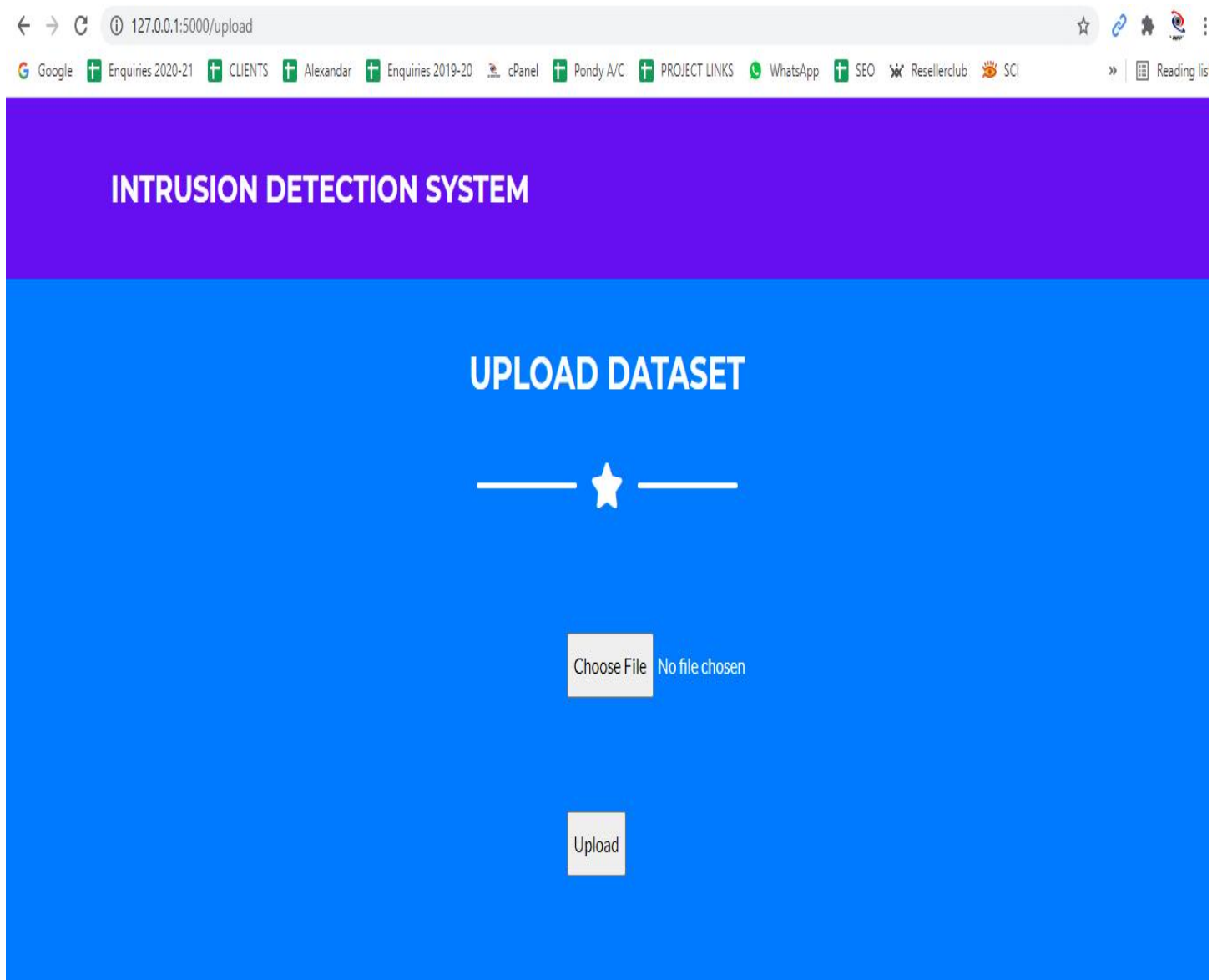


Lgin Page of APP

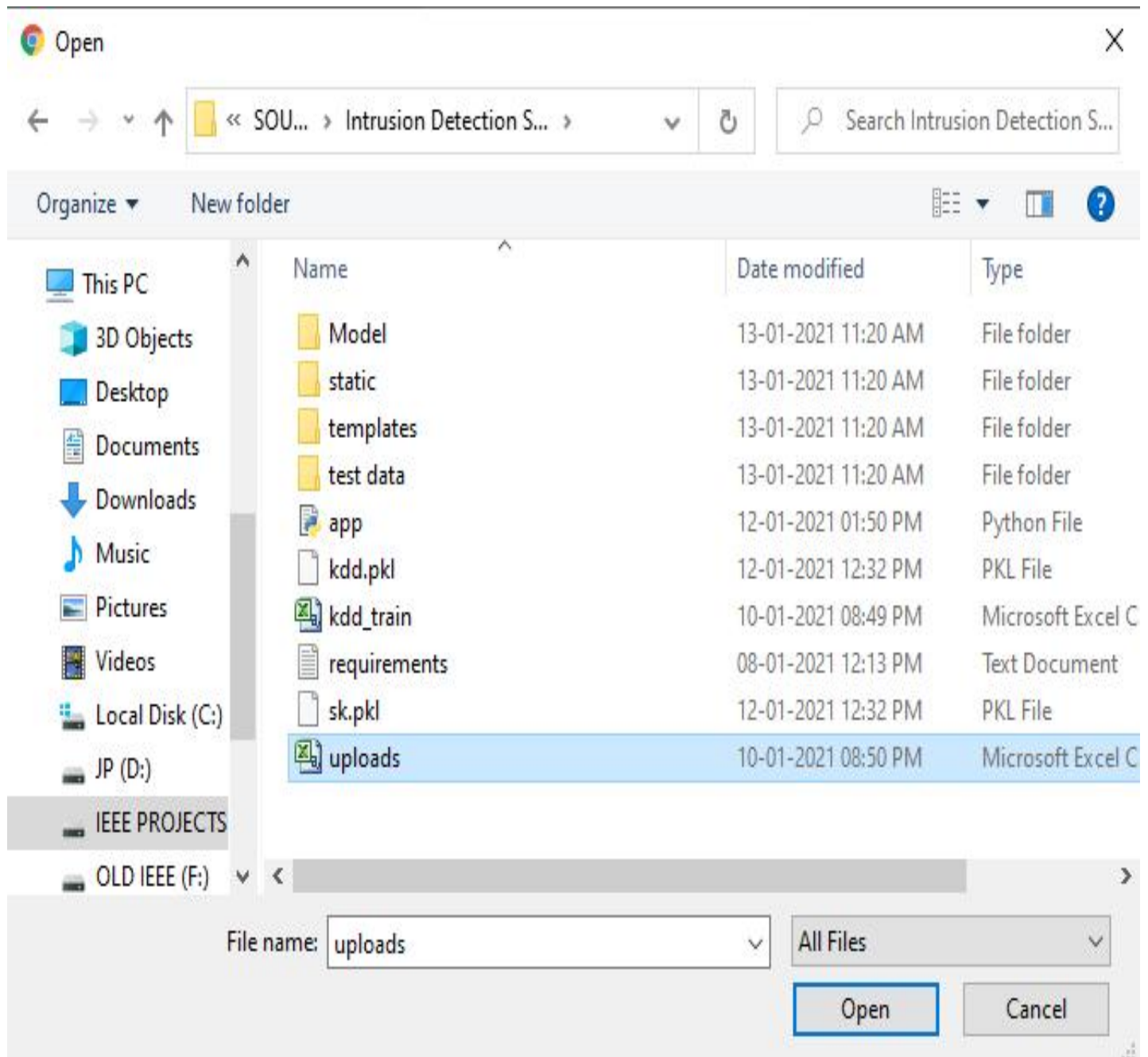
ABSTRACT

With the evolution in wireless communication, there are many security threats over the internet. The intrusion detection system (IDS) helps to find the attacks on the system and the intruders are detected. Previously various machine learning (ML) techniques are applied on the IDS and tried to improve the results on the detection of intruders and to increase the accuracy of the IDS. This paper has proposed an approach to develop efficient IDS by using the principal component analysis (PCA) and the random forest classification algorithm. Where the PCA will help to organise the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more efficiently in terms of accuracy as compared to other techniques like SVM, Naïve Bayes, and Decision Tree. The results obtained by proposed method are having the values for performance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %.

Abstact



CHOOSING DATASET TO UPLOAD



SELECTING DATASET FROM FOLDER

INTRUSION DETECTION SYSTEM

UPLOAD DATASET



Choose File uploads.csv

Upload

UPLOADING DATASET

INTRUSION DETECTION SYSTEM

PREVIEW



	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromise
Id													
1	0	tcp	ftp_data	SF	491	0	0	0	0	0	0	0	0
1	0	udp	other	SF	146	0	0	0	0	0	0	0	0

PREVIEW OF THE DATASET

localhost:5000/preview

Google Enquiries 2020-21 CLIENTS Alexandar Enquiries SEO Resellerclub SCI Reading list

localhost:5000 says
Training finished!

OK

INTRUSION DETECTION SYSTEM

1	0	udp	domain_u	SF	46	82	0	0	0	0	0	0	0
1	1985	udp	other	SF	147	105	0	0	0	0	0	0	0

Click to Train | Test

TRAINING THE DATASET

localhost:5000/home

Google Enquiries 2020-21 CLIENTS Alexandar Enquiries 2019-20 cPanel Pondy A/C PROJECT LINKS WhatsApp SEO Resellerclub SCI Reading list

Intrusion Detection System

Duration:

protocol_type:

src_bytes:

dst_bytes:

is_host_login:

is_guest_login:

diff_srv_rate:

srv_diff_host_rate:

flag:

INPUT TO CHECK THE INTRUSION

Intrusion Detection System

localhost:5000/predict

Intrusion Detection System

Duration:

protocol_type:

src_bytes:

dst_bytes:

is_host_login:

is_guest_login:

diff_srv_rate:

srv_diff_host_rate:

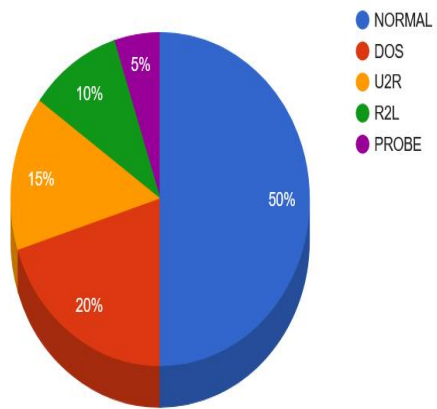
flag:

class is ['attacker']

[Analysis](#)

OUTPUT AFTER DETECTING INTRUSION

Intrusion Detection System(pie chart analysis)



Future

ANALYSIS CHART

PERFORMANCE ANALYSIS

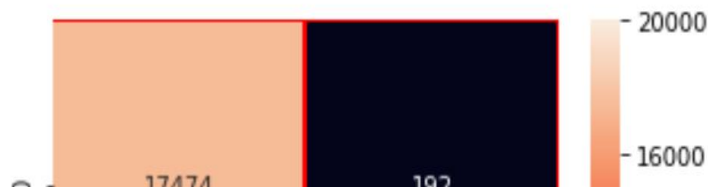
Precision and recall

Recall Precision

NORMAL 0.99 0.099

ATTACKER 0.98 0.095

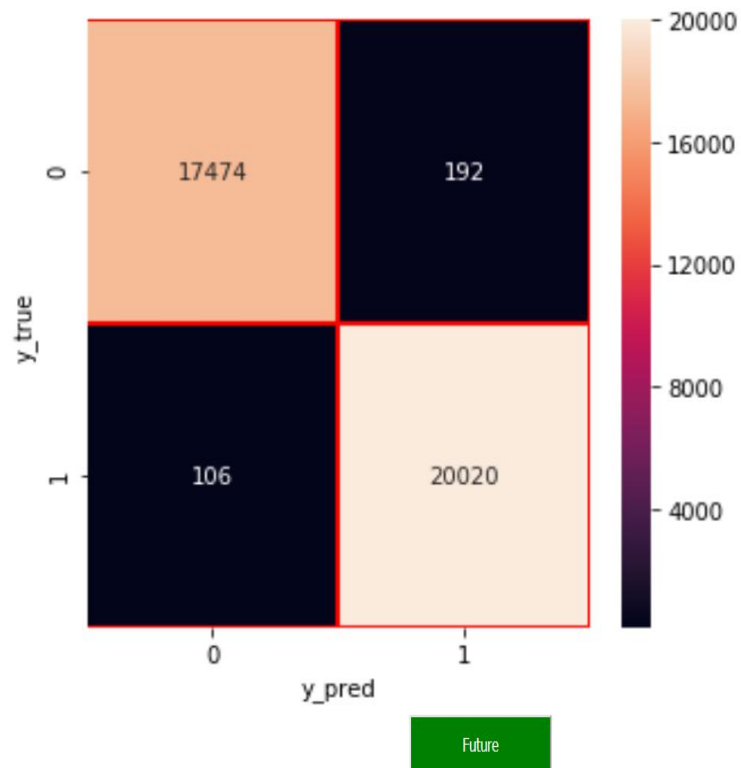
Confusion Matrix



ATTACKER

0.98 0.095

Confusion Matrix



Precision and recall

	Recall	Precision
NORMAL	0.99	0.099
ATTACKER	0.98	0.095

**ANOMALY DETECTION ON SELF-ORGANIZING NETWORKS: CONVENTIONAL
VERSUS CONTEMPORARY MACHINE LEARNING**

Sai Haneesh S

UG Scholar

Department of IT,

Sathyabama Institute of Science and
Technology, Chennai, India

Email: sangarajusaihaneeesh801@gmail.com

Jeyanthi P

Associate Professor

Department of IT,

Sathyabama Institute of Science and
Technology, Chennai, India

Nikhil K

UG Scholar

Department of IT,

Sathyabama Institute of Science
and Technology, Chennai, India

ABSTRACT

With the improvement of wi-fi communications at the Internet, there are many protection threats. An Intrusion Detection System (IDS) enables stumble on attacks on a system and locate intruders. Previously, diverse system studying (ML) techniques have been applied to IDS that have attempted to improve intruder detection effects and improve the accuracy of IDS. This article proposes a technique to implement IDS the use of Principal Component Analysis (PCA) and a random wooded area classification set of rules. Where PCA will help to organize the records by means of lowering the dimensional of the data and Random Forests will help within the category. The outcomes acquired show that the proposed method performs greater efficaciously in terms of accuracy as compared to other methods, consisting of SVM, Naive Bayes and Decision Tree. The results acquired by means of the proposed technique have values for the duration (min) of 3.24 minutes, accuracy (%) of 96.78% and accuracy (%) of 0.21%.

Keywords: Anomaly detection, deep learning, machine learning, mobile network communications, self organizing networks.

INTRODUCTION

Intrusion of a computing system is an attempt to break into or misuse it. An intrusion is any kind of action that compromises the integrity. An intrusion into a pc system is an attempted hacker abuse. Intrusion is any act that violates the integrity, confidentiality or availability of any statistics or pc gadget. Through a potentially overwhelming weak point or flaw within the device structure, an attacker tries to pass the authentication or authorization manner. With the massive boom of network service and facts being included in networks, network security is becoming an increasing number of vital than ever before. One option to this problem is to use a Network Intrusion Detection System (NIDS), which detects assaults by way of watching numerous activities at the community. Therefore, it is extra critical that such structures are extra accurate in detecting assaults, research quickly, and generate as few fake positives as viable. An Intrusion Detection System (IDS) detects malicious anomalies and facilitates defend your networks. Thus, IDS have turn out to be essential for computer networks. IDS requires matters: agility and dexterity. Security is at the top of all plans to save you any loss. A crucial feature of an IDS is to provide information approximately unusual hobby after which alert/screen to tell community directors and/or to inform a suspicious network connection. In addition, IDS can also distinguish between assaults into the organization (from its own personnel or clients or any others) and external attacks (assaults accomplished by way of hackers).

LITERATURE SURVEY

A Proposed Wireless Intrusion Detection Prevention and Attack System

AUTHORS: Jafar AboNada , Mohammad R asmi Al-Mosa

This electronic mail document is a "residing" template and already defines the factors of your article [title, text, headings, etc.] inside the fashion sheet. With the rapid deployment of wireless networks, the concept of network protection has confronted many risks. And therefore must offer security answers. Classical strategies of defensive networks from assaults are not suitable. For example, an intrusion detection gadget that works on stressed out networks is rendered useless on wi-fi networks. Wireless generation has opened a brand new subject for community customers. With its ease of use and customization, this technique has emerge as famous and is changing unexpectedly. But the concern of the earth, and the first worry. The reason for this is due to the decoration. With developing concern, you want to consider a security answer. This article proposes a new intrusion and attack prevention gadget for improving wi-fi networks. Therefore, the object will discuss the development of a wireless intrusion detection device, that's a wi-fi intrusion and assault prevention gadget "WIDPAS". It is based totally on 3 predominant obligations: tracking, evaluation and safety. With it, it monitors denial-of-provider or fake network assaults, then captures the assault and identifies the attacker, in addition to protects network users. Attack sorts for intrusion detection structures the Usage of gadget studying algorithms

Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm

AUTHORS: Kinam Park , Youngrok Song, Yun-Gyung Cheong

In this text, we show the results from our experiments to assess the efficacy of identifying unusual attacks. (e.G. IDS, malware and shell). We examine the act by the recipient of using making use of the

Random Forest algorithm to various statistics generated from the Kyoto 2006+ dataset, that is the present day network document statistics accrued for the development of intrusion detection structures. We finish with discussions and future research initiatives.

On the Selection of Decision Trees in Random Forests

AUTHORS: S. Bernard, L. Heutte , S. Adam

In this paper, we gift a take a look at of a circle of relatives of random wooded area (RF) matching strategies. In the "classical" RF induction system ,a set wide variety of selection bushes are induced to shape an ensemble. This kind of algorithm has two fundamental drawbacks:

- (i) the quantity of trees is fixed apriori
- (ii) the translation and evaluation possibilities which can be misplaced via the selection tree classifiers because of the precept of randomization.

Such a process, by which trees are delivered with out composition, does not guarantee that each one these trees work collectively effectively with in the identical fee. This thought increases two questions :Are there decision bushes in RF that lead to ugliness of the convergence impact? If so, is it viable to shape a more correct committee by removing the low efficiency decision bushes? The solution to these questions is solved as a sorting question. Thus, we show that optimal selection timber can be acquired even the use of a sub optimal classifier selection technique. This proves that the "classical" RF induction system, whereby random timber are randomly delivered to the ensemble, isn't always the exceptional approach for creating correct RF classifiers. We additionally present an hobby in RF development, via adding trees in a way this is extra dependent than in conventional "classical" RF induction algorithms. classifier selection technique. This proves that the "classical" RF induction system, whereby random timber are randomly delivered to the ensemble, isn't always the exceptional approach.

Random Forests Classifier at SMOTE and Feature Reduction for Intrusion Detection

AUTHORS:A.Tesfahun

,D.LalithaBhaskari

Intrusion detection structures (IDS) have turn out to be an important a part of laptop and community safety. The KDDCUP'ninety- nine dataset, which is an extended model of the NSL-KDD dataset for intrusion detection, was used to serve as an experimental instrument in this newsletter. Due to the inherent characteristics of intrusion detection, there's nonetheless a big imbalance among lessons inside the NSL-KDD dataset, which makes it hard for system learning within the discipline of intrusion detection. When considering rank inequality, this text applies the Synthetic Minority Sampling (SMOTE) method to the education dataset. An records- based feature choice approach is presented for the design of the NSL-KDD characteristic- decreased database. For tasks relating to intrusion detection, random forests are used as a classifier. The empirical findings demonstrate that a random forest classifier via SMOTE and feature selection that is entirely based on data acquisition provides the best overall performance in growing an IDS that is effective and strong for neighbourhood intrusion detection.

The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection

AUTHORS:Kim, H., Kang, H., and Le T. T. H.

An intrusion monitoring system is a device as well as software programme that displays networks or structures to malicious activity. (IDS). Standard IDSs no longer come across state-of-the-art cyber attacks along with low frequency DoS assaults and unknown assaults. Over the years, system studying has generated increasingly more hobby in over coming those barriers. In this article, we

proposed a brand new approach to enhance Gated Recurrent Unit (GRU) intrusion detection by means of incorporating the PCA- Scale with versions that are included in the GRU layer, such as PCA- Standardized and PCA-MinMax. The direction of the most variance is influenced by both complementary strategies' explicit application to the object maps under evaluation, thanks to their fantastic covariance. This technique may be carried out to the GRU version with little additional computational fee. We exhibit experimental results on two real tables, including the KDD Cup 99 along with NSL-KDD, illustrating the exquisite progress made by the GRU version taught with the PCA- Scaled method.

EXISTING SYSTEM

- Iftikhar Ahmad and colleagues examined a variety of systems to learn the algorithms for an intrusion detection machine.
- They compared various methods, such as SVM, Extreme Learning Machine, which is and Random Forest.
- The Extreme object learning technique works noticeably better than other algorithms, according to the results' investigators.
- B. Riyaz et al, labored here on enhancing the excellent of the facts set to offer an intrusion detection machine.
- Even though the characteristic selection method uses the policies to improve the dataset.
- Dynamic boom in IDS results were verified using the KDD dataset.
- Conventional methods for anomaly detection in self-organizing networks typically involve setting threshold values for certain network parameters, such as connection weights or activation levels. If these parameters fall outside of the expected range, an anomaly is detected.

On the other hand, contemporary machine learning methods for anomaly detection in self-organizing networks involve training a machine learning model to

recognize patterns in the data and identify anomalies based on these patterns.

DISADVANTAGES IN THE EXISTING SYSTEM

Systems running over the Internet are susceptible to numerous malicious activities. The primary hassle in this vicinity is intrusion into the information device.

- The existing consequences suggest that some improvements may be made in terms of accuracy, detection charge and fake fine charge.
- Some other methods can update the previous strategies which includes SVM and Naive Bayes.
- In addition, the examine established that positive facts may be advanced the usage of positive methods.
- Increase the enter quality of the mission.

PROPOSED SYSTEM

- The intrusion detection machine works to improve the gadget tormented by intruders.
- This machine can stumble on intruders.
- The proposed gadget attempts to cast off previous troubles with existing paintings.
- The proposed machine consists of two strategies: Principal Component Analysis and Random Forest.
- Principal thing evaluation is used to minimize the dimensional of the data set; with this approach the satisfactory of the dataset could be stepped forward, because the dataset can comprise the best attributes.
- Random woodland set of rules will then be used to locate the outliers, which gives a higher detection price and false high-quality fee than SVM.

ADVANTAGES OF PROPOSED SYSTEM

The blunders fee found in our

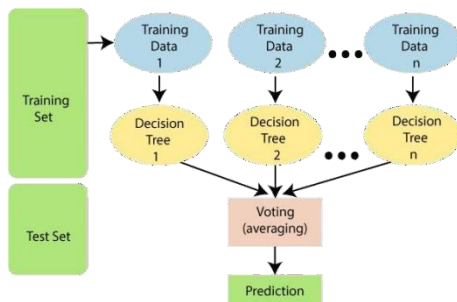
proposed method could be very low at 0.21%.

- In addition, the accuracy of the resulting algorithms is an awful lot higher than the previous one.
 - In addition, the execution time is less than other algorithms. **Proposed algorithms**
 - Principle component analysis
 - Random forest algorithm
 - Principle component analysis
 - Principal component analysis is an insatiable gaining knowledge of set of rules used to reduce dimensionality in gadget getting to know.
 - It is a statistical process that transforms observations of correlated records into fixed linear odds the use of an orthogonal transformation.
 - These newly converted functions are known as core components.
 - One of the most popular tools used for exploratory analysis and predictive exploratory fashions.
 - It is a technique of extracting clear styles from a given set of records by using reducing the variables.
 - PCA normally tries to discover a low-dimensional floor to undertaking excessive-dimensional information.
 - PCA works through inspecting the variance of each characteristic, due to the fact a
 - Some real packages of PCA are image processing, movie advice gadget, ultimate strength distribution in numerous conversation channels.
 - The PCA set of rules is based on positive mathematical ideas which includes:
 - Variance and covariance
 - Eigenvalues and eigenfactors
- Steps to the PCA algorithm**
- Getting the dataset
 - Data standardization
 - Calculation of covariance
 - Calculus of eigenvalues and eigenvectors

- Eigen vector Sorting
- Calculation of latest functions or pre dominant components
- Remove fewer or smaller characters from the new records set.
 - Randomforestalgorithm
 - Random Forest is a famous gadget getting to know algorithm for supervised studying techniques.
 - Can be used to solve regression and classification issues in ML.
 - It is primarily founded on the concept of ensemble learning, known as a technique that utilizes multiple classifications to address a challenging problem and enhance the model's performance.
 - The random wooded area algorithm is composed of many decision trees.
 - Clustering, clustering, and more clustering are used to create the "woodland" produced by a random programme. Device mastery algorithms' accuracy is increased by the ensemble meta-algorithm known as bagging.

3 Random Forest, as its name suggests, is a classifier which utilizes a mean to improve predictions for unique datasets and contains a lot of selection wood for those datasets.

4 Random Forest predicts the end outcomes based on a few voter predictions rather than relying solely on one unique choice tree.



SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:

System: Pentium IV 2.4GHz.

HardDisk : 40GB.

FloppyDrive : 1.44Mb.

Monitor : 15VGAColour.

Mouse : Logitech.

Ram : 512Mb.

SOFTWARE REQUIREMENTS:

Operating system : Windows7.

Coding Language : Python

Database : MYSQL

SYSTEM DESIGN SYSTEM ARCHITECTURE:

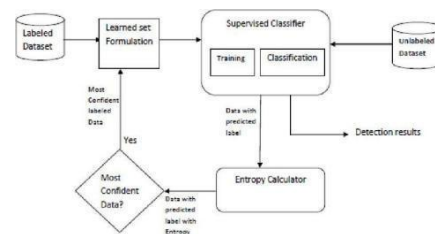


Fig.1

METHODOLOGY MODULE DESCRIPTION: METHODOLOGY

Data Collection:

This is the first real move in unquestionably creating a device learning model, facts series. This crucial stage establishes how accurate the version will be. The more data we collect, the better our algorithm will perform.

There are many methods for collecting data, such as internet feeds, manual intervention, and others

Data set on this intrusion detection system database taken from the kdd hyperlink: <http://kdd.lcs.Uci.Edu/databases/kddcup99/kddcup99.Html>.

Dataset:

There are 125974 characters in the collection. The dataset contains 42 columns, each of which is explained below.

Data Preparation:

Information is being transformed by us. By

1. Duration	length (number of seconds) of the connection
2. Protocol_type	type of the protocol, e.g. tcp, udp, etc.
3. Src_bytes	number of data bytes from source to destination
4. Dst_bytes	number of data bytes from destination to source
5. Is_hot_login	1 if the login belongs to the "hot" list; 0 otherwise
6. Is_guest_login	1 if the login is a "guest" login; 0 otherwise
7. Diff_srv_rate	% of connections to different services
8. Srv_diff_host_rate	% of connections to different hosts
9. Flag	normal or error status of the connection
10. Labels	Normal or attacker

delaying a few columns, missing data were deleted. We can first make a list of the names of the columns we want to save or retain.

After that, we remove or delete all columns other than the ones we want to keep. Finally, we remove or mark off rows in the dataset that have missing numbers. We categorise evaluation and instruction into illnesses.

Model Selection:

- The principal component evaluation technique is used to lessen the level of detail of an information set. Principal component analysis provides the intended outcomes as one of the most precise and efficient techniques for discounting spatial information. This strategy lessens some elements of a information set to a preferred wide variety of attributes, which can be referred to as principals. This method creates a dataset from all of the inputs that contains a huge number of attributes, possibly making the dataset's dimensions quite large. The facts points are transferred to 1 axis and the primary elements are ordered for execution. PCA may be accomplished the usage of the subsequent steps:

- Obtain a dataset containing all observations.
- For each value, add the mean vector.
- Include the complete dataset's covariance matrix.
- Calculate the eigenvalues ($v_1, v_2, v_3, \dots, v_n$) and eigenvector ($e_1, e_2, e_3, \dots, e_n$).
- Set the n eigenvectors and decreasing eigenvalues so that a matrix of M eigenvalues

is obtained.

- Make a fresh sample area in the shape of a M using this.
 - The significant areas are the most
- Information is being transformed by us. Random forest is one of the more important machine learning methods for solving scholastic issues. The method for supervised classification as es group includes random forest. These rules are implemented in a number of stages, one of which entails expanding the forest's data set and the opposite entails pronouncing a category.

Analyze and Prediction: In the actual dataset, we chose only 9 features **Accuracy on test set:** In the test set we got an accuracy of 99.1%.

Saving the Trained Model:

When you are confident sufficient to have the model trained and examined in a production environment, step one is to store it as a .H5 or .Pkl file the usage of a library like muria.

Make positive you have it stored for your surroundings. Then we import a replica of the module and dump it in .PKL document

TRAINING THE MODEL

We used a dataset produced by a SON model in this study that had been previously examined in [8] and was made publicly accessible for further study [23]. The dataset consists of four different application situations, and data is gathered from a drive test record that has been minimized at frequently (with a rate of sampling of 5 kHz) [24]. The summary includes details on mobile users as well as the activities they engaged in near bases (cells) during particular measurement intervals. Four distinct sets for data, each with a different amount of users and use cases, are included in the dataset.

Datasets 2 are made up of two different use cases, each with 114 features and an extended recording period (80 seconds), creating a much bigger

observation base. (42,000). Similar in datasets 1 you show different anomaly rates, at dataset 3 having alot more anomalous measurements than data

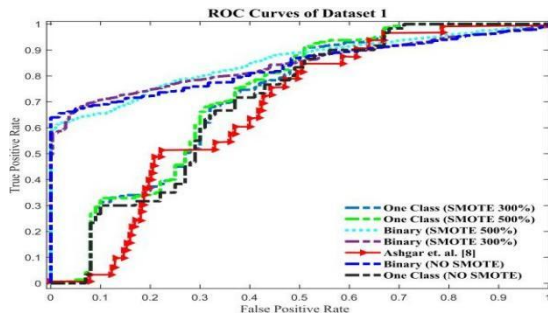


Fig.A

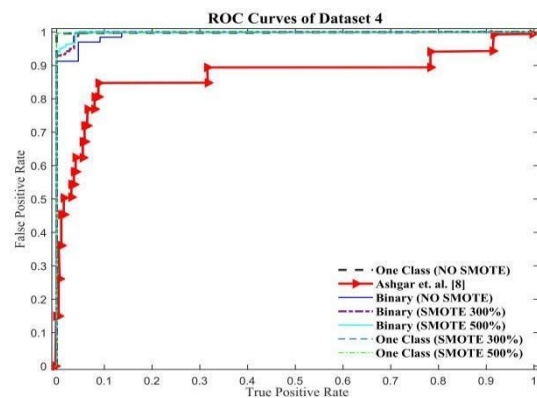


Fig.B

CONCLUSION

As the use of structures on the Internet is increasing hastily, so are security concerns. The proposed approach allows you to correctly stumble on intruders at the Internet. When compared to earlier algorithms like SVM, Naive Bayes, and Decision Tree, the suggested algorithm predicts the right outcomes. Using their proposed approach, it is viable to noticeably enhance the detectionand fake blunders prices. Here, a

knowledge discovery dataset has been used.

REFERENCES

1. JafarAbo Nada; Mohammad Rasmi Al-Mosa, 2018 International Arab Conference on Information Technology (ACIT),A Proposed Wireless IntrusionDetectionPreventionandAttackSystem
2. Kinam Park; Youngrok Song; Yun-Gyung Cheong, 2018 IEEE Fourth InternationalConferenceonBigDataComputingServiceand Applications(BigDataService),ClassificationofAttackTypesforIntrusionDetectionSystemsUsingaMachineLearning Algorithm
3. S. Bernard, L. Heutte and S. Adam “On the Selection of Decision Trees in RandomForests” Proceedings of International Joint Conference on Neural Networks, Atlanta,Georgia,USA,June14-19,2009,978-1-4244-3553-1/09/\$25.00©2009IEEE
4. A.Tesfahun,D.LalithaBhaskari,“ Intrusion Detection using Random Forests Classifierwith SMOTE and Feature Reduction” 2013 International ConferenceonCloud& Ubiquitous Computing& Emerging Technologies,978-0-4799-2235-2/13\$26.00©2013IEEE
5. Le, T.-T.-H., Kang, H., & Kim, H. (2019). The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection .2019 International Conference on Platform Technology and Service (PlatCon) . Doi: 10.1109 /platcon . 2019. 8668960
6. AnishHalimaaA,DrK.Sundarakantham:Procees dings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) 978-1-5386-9439-8/19/\$31.00 ©2019 IEEE “MACHINE LEARNING BASEDINTRUSIONDETECTIONSYSTEM.”
7. Mengmeng Ge, Xiping Fu, Naeem Syed, Zubair Baig, Gideon Teo, AntonioRobles Kelly (2019).Deep Learning Based Intrusion Detection for IoT Networks ,2019 IEEE 24th Pacific Rim International

Symposium on Dependable Computing (PRDC), pp.256-265, Japan.

8. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, "An Investigation on Intrusion Detection System Using Machine Learning" 978-1-5386-9276-9/18/\$31.00 © 2018 IEEE.

9. Rohit Kumar Singh Gautam, Er. Amit Doegar; 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms."

10. Kazi AbuTaher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahma, 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection."

11. L. Haripriya, M.A. Jabbar, 2018 Second International Conference on Electronic, Communication and Aerospace Technology (ICECA) "Role of Machine Learning in Intrusion Detection System : Review"

12. Nimmy Krishnan, A. Salim, 2018 International CET Conference on Control, Communication, and Computing (IC4) "Machine Learning-Based Intrusion Detection for Virtualized Infrastructures"

13. Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) "Feature extraction using Deep Learning for Intrusion Detection System."

14. Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, Rashmi Bhattad, 2019 3rd International Conference on Computing

Methodologies and Communication (ICCMC) "A Review of Machine Learning Methodologies for Network Intrusion Detection."

15. Iftikhar Ahmad, Mohammad Basher, Muhammad Javed Iqbal, Aneel Rahim, IEEE Access (Volume:6) Page(s): 33789–33795 "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection."

