ANALYSIS OF DECISION TREE CLASSIFIER, NOVEL TREE SPECIFIC RANDOM FOREST CLASSIFIER, SUPPORT VECTOR MACHINE ALGORITHM WITH K-NEAREST NEIGHBOR FOR DETECTING SPAM SMS.

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 the Decision Tree Classifier, Novel Tree Specific Random Forest Classifier, Support Vector Machine 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 non spam 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 Decision Tree Classifier(95%), Novel Tree Specific Random Forest Classifier(97.3%), Support Vector Machine(98%) Algorithm with a minimum mean error when compared with K-Nearest Neighbor (93%). There is a significant difference of 0.010 between the classifiers which infers the groups are significant.
Conclusion: The study proves that the Decision Tree Classifier Algorithm exhibits better accuracy than the K-Nearest Neighbor in Classification of SMS spam detection.

Keywords: Decision Tree Classifier, SMS, Message, Machine learning, K-Nearest Neighbor, Dataset, Spam, Ham, Novel Tree Specific Random Forest Classifier, Support Vector Machine.

[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. 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 (Yadav et al. 2012). Hackers try to intrude in mobile computing devices and SMS support for mobile devices has become vulnerable, attack tries to intrude the system by sending unwanted links, with which by clicking on those links the attacker can gain remote access over the computing devices. Responding to the text message can allow malware to be installed that will silently collect personal information from your phone. If they don't use your information themselves, the spammers may sell it to marketers or other identity thieves. You might end up with unwanted charges on your cell phone bill. In this technique, machine learning classifiers such as Logistic regression (LR), K-nearest neighbour (K-NN), and Decision tree (DT) are used for classification of ham and spam messages in mobile device communication. The SMS spam collection data set is used for testing the method.(Baaqeel and Zagrouba 2020)(Nuruzzaman et al. 2012). With the SMS spam data set, the approach is put to the test. The proposed method is effective in detecting spam SMS and distinguishing valid from garbage SMS. Spam and ham transmissions are identified using a variety of machine learning algorithms.

Most referred articles similar to this work have been explored (Mashael, Jalal, and Abdelouahid 2015). Most referred articles similar to this work have been explored. (Rudnitskaya 2009; Hatton 2011). The purpose is to explore the results of applying machine learning techniques to detect message spam detection. In that, they are going to make a version to classify a message as an unsolicited message or ham. In that model, they trained and tested data using different machine learning algorithms and found out which algorithm works best in the dataset.(Cormack 2008a) In this, we will be using classification algorithms like Logistic Regression, KneighborsClassifier, Novel Tree Specific Random Forest, Decision Tree Classifier, and Support Vector Machines. It achieves an average classification accuracy of 97.20% and outperforms all other feature representations and histograms of oriented gradients using the same classifier on the dataset (Abdulhamid et al. 2017).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 Random Forest and 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(Koujalagi 2019).

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 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.

Novel Tree Specific Random Forest Classifier (RF)

The random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Random forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

Novel Tree Specific Random Forest Classifier works in two-phase first is to create the random forest by combining the N decision tree, and the second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Input: SMS spam dataset
Output: Accuracy

Step 1: Select random K data points from the training set.

Step 2: Build the decision trees associated with the selected data points (Subsets).

Step 3: Choose the number N for decision trees that you want to build.

Step 4: Repeat Step 1 & 2.

Step 5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Support Vector Machine (SVM)

SVM stands for Support Vector Machine. SVM is a supervised machine learning algorithm that is commonly used for classification and regression challenges. Common applications of the SVM algorithm are Intrusion Detection System, Handwriting Recognition, Protein Structure Prediction, Detecting Steganography in digital images, etc. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

The Working process can be explained in the below steps and diagram:

Input: SMS spam dataset
Output: Accuracy

Step 1: Load Pandas library and the dataset using Pandas

Step 2: Define the features and the target

Step 3: Split the dataset into train and test using sklearn before building the SVM algorithmmodel

Step 4: Import the support vector classifier function or SVC function from the Sklearn SVM module. Build the Support Vector Machine model with the help of the SVC function

Step 5: Predict values using the SVM algorithm model

Step 6: Evaluate the Support Vector Machine model.

K-Nearest Neighbor (KNN)

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 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 ((Spregelburd 2016; Castro 2006; Rafat et al. 2022)).

3. Results

The proposed algorithm Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine, and 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), 97.3%(RF), 98%(SVM), and 93%(KNN) as a result. On comparing all we come to know that the SVM has higher accuracy than, Novel Tree Specific Random Forest Classifier, Support Vector Machine, Statistical Analysis of Mean, Standard deviation and Standard Error and Sensitivity of Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine, and K-Nearest Neighbor is done. There is a statistically significant difference in Accuracy values between the algorithms. The Support Vector Machine Algorithm had the higher Accuracy and Sensitivity compared with Decision Tree, Novel Tree Specific Random Forest Classifier, and K-Nearest Neighbor. The Standard error is also less in KNN in comparison to the Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine Algorithm as in Table 2. Comparison of the significance level for Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine, and KNN algorithms with value p = 0.010 is done. Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine, and KNN have a significance level less than 0.10 with a 95% confidence interval as mentioned in Table 3.

4. Discussion

The work proves that SVM, RF, DT is better than KNN in detecting spam SMS in terms of accuracy and precision. (Trần 2018) However, the mean error of SVM, RF, DT seems to be higher than KNN. Experimental work was done among 2 groups SVM, RF, DT, and KNN by varying the test size. From the experimental results done in jupyter, the accuracy of SVM is 98%, RF is 97.3%, DT is 95% Whereas KNN provides the accuracy to be 93%. This depicts that SVM, RF, DT is better than KNN. The various parameters like Precision, Recall, F1-measure are also compared. From the SPSS graph, the proposed SVM, RF, DT perform better in terms of accuracy (98%)(97.3%)(95%) compared with the KNN algorithm. Experiments were conducted among the study groups KNN and Support Vector Machine, RF, DT by varying sample sizes. From the experiments, it is observed that the proposed Support Vector Machine, RF, DT performed better in terms of classification of SMS spam detection by achieving better accuracy and less error rate compared to the KNN algorithm.(Cormack 2008b)(Brunton 2015) This experiment consists of 4 groups KNN, SVM, RF, and DT. To collect all the information, the research involved carefully studying the different filtering algorithms and existing anti-spam tools. (Nuruzzaman et al. 2012)These large-scale research papers and existing software programs are one of the sources of inspiration behind this project work.The whole project was divided into several iterations.(Dhanaraj and Karthikeyani 2013) Each iteration was completed by completing four phases: inception, where the idea of work was identified; elaboration, where the architecture of the part of the system is designed; construction, where existing code is implemented; transition, where the developed part of the project is validated. All these algorithms are compared to each other and concluded the best algorithm is SVM. (Gonsalves et al. 2019)(Hossain et al., n.d.)(Spregelburd 2016; Castro 2006; Rafat et al. 2022) Although the proposed methodology attained satisfactory results, the limitation in the proposed approach is that there needs to be improved detection of overlapping cells. In the future, this can be eliminated by using high-accuracy techniques with Support Vector Machine. Compared to the result of previous work, our classifier increases the overall accuracy. 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.

5. Conclusion

In this paper, the recent advances in SMS spam filtering, mitigation, and detection techniques as well as their limitations and future research direction, Many different SMS spam techniques, used datasets and comparisons are discussed. We have also developed a taxonomy of the techniques and identified the established results. The results show that the proposed Decision Tree, RF, and SVM outperform KNN in terms of Accuracy. The Proposed Decision Tree, RF, SVM proved with better accuracy (95.7%),(97%),(98%) 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 collections, data 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

Abdulhamid, Shafi'i Muhammad, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I. Abubakar, and Tutut Herawan. 2017. "A Review on Mobile SMS Spam Filtering Techniques." IEEE Access. https://doi.org/10.1109/access.2017.2666785.

Baaqeel, Hind, and Rachid Zagrouba. 2020. "Hybrid SMS Spam Filtering System Using Machine Learning Techniques." 2020 21st International Arab Conference on Information Technology (ACIT). https://doi.org/10.1109/acit50332.2020.9300071.

Brunton, Finn. 2015. Spam: A Shadow History of the Internet. MIT Press.

Castro, Francisco. 2006. SPAM. Editorial Galaxia.

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

2008b. Email Spam Filtering: A Systematic Review. Now Publishers Inc.

Dhanaraj, S., and V. Karthikeyani. 2013. "A Study on E-Mail Image Spam Filtering Techniques." 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering. https://doi.org/10.1109/icprime.2013.6496446.

Gonsalves, Lianne, Winnie Wangari Njeri, Megan Schroeder, Jefferson Mwaisaka, and Peter Gichangi. 2019. "Research and Implementation Lessons Learned From a Youth-Targeted Digital Health Randomized Controlled Trial (the ARMADILLO Study)." JMIR mHealth and uHealth 7 (8): e13005.

Hatton, Les. 2011. E-Mail Forensics: Eliminating Spam, Scams and Phishing.

Hossain, Syed Md Minhaz, Khaleque Md Aashiq Kamal, Anik Sen, and Iqbal H. Sarker. n.d. "TF-IDF Feature-Based Spam Filtering of Mobile SMS Using Machine Learning Approach." https://doi.org/10.20944/preprints202109.0251.v1.

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.

Koujalagi, Ashok. 2019. "Mobile SMS Spam Recognition Using Machine Learning Techniques with the Help of Biasian and Spam Filters." International Journal of Computer Sciences and Engineering. https://doi.org/10.26438/ijcse/v7i4.540542.

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.

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.

Nuruzzaman, M. Taufiq, M. Taufiq Nuruzzaman, Changmoo Lee, Mohd Fikri Azli Abdullah, and Deokjai Choi. 2012. "Simple SMS Spam Filtering on Independent Mobile Phone." Security and Communication Networks. https://doi.org/10.1002/sec.577.

Rafat, Khan Farhan, Qin Xin, Abdul Rehman Javed, Zunera Jalil, and Rana Zeeshan Ahmad. 2022. "Evading Obscure Communication from Spam Emails." Mathematical Biosciences and Engineering: MBE 19 (2): 1926–43.

Rudnitskaya, Alena. 2009. The Concept Of Spam In Email Communication. GRIN Verlag.

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.

Spregelburd, Rafael. 2016. Spam.

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.

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.

Yadav, Kuldeep, Swetank K. Saha, Ponnurangam Kumaraguru, and Rohit Kumra. 2012. "Take Control of Your SMSes: Designing an Usable Spam SMS Filtering System." 2012 IEEE 13th International Conference on Mobile Data Management. https://doi.org/10.1109/mdm.2012.54.

Tables and Figures

Table 1. Comparison of Test_size and accuracy achieved during the evaluation of Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine, 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%
SVM	0.20	98.21%
SVM	0.25	97.85%
SVM	0.30	98.09%
SVM	0.35	98.15%

SVM	0.40	98.03%
RF	0.20	97.40%
RF	0.25	96.91%
RF	0.30	97.25%
RF	0.35	97.33%
RF	0.40	97.31%

Table 2. Statistical Analysis of Mean, Standard deviation, and Standard Error of and Sensitivity of Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine and KNN. There is a statistically significant difference in Accuracy and Sensitivity values between the algorithms. Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine had the highest Accuracy (95%),(97.3%),(98%) and Sensitivity (93.0%) compared with KNN. The Standard error is also less in KNN in comparison to the Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine.

ANOVA

ACCURACY

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	210.308	3	70.103	186.884	.010
Within Groups	13.504	36	.375		
Total	223.812	39			

Table 3. Comparison of the significance level for Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine and KNN algorithms with value p = 0.05. Both Decision Tree, Novel Tree Specific Random Forest Classifier, Support Vector Machine and KNN have a significance level less than 0.002 in terms of accuracy with a 95% confidence interval.

Oneway

Descriptives

ACCURACY

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
RF	10	95.7580	.46557	.14722	95.4250	96.0910	94.70	96.34
DT	10	95.7580	.46557	.14722	95.4250	96.0910	94.70	96.34
SVM	10	97.9850	.21793	.06892	97.8291	98.1409	97.60	98.25
KNN	10	91.6390	1.00968	.31929	90.9167	92.3613	89.76	92.91
Total	40	95.2850	2.39557	.37877	94.5189	96.0511	89.76	98.25

Multiple Comparisons

Dependent Variable: ACCURACY

Bonferroni

(I) GROUPS	(J) GROUPS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
RF	DT	.00000	.27390	1.000	-.7647	.7647
	SVM	-2.22700*	.27390	.000	-2.9917	-1.4623
	KNN	4.11900*	.27390	.000	3.3543	4.8837
DT	RF	.00000	.27390	1.000	-.7647	.7647
	SVM	-2.22700*	.27390	.000	-2.9917	-1.4623
	KNN	4.11900*	.27390	.000	3.3543	4.8837
SVM	RF	2.22700*	.27390	.000	1.4623	2.9917
	DT	2.22700*	.27390	.000	1.4623	2.9917
	KNN	6.34600*	.27390	.000	5.5813	7.1107
KNN	RF	-4.11900*	.27390	.000	-4.8837	-3.3543
	DT	-4.11900*	.27390	.000	-4.8837	-3.3543
	SVM	-6.34600*	.27390	.000	-7.1107	-5.5813

*. The mean difference is significant at the 0.05 level.

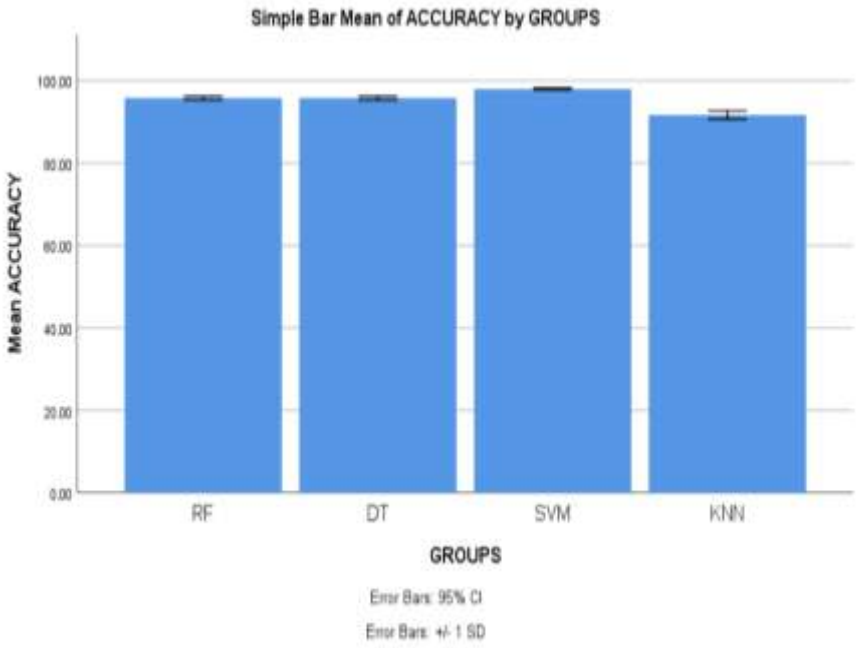

Fig. 1. Comparison of mean accuracy of KNN and Decision Tree, SVM, RF algorithms. The standard errors appear to be less in the Decision Tree, SVM, RF compared to KNN. Decision Tree appears to produce more consistent results with higher accuracy. X-Axis: KNN vs Decision Tree Algorithm. Y-Axis: Mean accuracy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