Anomaly Detection System based on Computing Intelligence

1. Abstract:

Our project is purely based on artificial intelligence algorithms and machine learning models, stacking models to be precise. Stacking is an algorithm which takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction.

Detecting and diagnosing the root cause of network traffic log problems is a slow and tedious process, particularly for previously unseen failure mode but in the context of troubleshooting anomalies in the network traffic logs our project is developed on stacking method to find out the malicious logs from the Advanced Security Network Metrics (ASNM) datasets.

There have been basically three orthogonal approaches to building intrusion detectors according to the input training data: (1) Knowledge-based detection, which models characteristics of malicious intrusions and then match them (2) anomaly-based detection, which models normal behaviors and detects deviations from them. (3) classification-based detection, which models both malicious and legitimate behaviors at the same time.

The problems with these approaches is it showed a high false negative rate in case of evasions by unknown or zero day attacks, long time required for training and profiling and susceptibility.

To overcome all the drawbacks our project is completely based on a stacking model where we have taken four machine learning algorithms and among these algorithms at a time one algorithm is used at level 1 and remaining at level 0 for better testing accuracy rate. The four algorithms being used are: K-Nearest Neighbor, Naïve Bayes, Support Vector Machine and Decision tree.

The performance of these algorithms are quite similar with the most effective being naïve bayes when it is kept at level 1 followed by support vector machine, Decision Tree and K-Nearest Neighbor at level 0.

1.1 Introduction:

These days developed & bigger companies are improving their network security architecture to defend and manage the cyberattacks and network security threats. As the threat of being possibly attacked by a hacker has dramatically increased , even emerging technologies such as Cloud computing, Fog, Edge computing and Internet of Things (IoT) etc are susceptible to such attacks.

Once these attackers penetrate network-related environments and get into the network of the company or an organization the damages they can cause are beyond imagination. They can cause economic damage, steal critical information, and can continue to stay in the network without being detected and do further damage. Network anomaly detection systems (NADSs) play an essential role in every network defense system as they monitor network packets to prevent potential threats and user behavioral abnormalities. Anomaly detection is basically finding abnormal or unusual patterns in a network with the help of specified algorithms. Anomaly detection helps in numerous ways as it tells an organization where they are vulnerable for attacks and also gives them the time to improve upon them.

Anomaly describes any change in the specific established standard communication of a network. An anomaly may include both malware and cyberattacks, as well as faulty data packets and communication changes caused by network problems, capacity bottlenecks, or equipment failures. Anomalies are been detected in various organizations by the various methods such as : Regular checks for Network Intrusion , i.e. NIDS , Abnormal Finance activities detection, Advanced Penetration Detection,Protecting Web based business.

1.2 Real Life Implications:

Artificial Intelligence is now broadly utilized in every and each element of lifestyles and supplies advanced lifestyles stories and higher results. AI structures help in spotting and combat cyberattacks and cyber threats primarily based totally at the non-stop center of data, recognising styles and backtracking the attacks, These days AI is being used in each and every technology like self driving cars etc.

Machine Learning enables us to enhance commercial enterprise decisions, increase productivity, stumble on disease, forecast the weather, and lots more. Basically, a gadget learns routinely from the inputs. Some of the nice gadget gaining knowledge of examples are stated traffic alert , image recognition etc.Anomaly detection is normally understood to be the identity of uncommon items, occasions or observations which deviate substantially from the bulk of the facts and do now no longer agree to a properly described perception of ordinary behavior

1.3 Related Works:

Paper 1: Dynamic Network Anomaly Detection System by Using Deep Learning Techniques.

In the above stated paper the author has used the LSTM (Long Short Term Memory) approach, About LSTM = LSTM is a special recurrent neural network structure, which is proposed to solve the problem of long-term dependence. The forget gate lets the neural network forget the useless information, the input gate adds new content to the neural network and the output gate determines the final output of the current node.

- And has also used Attention Mechanism, The Attention Mechanism (AM deep learning) is actually imitating the attention mechanism of the human brain. When reading a piece of text, we usually focus on some keywords so that we can quickly summarize the main content of the text same is with AM, in anomaly detection, the role of AM is to calculate the impacts of each network traffic on the last network traffic , The above said algorithm has an accuracy of 93%. The advantages of using the above method is that the

author has used loss function to fine tune the efficiency of the algorithm (in the given function we calculate the loss).

Paper 2: Insider Threat Detection Based on User Behavior Modeling and Anomaly Detection Algorithms

Insider threat is a security issue that arises from persons who have access to a corporate network, systems, and data, such as employees and trusted partners ,Although insider threats do not occur frequently, but the magnitude of damage is greater than from external intrusions, The proposed paper has an accuracy of 53.77%.

– Assumptions & data used :

All activity logs of individual users recorded in the corporate system are collected. Then, candidate features are extracted by summarizing specific activities. For example, if the system logs contain information on when a user connects his/her personal USB drive to the system, the total number of USB connections per day can be extracted as a candidate variable, and user-generated contents, such as the body of an e-mail, to create candidate features are also taken into account .

–Proposed method:

The author has after taking all the above data into consideration has constructed an algorithm which finds insider threats in an organization.

Paper 3: A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects

Anomaly detection via Mahalanobis-squared distance (MSD) is a famous unsupervised algorithm . Despite the recognition and excessive applicability of the MSD-primarily based totally anomaly detection

technique, a few primary hard troubles and obstacles which includes environmental variability, dedication of an beside the point threshold limit, estimation of an erroneous covariance matrix, and non-Gaussianity of schooling statistics can also additionally result in fake alarms and inaccurate effects of harm detection. The most important goal of this newsletter is to recommend a unique anomaly detection technique primarily based totally on adaptive Mahalanobis-squared distance and one-elegance kNN rule referred to as AMSD-kNN for SHM below various environmental situations. The critical concept at the back of the proposed technique is to discover enough nearest pals of schooling and checking out datasets in a -degree technique for putting off the environmental variability situations and estimate nearby covariance matrices. A powerful method primarily based totally on a multivariate normality speculation check is proposed to discover enough nearest pals that assure the estimate of well-conditioned nearby covariance matrices. The extraordinary novelty of the proposed AMSD-kNN technique is to create a unique unsupervised getting to know method for SHM through a brand new multivariate distance degree and one-elegance kNN rule. Generalized excessive cost distribution modeling through the block maxima (BM) technique is offered to decide an correct threshold limit. Due to the significance of selecting ok blocks withinside the BM technique, a goodness-of-in shape degree thru the Kolmogorov-Smirnov speculation check is carried out to pick an ideal block number. The overall performance and effectiveness of the proposed techniques are established through famous benchmark structures. Several comparative research also are performed to illustrate the prevalence of the proposed techniques over a few ultra-modern techniques. Results display that the proposed AMSD-kNN and BM techniques exceptionally achieve detecting harm below environmental variability situations

PAPER 4: Log-Based Anomaly Detection with the Improved K-Nearest Neighbor

In this research paper about Log based Anomaly Detection using K-Nearest Neighbor Logs play a crucial position withinside the protection of huge-scale systems. The wide variety of logs which suggest everyday (everyday logs) differs significantly from the wide variety of logs that suggest anomalies (bizarre logs), and the 2 varieties of logs have certain differences. To mechanically achieve faults with the aid of using the K-Nearest Neighbor (KNN) set of rules, an outlier detection approach with excessive accuracy, is a powerful manner to come across anomalies from logs. However, logs have the traits of huge scale and really choppy samples, to be able to have an effect on the outcomes of KNN set of rules on log-primarily based totally anomaly detection. Thus, we advocate a progressed KNN set of rules-primarily based totally approach which makes use of the present mean-shift clustering set of rules to successfully pick the education set from big logs. Then we assign unique weights to samples with unique distances, which reduces the terrible impact of unbalanced distribution of the log samples at the accuracy of KNN set of rules. By evaluating experiments on log units from 5 supercomputers, the outcomes display that the approach we proposed may be efficiently carried out to log-primarily based totally anomaly detection, and the accuracy, take into account fee and F degree with our approach are better than the ones of conventional key-word seek approach.

Paper 5:Deep Learning for Anomaly Detection

In this research paper the author wants to present a complete expertise of deep learning-primarily based totally anomaly detection strategies in numerous software domains. First, it introduces what is the paradox detection problem, the techniques taken before the deep version generation and the demanding situations it faced. Then it surveys the modern day deep mastering fashions considerably and discusses the strategies used to triumph over the restrictions from traditional algorithms. Second to last, it researches deep version anomaly detection strategies in actual global examples from LinkedIn production systems. The paper concludes with a dialogue of destiny trends. Then we become conscious of introducing the modern day deep anomaly detection algorithms. In deep version anomaly detection strategies, we cover essential tasks:

1) learning regular representations from complicated data, wherein RNN, LSTM, Auto-Encoder,

2) Detecting anomalies, whilst we summarize the strategies used to correctly discover anomalies primarily based totally on reconstruction errors, reconstruction probabilities and the use of one class NN

Paper 6: Study and Analysis of Decision Tree Based Classification Algorithms

Machine studying is to research devices on the premise of diverse schooling and trying out statistics and determining the consequences in each situation without specific programmed. One of the strategies of device studying is Decision Tree. Different fields used Decision Tree algorithms and used it for their respective applications. These algorithms may be used as to locate statistics in
alternative statistical procedures, to extract text, scientific licensed fields and additionally in seek engines. Different Decision tree algorithms had been constructed in keeping with their accuracy and fee of effectiveness. To use the great set of rules in each situation of choice making may be very essential for us to know. This paper consists of 3 distinctive algorithms of Decision Tree which are
ID3, C4.five and CART.he Decision Tree algorithms ID3 C4.five and CART were carried out at the dataset. Decision tree outperforms others in phrases of accuracy, time and precision. It is based on the set of rules used for advice to locate interesting resources. At last, the complete look at is completed about choice tree algorithms and this paper concludes that CART is the set of rules for this dataset that may be very specific and most correct some of the others.

Paper 7 : An Improved KNN-Based Efficient Log Anomaly Detection Method with Automatically Labeled Samples

In this research paper the author determines that Logs which file unusual log states (anomaly logs) may be seen as outliers, and the k-Nearest Neighbor (kNN) set of rules has particularly excessive accuracy in outlier detection methods. Therefore, we use the kNN set of rules to discover anomalies withinside the log information. However, there are a few troubles while the usage of the kNN set of rules to discover anomalies, 3 of which

are: immoderate vector measurement ends in inefficient kNN set of rules, unlabeled log information can't assist the kNN set of rules, and the imbalance of the range of log information distorts the type selection of kNN set of rules. In order to clear up those 3 troubles, we recommend an green log anomaly detection approach primarily based totally on an stepped forward kNN set of rules with a mechanically categorized pattern set. This approach first proposes a log parsing approach primarily based totally on N-gram and common sample mining (FPM) approach, which reduces the measurement of the log vector transformed with Term frequency.Inverse Document Frequency (TF-IDF) technology. Then we use clustering and self-schooling to get categorized log information patterns set from ancient logs mechanically.

To routinely obtain categorized log statistics pattern set, we use clustering and self-training

approach for peculiar logs have the traits of small amount and lengthy distance from normal

logs. Finally, we use common weighting distance to enhance kNN algorithm, decreasing the negative

results of log pattern imbalance at the accuracy of kNN algorithm. Through experiments on log

units generated through six datasets with differing types and the contrast with 3 different log-primarily based totally

anomaly detection methods, the consequences display that our approach can enhance the effectiveness of log

primarily based totally anomaly detection with kNN algorithm, and make sure the accuracy on the identical time.

Paper 8: A Lightweight Anomaly Detection Model using SVM for WSNs in IoT, through a Hybrid Feature Selection Algorithm based on GA and GWO

Anomaly or intrusion detection system (IDS) is an efficient protection mechanism, technology for stressed networks, contemporary technology with excessive computational complexity are mistaken for aid-restrained WSNs in IoT and additionally they fail to locate new WSN attacks. Furthermore, managing the big quantity of intrusion wi-fi site visitors gathered through sensors, inflicting gradual detecting process, better aid utilization and faulty detection.
Hence, thinking about WSN barriers for growing an IDS in IoT, establishes a great task for protection researchers. This paper proposes a brand new version to increase a support vector machine (SVM)-primarily based totally light-weight IDS (LIDS) the usage of combination ideas of genetic algorithm (GA) and mathematical equations of gray wolf optimizer (GWO) that is referred to as GAB GWO. The GABGWO via applying new crossover and mutation operators attempts to locate the maximum applicable site visitors functions and get rid of nugatory ones, with a view to boost the overall performance of the LIDS. The overall performance of LIDS is evaluated by the usage of AWID real-world wi-fi dataset beneath eventualities with and without the usage of GAB GWO. The effects confirmed a promising conduct of the proposed GAB GWO set of rules in deciding on most desirable traffics, reducing the computational fees and providing excessive accuracies for LIDS. The hybrid set of rules is likewise in comparison to natural GA and GWO and different latest techniques and its miles determined that its overall performance is higher than them.

Paper 9 : Study and Analysis of Decision Tree Based Classification Algorithms

Machine getting to know is to analyze devices on the premise of diverse education and trying out statistics and determine the effects in each circumstance without expressly programmed. One of the strategies of

device getting to know is the Decision Tree. Different fields used Decision Tree algorithms and used it for their respective applications. These algorithms may be used to discover statistics in alternative statistical procedures, to extract text, clinical licensed fields and additionally in seek engines. Different Decision tree algorithms were constructed in line with their accuracy and fee of effectiveness. To use the pleasant set of rules in each situation of choice making may be very crucial for us to know. Decision tree outperforms others in phrases of accuracy, time and precision. It is based on the set of rules used for advice to discover interesting resources. At last, the complete take-a look is accomplished about choice tree algorithms and this paper concludes that CART is the set of rules for this dataset that may be very specific and most correct a number of the others.

Paper 10: A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks

On account of the reality that those networks can not be supervised, this paper, therefore, offers the trouble of anomaly detection. First, the 3 functions of temperature, humidity, and voltage are extracted from the community traffic. Then, community facts are clustered the usage of the density-primarily based totally spatial clustering of packages with noise (DBSCAN) set of rules. It additionally analyzes the accuracy of DBSCAN set of rules to enter facts with the assistance of density-primarily based totally detection techniques. This set of rules detects the factors in areas with low density as anomalies. By the usage of everyday facts, it trains to help vector machines. And, finally, it gets rid of anomalies from community facts. The proposed set of rules is evaluated through the usual and standard facts set of Intel Berkeley Research lab (IRLB). In this paper, we ought to obliterate DBSCAN's trouble in deciding on enter parameters through profiting from coefficient correlation. The benefit of the proposed set of rules over preceding ones is in the usage of tender computing

methods, easy implementation, and enhancing detection accuracy via simultaneous evaluation of these 3 functions.

Paper 11: On-Line Anomaly Detection With High Accuracy

In this paper the author discussesses about traffic anomaly detection is crucial for superior Internet management. Existing detection algorithms commonly convert the excessive-dimensional records to a protracted vector, which compromises the detection accuracy because of the lack of spatial statistics of records. Moreover, they're commonly designed primarily based totally at the separation of ordinary and anomalous records in a time period, which now no longer simplest introduces excessive garbage and computation value, however additionally prevents well timed detection of anomalies. Online and correct site visitors anomaly detection is crucial however tough to help. To cope with the challenge, this paper at once fashions the tracking records in whenever slot as a 2-D matrix, and detects anomalies withinside the new time slot primarily based totally on bilateral major thing analysis (B-PCA). We advise numerous novel strategies in OnlineBPCA to help brief and correct anomaly detection in actual time, which includes a unique B-PCA-primarily based totally anomaly detection precept that collectively considers the variant of each row and column major instructions for extra correct anomaly detection, an approximate set of rules to keep away from the use of generation process to calculate the major instructions in a close-form, and a sequential anomaly set of rules to quick replace major instructions with low computation and garage value whilst receiving a brand new records matrix at a time slot. To the pleasant of our knowledge, that is the primary painting that exploits 2-D PCA for anomaly detection. We have performed enormous simulations to examine our OnlineBPCA with the state-of-artwork anomaly detection algorithms the use of actual site visitors strains Abilene and GÈANT. Our simulation outcomes show that, in comparison with different algorithms, our OnlineBPCA can reap extensively higher detection overall performance

with low fake effective rate, excessive authentic effective rate, and coffee computation value.

1.3 Proposed Solution:

Here we propose a novel technique to determine the malicious intent of a network request. It includes various methods relating to machine learning modules and methods like stacking,, . In this model we combine different algorithms like KNN, SVM, Naive Bayes and decision trees to fit them in different tiers of the stack like tier 0,1,2. After this we dry run the code on the data sets to attain utmost accuracy in all the three datasets.

1.4 Research Contribution

The method i.e stacking model which we have used is efficient and also do have less false alarm rate , The contribution of our work in the field of anomaly detection would be that our algorithm can help organizations to know the anomalous behavior in their network one of the benefit of knowing the anomaly in the network is that the organization can work on it and make itself immune of the attacks or breaches via that particular anomaly , Also what we have focused more on is the less number of false alarm rate by the algorithm as higher number of these can cause problems in the functioning of the network as well as a algorithm with higher false rate can't' be trusted. Hence what we are contributing is an algorithm with less false alarm , More efficiency and with that which is also cost effective.

2. Materials and Method

2.1 Experimental Methods:

The Machine utilized for implementing the research have the following requirements:

` **Table 1.** Tested Environment

OS	Windows 10
RAM	8 GB
GPU	4 GB
IDE	Visual Studio Code (Python)

The entire algorithm is implemented with the above configuration. As we can observe from the above table, the physical requirements are negligible and the proposed model can be compatible with almost any physical machine.

2.2 Data Input:

At first, we are importing all the relevant libraries and they are:

1. Pandas: It is an open-source python package that is most widely used for data analysis and Machine learning tasks, here it is used for data cleaning and analysis.

2. NumPy: A library in python language adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

3. Sklearn: It is widely used for statistical modeling including classification, regression, clustering and dimensionality reduction.

4. Matplotlib: It is a numerical extension of NumPy library, specifically used for cross-platform, data visualization and graphical plotting library for python.

2.3 Pre-Processing:

2.3.1 ***Importing the Dataset:*** Here we have three sub-dataset of ASNM dataset namely, ASNM-CDX-2009, ASNM-NBPOv2 and ASNM-TUN subsequently we will be importing them one by one for 12 programs each.

2.3.2 ***Segmentation and Masking:*** Since our dataset is not organized we need to make it neat and clean so for that we need to separate and label features and store them into two different arrays.

2.3.3 Resizing: Here all the network traffic data which was in the form of IP address, MAC address, port number, VLAN Id is converted into binary form for better calculation and effective output.

2.3.4 Standardization: All the entries of the CSV dataset were applied for standardization so that the resulting average is calculated as 0, and a unit standard deviation can be observed.

2.4 Proposed Model Architecture:

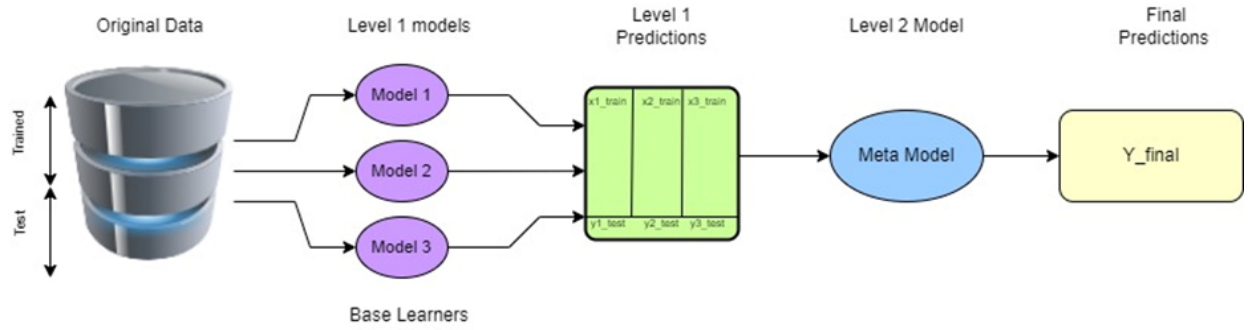


Figure. 1. The Stacking Method Architecture Diagram

2.5 Data Source

The publicly available data source was collected from center of Excellence IT4Innovations, Faculty of Information Technology, Brno University of Technology, 612 00 Brno, Czech Republic, which consists of three datasets that have been built from network traffic traces using ASNM (Advanced Security Network Metrics) features. Each of this dataset consists of 5000-6000 rows and 50-60 columns.

Our Dataset comprises of three sub-dataset:

(i.) ASNM-CDX 2009: This dataset consists of ASNM features extracted from tcpdump capture of malicious and legitimate TCP communications on network services which are vulnerable to buffer overflow attacks and are included in CDX-2009 dataset of network traffic dumps.

The final composition of the dataset is depicted in table ASNM-CDX-2009 dataset contains two types of labels that are enumerated by increasing order of their granularity in the following:

Label_2: Is a two-class label, which indicates whether an actual sample represents a network buffer overflow attack or legitimate traffic.

· **Lebel_poly:** is composed of two parts that are eliminated be a separator:

(a) A two-class label where legitimate and malicious communications are represented by symbols 0 and 1 respectively.

(b) An acronym of network service. This label represents the type of communication on a particular network service.

(ii.) ASNM-NPBO v2: This dataset contains non-payload-based obfuscation techniques applied onto malicious traffic and onto several samples of legitimate TCP communications on selected vulnerable network services. The selection of vulnerable services was aimed on high severity of their successful exploitation leading to remote shell code execution through established backdoor communication.

legitimate representatives of the dataset were collected from two sources:

a) Legitimate traffic simulation in our virtual network architecture and also employed non-payload- based obfuscations for the purpose of real network simulation.

b) Common usage of all selected services was captured in the campus network, and all traffic was anonymized and further filtered on high severity alerts by signature-based NIDS Suricata and Snort through virus total API.

(iii.) ASNM-TUN: It consists of ASNM features extracted from tcpdump capture tunneling obfuscation techniques applied onto malicious traffic created with the intention to evade and improve machine learning classifiers and besides legitimate network traffic samples.

ASNM-TUN dataset contains four types of labels that are listed by increasing order of their level in the following:

· **Label_2:** It is a two-class label, which indicates whether an actual sample represents a network buffer overflow attack or a legitimate communication

· **Lable_3:** It is a three-class label, which distinguishes among legitimate traffic, direct attacks and obfuscated network attacks.

· **Label_poly:** It is a label that is composed of two parts:

(a) A three-class label

(b) An acronym of a network service

· **Label_poly_s:** It is composed of three parts:

(a) A three-class label

(b) An acronym of network service

(c) A network modification technique involved.

2.6 Parameters and customization for computer visions Models

Table 1 shows the various parameters setting for Computer Vision Models.

2.6.1 Stacking Model:

Stacking is an ensemble learning technique that uses meta-learning for generating predictions. Here the original data is splitted into n-folds (train data and test data) further, these data are embedded into the different models (we have used KNN, SVM, Naïve bayes and Decision tree models in this research) and they result with some predictions which then added back to the level 2 to yield the final result.

2.6.2: K Nearest Neighbor (KNN):

The K Nearest Neighbor algorithm falls under the supervised learning category and it is used for classification. It considers K Nearest Neighbors (Data Points) to predict the class or continuous value for the new datapoint.

2.6.3: Support Vector Machine (SVM):

Support Vector Machine is a supervised machine learning algorithm that can be used for both classification and regression challenges. In this algorithm we plot each data item as a point in N-Dimensional space (where N is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes.

2.6.4: Naïve Bayes:

Naïve Bayes is a classification technique and is easy to build and particularly useful for large data sets, which assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

2.6.5 Decision Tree:

Decision trees can be used for classification and regression problems. The name suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

3. Results and Discussions:

Here the dataset was split into two distinct subarrays by the train-test split approach. The training component comprised 75% of the original data, while the testing component accounted for 25% of the original dataset. After processing the details of all the models, an investigation using performance was conducted to identify the best model out of 4 models. In this comparison analysis, the parameter metrics utilized for distinguishing models were accuracy, precision, F1 score and recall, which can be calculated from the confusion matrix. All the models' output is depicted via the likelihood of belonging to a specific class. Which can be in the form of a fraction between 0 and 1. Here we have three datasets i.e.

- ASNM-CDX 2009 dataset
- ASNM-NBPOv2 dataset
- ASNM-TUN dataset

And four algorithms are being used and they are as follows:

- K Nearest Neighbor (KNN)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- Naïve Bayes (NB)

We have used a stacking method to find the best algorithm which finds better accuracies in our dataset so we have 4 programs for each dataset and we have 3 dataset and in total we have 12 programs. Below is the split of the four programs for each dataset:

1. Level 0: DT, SVM, NB **Level 1:** KNN

2. Level 0: KNN, SVM, NB **Level 1:** DT

1. Level 0: KNN, DT, NB **Level 1:** SVM

1. Level 0: KNN, DT, SVM, **Level 1:** NB

Figure. 2. Comparison of metrics – (Precision, Recall, F1-score) for ASNM-CDX 2009 Dataset

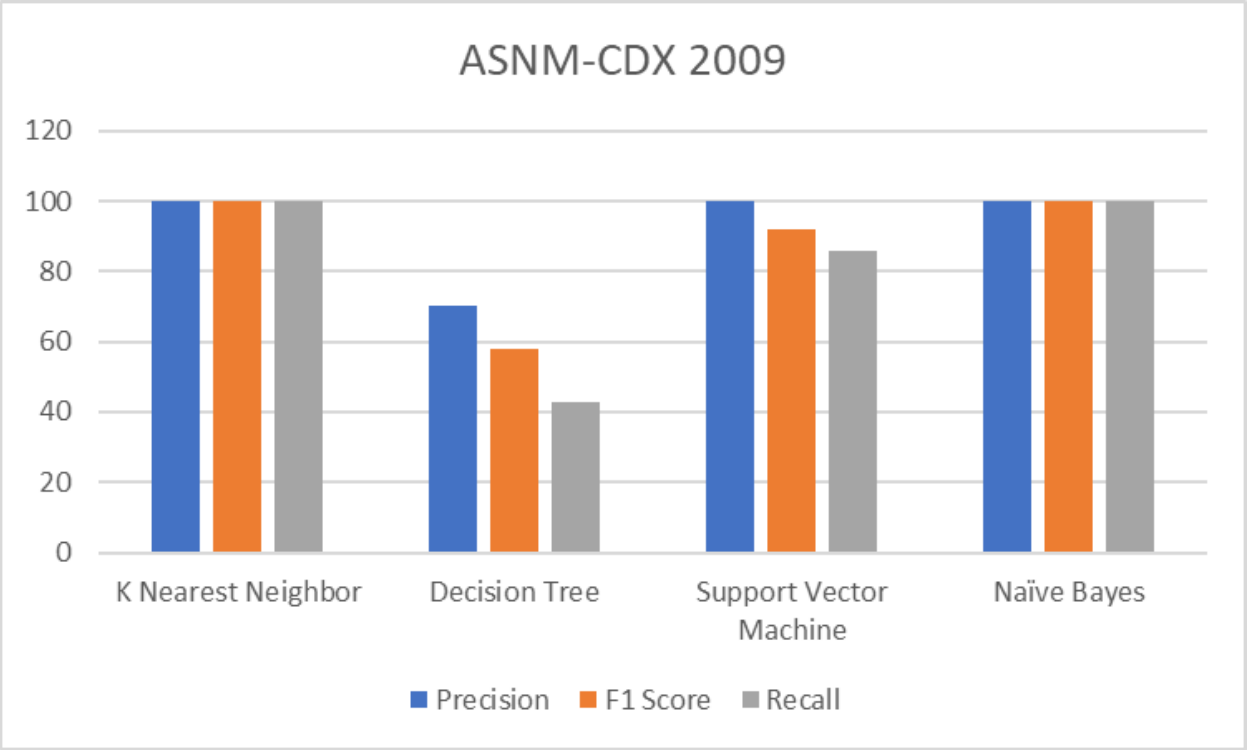
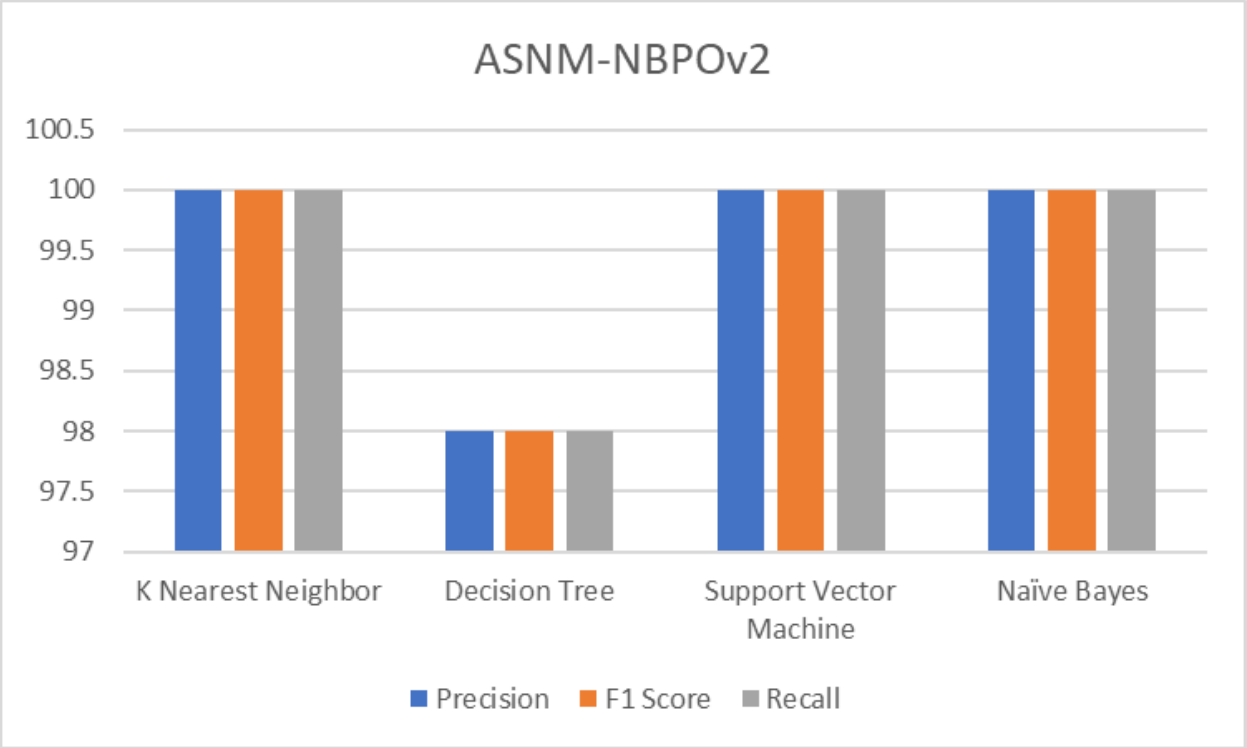


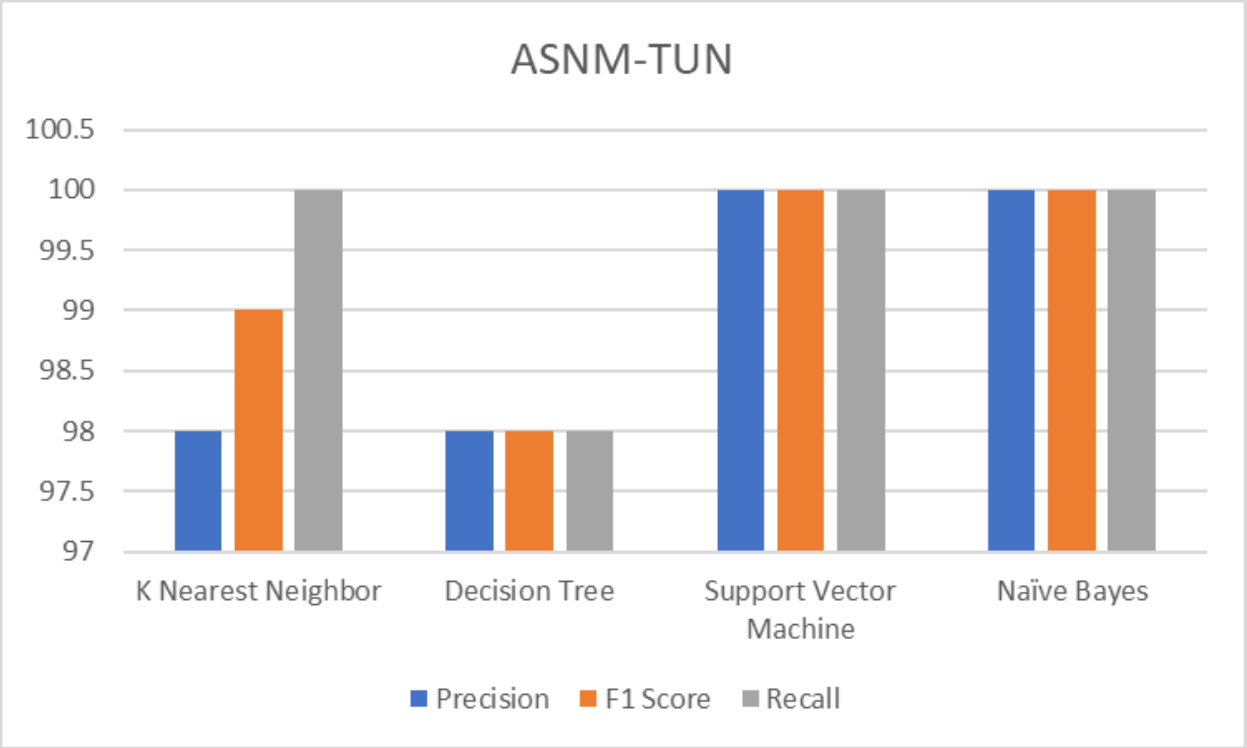
Figure 2. depicts the performance shown by four algorithms on ASNM-CDX 2009 dataset. K Nearest Neighbor algorithm and Support Naïve Bayes algorithm have a perfect score of metrics. The Support Vector Machine model has a high precision score only while least metrics are shown by the decision tree model.

Figure. 3 Comparison of metrics – (Precision, Recall, F1-Score) for ASNM-NBPOv2 Dataset.



The graph of figure 3 shows the performance of all AI models on ASNM-NBPOv2 dataset. Support Vector Machine, Naïve Bayes and K Nearest Neighbor show high values in all the three metrics while Decision Tree accounted for 98%.

Figure. 4 Comparison of Metrics – (Precision, Recall, F1-Score) for ASNM-Tun dataset



The bar graph of figure 4 shows the performance of the 3 Artificial Intelligence models on ASNM-TUN dataset. Support Vector Machine and Naïve Bayes display all the metrics with utmost percentage. The Precision of K Nearest Neighbor is 100% while the decision tree has equal AMOUNT OF proportion in all the three domains.

Tabel. 2. Confusion Matrix of **K-Nearest Neighbor Algorithm** at level 1 with all three dataset of ASNM dataset.

1. For ASNM-CDX 2009 Dataset:

1429	0
0	14

2. For ASNM-NBPOv2 Dataset:

2691	2
2	167

3. For ASNM-TUN dataset:

43	0
0	55

Tabel. 3. Confusion Matrix of **Decision Tree Algorithm** at level 1 with all three dataset of ASNM dataset.

1. For ASNM-CDX 2009 Dataset:

1425	4
8	6

2. For ASNM-NBPOv2 Dataset:

2689	4
3	166

3. For ASNM-TUN dataset:

42	1
2	53

Tabel. 4. Confusion Matrix of **Support Vector Machine Algorithm** at level 1 with all three dataset of ASNM dataset.

1. For ASNM-CDX 2009 Dataset:

1429	0
0	14

2. For ASNM-NBPOv2 Dataset:

2693	0
0	169

3. For ASNM-TUN dataset:

43	0
0	55

Table. 5. Confusion Matrix of **Naïve Bayes Algorithm** at level 1 with all three dataset of ASNM dataset.

1. For ASNM-CDX 2009 Dataset:

1429	0
2	12

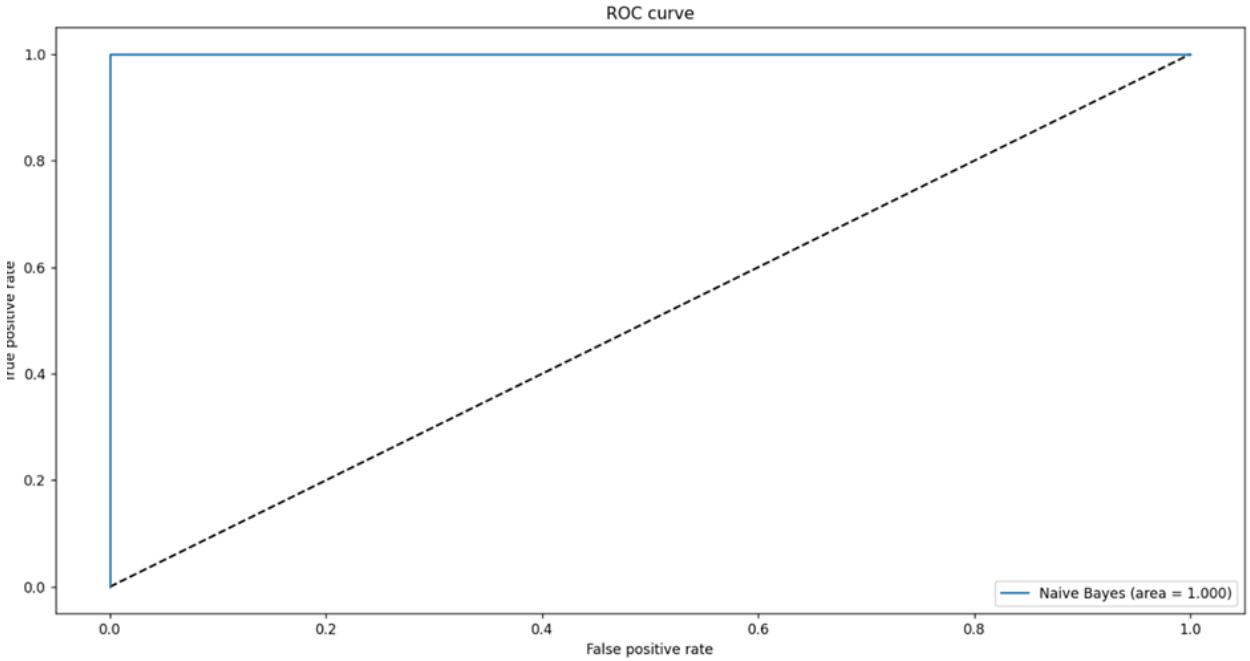
2. For ASNM-NBPOv2 Dataset:

2693	0
0	169

3. For ASNM-TUN dataset:

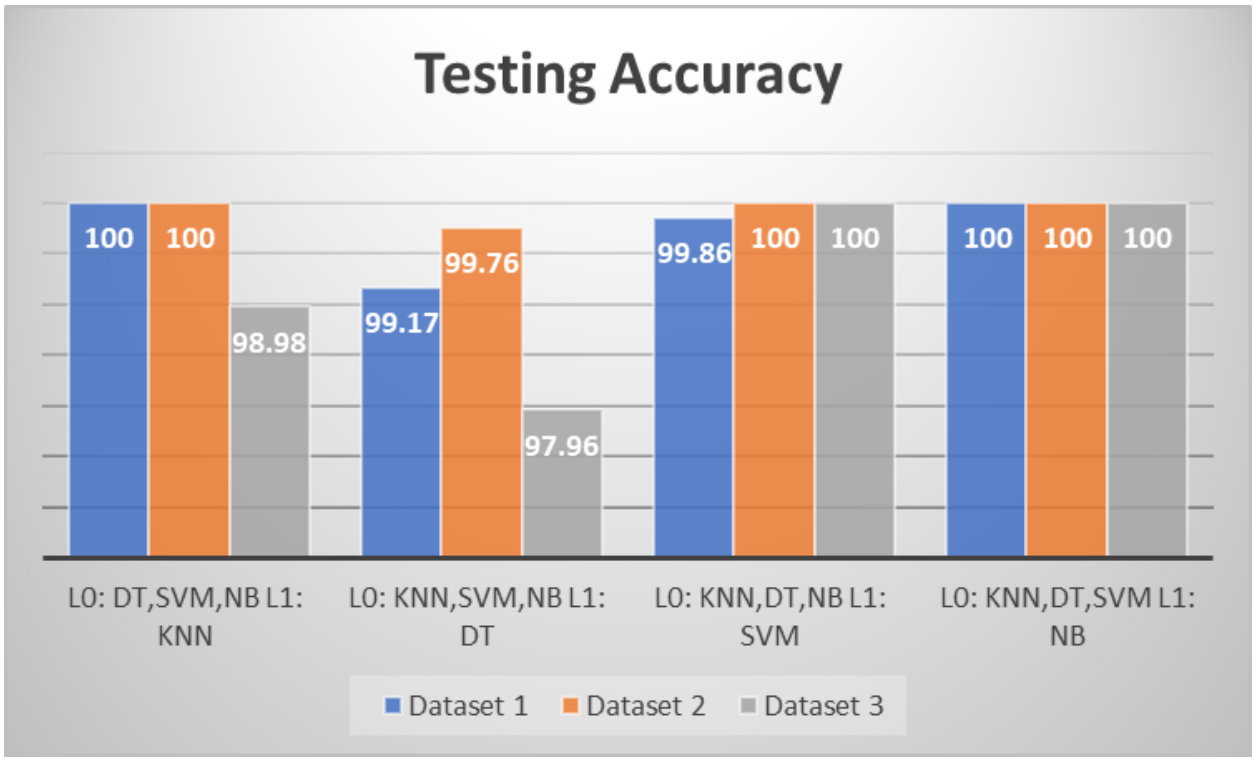
43	0
0	55

Figure. 5. ROC curve for KNN, Decision tree, SVM at level 0 and naïve bayes at level 1 model.



The above model of KNN, SVM and Decision Tree at level 0 and Naïve Bayes at level 1 model shows utmost true positive to false positive ratio with area under curve of 1.000.

After a through comparison and contrast, it is empirical that the proposed model of Naïve Bayes at level 1 and other three algorithms at level 0 has surpassed all the other techniques, with an accuracy of 100%, and the side metrics of Precision, F1 Score and Recall score were found to be 100% in all the sub dataset of ASNM dataset. **(Figure 6)**.



From the above graph it is clearly visible that the model in which Naïve Bayes is at level 1 has 100% accuracy in all the three datasets.

Anomalies being a disturbance to any result or work make the output deviate between observed data and the normal state but with the help of this project idea we will be able to identify the best algorithms for anomaly detection with highest accuracy by using Machine learning and deep learning models we can implement it and

4. Conclusion:

In conclusion of the outlier analysis , we have determined that the hybrid model of algorithms for the detection in logs shows higher accuracy percentage compared to that of the normal anomaly detection method and hence gives better results. After combining various algorithms we determined that **L0: KNN, DT, SVM and L1: NB** this model of the hybrid branch shows the best accuracy and hence will be preferred for resolving issues in datasets/logs for detecting these unusual patterns, i.e. Anomalies.

4.1 Future Work :

With the help of our idea in the research paper we can implement in future projects for better accuracy and precision from Anomaly detection in data sets. The hybrid model of algorithms gives better results and accuracy compared to that of singular algorithms used for detection of anomalies.

We can increase the accuracy of our findings by improving the accuracy of our algorithm thus resulting in findings which will be closer to actual value and be precise. This will lead to projects being more accurate and more reliable.

5 References:

- https://ieeexplore.ieee.org/document/9115004
- http://www.fit.vutbr.cz/~ihomoliak/asnm/
- https://www.edureka.co/blog/what-is-a-neural-network/
- https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207#:~:text=Artificial%20Neural%20Network(ANN)%20uses,complex%20patterns%20and%20prediction%20problems.
- https://www.ijitee.org/wp-content/uploads/papers/v8i9/I7914078919.pdf
- https://www.mdpi.com/1099-4300/23/5/529
- https://www.analyticssteps.com/blogs/8-applications-neural-networks
- https://www.xenonstack.com/blog/artificial-neural-network-applications
- https://towardsdatascience.com/building-our-first-neural-network-in-keras-bdc8abbc17f5

- https://link.springer.com/article/10.1007/s11277-017-4961-1
- https://www.worldscientific.com/doi/abs/10.1142/S0218194020500114
- https://dl.acm.org/doi/abs/10.1145/3336191.3371876
- https://dl.acm.org/doi/abs/10.1145/3441448
- https://www.researchgate.net/profile/Purvi-Prajapati/publication/330138092_Study_and_Analysis_of_Decision_Tree_Based_Classification_Algorithms/links/5d2c4a91458515c11c3166b3/Study-and-Analysis-of-Decision-Tree-Based-Classification-Algorithms.pdf
- https://jcomsec.ui.ac.ir/article_24558_4491.html