AN ACCURATE DETECTION OF SPAM SMS USING DECISION TREE CLASSIFIER ALGORITHM COMPARED WITH K-NEAREST NEIGHBOR

N. Srinivasulu[1], R. Sabitha[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The proposed study aims to detect Spam SMS using a Novel Attribute Selection Measure in Decision Tree Classifier Algorithm in comparison with K-Nearest Neighbor.

Materials and Methods: The dataset considered in the current research is available on Kaggle, a machine learning repository. The dataset “SMS spam collection dataset” contains 5572 instances and two attributes v1 and v2. The v2 is the input messages which are either spam or nonspam. The predicted label v1 has two classes: 0 = nonspam and 1 spam. In the data, 4900 are nonspam samples and 672 are spam samples. The sample size was calculated using G Power(95%). The accuracy and sensitivity of the classification of SMS spam detection were evaluated and recorded.

Results: The accuracy was maximum in the classification of SMS spam detection using the Decision Tree Classifier Algorithm (95%) which uses a Novel Attribute Selection Measure with a minimum mean error when compared with K-Nearest Neighbor (93%). There is a significant difference of 0.12 between the classifiers. **Conclusion:** The study proves that the Decision Tree Classifier Algorithm which uses a Novel Attribute Selection Measure exhibits better accuracy than the K-Nearest Neighbor in the Classification of SMS spam detection.

Keywords: Decision Tree Classifier, Junk, SMS, Machine learning, K-Nearest Neighbor, Spam, Mobile, Novel Attribute Selection Measure.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

1. Introduction

SMS is one of the most effective forms of communication. It is based on cellular communication systems, just the phone must have a proper network connection to send or receive the messages. Spam is considered to be one of the serious problems in email and instant message services.(Mashael, Jalal, and Abdelouahid 2015) Nowadays usage of smartphones is increasing, so the number of spam messages is also increasing. Spam messages are defined as unwanted or junk messages. Spam may result; in the leaking of personal information, invasion of privacy, or accessing unauthorized data from mobiles.(Cormack 2008) Hackers try to intrude in mobile computing devices and SMS support for mobile devices has become vulnerable, attack tries to intrude into the system by sending unwanted links, and by clicking on those links the attacker can gain remote access over the computing devices.In this technique, machine learning classifiers such as Logistic regression (LR), K-nearest neighbour (K-NN), and decision tree (DT) are used for the classification of ham and spam messages in mobile device communication. The SMS spam collection data set is used for testing the method.(Salehi 2011)(Mishra and Soni 2021).The method is put to the test with the SMS spam data set. The proposed method is quite useful in detecting Spam SMS and distinguishing between legitimate and garbage SMS. Different machine learning algorithms are used to identify spam and ham transmissions.

Most referred articles similar to this work have been explored (Mashael, Jalal, and Abdelouahid 2015). Around 45 related articles published in IEEE Xplore were published related to this work in google scholar. pam are unsolicited and unwanted messages sent electronically whose content may be malicious. (Wang and Katagishi 2014). Email spam is sent/received over the Internet while SMS spam is typically transmitted over a mobile network. We’ll refer to users that sent spam as spammers. SMS messages are usually very cheap (if not free) for the user to send, making it appealing for unrightful exploitation. This is further aggravated by the fact that SMS is usually regarded by the user as a safer, more trustworthy form of communication than other sources, e. g., emails. (Akinyelu 2021)The dangers of spam messages for the users are many: undesired advertisement, exposure of private information, becoming a victim of a fraud or financial scheme, being lured into malware and phishing websites, involuntary exposure to inappropriate content, etc. For the network operator, spam messages result in an increased cost in operations. (Mishra and Soni 2021; Mashael, Jalal, and Abdelouahid 2015).Our team has extensive knowledge and research experience that has translated into high quality publications(K. Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Yaashikaa, Senthil Kumar, and Karishma 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; H. Mohan et al. 2022)

The research gap identified from the literature survey is that the classification model adopting KNN requires lots of training data. The limitation of this study is that the community has long invested efforts in developing spam SMS susceptibility models. However, no clear standards are still in place with respect to some key parts of the analyses. The research gap identified from the literature survey is that classification models adopting KNN require lots of training data. The existing approaches have poor accuracy. The aim of this study is to implement a Novel Decision Tree and improve the classification accuracy by incorporating Decision Tree and comparing the performance with KNN.

2. Materials and Methods

The research work was performed in the Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The work was carried out on 300 records taken from a Kaggle dataset. The accuracy in predicting SMS spam detection was performed by evaluating two groups. A total of 10 iterations was performed on each group to achieve better accuracy.This work is carried out in the Department of Computer Science and Engineering at Saveetha School of Engineering Chennai. The accuracy in SMS spam detection was performed by evaluating two groups. A total of 10 iterations were performed on each group to achieve better accuracy. It was implemented using jupyter, and the hardware configuration required is an intel i5 processor, 512 GB HDD, 4GB Ram, and the software configuration required is a Windows OS. The work was carried out on 5572 rows × 2 columns records from a data-master dataset. The Study uses a dataset downloaded from Kaggle.

Decision Tree (Dt)

The Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

DT Algorithm

```
Input: SMS spam dataset
Output: Accuracy
Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
Step-2: Find the best attribute in the dataset using the Novel Attribute Selection Measure (ASM).
Step-3: Divide the S into subsets that contain possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3.
Continue this process until a stage is reached where you cannot further classify the nodes and call the final node as a leaf node.
```

K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well-suited category by using K- NN algorithm.

KNN Algorithm

```
Input: SMS spam dataset
Output: Accuracy
Step 1: Load the data.
Step 2: Initialize K to your chosen number of neighbors.
Step 3: For each example in the data
    3.1 Calculate the distance between the query example and the current example from the data.
    3.2 Add the distance and the index of the example to an ordered collection.
Step 4: Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
Step 5: Pick the first K entries from the sorted collection.
Step 6: Get the labels of the selected K entries.
Step 7: If regression, return the mean of the K labels.
Step 8: If classification, return the mode of the K labels.
```

Statistical Analysis

The SPSS statistical software was used in the research for statistical analysis. In this machine learning algorithm, the dependent variable is categorical and measures the relationship between the independent variable and categorical dependent variable using the logistic function. The independent variable is messages. Group statistics and independent sample t-tests were performed on the experimental results and the graph was built for two groups with two parameters under study. The independent variables are useless content, spam information. The dependent variables that affect the output are Accuracy and Precision(Frehner 2008).

3. Results

The proposed algorithm Decision Tree which uses a Novel Attribute Selection Measure and the existing algorithm KNN were run at a time in jupyter using python code. In executing all the commands we get the best significant values. From simulation results, we get an accuracy of 95% (DT) and 93% (KNN) as a result. On comparing both we come to know that the Decision Tree has higher accuracy than KNN. Statistical Analysis of Mean, Standard deviation and Standard Error, and Sensitivity of Decision Tree and KNN is done. There is a statistically significant difference in Accuracy values between the algorithms. The Decision Tree Algorithm had the higher Accuracy and Sensitivity compared with KNN. The Standard error is also less in KNN in comparison to the Decision Tree Algorithm as in Table 2. Comparison of the significance level for Decision and KNN algorithms with value p = 0.12 is done. Both Decision Tree and KNN have a significance level of less than 0.12 with a 95% confidence interval as mentioned in Table 3.

4. Discussion

The work proves that DT is better than KNN in detecting Spam SMS in terms of accuracy and precision. However, the mean error of DT seems to be higher than KNN. Experimental work was done among 2 groups DT and KNN by varying the test size. From the experimental results done in jupyter, the accuracy of DT is 95%, Whereas KNN provides the accuracy to be 93%. This depicts that DT is better than KNN. The various parameters like Precision, Recall, F1-measure are also compared. From the SPSS graph, the proposed DT performs better in terms of accuracy (95%) compared with the KNN algorithm. The research work involved a careful study of the different filtering algorithms and existing anti-spam tools to collect all the information.(Hofmann and Klinkenberg 2016; Wei and Nguyen 2020) These large-scale research papers and existing software programs are one of the sources of inspiration behind this project work. (Agarwal, Kaur, and Garhwal 2015)The whole project was divided into several iterations. Each iteration was completed by completing four phases: inception, where the idea of work was identified; elaboration, where the architecture of the system is designed; construction, where existing code is implemented; transition, where the developed part of the project is validated. Adding meaningful features such as the length of messages in a number of characters, adding certain thresholds for the length, and analyzing the learning curves and misclassified data have been the factors that contributed to this improvement in results.(Akbari and Sajedi 2015; Alzahrani and Rawat 2019). When compared to the decision tree there are a few more best algorithms to get the best result. (Popovac et al. 2018)(Trần 2018) Despite the fact that the presented methodology yielded good results, the limitations are in this approach's weakness is the necessity for enhanced identification of overlapping cells. This may be avoided in the future by combining high-accuracy approaches with a Decision Tree that use a Novel Attribute Selection Measure.

5. Conclusion

In this paper the compiled list of the most current developments in SMS spam filtering, mitigation, and detection approaches, as well as their drawbacks and future research directions. There are several SMS spam strategies, datasets, and comparisons explored. We have also developed a taxonomy of the techniques and identified the established results. The results show that the proposed Decision Tree outperforms KNN in terms of Accuracy. The Proposed Decision Tree which uses a Novel Attribute Selection Measure proved with better accuracy (95.7%) when compared with KNN (93.2%).

Declarations

Conflicts of Interest

No conflicts of interest in this manuscript.

Author Contributions

Author NSV was involved in data collection, analysis, algorithm framing, implementation, and manuscript writing. Author RS was involved in designing the workflow, guidance, and reviewing the manuscript.

Acknowledgments

The authors would like to impress their graduates towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formally known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Mass Datta Developers, Chennai, India.
2. Saveetha University.
3. Saveetha Institute of Medical And Technical Sciences.
4. Saveetha School of Engineering.

6. References

Agarwal, Sakshi, Sanmeet Kaur, and Sunita Garhwal. 2015. “SMS Spam Detection for Indian Messages.” 2015 1st International Conference on Next Generation Computing Technologies (NGCT). https://doi.org/10.1109/ngct.2015.7375198.

Akbari, Fatemeh, and Hedieh Sajedi. 2015. “SMS Spam Detection Using Selected Text Features and Boosting Classifiers.” 2015 7th Conference on Information and Knowledge Technology (IKT). https://doi.org/10.1109/ikt.2015.7288782.

Akinyelu, Andronicus A. 2021. “Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-Based and Nature-Inspired-Based Techniques.” Journal of Computer Security. https://doi.org/10.3233/jcs-210022.

Alzahrani, Amani, and Danda B. Rawat. 2019. “Comparative Study of Machine Learning Algorithms for SMS Spam Detection.” 2019 SoutheastCon. https://doi.org/10.1109/southeastcon42311.2019.9020530.

Cormack, Gordon V. 2008. Email Spam Filtering: A Systematic Review. Now Publishers Inc.

Frehner, Carmen. 2008. Email, SMS, MMS: The Linguistic Creativity of Asynchronous Discourse in the New Media Age. Peter Lang Pub Incorporated.

Hofmann, Markus, and Ralf Klinkenberg. 2016. RapidMiner: Data Mining Use Cases and Business Analytics Applications. CRC Press.

Jayabal, Ravikumar, Sekar Subramani, Damodharan Dillikannan, Yuvarajan Devarajan, Lakshmanan Thangavelu, Mukilarasan Nedunchezhiyan, Gopal

Kaliyaperumal, and Melvin Victor De Poures. 2022. "Multi-Objective Optimization of Performance and Emission Characteristics of a CRDI Diesel Engine Fueled with Sapota Methyl Ester/diesel Blends." Energy. https://doi.org/10.1016/j.energy.2022.123709.

Jayakodi, Santhoshkumar, Rajeshkumar Shanmugam, Bader O. Almutairi, Mikhlid H. Almutairi, Shahid Mahboob, M. R. Kavipriya, Ramesh Gandusekar, Marcello Nicoletti, and Marimuthu Govindarajan. 2022. "Azadirachta Indica-Wrapped Copper Oxide Nanoparticles as a Novel Functional Material in Cardiomyocyte Cells: An Ecotoxicity Assessment on the Embryonic Development of Danio Rerio." Environmental Research 212 (Pt A): 113153.

Kotteeswaran, C., Indrajit Patra, Regonda Nagaraju, D. Sungeetha, Bapayya Naidu Kommula, Yousef Methkal Abd Algani, S. Murugavalli, and B. Kiran Bala. 2022. "Autonomous Detection of Malevolent Nodes Using Secure Heterogeneous Cluster Protocol." Computers and Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2022.107902.

Krishnan, Anbarasu, Duraisami Dhamodharan, Thanigaivel Sundaram, Vickram Sundaram, and Hun-Soo Byun. 2022. "Computational Discovery of Novel Human LMTK3 Inhibitors by High Throughput Virtual Screening Using NCI Database." Korean Journal of Chemical Engineering. https://doi.org/10.1007/s11814-022-1120-5.

Mashael, Al-Omany, Al-Muhtadi Jalal, and Derhab Abdelouahid. 2015. Detection of SMS Spam Botnets in Mobile Devices: Design, Analysis, Implementation. LAP Lambert Academic Publishing.

Mishra, Sandhya, and Devpriya Soni. 2021. "DSmishSMS-A System to Detect Smishing SMS." Neural Computing & Applications, July, 1–18.

Mohan, Harshavardhan, Sethumathavan Vadivel, Se-Won Lee, Jeong-Muk Lim, Nanh Lovanh, Yool-Jin Park, Taeho Shin, Kamala-Kannan Seralathan, and Byung-Taek Oh. 2022. "Improved Visible-Light-Driven Photocatalytic Removal of Bisphenol A Using V2O5/WO3 Decorated over Zeolite: Degradation Mechanism and Toxicity." Environmental Research. https://doi.org/10.1016/j.envres.2022.113136.

Mohan, Kannan, Abirami Ramu Ganesan, P. N. Ezhilarasi, Kiran Kumar Kondamareddy, Durairaj Karthick Rajan, Palanivel Sathishkumar, Jayakumar Rajarajeswaran, and Lorenza Conterno. 2022. "Green and Eco-Friendly Approaches for the Extraction of Chitin and Chitosan: A Review." Carbohydrate Polymers 287 (July): 119349.

Popovac, Milivoje, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. 2018. "Convolutional Neural Network Based SMS Spam Detection." 2018 26th Telecommunications Forum (TELFOR). https://doi.org/10.1109/telfor.2018.8611916.

Salehi, Saber. 2011. A Comparative Evaluation of Machine Learning Approaches in SMS Spam Detection.

Saravanan, A., P. Senthil Kumar, B. Ramesh, and S. Srinivasan. 2022. "Removal of Toxic Heavy Metals Using Genetically Engineered Microbes: Molecular Tools, Risk Assessment and Management Strategies." Chemosphere 298 (July): 134341.

Sathish, T., R. Saravanan, V. Vijayan, and S. Dinesh Kumar. 2022. "Investigations on Influences of MWCNT Composite Membranes in Oil Refineries Waste Water Treatment with Taguchi Route." Chemosphere 298 (July): 134265.

Trần, Hữu Trung. 2018. SMS Spam Detection for Vietnamese Messages: Graduation Thesis for the Honor Degree of Information Technology.

Vivek, J., T. Maridurai, K. Anton Savio Lewise, R. Pandiyarajan, and K. Chandrasekaran. 2022. "Recast Layer Thickness and Residual Stress Analysis for EDD AA8011/h-BN/B4C Composites Using Cryogenically Treated SiC and CFRP Powder-Added Kerosene." Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-022-06636-5.

Wang, Jianyi, and Kazuki Katagishi. 2014. "Image Content-Based 'Email Spam Image' Filtering." Journal of Advances in Computer Networks. https://doi.org/10.7763/jacn.2014.v2.92.

Wei, Feng, and Trang Nguyen. 2020. "A Lightweight Deep Neural Model for SMS Spam Detection." 2020 International Symposium on Networks, Computers and Communications (ISNCC). https://doi.org/10.1109/isncc49221.2020.9297350.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Yaashikaa, P. R., P. Senthil Kumar, and S. Karishma. 2022. "Review on Biopolymers and Composites – Evolving Material as Adsorbents in Removal of Environmental Pollutants." Environmental Research. https://doi.org/10.1016/j.envres.2022.113114.

Tables and Figures

Table 1. Comparison of Test_size and accuracy achieved during the evaluation of Decision Tree and KNN models for classification with different iterations.

Algorithm	**Test_size**	**Accuracy**
KNN	0.20	92.91%
KNN	0.25	92.32%
KNN	0.30	93.01%
KNN	0.35	92.06%
KNN	0.40	92.19%
DT	0.20	96.05%
DT	0.25	96.34%
DT	0.30	95.87%
DT	0.35	95.64%
DT	0.40	95.92%

Table 2. Statistical Analysis of Mean, Standard deviation, and Standard Error of and Sensitivity of Decision Tree and KNN. There is a statistically significant difference in Accuracy and Sensitivity values between the algorithms. Decision Tree had the highest Accuracy (97.3%) and Sensitivity (93.0%) compared with KNN. The Standard error is also less in KNN in comparison to the Decision Tree.

GROUP		N	Mean	Std. Deviation	Std. Error Mean
Accuracy	Decision Tree	5	95.7580	0.46557	.14722
	KNN	5	91.6000	1.07497	.33993

Table 3. Comparison of the significance level for Decision Tree and kNN algorithms with value p = 0.05. Both Decision Tree and KNN have a significance level less than 0.002 in terms of accuracy with a 95% confidence interval.

	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig.	t	df	Sig(2-tailed)	Mean Difference	Std. Error Difference	95% confidence interval of the Difference	
								Lower	Upper

Accuracy	7.761	.012	11.224	18	.000	4.15800	.37045	3.37972	4.93628
			11.224	12.26	.000	4.15800	.37045	3.35277	4.96323

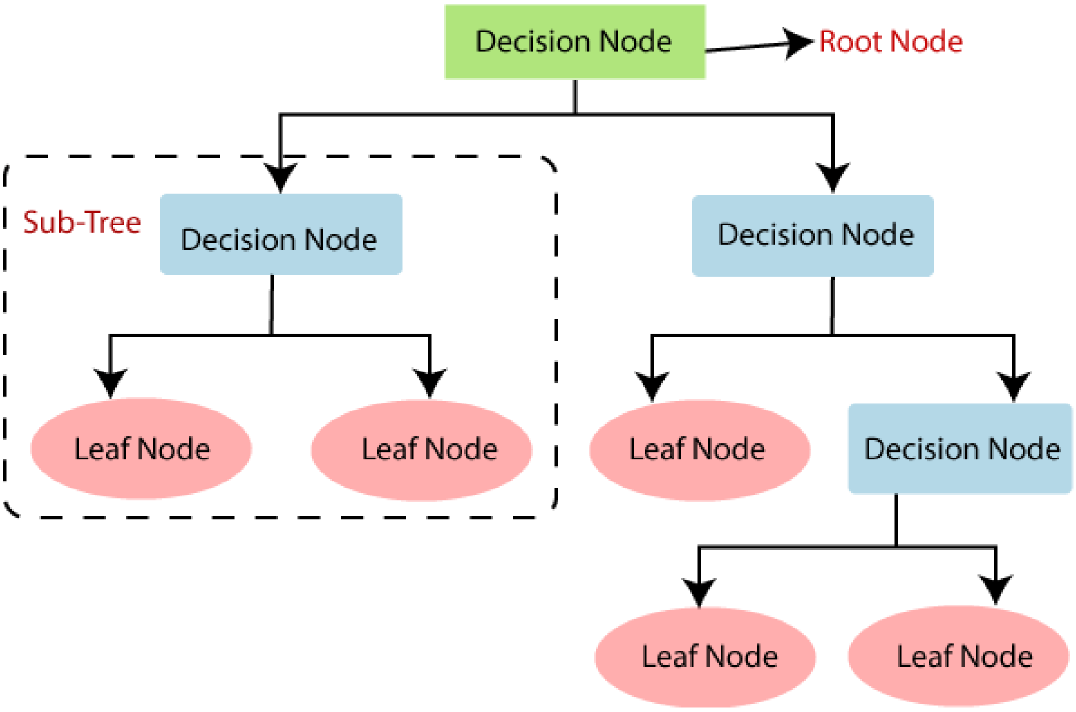

Fig. 1. Flowchart for Decision Tree

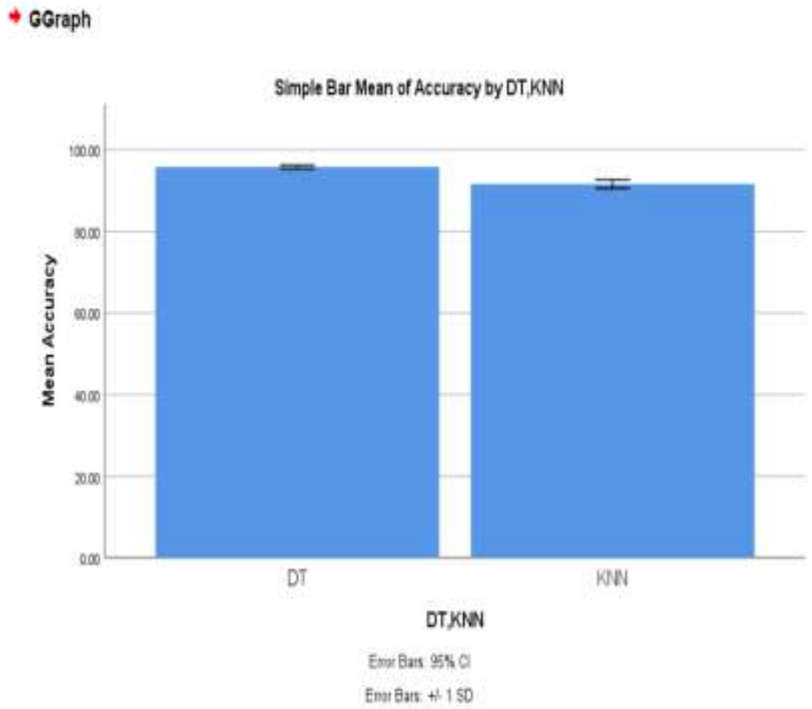

Fig. 2. Comparison of mean accuracy of KNN and Decision Tree algorithms. The standard errors appear to be less in the Decision Tree compared to KNN. Decision Tree appears to produce more consistent results with higher accuracy. X-Axis: KNN vs Decision Tree Algorithm. Y-Axis: Mean accuracy of detection +/- 1 SD.