DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 11(3) WEBSITE: https://fugus-ijsgs.com

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL SUSTAINABILITY
(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

Development of K-Neareast Neighbor Based Model for Network Intrusion Detection Using Random Forest for Feature Selection

[*1]Ogah Muhammad Usman., [2] Tahir Abdulhakim [3] Ayinla Munirat Tope,

[1]Department of Computer Science, Nasarawa State University Keffi, Nasarawa State, Nigeria
[2]Department of Computer Science, Nasarawa State University Keffi, Nasarawa State, Nigeria
[3]Department of Computer Science, Nasarawa State University Keffi, Nasarawa State, Nigeria
Keffi, Nasarawa State.

***Corresponding Author's Email Address:** mohammedogah@gmail.com, abdullahitahir@nsuk.edu.ng, ayinlamunirat@gmail.com

Received on: August, 2025 **Revised and Accepted on:** September, 2025 **Published on:** October, 2025

ABSTRACT

Networked monitoring has emerged as one of the primary security technological advances, and network security challenges have gained increasing attention due to the extensive usage of the web. As a result, conventional signature-based and anomaly-centered intrusion detection systems are not effective enough. The efficacy and preciseness of network security measures are significantly impacted by the beginning information source's substantial dimension and volume of content. An inventive method for choosing the key attributes and reducing the overall length of information was the random forest-based feature selection, which simultaneously increased the Ture Positive Rate and decreased the False Positive Rate. Consequently, the classifier and selection of features both contribute significantly to improving network monitoring performance. On the simulated dataset, the suggested K-nearest neighbor and Random Forest perform admirably and successfully ensure the kNN classifier's progress. Simulations and experiments were conducted using the R programming language, and the results indicate that the model we developed reaches 99% accuracy. The experimental findings demonstrate the effectiveness

Keywords: Network Intrusion Detection System, Machine learning, K-nearest neighbor, Random Forest performs, R programming language.

I.0 INTRODUCTION

Security and defense against multiple cyberattacks have become urgent issues in recent years (Mubarak *et al.*, 2024). The primary cause of this is the enormous expansion of computing infrastructure and the numerous pertinent apps that people and organizations utilize for reasons that are personal or professional, particularly since the Worldwide Web of Things (IoT) gained traction (Chou and Jiang, 2021). Cyberattacks seriously harm enormous networks and result in substantial financial losses (Lv *et al.*, 2020

They look for abnormal activity or illegal usage and try to prevent people, machines, or parts of systems from accessing these amenities (Ajdani and Ghaffary, 2021) utilized Adaboost based on network intusion detection system with data enhancement techniques such as SMOTE , oversampling and cost sensitive methods.
Feature selection is utilized as a means to improve the outcomes of classication. Wang *et al.* (2023) conducted NIDS but their accuracy can be improved by using Random forest feature importance to enhance the model. Also they didn't train their model on other intrusion detection dataset to have a base performance.

One of the main purpose of using Random Forest for NIDS is its accuracy, speed, as the capability of detecting intrusions in real-time is essential, perform importance features extraction through clustering over data.The resistance of Random Forest to noise (Liu *et al.*, 2022) .

Machine learning is a scientific study of algorithms through the application of AI that was translate and improve from classifiers otherwise known as experience without explicitly being programmed (Rupa and Srinivasu, 2022) (Khorram and Akhan, 2021). This research work is subjected to develop an efficient K-Nearest Neighbor (KNN) based on Feature selection (FS) of Random forest (RF) importance variable. This research is limited to Anomaly detection based IDS using Unswnb-15 dataset. The study was classified as follows: section 1, introduction provides a rationale for the current research; section 2 offers a relevant literature review and the research framework; section 3 details the methodology employed in the current research; and section 4 offers results and discussion, while section 5 makes a conclusion and recommendation

1.2 Literature Review

Several NIDS have been implemented curbs attack (Mubaraq *et al.*, 2024). These NIDSs are widely used to safeguards systems and network from been compromised (Ahmad *et al.*, 2022) . The systems for detecting intrusions IDS may be divided into two primary categories: anomaly-based (AIDS) and

DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118

signature-based (SIDS). IDS can be categorized based on how it is deployed or how it detects threats. Figure 1 provides a taxonomy of categorization. IDS is further divided into host-based-IDS (HIDS) and NIDS from the deployment-based IDS perspective (Hang et al., 2023).
).

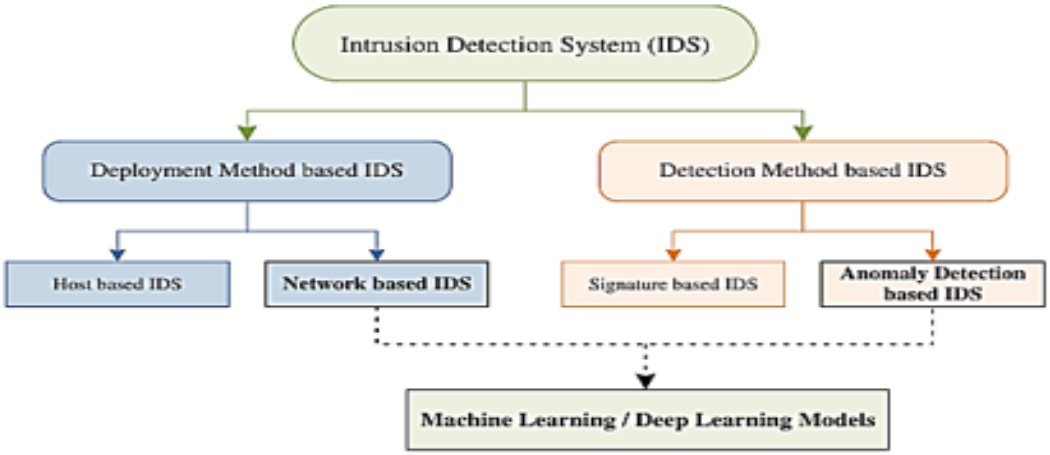


Figure 1 : Classification Taxonomy (Wang et al. 2023)

Figure .1 indicate the classification taxonomy of intrusion detection system and its classification in deployment method based IDS and detection method based IDS. Accordng to Ahmad *et al.* (2022) they introduced an improve accuracy or performance with experience in NIDS.

Zhiqiang and Yucheng (2022) proposed PSO-KNN for NIDS, only KNN is considered towards improving the performance of network intrusion detection. However, SVM and naive Bayes might also be selected for this purpose, in view of the problem that the PSO converges too fast and easily falls into local optimum an integrated optimization algorithm with efficiency and accuracy of 90.%.

2.0 METHODOLOGY

In this study, the methods that were used to realize the aim and objective of this research work. figure 2 shows the methodology deployed to achieve each objective. It starts with

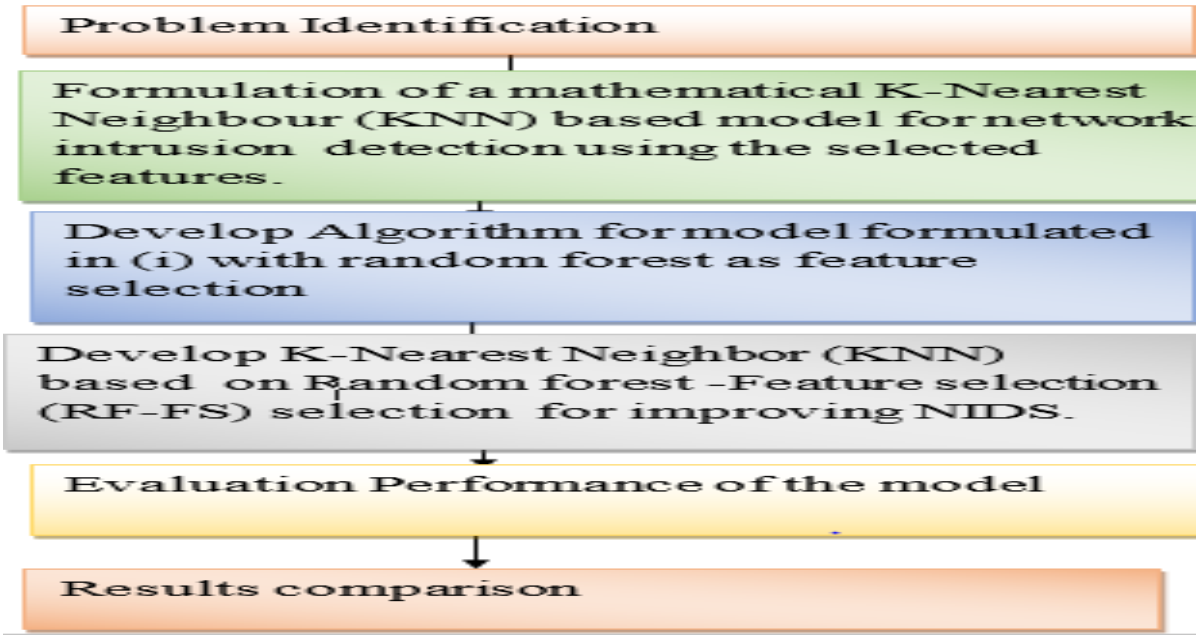


Figure 2. Flow Chart of Research Methodology

A Problem Identification

Figure 2 shows the methodology deployed to achieve each objective. It starts with formulation of a mathematical K-Neareast Neigbour (KNN) based model for Network intrusion detection detection using the selected features, Development of K-Neareast Neigbour (KNN) based Network intrusion detection detection model using the selected features, thereby,

DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118

Develop an algorithm for model formulated in "ii" with Random forest as feature selection by adding the component of random forest to KNN algorithm it has enhance the model and performance the evaluation of the model develop K-Neareast Neigbour (KNN) using Let:

Di be $IjIj$ to each feature jj be Random Forest assigns importance scores

Let RF(X) represent the function that selects the top kk features using Random Forest based on their importance scores.

Let kNN(x,D) represent the kNN algorithm which predicts the class label for the data point xx using the dataset DD

The mathematical model can be represented as follows:

$$d_j^i = \max\left[\sqrt{p_j^2 - \left(C_j^i\right)^2}\right] \forall P \& c_i \quad (1)$$

$$D^i = d^i - \sum_{j-1}^{n} \frac{\sqrt{test\_sample_j^2 \left(C_j^i\right)^2}}{d_j^i} \times d_j^i \quad (2)$$

Where Di is the approximate distance of the test sample from the i[th] cluster border, and d i is the Euclidean Accuracy, False positive rate, Precision, Recall, F-measure and Error-rate.

B Formulation of Mathematical Model of the K-Nearest Neighbor (KNN)

classification.

distance between the test sample and the center of the i[th] cluster (C i).

(3)

The combined model can be represented as:

Combined(x,D)=kNN(RF(X),D)Combined(x,D)=kNN (RF(X),D) (4)

This function takes a new data point xx and the dataset DD as input, selects the top kk features using Random Forest (RF(X)RF(X)), and then predicts the class label for xx using kNN (kNNkNN).

C Model saving

The trained kNN model is saved as a serialized joblib file ('nids_detection_model.pkl') for future use.

Figure 3. Implementation of Development of KNN

a.Training. The model is trained for a specified number of 9 (nine)iterations

b. During training, the model learns to make predictions that minimize the loss. Prediction: After training, the KNN model is used to predict whether network traffic represents anattack or not.These predictions are stored in the 'knn_predictions'test label.

c.The validation set accuracy reaches its peak at k=9, where it achieves an impressive accuracy of approximately 99.92%. This indicates that the new model performs optimally when considering nine nearest neighbors for classification

DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOL. 11(3) WEBSITE: https://fugus-ijsgs.com

Figure 4: RF-KNN Network Intrusion Detection Flowchart

Figure 4 is the flowchart of RF-KNN Network intrusion detection, from the beginning theUNSWNB-15 serve as the input before the Data-processing, if Not go back to RF feature Selection. The particle length or problem dimension maximum number of iterations is 9, the population size is 10, the use of a rectangular weight function ensures equal treatment of all neighbors within the distance threshold the validation set accuracy reaches its peak at k=9,

F Performance Evaluation

The selected metrics for carrying out the performance analysis are outlined as follows:

a) **Precision:** Precision can be represented as follows:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

b) **Accuracy**: The percentage of instances the model properly predicted

$$Accuracy \frac{TP+TN}{TP+FN+TN+FP} \quad (6)$$

c) **Recall:** Recall is the percentage of the two samples that are all also classified as positive

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

d) **AUC-ROC** :Instead, we will evaluate the ROC AUC, which ranges from 0 (the worst) to 1 (the best), with arbitrary conjecture scoring 0.5.

e) **Receiver Operating Characteristic (ROC):** performs bias criteria is changed.

$$ROC = \frac{Sensitivity+Specificity}{2} \quad (8)$$

f) **F1 Score:** It can also be seen as a performance measure that blends precision and recall.

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \quad (9)$$

3.0 RESULTS AND DISCUSSION

This experiment utilized R statistical package, trained on 80% of the dataset and tested on the remaining 20% to be able to access the model functionalities**,** The UNSW-nb15 dataset was selected for its extensive coverage of various network attacks and its substantial size, comprising over 2 million records. Obtained from the provided link,

A Missingness analysis

The missingness analysis showed on the UNSW-nb15 dataset for the development of a KNN-based model for network intrusion detection revealed that there were 6,426,675 missing values across the 49 columns, accounting for approximately 5.16% of the entire dataset.

Table 2. Missingness Analysis of the Data

Characteristics	Missing data
Total missing values	6426675
Percentage of missing values	5.16%

DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 11(3) WEBSITE: https://fugus-ijsgs.com

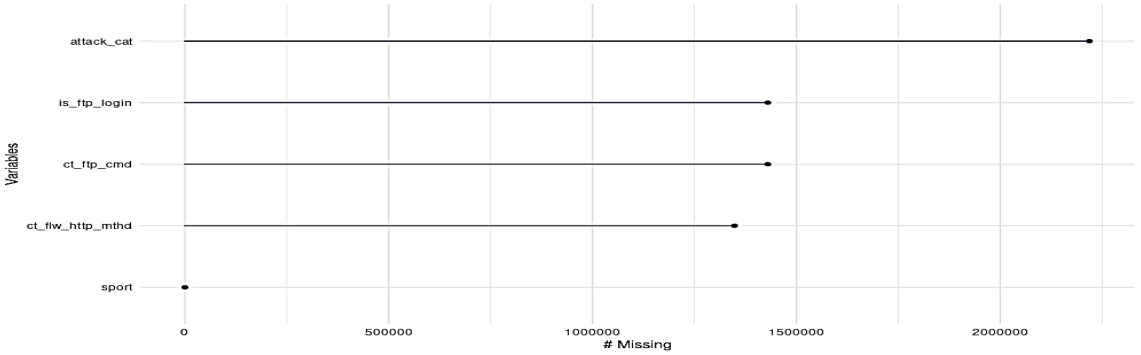


Figure 5. Missingness analysis

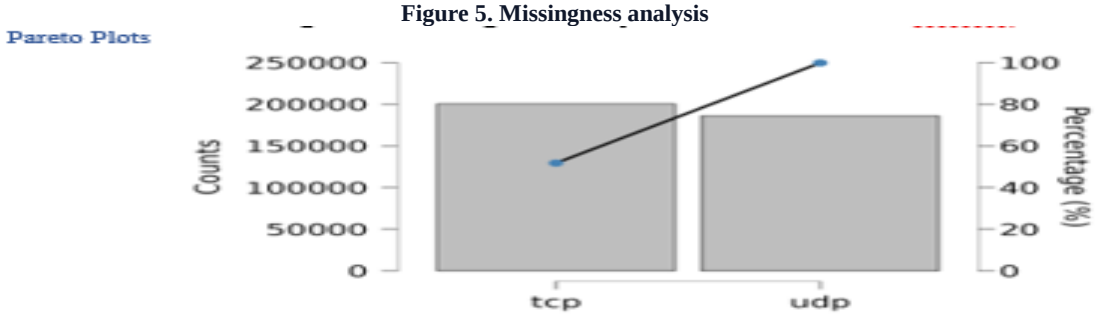


Figure 6: Pareto Plot Showing the Distribution of Proto

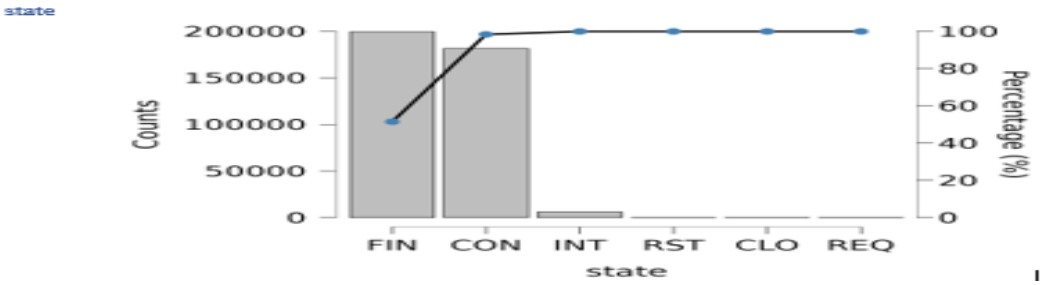


Figure 7: Pareto Plot showing The Distribution of State

Figure 6 to 7 presents the descriptive analysis of categorical data through pareto plots, offering valuable insights into the distribution of various categorical variables within the UNSW-NB15 dataset.

The examination of frequency tables for binary data within the data offers valuable insights techniques distribution , characteristics of these variables (Table

Table 3. Frequency Tables of Binary data

Frequencies for is_sm_ips_ports			is_ftp_login	
Values	**Frequency**	**Percent**	**Frequency**	**Percent**
0	387503	100	347502	89.677
1	0	0	40001	10.323
Missing	0	0	0	0
Total	387503	100	387503	100

Table 4: Random Forest Classification

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
30	7	248002	62001	77500	1.000	1.000	0.989

Note. The model is optimized with respect to the out-of-bag accuracy .

The model configuration details reveal key parameters and metrics of a random forest classification model designed for network traffic analysis (Table 4). Furthermore, the model's evaluation on a test set containing 77,500 samples reinforces its capacity for accurate classification tasks

DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 11(3) WEBSITE: https://fugus-ijsgs.com

Table .5: K-Nearest Neighbors Classification

Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
9	Rectangular	Euclidean	248002	62001	77500	0.999	0.999

Note. The model is optimized with respect to the validation set accuracy .

The K-Nearest Neighbors (KNN) model was fine-tuned with 9 nearest neighbors, utilizing rectangular weights and the Euclidean distance metric for similarity measurement. Trained on 248,002 instances, validated on 62,001, and tested on 77,500, the model achieved outstanding validation and test accuracies of 99.9% (Table 5). Leveraging the UNSW-nb15 dataset, comprising 2,540,044 network traffic instances collected over nine weeks, the KNN algorithm classified traffic as normal or malicious. The k value of 9 was optimized for maximum validation accuracy. Validation accuracy reached 99.9%. This rigorous approach ensures robust classification in network security scenarios, advancing threat detection capabilities effectively

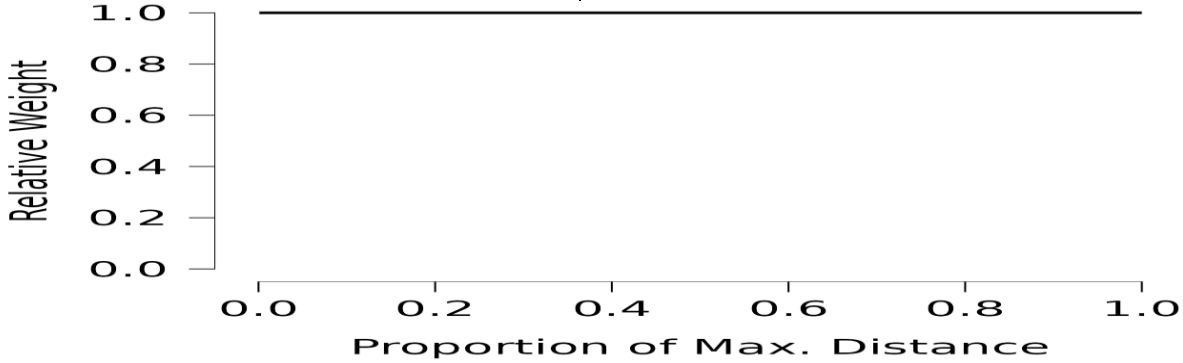


Figure 9. Rectangular Weight Function

Figure 9 provides a visual representation of the rectangular weight function employed in the K-Nearest Neighbors (KNN) model.

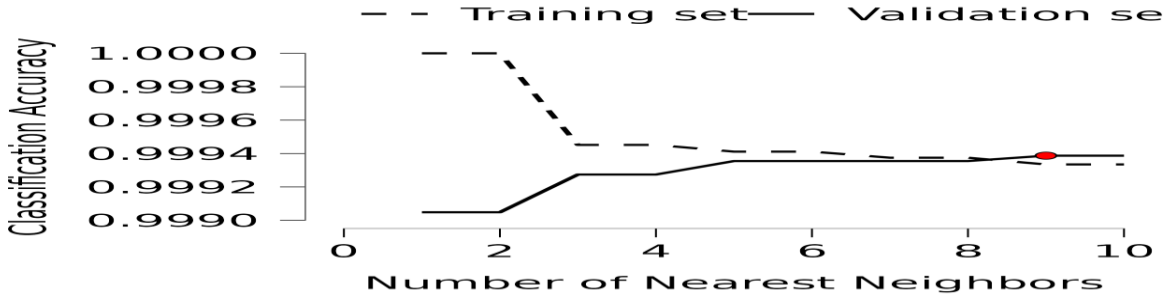


Figure 10. Classification Accuracy Plot

Figure 10 presents a classification accuracy plot representing the performance of the K-Nearest Neighbors (KNN) model. This plot illustrates how both the training set accuracy and validation set accuracy vary with the number of nearest neighbors (k) considered in the classification process.

Upon examination of the plot, it becomes apparent that the validation set accuracy reaches its peak at k=9, where it achieves an impressive accuracy of approximately 99.92%. This indicates that the model performs optimally when considering nine nearest neighbors for classification. However, it's notable that the training set accuracy remains consistently high, hovering around 100%, across different values of k.

This observation suggests that while the model achieves near-perfect accuracy on the training set, its performance on unseen data, as represented by the validation set, peaks at k=9..

DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOL. 11(3) WEBSITE: https://fugus-ijsgs.com

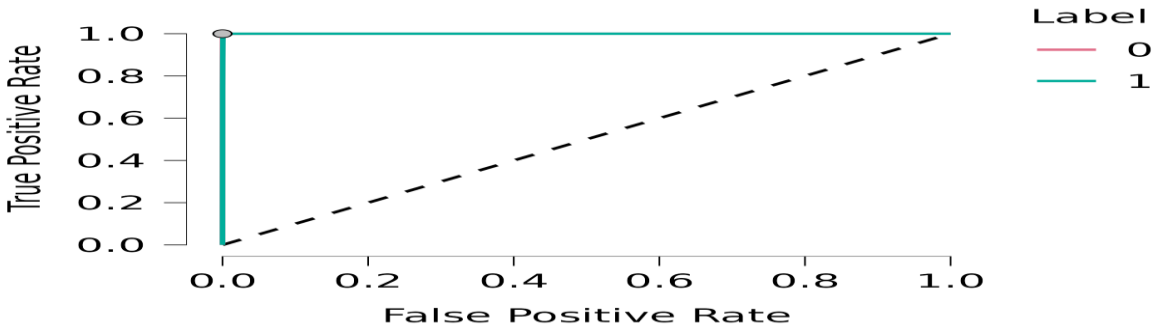


Figure 10. ROC Curves Plot

Figure 10 illustrates the True Positive Rate (TPR) versus False Positive Rate (FPR) trade-off for the KNN model, commonly represented to balance the true positive rate (correctly identifying positive instances) against the false positive rate (incorrectly classifying negative instances as positive). In this plot, the KNN model achieves an exceptional TPR of 1.0 for the positive class (Label = 1) when the FPR is approximately 0.008 or 0.8%. This indicates that the model maintains an impressively low false positive rate while sustaining a high true positive rate, demonstrating remarkable discrimination ability.

By synthesizing insights from the three plots, several key inferences can be drawn regarding the KNN model's performance. Firstly, the validation set accuracy reaches its peak at k=9, suggesting that this value maximizes the model's performance on unseen data. Secondly, the use of a rectangular weight function ensures equal treatment of all neighbors within the distance threshold, which may be advantageous for this particular problem. Finally, the minimal gap between training set accuracy and validation set accuracy implies that the model does not overfit to the training data, indicating robust generalization to unseen instances.

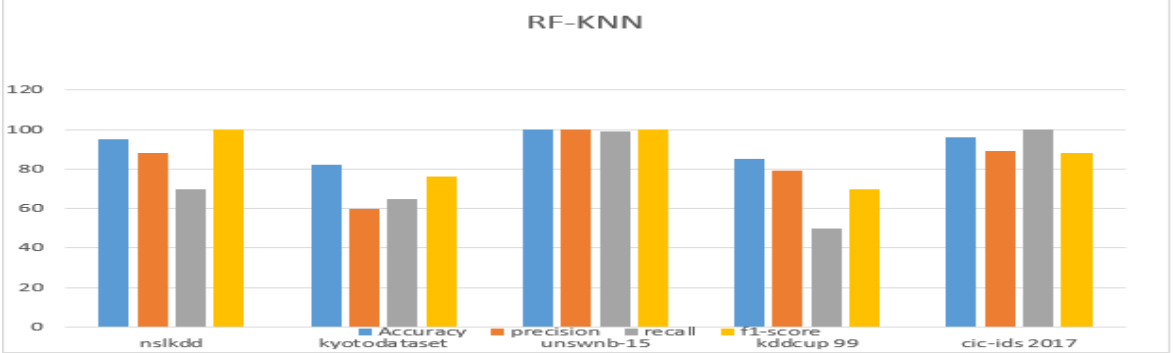
Overall, the KNN classification model exhibits outstanding performance on the provided dataset, achieving near-perfect accuracy and impressive evaluation metrics. Its efficacy in handling class imbalance, coupled with high precision and recall scores, renders it a valuable tool for the classification task at hand. However, it's crucial to acknowledge that the model's performance may vary based on dataset characteristics and problem domains, underscoring the importance of contextual interpretation and application.

Table 7. Performance Validation

Reference	Approach	Accuracy %
Huang *et al*. (2023)	Machine Learning Algorithm (Knn and NN)	96.7%
Wang *et al*. (2023)	KNN	92.68%
Thakkar and Lobiya (2022)	Deep learning	89.03%
Zhiqiang and Yucheng (2022)	KNN +PSO	90%
This study		
	KNN+RF	**99.9%**

Table 7 shows the accuracy when compared with existing models. Figure 4.9 shows the Performance comparison of RF-KNN on different intrusion detection dataset. In other to have a base performance, the algorithm was run on different dataset used by previous researcher

.



DOI: https://doi.org/10.57233/ijsgs.v11i3.919 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 11(3) WEBSITE: https://fugus-ijsgs.com

Figure 11. Performance Comparison of Random Forest -K-Neast Neighbour on Different Intrusion Detection Dataset.

4.0 CONCLUSION

Most of these techniques presently used for Network intrusion detection system involves traditional techniques which is not secure enough to prevent intruder from perpetuating attacks. In this study, we have evaluated and made comparisons of different intrusion detection dataset used by previously researcher inorder to have base performance. The UNSWNB-15 dataset was used, feature selection using the selected features of the unswnb-15 datasets This study successfully developed a model that is based on an improved k-neareast network and supports other security and mathematical model that improves accuracy, precision, true positive rate (TPR), F1-Score, pareto plot and has given much fair level of higher accuracy compared to baseline literatures.

5.0 RECOMMENDATIONS

However, it is recommended here that future advance research and broad studies are required in order to achieve low level of false positive r (FPR) that can guarantee security of system, network their data and information specially as a result of using deep learning algorithm.

REFERENCES

Ahmed, N., Ngadi, A. B., Sharif, J. M., Hussain, S., Uddin, M., Rathore, M. S., Iqbal, J., Abdelhaq, M., Alsaqour, R., Ullah, S. S., & Zuhra, F. T. (2022).Networkthreat detection using machine/deep learning in SDN-based platforms: A comprehensive analysis of state-of-the-art solutions, discussion, challenges,and future research direction.Sensors,22(20), 7896

Ajdani, M. & Ghaffary, H. (2021) Design network intrusion detection system using support vector machine. Int J Commun Syst 34(3): e4689

Alamed, H. (2020). A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. Expert Syst. Appl. 2020, 148, 113249.

Chou, D. & Jiang, M. (2021). A Survey on Data-driven Network Intrusion Detection. ACM Comput. Surv. (CSUR) 2021, 54, 1–36

Dini, P., Elhanashi, A., Begni, A., Saponara, S., Zheng, Q., & Gasmi, K. (2023). Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity. Applied Sciences, 13(13), 7507.

Huang, H. C., Liu, I. H., Lee, M. H., & Li, J. S. (2023). Anomaly Detection on Network Traffic for the Healthcare Internet of Things. Engineering Proceedings, 55(1), 3.

Khorram, K., K. & Akhan, K. (2021). Spam Detection using KNN, Back Propagation and Recurrent Neural Network.International Journal of Engineering Research & Technology 4 (09), 559 – 564.

Liu, Z., & Shi, Y. (2022). A hybrid IDS using GA-based feature selection method and random forest. Int. J. Mach. Learn. Comput, 12(2), 43-50..

Lv, L. Wang, W. Zhang, Z. & Liu, X. (2020). A novel intrusion detection system based on optimal hybrid kernel extreme learning machie ", Knowledge based systems, 3(195), 1-17

Mubaraq, S. (2024). Sensors and machine learning and AI operation-constrained process control method for sensor-aided industrial internet of things and smart factories. Measurement: Sensors, 25, 100668.

Ruba, M., & Srinivasin, P. (2022). A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing. Journal of Network and Computer Applications, 151, 102507. https://doi.org/10.1016/j.jnca.2019.102507

Sivagaminathan, V., Sharma, M., & Henge, S. K. (2023). Intrusion detection systems for wireless sensor networks using computational intelligence techniques. Cybersecurity, 6(1), 27.

Tufan, B. A., Comuzzi, M., & Rhee, K. H. (2021). TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. IEEE access, 7, 94497-94507.

.(2023). Fusion of statistical importance for feature selection in Deep Neural Network-based Intrusion Detection System. Inf. Fusion 2023, 90, 353–363.

Wang, Y., Liu, Z., Zheng, W., Wang, J., Shi, H., & Gu, M. (2023). A Combined Multi-Classification Network Intrusion Detection System Based on Feature Selection and Neural Network Improvement. Applied Sciences, 13(14), 8307.

Zhiqiang, P., & Yucheng, X. (2022). Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion CNN-LSTM-based approach. Sensors, 21(2), 626.