International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

# K Nearest Neighbor Based Model for Intrusion Detection System

**M.Nikhitha, M.A.Jabbar**

***Abstract**: Network security has become more important in this digital era due to the usage of information and communications technology (ICT). Data security is also one of the major issues in today's world. Due to the usage of this ICT technologies threat to network is also increasing. So in order to solve these problems the researchers has developed IDS that deals with network traffic to identify the harmful users and hackers in the computer. In this paper, we designed a model for IDS for classification of attacks using K-Nearest Neighbor classifier algorithm. KNN is a supervised and lazy machine learning classifier, it shows its best performance in terms of accuracy and classifications. Experimental analysis was conducted on ISCX dataset to judge the implementation of model. The Experimental outcome shows that our suggested model recorded an improved accuracy of 99.96%.*

***Index Terms**: Network Security, Intrusion detection system, Data Security, k nearest neighbor, Machine learning*

## I. INTRODUCTION

Day by day, Internet usage has been progressively increasing with the rapid growth of network and technology. This extreme and fast growth has given rise to new threats and vulnerabilities to networks. Intruders are the attackers or malicious users, designing new ways for the network intrusion. Previously the traditional approach like firewalls, encryption, authentication and VPN are used in order to secure the network infrastructure from intruders [9]. Intrusion Detection System is an upgraded version of these technologies, which is mainly used to identify attacks in the network and warns the system if any intruder has invaded into the system [3].

KNN is simplest among all the algorithms in machine learning. KNN is a lazy learning and also known as instance-based learning [2]. KNN is an algorithm that does not give any information about the structure of data which is a non-parametric. KNN algorithm is widely used for classification problem even it can be used for both classification and regression. KNN classifier shows the best in accuracy and produces better performance than others.

In this article we suggest IDS using k nearest neighbor classifier to improve accuracy of classifier in classification of different attack types. Section 2 describes about Literature Review and Related Work is narrated in Section 3. Our proposed work is explained in Section 4 and in Section 5 we analysis Experimental Results. Finally in section 6 we conclude.

**Revised Manuscript Received on 30 July 2019.**
***** Correspondence Author

**M.Nikhitha***, Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India.
**Dr.M.A.Jabbar**, Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India.

## II. LITERATURE REVIEW

### A. IDS (Intrusion detection system)

Intrusion is a type of attack or an intervention occurs within a system. IDS is a software or an application is for observing and analyzing the traffic within the system network and protecting it from intruders. Thes primary objective of IDS is to detect intrusions and identify various types of attacks.

**Attack Types**:

IDS plays crucial role in detecting the attacks.IDS is categorized into different attacks like DOS, Probe, R2L and U2R [4][5].

1. DOS attack: In this attack, the attacker avoids the authorized user from accessing the network or making the services unavailable to them. Ex: Smurf, Teardrop, Neptune.
2. Probe attack: In these types of attacks, before initiating the attack the attacker will gather all the required information of the target system. Ex: Satan, Ipsweep and Nmap.
3. User to Root (U2R) attack: The attacker starts as a normal account user then slowly exploits vulnerabilities to obtain illegal root access of the computer. Ex: Perl, Eject and load module.
4. Remote to Local (R2L) attack: In these, the trespasser wants to send packets to target machine remotely to expose vulnerabilities and obtain access of local target machine. Ex: multihop, send-mail and Imap

### B. KNN (K-Nearest Neighbor)

K-Nearest Neighbor is a data mining classifier. KNN is a supervised classifier, proposed by Fix and Hodges in 1951 [7].The output of the target variable is predicted by finding the k closest neighbor, by calculating the Euclidean Distance. It is a non-parametric classification technique which does not make any assumptions about underlying data [6]. The advantages of KNN [8] are:

i. .Easy to implement and understand.
ii. It is very effective and efficient if training data is very large.
iii. It is robust for noisy data.
iv. It constantly evolves and easily adapts to new environment.
v. Easy to implement for multi-class problem.

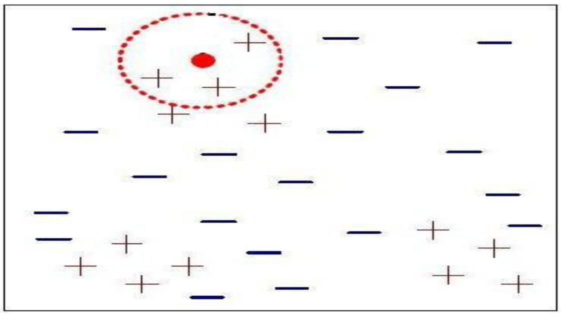


**Fig 1: KNN Classification of Data Instances [19]**

### C. KNN Algorithm

Step 1: Load the train and test dataset.

Step 2: Choose k value as the number of neighbors.

Step 3: For each data sample in test data

- Calculate Distance between the selected sample and its neighbors.
- Store the distances and sort them in ascending order.
- List out the first k entries.
- Assign a class to that new sample based on the majority of the classes present in the neighbor points.

Step 4: Record the accuracy.

## III. RELATED WORK

In 2018 L. Haripriya and M. A.Jabbar proposed a novel IDS using ANN and feature subset selection [5].Authors adapted back propagation algorithm to classify the attacks along with feature subset selection on KYOTO dataset. The experiment result showed an increase in accuracy, precision, recall, and F-measure. Their method recorded an accuracy of 98.66%.

In 2016 Nabila Farnazz and M.A.Jabbar proposed IDS [4] using Random Forest classifier. Feature selection technique is used to reduce data dimensionality. The experimental result was done on dataset NSL-KDD and the model is efficient for higher classification accuracy and DR. Their proposed random forest model recorded an accuracy of 99.67%.

Md Al Mehedi Hasan *et al* [16] proposed IDS using RF and SVM. Authors established two models based on these classifiers. The implementation of these models was compared based on their values of accuracy, false positive rate, f-value and detection rate.

Amreen sultana and M A jabbar proposed intelligent NIDS using data mining techniques. The authors used Average One Dependence Estimator (AODE) classification algorithm for detection type of attacks [17]. The experimental analysis was done on NSL KDD dataset. The outcome shows increase in accuracy with 97% and detection rate with 98%.

In 2017, a novel ensemble IDS was developed [18]. The authors used composite classifier (RFAODE) for IDS. The classifier is combination of RF (Random forest) and Average One-Dependence Estimator (AODE) algorithms. The implementation evaluation is done on KYOTO dataset. The proposed model increased the accuracy to 90.51% and FAR to 0.14.

## IV. PROPOSED WORK

This Section discuss about proposed KNN based model for IDS.

**Algorithm**: Intrusion Detection System using KNN.
**Input**: ISCX dataset.
**Output**: Attack Classification.
Step 1: Load ISCX dataset.
Step 2: Try Cross Validation (5 & 10 fold).
Step 3: Build the model using KNN.
Step 4: Test data is given to KNN for classification.
Step 5: Calculate Accuracy, Precision, Recall and F-measure.

For Evaluation we used ISCX dataset which is in CSV format. The proposed model performed the execution in python and calculated the Euclidean distance based on various K values. Proposed model also applied cross validation technique for classification with K-folds as 5 and 10. Features of ISCX dataset is shown in table 1 below:

Table 1: Specifications of ISCX data set.

| | |
|---|---|
| Number Of Attributes | 78 |
| Number Of Instances | 65536 |
| Missing Values | NO |
| Number Of Classes | 2 |

**CROSS VALIDATION**: Data scientist uses validation as one of the important statistical technique to estimate the stability or skill of the model to see how it will react to new data [10]. Cross validation is a re-sampling method to limit data by evaluating the models [11]. It is popular because it is easy to understand and results are less optimistic or biased [10]. It restricts the problems like over fitting and under fitting.

One round of cross validation includes dividing some part of data into corresponding subsections, performing the analysis on one subsection and validating the analysis on other subsections [12]. If data is splitted into *n* equal sized subparts then out of which one subpart is for testing and remaining *n-1* subparts is used for validating the model. The procedure is repeated for n times and n results are obtained which are averaged to obtain single estimation.

## V. EXPERIMENTAL RESULT

Performance measure of the classifier is calculated based on error matrix through which we can derive values of accuracy, Precision, Recall and F-measure etc.

CONFUSION MATRIX: It is also called as error matrix which describes about the classifiers or models performances on test data. It also allows the algorithms performance conceptualization [13][15]. Basic confusion matrix is shown in Table 2.

**Table 2: Basic Confusion Matrix**

| | | **Prediction** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Actual** | **Positive** | True Positive(TP) | False Negative(FN) |
| | **Negative** | False Positive(FP) | True Negative(TN) |

ACCURACY: Accuracy is evaluated as

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

PRECISION: It is defined from confusion matrix as

Precision=TP/FP+TP

RECALL: Recall is explained as

Recall=TP/TP+FN

F-MEASURE: f-measure is explained as

F=2* recall* precision / recall + precision

KNN algorithm is applied on the dataset. Cross validation for k-folds values as 5 and 10 is applied for classification.

The K value ranges from K=2, 3, 4, 5, 6, 7, 8, 9 and 10. Fig 2 describes about the accuracy for full training set and Fig 3 shows the classification of accuracy by applying 5 & 10 cross validation. The result of our model tabulated in Table [3].

For training the dataset is splits as follows:

i. Training=80% and Testing=20% then accuracy=99.96%.
ii. Training=70% and Testing=30% then accuracy=99.96%.
iii. Training=60% and Testing=40% then accuracy=99.96%.

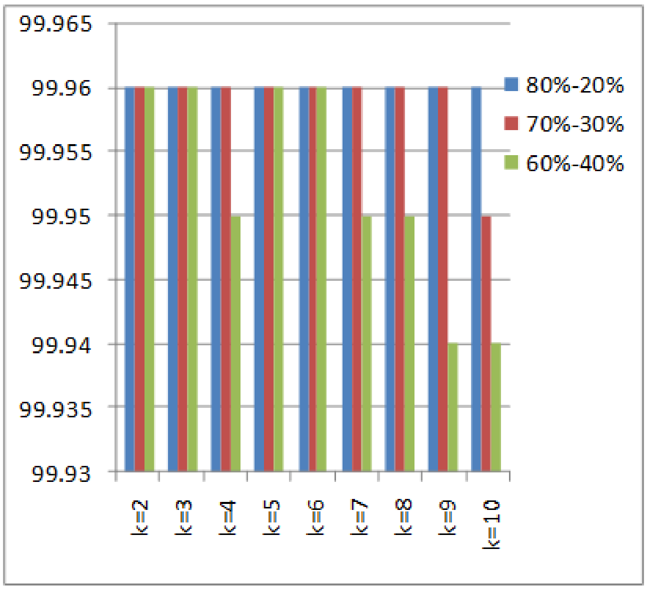


**Fig 2: Classifying accuracy for full training set.**

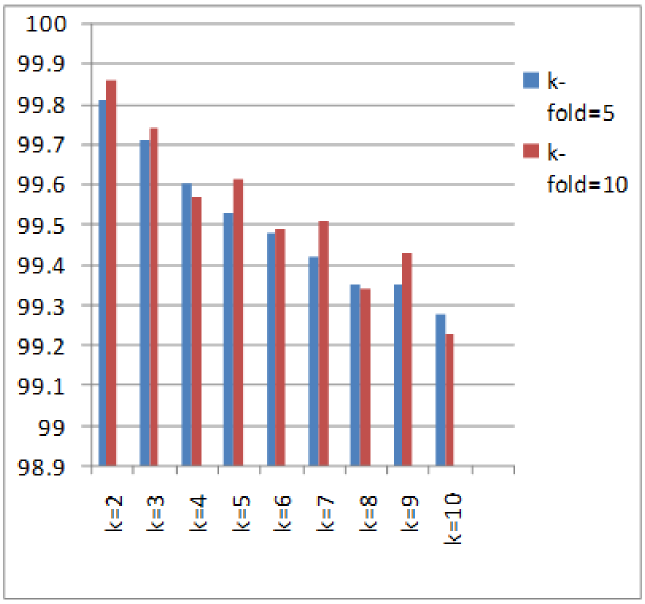


Fig 3: Accuracy classification for 5 & 10 cross validation technique.

Here, we compared our proposed method with SVM and Decision Tree classifiers which obtained an accuracy of 99.66% and 99.94%. The comparisons of our approach with different models are shown in following fig 4 and table 4:

**Table 4: Differentiating proposed model with others**

| S NO | Approach | Accuracy |
|---|---|---|
| 1 | SVM | 99.66% |
| 2 | Decision Tree | 99.94% |
| 3 | KNN | 99.96% |

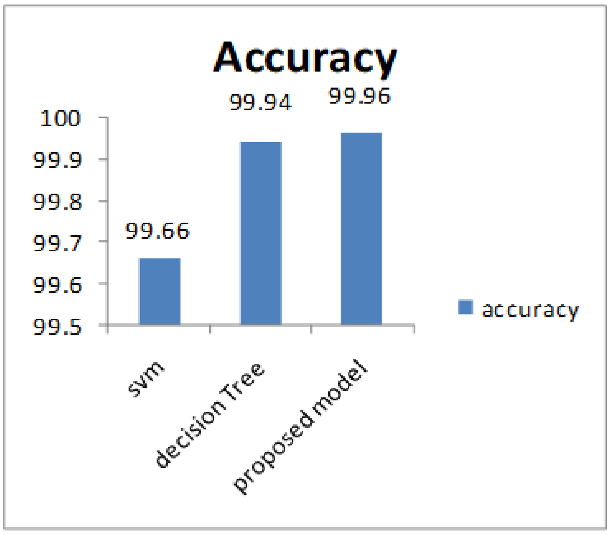


**Fig: 4 Comparison of our approach with different models.**

## VI. CONCLUSION

In this paper, we applied the k nearest neighbor algorithm to IDS data set to classify the type of attacks like Dos, probe, U2R and R2L. Proposed method is validated by 5 and 10 cross validation for the classification. The Experimental analysis shows that when compared to other classification methods, proposed model have increased the accuracy, precision, recall and f-measure values. In future, we plan to apply optimization techniques for the classification of IDS dataset.

## REFERENCES:

1. K.KanakaVardhini et.al, "Enhanced Intrusion Detection System using Data Reduction: An Ant Colony Optimization Approach", In-ternational Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 9 (2017) pp.1844-1847.
2. http://www.scholarpedia.org/article/K-nearest_neighbor
3. https://searchsecurity.techtarget.com/definition/intrusion-detection-system
4. Nabil Farnaaz and M.A.Jabbar, "Random Forest Modeling for Network Intrusion Detection System", ELSEVIER, Science Direct 2016.
5. L.haripriya and M.A.Jabbar," A Novel intrusion detection system using ANN and feature subset selection", international journal of engineering and technology, 2018
6. https://medium.com/datadriveninvestor/knn-algorithm-and-implementation-from-scratch-b9f9b739c28f
7. http://www.scholarpedia.org/article/K-nearest_neighbor
8. https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/
9. Mr MohitTtiwari Raj Kumar, et al, proposed "Intrusion detection system ", International Journal of Technical Research and Applications in april 2017
10. https://towardsdatascience.com/cross-validation-70289113a072
11. https://machinelearningmastery.com/k-fold-cross-validation/
12. https://en.m.wikipedia.org/wiki/Cross-validation_(statistics)
13. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
14. https://machinelearningmastery.com/an-introduction-to-feature-selection/
15. https://classeval.wordpress.com/introduction/basic-evaluation-measures/

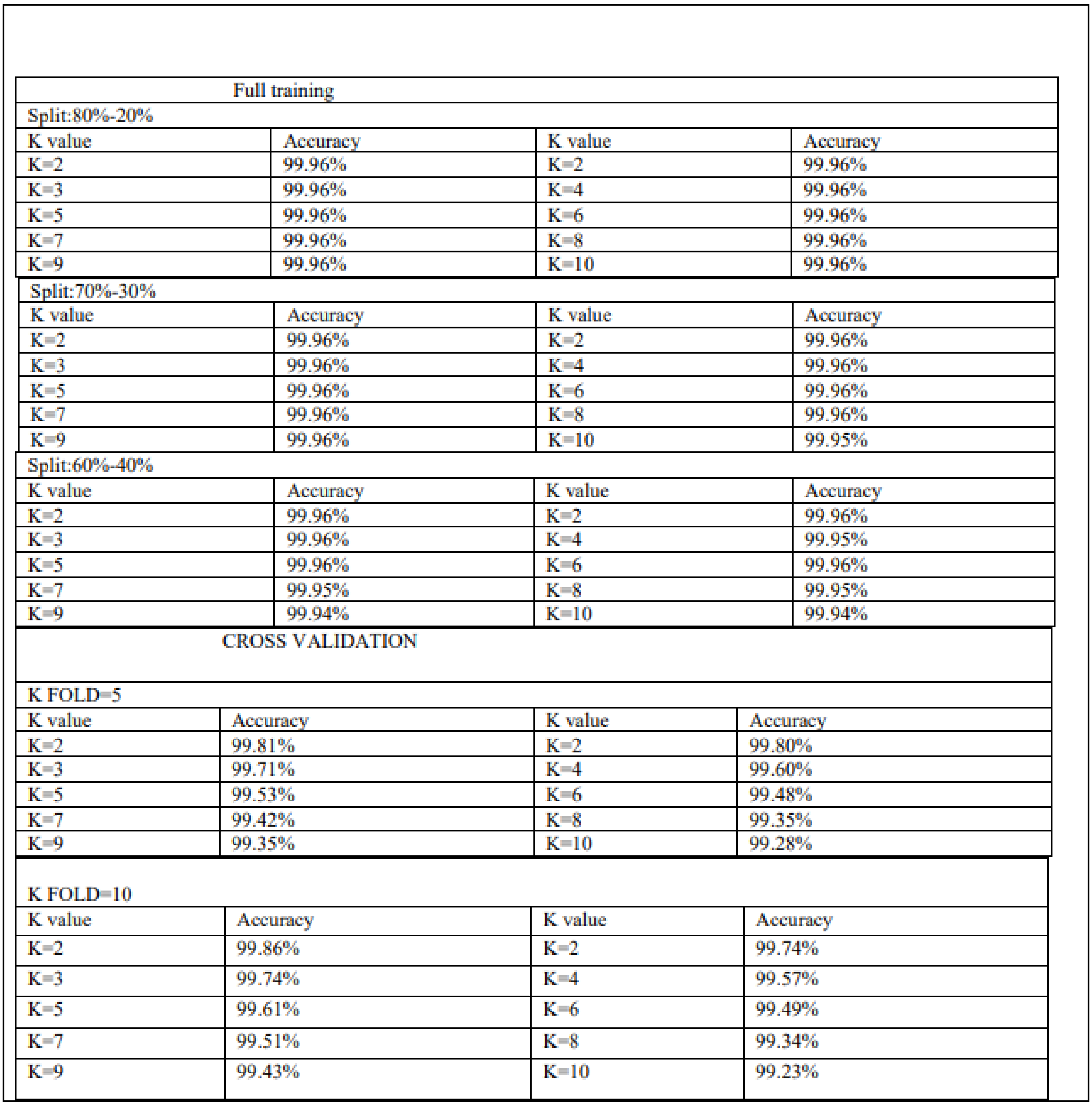
16. Md. Al Mehedi Hasan, Mohammed Nasser, Biprodip and Shamim Ahmad, Support Vector Machine and Random Forest Modeling for IDS,*JILSA*, pp. 45–52, (2014).
17. Amreen Sultana and MA.Jabbar,” intelligent network intrusion detection system using data mining techniques” IEEE explore 2017.
18. MA. Jabbar, Rajanikanth Aluvalu, Sai Satyanarayana Reddy S”, “ RFAODE: A Novel Ensemble Intrusion Detection System”, Elsevier, ICACC-2017, 22- 24 August 2017.
19. M.A.Jabbar, B.A.Deekshatulu, p.chandra, “Heart Disease classification using nearest neighbor classifier using Feature subset selection”, Anale. Seria Informatică. Vol. XI fasc. 1 – 2013.

## AUTHORS PROFILE

**M.Nikhitha** is a research scholar at the Computer Science and Engineering Department, Vardhaman College of Engineering, Hyderabad, Telangana, India .

**Dr. M.A.JABBAR** is a **Vice chair, IEEE CS chapter** ,**Hyderabad Section** and Professor and Centre Head at the Computer Science and Engineering Department, Vardhaman College of Engineering, Hyderabad, Telangana, India. He has been teaching for more than 19 years. He obtained Doctor of Philosophy (Ph.D.) from JNTUH. He published more than 50 papers in various journals and conferences. He is Reviewer for Scopus and SCI journals like Springer, Elsevier, and IEEE Transactions on Systems Man and Cybernetics, Wiley. He served as a technical committee member for more than 40 international conferences. He has been Editor for 1st ICMLSC 2018 international conference held during 22nd and 23rd June 2018 at Hyderabad.

Table 3: Classification Accuracy Using Full Training and Cross validation.



| Full training | | | |
|---|---|---|---|
| Split:80%-20% | | | |
| K value | Accuracy | K value | Accuracy |
| K=2 | 99.96% | K=2 | 99.96% |
| K=3 | 99.96% | K=4 | 99.96% |
| K=5 | 99.96% | K=6 | 99.96% |
| K=7 | 99.96% | K=8 | 99.96% |
| K=9 | 99.96% | K=10 | 99.96% |
| Split:70%-30% | | | |
| K value | Accuracy | K value | Accuracy |
| K=2 | 99.96% | K=2 | 99.96% |
| K=3 | 99.96% | K=4 | 99.96% |
| K=5 | 99.96% | K=6 | 99.96% |
| K=7 | 99.96% | K=8 | 99.96% |
| K=9 | 99.96% | K=10 | 99.95% |
| Split:60%-40% | | | |
| K value | Accuracy | K value | Accuracy |
| K=2 | 99.96% | K=2 | 99.96% |
| K=3 | 99.96% | K=4 | 99.95% |
| K=5 | 99.96% | K=6 | 99.96% |
| K=7 | 99.95% | K=8 | 99.95% |
| K=9 | 99.94% | K=10 | 99.94% |
| CROSS VALIDATION | | | |
| K FOLD=5 | | | |
| K value | Accuracy | K value | Accuracy |
| K=2 | 99.81% | K=2 | 99.80% |
| K=3 | 99.71% | K=4 | 99.60% |
| K=5 | 99.53% | K=6 | 99.48% |
| K=7 | 99.42% | K=8 | 99.35% |
| K=9 | 99.35% | K=10 | 99.28% |
| K FOLD=10 | | | |
| K value | Accuracy | K value | Accuracy |
| K=2 | 99.86% | K=2 | 99.74% |
| K=3 | 99.74% | K=4 | 99.57% |
| K=5 | 99.61% | K=6 | 99.49% |
| K=7 | 99.51% | K=8 | 99.34% |
| K=9 | 99.43% | K=10 | 99.23% |