Hybrid Machine Learning Model to Enhance Cybersecurity: An Integration of KNN, RF and XGBoost

U. Rashid [1], M. Qadir [2], M. Alam [3], S. Farid [4]

[1,4]*Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan.*
[2]*Department of Information Security, The Islamia University of Bahawlapur, Pakistan*
[3]*ICCC, Informatics Complex, H-8, Islamabad, Pakistan*

[4]shahidfarid@bzu.edu.pk

Abstract- Living in the digital age has changed the way we do the majority of things, including living, working, and interacting, and positively impacted living in unprecedented ways when it comes to convenience, connectivity, and accessibility. Today, in the digital age, cybersecurity is the major concern. Cyber-attacks have become more sophisticated and frequent, there is a need for developing advanced approaches to protect computer systems and networks from being compromised. In recent years, machine learning has become one of the most powerful tools that can help to bring improvement in the cybersecurity field. However, current machine learning models aren't able to detect sophisticated cyber-attacks, like zero day and targeted attacks. In this paper, we investigate the limitations of current ML models in the field of Cybersecurity, and propose a framework for their improvement. The proposed framework includes the hybrid machine learning method to detect cyber-attacks. Performance evaluation has been done on KDD dataset than experiments show that the proposed hybrid model is superior to conventional machine learning model in terms of accuracy, precision, recall rate and F1-measure. Overall attack detection rates are improved by the integration. Overall proposed approach solutions to strengthen the cybersecurity appear promising using machine learning. It will assist organizations to detect and prevent cyber-attacks better in order to secure their computer systems and networks.

Keywords- Machine learning, cyber security, KNN, XGBoost and random forest

I. INTRODUCTION

Cyberspace has grown with the increase and accessibility of the internet throughout the world. The occurrences of cyber-attacks are becoming more probable in the world of the internet. Cybersecurity has become an important factor for business organizations due to the existing and increasing dependency on information technology for storage, processing, and transferring of valuable information. Attackers are always seeking for rich information they can exploit. It has been highlighted that cyber security is crucial concern as new types of cyber threats are emerging. Attackers are people who unauthorized get into some system and then purposefully sent malicious packets to users system trying to get hold of, modify or corrupt the information in a prohibited manner are involved in an unlawful activity [1]. Attacks refered to deliberate attempts to steal data or gain access to systems and networks that are not authorized to use. According to NTT's Global Threat Intelligence Report for 2024[2], new vulnerabilities are spawning constantly, driven by a powerful combination of factors, changing threat landscapes, technological advancements and the continuing upsurge of sophisticated cybercriminal activities. The report highlights that vulnerabilities in widely-used software, particularly in cloud-based applications, are frequently targeted. the top five most attacked sectors represent a large portion of the total cyberattacks. There are also technology, manufacturing, finance, transport & distribution and public sectors. Together, these sectors accounted for 74% of all global attacks. It also points to the continued targeting of these industries because of their central role in global operations and of the value of their data.

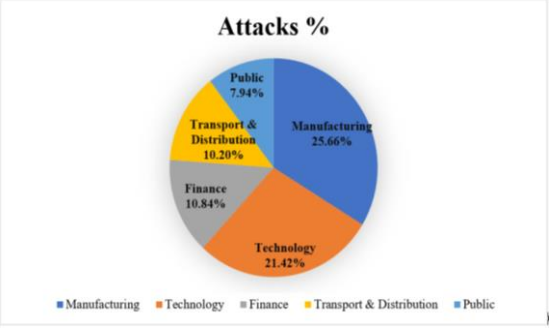


Figure 0 Global Threat Intelligence Report 2024

These Attacks are frequently uncommon in nature, making it necessary to create intelligent algorithms in order to recognize them. The network traffic is watched by an intrusion detection system (IDS), which reports any violations depending on a predefined security level [3]. IDS are often divided into a variety of types, such as host-based (HIDS) or network-based (NIDS), depending on the user's perspective. (NIDS) are used to monitor network activity and network-connected devices in order to identify any unwanted access. However, it is important to note that (HIDS) specifically focuses on addressing the risks associated with the host system [4]. Two techniques may be used to keep tabs on harmful activities. One approach is signature-based detection, which analyzes the malware's unique digital traces or signatures to identify it. In contrast, anomaly-based detection detects unknown, suspicious actions by identifying deviations from regular patterns. Rather than relying on machine learning to identify recognized risk, an anomaly-based detection system uses a normalized baseline to navigate the detection process and differentiate normal behavior from abnormal patterns. Every network transaction is contrasted with the baseline, which represents the system's usual behavior [5].

Machine learning (ML) techniques have grown into an efficient tool for detecting the digital threats. However, due to their complexity and dynamic nature, classic ML models frequently find it difficult to identify and categories new types of cyber-attacks. This study explores a machine learning model that merges K nearest neighbors and Random Forest approach to enhance cybersecurity measures effectively and efficiently. The KNN algorithm is well known for its simplicity and effectiveness, in categorizing data, by comparing it to existing data points to identify outliers and anomalies. On the other hand the Random Forest technique is a learning method that creates multiple decision trees to deliver precise results and resilience when handling vast and intricate datasets. By combining the strengths of both KNN and Random Forest algorithms in a model, for cybersecurity threat detection enhances the systems efficiency significantly. The hybrid model leverages KNNs anomaly detection capabilities along with random forests classification accuracy and efficient management of datasets to provide a holistic solution, to the increasing cyber threat issues. Different performance metrics have been used to evaluate the model performance. The findings indicate that, when compared to the current model, our hybrid model is more effective in identifying intrusions.

The structure of this paper is organized as follows: Section 2 provides an overview of related work. Section 3 outlines the suggested methodology in detail. In Section 4, the paper presents the specifics of the datasets used, along with the experiments conducted and the corresponding results. Finally, Section 5 encompasses the conclusion, discussion, and potential future work.

II. Related Work

In order to identify network threats by building efficient models, several researchers support the integration of machine learning (ML) technologies for intrusion detection. An effort has been made by authors [6] to suggest using naive Bayes and contrast it with decision trees for the purpose of identifying anomalous networks. Bayesian analysis is known for its efficiency, in computing due its simple structure and ability to be updated incrementally in real time cybersecurity scenarios. This makes it a practical choice for real time cybersecurity uses. These discoveries highlight the effectiveness of Naive Bayesian analysis in detecting intrusions quickly and reliably making it a valuable asset, in the field of cybersecurity. The study [7] has focused on using a genetic algorithm (GA) to enhance IDS that are based on support vector machines (SVM). This new approach combines the features of genetic algorithm (GA) and support vector machines (SVM) providing a strong and effective answer, for tackling cybersecurity issues. The research article [8] presents the design architecture of intrusion detection systems utilizing neural network self-organizing maps. Furthermore, it explores the use of user anomalous behavior as a basis for intrusion detection. In the research study [9] authors propose two hybrid modeling techniques for Intrusion Detection Systems (IDS). The first one is based on a hierarchical hybrid intelligent system model DT-SVM, where Decision Trees (DT) and Support Vector Machines (SVM) are combined. The second one, an ensemble method which combines DT and SVM as base classifiers to improve detection accuracy by using the strength of both algorithms. The study [10] proposed to use such data mining algorithms as random forests to enhance IDSs in misuse, anomaly, and hybrid networks detection. Random forests build intrusion patterns from training data automatically for misuse detection and detect anomalies in the process of identifying outliers for anomaly detection. The study [11] suggested IDS-NNM, the Intrusion Detection System using a neural network model to enhance security. In its implementation, it makes use of a specially designed window-based feature extraction technique, parameterized using actual network data from critical infrastructures, normality modeling through error back-propagation and the Levenberg-Marquardt algorithms. A host-based IDS was developed by the authors of [12] by combining of log file analysis for misuse detection and Backpropagation (BP) neural networks for anomaly detection. A technique based on fuzzy logic was developed by the authors of [13] for efficiently

recognizing network intrusion activities. Fuzzy logic technique uses automated rule generation from frequent items to increase detection capability regarding intrusions. The Authors of [14] suggested an IDS that effectively detects various forms of network intrusions by utilizing genetic algorithm (GA). GA parameters and evolution processes are thoroughly examined and put into practice. Using a naive Bayesian classifier to detect potential intrusions, the authors of [15] also proposed a multi-layer Bayesian-based IDS. The Authors in [16] presented a novel intrusion detection method that increases the detection rate of unidentified assaults and decreases time complexity. which uses the C4.5 decision tree technique firstly, split the network information to segments that are smaller and then for the subsets generate several SVM models. In study [17] , the authors employ RNN networks for botnet anomaly detection. They leverage the effectiveness of RNN networks specifically on timing features to enhance the accuracy of classification. By integrating K-MEANS clustering with the KNN classifier, the authors in research article [18] increase the existing KNN classifier's accuracy of detection. SVM and GA are used by the authors in study [19] to improve the accuracy of network attack identification by optimizing the selection, parameters, and weights of SVM features. The Authors in [20] used a variety of ML methods to spot incursion on KDDCUP'99 dataset in order to assess how successful these classifiers were. To increase performance, the training and testing data of the dataset were separated using k-fold cross-validation (k=10). They achieved a DT accuracy rate of 94.00%, which was the highest of any other methods. The Authors in study [21] suggested utilizing deep learning architectures to create a network intrusion detection system (IDS) that is durable and adaptable for identifying and categorizing network intrusions. The focus is on how DNNs might provide adaptable IDS with learning power to identify known and novel or zero-day network behavioral traits, hence expelling the systems invader and lowering the risk of penetration. The study [22] proposed machine learning-based Intrusion Detection Systems specifically meant for IoMT. Techniques like Multinomial Naive Bayes, Logistic Regression, Decision Tree, Ensemble Voting Classifier, and different boosting methods were studied. Notably, the Adaptive Boosting algorithm proved to outperform the existing models from all angles of consideration related to accuracy, precision, recall, F1-score, False Detection Rate (FDR), and False Positive Rate (FPR) used in ToN-IoT based IDS models. After literature analysis, we proposed a framework to accomplish our goal and the next section will describe methodology process.

III. METHODOLOGY

The most crucial task in any kind of research or study is to choose research method that is most appropriate to conclude the research study we have used the quantitative methodology for our research study. Some of major reasons for adaptation of are, quantitative methodology is well-suited to check causal relationship between variables, cause and effect of relationship between variables and shows the impact of variables on each other. Preprocessing of data has been carried out for data cleaning, feature scaling, and label encoding. Normalized reduced dataset has been taken through feature selection technique. Training and testing of model has performed on training and testing data during training and evaluation phase. The entire procedure followed during the study is depicted in fig 2.

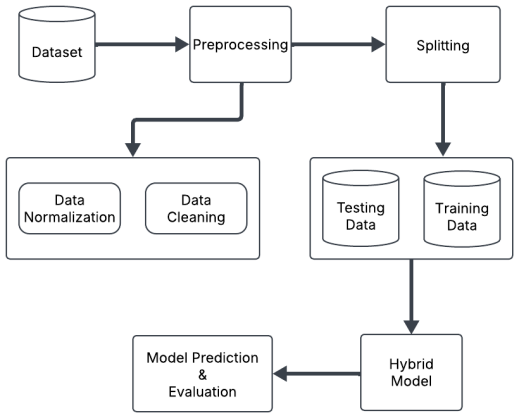


Figure 2 A Hybrid Model-Based Approach

Dataset

Dataset used in our research study to perform Experimentation is KDD CUP 1999 that we got from UCI Machine Learning Repository [23] . KDD CUP 1999[24] is well-known benchmarked dataset having 4000000 instances and 42 attributes and is divided into two categories either Normal or Intrusion, include various type of attacks Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L), Probing Attack. Since 1999, KDD'99 has been mostly used data set in the field of network intrusion detection and cybersecurity.

Exploratory Data Analysis (EDA)

When it comes to data-oriented science, Exploratory Data Analysis (EDA) is an essential step in the procedure with the primary goal of identifying patterns, describing characteristics, diagnosing issues and evaluating the validity of implied features. Data preparation and pre-analysis involving the handling of missing data, feature standardization, and cleansing of the dataset, was the first step of study which carried out on KDD dataset.

The EDA started with an extensive analysis of the data featuring summary statistics like the mean, median and the standard deviation for the dataset. EDA helps us in developing a better understanding of the structure of the data, but it also assisted in constructing the more appropriate machine learning algorithms with greater efficiency. As a result of this thorough investigation, we found out the variables that mattered most in the detection system. In examining the dataset, we found it appropriate to redefine our hybrid model to target higher accuracy. According to the graph shown in Fig. 3.2 illustrate correlation between different categories of data presented in dataset.

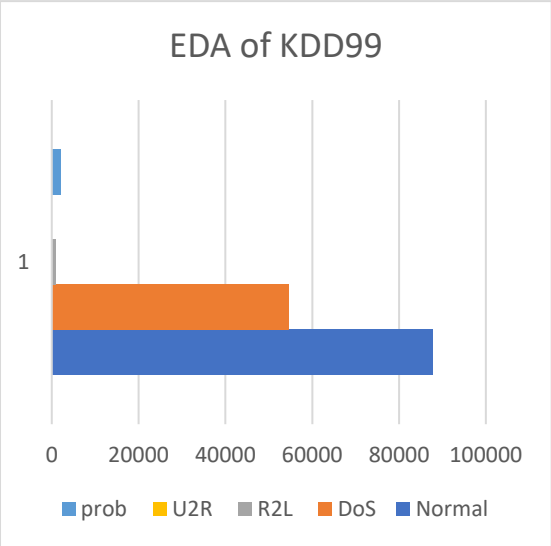


Figure 0 Exploratory Data Analysis

Preprocessing

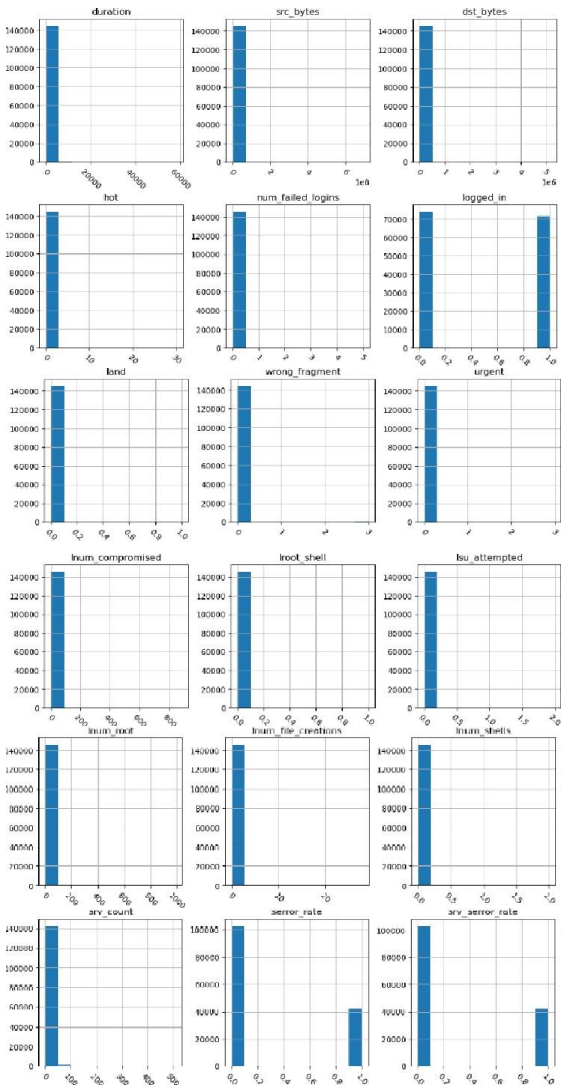
Preprocessing is the technique used to convert raw data into usable data form. That's often carried out to handle missing values, normalize features, and removing duplicate records to mitigate data imbalance and enhance training process. The goal of preprocessing is to make the training/testing process easier by suitable conversion and scaling the entire dataset [25].

Data cleaning

Data cleaning is fundamental and crucial step because the quality of the data directly impacts predictions obtained from the data. The approach and techniques used in data cleaning of handles missing values by e.g. Dismissal of rows and columns containing missing values for example if they are not vital for the analysis, or asking them to be filled up with statistical methods (mean, median or mode). Identifying and eliminating duplicate entries from the dataset, which can cause redundancy and overshadow results. Detection and Handling Outlier, usually defined as the data values that are significantly deviate from other values of data set. These values put their impact on data analysis or made some difficulty for model to predict accurately. Detection and Handling Outlier, usually defined as the data values that are significantly deviate from other values of data set. these values put their impact on data analysis or made some difficulty for model to predict accurately. Original data set has 42 attributes after performing preprocessing reduced to 30, focusing on most relevant feature and improving the model.

Feature Scaling

Feature scaling is one of the preprocessing steps which practices for rescaling range there by normalizing the different features or independent variable data. This enables each of our features to make an equivalent contribution when penalizing the model [26]. Two techniques for feature scaling are common, Normalization: the values are scaled to a range between 0 and 1. Standardization: Scale the values to have a mean of 0 and a standard deviation of 1. Proper feature scaling has the ability to increase both the accuracy and speed at which machine learning models are capable of making predictions.



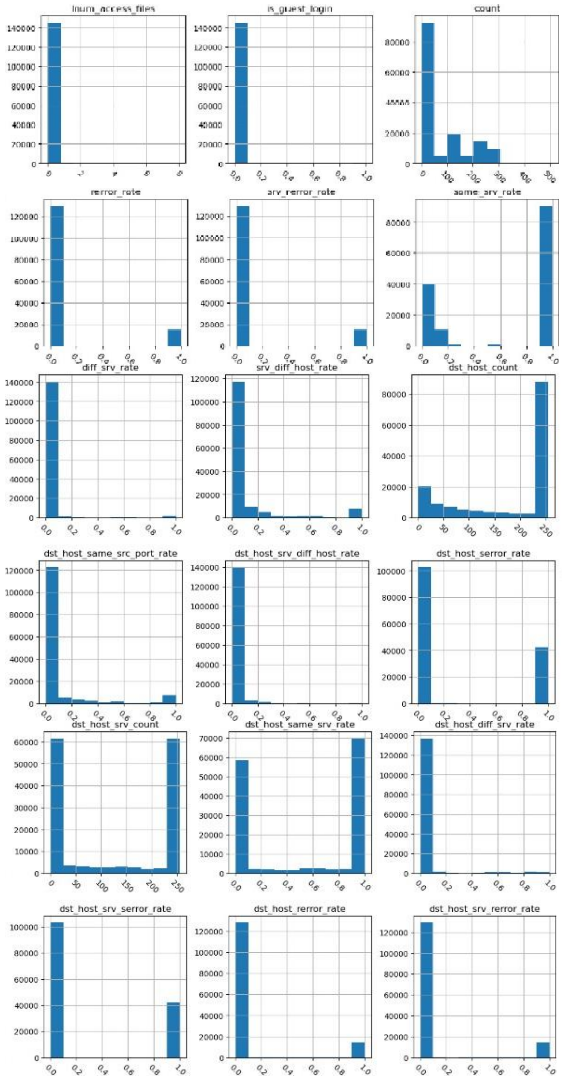


Figure 4 Distribution of Numeric Feature graph

Features Selection

The most important and crucial aspect of preprocessing data is features selection in order to perform classification. The most important and crucial aspect of preprocessing data is features selection in order to perform classification. It minimizes the features as compared to target classes, as well as unnecessary features, and redundant data that lead to target class errors. The basic goal of selecting a feature is selecting the most optimal subset of features for the training of model [27].

This study employed Correlation-based Feature Selection (CFS). CFS is one very known and used filters. Correlation based Feature Selection (CFS), is a simple filter-based algorithm that rates feature subsets depending on the correlation between features according to heuristic evaluation function. In general, the evaluation function favors sets that contain features which correlate well with class and poorly with one another. Non relevant features will be given importance as they will have low correlation with the class. Such redundant features should be reduced as they will have high correlation with one or more other features.

Table 1 Distribution of Testing Dataset

Attack Category	Number of Sample
Normal	87831
DoS	54572
R2L	999
U2R	52
prob	2131

IV. Proposition of The Hybrid Model

The K-Nearest Neighbors (KNN), Random Forest (RF) and XGBoost models have been integrated with the aim of improving the precision of cybersecurity threat detection. The model is a hybrid that uses a stacking approach to use the KNN algorithm's predictions as a input to the random forest classifier. In this case, the KNN algorithm is treated as the first level learner. This is because KNN identify patterns in the known features of dataset, which is particularly helpful for anomaly detection. The Random Forest, treated as the second level learner, use prediction of KNN model as input to RF this can assist RF in learning from KNN predictions and identify more complicated patterns in dataset. Different approaches like Stacking, Feature Engineering and Ensemble of Model are to use to combine algorithms effectively [28]. The KNN stacking strategy used in this research paper empowers KNN to take the detection of anomalies further by converting its output into features that Random Forest will exploit. Random Forest is known to be powerful in multivariate analysis owing to its structure of being an ensemble of multiple decision trees, it can handle large datasets. Stacking is able to take the predictions made by KNN and refine the detection process and therefore increase the prediction accuracy and reduce false positives. The hybrid system was able to use the best features of both algorithms. In order to improve the performance of the model even further, XGBoost is implemented as the third-level learner in hierachy, aiming to refine the overall predictive ability and error in the system. While XGBoost aims at enhancing the prediction accuracy of the Random Forest approach by reducing the false alarms it provides, by its nature of learning complex interactions and sparsity. The combination merges the abilities of all three algorithms: KNN for fast execution of odd noise detection, Random Forest for complex tasks and XGBoost for refinement and accuracy. The ensemble method guarantees that the hybrid has the optimum degree of both generalization and specialization, which improves the quality of detection of the cyber security threats. Fig 5 Illustrates Stacking Approach

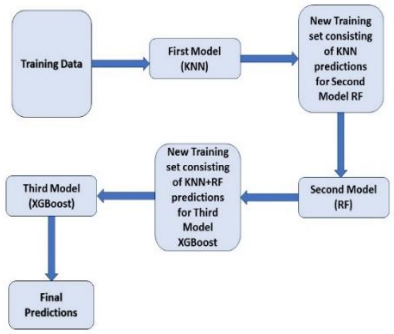


Figure 5 Stacking Approach

Evaluation

Performance evaluation of the proposed hybrid model measured through critical metrics to check its efficacy and robustness. Metrics used to evaluate are Confusion matrix, Accuracy (precision), Detection Rate (Recall) and F-measure.

Confusion matrix

The Confusion Matrix is actually a very strong validation metrics. it provides the comprehensive True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) record. In fact, all other important measures including the accuracy, the precision, the recall and even the F1-score can be derived from confusion matrix [29].

Table 2 Confusion Matrix

Actual Label	Predicted Label		
		Yes	No
	Yes	TP	FN
	No	FP	TN

• *True Positive (TP):* total number of attacks predicted positive and actual were positive.
• *False Negative (FN):* total number of attacks predicted negative and actually were positive.
• *False Positive (FP):* total number of attacks predicted positive and actual were negative.
• *True Negative (TN):* total number of attacks predicted negative and actual were negative.

Accuracy

Accuracy is crucial performance metric for classification model, the percentage of correctly classified instances over the sets of all instances. That means the model is being stellar in both the positive and negative classes [30]

$$Accuracy = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

Precision

A model's precision is a performance measurement that quantifies how closely positive predictions are predicted by the model. A high precision model has a low false positive rate, and is reliable at making positive classifications [30].

$$Precision = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

Recall

recall or sensitivity, it actually measures how well a model can predict all actual positive instances, since it is the proportion of actual high instances that it has correctly identified. A high recall promotes models that have eyes to spot many positive cases and as a result, the chance of missing a real positive is reduced[30].

$$Recall = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

Error Rate

Error rate is a method that measures the performance of a classification model as the ratio of incorrect predictions, to the total predictions. The less accurate a model, the higher the error rate is, so minimizing errors is highly important for reliability of the model.

$$Error\ Rate = \frac{(\text{FP} + \text{FN})}{(TP + TN + FP + FP)}$$

F-Measure

F-measure is the important performance evaluation metric that provide single evaluation value by balancing the recall and precision [30].

$$F - Measure = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)}$$

V. RESULTS

The results elaborate the better performance of the hybrid model in terms of accuracy, precision, recall, and F1-score as compared to single models such as DT, Bayes Net, Random Tree and SVM. Such comparison also explains that hybrid approach is good in terms of reducing false positive and improve detection rate in the case of cyber threats. The Stacked model is better than KNN or RF operated alone because KNN is more accurate in outlier detection while RF can manage complex classification tasks and large data sets for more generic threat detection with few complexities. Results of the research provide some insights on how hybrid models could be used to strengthen cyber security in the future as they provide a better and accurate detection system. The comparison of results is given in table 3.

Table 3 Performance Evaluation Results

Experimental Analysis	Accuracy	Recall	Precision	Error	F-Measure
KNN+RF+ XGBoost	0.9989	0.9985	0.9983	0.001099	0.9986
DT	0.9710	0.9658	0.9612	0.028999	0.9635
Bayes Net	0.9323	0.9424	0.8906	0.067603	0.9170
Random Tree	0.9830	0.9832	0.9741	0.016994	0.9786
SVM	0.9781	0.9782	0.9670	0.021877	0.9725

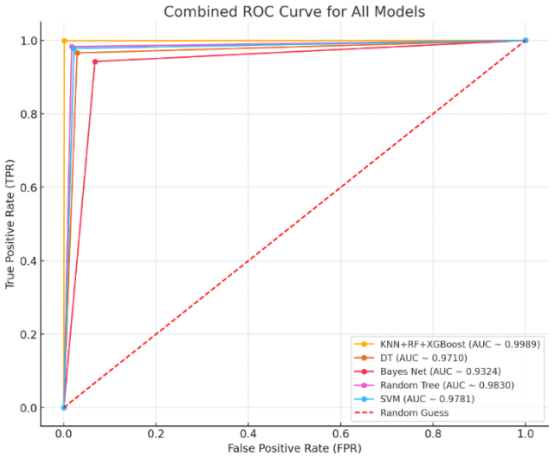


Figure 6 Combined ROC Curve for All Models

VI. Conclusion

Study proposed a hybrid machine learning model which is based on K-Nearest Neighbors and Random Forest to boost the detection of cybersecurity threats. The combination of these two algorithms enhances the detection procedure because KNN provides the capability of detecting anomalies by identifying outliers when doing proximity-based classification and RF is strong and accurate with large and complex datasets. The integrated model tackles the weaknesses of each single algorithm, thus an enhanced system is developed for cyber threats to be detected in real-time. It was determined by the performance evaluation using various evaluation metrics which include accuracy, precision, recall and F1 score that the hybrid model performed comparatively better than the standalone models. The stacking approach of KNN where KNN predictions act as inputs for RF enhanced pattern recognition and consequently more detection and decrease in false positive rate. This makes the model optimal in Cybersecurity situations because timely and accurate detection of the threats is a priority. In future, we will tend to apply this proposed hybrid model on other datasets and evaluate its performance to enhance it further.

References

[1] M. A. Talukder *et al.*, "A dependable hybrid machine learning model for network intrusion detection," *Journal of Information Security and Applications*, vol. 72, p. 103405, Feb. 2023, doi: 10.1016/J.JISA.2022.103405.

[2] G. Garten, "Forward from CTO, NTT SECURITY HOLDINGS."

[3] L. Ashiku and C. Dagli, "Network Intrusion Detection System using Deep Learning," *Procedia Comput Sci*, vol. 185, pp. 239–247, Jan. 2021, doi: 10.1016/J.PROCS.2021.05.025.

[4] "Difference between HIDs and NIDs - GeeksforGeeks." Accessed: May 05, 2023. [Online]. Available: https://www.geeksforgeeks.org/difference-between-hids-and-nids/

[5] "Intrusion Detection System (IDS): Signature vs. Anomaly-Based - N-able." Accessed: May 05, 2023. [Online]. Available: https://www.n-able.com/blog/intrusion-detection-system

[6] N. Ben Amor, S. Benferhat, and Z. Elouedi, "Naive Bayes vs decision trees in intrusion detection systems," 2004, doi: 10.1145/967900.967989.

[7] D. S. Kim, H. N. Nguyen, and J. S. Park, "Genetic algorithm to improve SVM based network intrusion detection system," *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, vol. 2, pp. 155–158, 2005, doi: 10.1109/AINA.2005.191.

[8] L. Vokorokos, A. Baláž, and M. Chovanec, "INTRUSION DETECTION SYSTEM USING SELF ORGANIZING MAP," *Acta Electrotechnica et Informatica No. 1*, vol. 6, p. 1, 2006.

[9] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 114–132, Jan. 2007, doi: 10.1016/J.JNCA.2005.06.003.

[10] J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 38, no. 5, pp. 649–659, 2008, doi: 10.1109/TSMCC.2008.923876.

[11] O. Linda, T. Vollmer, and M. Manic, "Neural Network based intrusion detection system for critical infrastructures," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1827–1834, 2009, doi: 10.1109/IJCNN.2009.5178592.

[12] Y. Lin, Y. Zhang, and Y. J. Ou, "The design and implementation of host-based intrusion detection system," *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010*, pp. 595–598, 2010, doi: 10.1109/IITSI.2010.127.

[13] R. Shanmugavadivu, "NETWORK INTRUSION DETECTION SYSTEM USING FUZZY LOGIC".

[14] M. S. Hoque, A. Mukit, M. Abu, and N. Bikas, "AN IMPLEMENTATION OF INTRUSION DETECTION SYSTEM

USING GENETIC ALGORITHM," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 4, no. 2, 2012, doi: 10.5121/ijnsa.2012.4208.

[15] H. Altwaijry, "Bayesian based intrusion detection system," *Lecture Notes in Electrical Engineering*, vol. 170 LNEE, pp. 29–44, 2013, doi: 10.1007/978-94-007-4786-9_3/COVER.

[16] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Syst Appl*, vol. 41, no. 4, pp. 1690–1700, Mar. 2014, doi: 10.1016/J.ESWA.2013.08.066.

[17] P. Torres, C. Catania, S. Garcia, and C. G. Garino, "An analysis of Recurrent Neural Networks for Botnet detection behavior," *2016 IEEE Biennial Congress of Argentina, ARGENCON 2016*, Oct. 2016, doi: 10.1109/ARGENCON.2016.7585247.

[18] H. Shapoorifard and P. Shamsinejad, "Intrusion Detection using a Novel Hybrid Method Incorporating an Improved KNN," *Int J Comput Appl*, vol. 173, no. 1, pp. 975–8887, 2017.

[19] P. Tao, Z. Sun, and Z. Sun, "An Improved Intrusion Detection Algorithm Based on GA and SVM," *IEEE Access*, vol. 6, pp. 13624–13631, Mar. 2018, doi: 10.1109/ACCESS.2018.2810198.

[20] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. Minhaz Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques," *Communications in Computer and Information Science*, vol. 1235 CCIS, pp. 121–131, 2020, doi: 10.1007/978-981-15-6648-6_10/FIGURES/2.

[21] L. Ashiku and C. Dagli, "Network Intrusion Detection System using Deep Learning," *Procedia Comput Sci*, vol. 185, pp. 239–247, Jan. 2021, doi: 10.1016/J.PROCS.2021.05.025.

[22] P. Kulshrestha and T. V. Vijay Kumar, "Machine learning based intrusion detection system for IoMT," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 5, pp. 1802–1814, May 2024, doi: 10.1007/S13198-023-02119-4/FIGURES/12.

[23] "Home - UCI Machine Learning Repository." Accessed: Jul. 26, 2023. [Online]. Available: https://archive.ics.uci.edu/

[24] "KDD Cup 1999 Data - UCI Machine Learning Repository." Accessed: Jul. 26, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/130/kdd+cup+1999+data

[25] K. C. Santos, R. S. Miani, and F. de Oliveira Silva, "Evaluating the Impact of Data Preprocessing Techniques on the Performance of Intrusion Detection Systems," *Journal of Network and Systems Management*, vol. 32, no. 2, pp. 1–54, Apr. 2024, doi: 10.1007/S10922-024-09813-Z/FIGURES/24.

[26] H. Alshaher, "Studying the Effects of Feature Scaling in Machine Learning".

[27] U. Rashid, M. Faheem Saleem, S. Rasool, A. Abdullah, H. Mustafa, and A. Iqbal, "Anomaly Detection using Clustering (K-Means with DBSCAN) and SMO," *Journal of Computing & Biomedical Informatics*, vol. 07, no. 02, Sep. 2024, doi: 10.56979/702/2024.

[28] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front Comput Sci*, vol. 14, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/S11704-019-8208-Z/METRICS.

[29] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, pp. 83–106, Jul. 2021, doi: 10.1016/B978-0-12-818366-3.00005-8.

[30] A. Meryem, "Hybrid intrusion detection system using machine learning," 2020. [Online]. Available: www.idg.com/tools-for-