COMPARISON OF SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBOR IN DETECTING SPAM SMS FOR IMPROVED ACCURACY

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 Kernel-based Technique in Support Vector Machine 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 Support Vector Machine (98%) which uses Novel Kernel-based Technique with a minimum mean error when compared with K-Nearest Neighbor (93%). There is a statistically significant difference of 0.001 between the classifiers.
Conclusion: The study proves that Support Vector Machine which uses a Novel Kernel-based Technique exhibits better accuracy than K-Nearest Neighbor in Classification of SMS spam detection.

Keywords: Novel Kernel-based Technique, Support Vector Machine, Ham, Spam, Message, K-Nearest Neighbor, SMS, Spam, Machine Learning.

[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

Mobile messages could be a way for people to communicate, as billions of people use mobile devices to send and receive messages. ((Sridevi Gadde; A. Lakshmanarao; S. Satyanarayana "SMS Spam Detection Using Machine Learning and Deep Learning Techniques" n.d.2021)However, due to the lack of proper message filtering methods, such a wide range of communication is insecure. Spam is one of the reasons for this weakness, and it makes mobile message communication insecure. As the number of people using cell phones grows, so does the number of spam texts. However, the majority of the received messages will be spam, with only a few of them being ham or necessary messages.(Hackeling 2017) SMS Spamming is a huge annoyance for mobile users since they receive a lot of trash messages instead of legal ones. The messages are classified as electronic, and spontaneous, and the business is in danger for the most part because of the following factors, particularly the availability of low-cost bulk SMS, unwavering quality, overall execution, and the likelihood of receiving a response from the recipient. In this technique, machine learning classifiers such as Logistic regression (LR), K-nearest neighbor (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. The method is put to the test with the SMS spam data set. The proposed approach is quite beneficial in identifying Spam SMS and distinguishing between legitimate and garbage SMS. Different machine learning methods are used to identify spam and ham transmissions.(Mishra and Soni 2021)(Salehi 2011)(Hackeling 2017)

Most referred articles similar to this work have been explored. (Cruz et al. 2017). 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 2008; Kigerl 2018) In this, classification algorithms like Logistic Regression, K neighbors Classifier, Novel Tree Specific Random Forest, Decision Tree Classifier, and Support Vector Machine will be used. 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)(Koujalagi 2019).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 classification models adopting KNN require lots of training data. The existing approaches have poor accuracy. The aim of this study is to implement a Support Vector Machine that uses a Novel Kernel-based technique and improve the classification accuracy by incorporating a Support Vector Machine 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 Study 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 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. The Study uses a dataset downloaded from Kaggle.

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. The proposed method uses a Novel Kernel-based Technique. Common applications of the SVM algorithm are Intrusion Detection systems, 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 algorithm model
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 which uses a Novel Kernel-based Technique.
Step 5: Predict values using the SVM algorithm model
Step 6: Evaluate the Support Vector Machine model.

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 good suite 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 and spam information. The dependent variables that affect the output are Accuracy and Precision (Baaqeel and Zagrouba 2020).

3. Results

The proposed algorithm Support Vector Machine that implements a Novel Kernel-based technique and 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 98%(SVM) and 93%(KNN) as a result. On comparing both it is known that the Support Vector Machine which uses a Novel Kernel-based Technique has higher accuracy than KNN. Statistical Analysis of Mean, Standard deviation Standard Error, and Sensitivity of Support Vector Machine and KNN is done. There is a statistically significant difference in Accuracy values between the algorithms. Support Vector Machine Algorithm had the higher Accuracy and Sensitivity compared with KNN. The Standard error is also less in KNN in comparison to the Support Vector Machine Algorithm as in Table 2. Comparison of the significance level for Support Vector Machine and KNN algorithms with value p = 0.001 is done. Both Support Vector Machine and KNN have a significance level of less than 0.001 with a 95% confidence interval as mentioned in Table 3.

4. Discussion

The work proves that SVM is better than KNN in detecting spam SMS in terms of accuracy and precision. (Trần 2018) However, the mean error of SVM seems to be higher than KNN. Experimental work was done among 2 groups SVM and KNN by varying the test size. From the experimental results done in jupyter, the accuracy of SVM is 98%,

Whereas KNN provides the accuracy to be 93%. This depicts that SVM is better than KNN. The various parameters like Precision, Recall, and F1-measure are also compared. From the SPSS graph, the proposed SVM performs better in terms of accuracy (98%) compared with the KNN algorithm. Experiments were conducted among the study groups KNN and Support Vector Machine with varying sample sizes. The experiments show that the proposed Support Vector Machine performed better in terms of classification of SMS spam detection by achieving better accuracy and less error rate than the KNN algorithm.(Cormack 2008) In this experiment, the research work involved a careful study of the different filtering algorithms and existing anti-spam tools. These large-scale research papers and existing software programs are one of the sources of inspiration behind this project work.(Yadav et al. 2012) 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 system is designed; construction, where existing code is implemented; transition, where the developed part of the project is validated. (Gonsalves et al. 2019)(Hossain et al., n.d.)(Nuruzzaman et al. 2012)(Baaqeel and Zagrouba 2020). Compared to the Support Vector Machine there are a few more algorithms where we have better accuracy than SVM. (Sridevi Gadde; A. Lakshmanarao; S. Satyanarayana "SMS Spam Detection Using Machine Learning and Deep Learning Techniques" n.d.2021)). Despite the fact that the presented methodology yielded promising results, the limitation of this approach is the necessity for enhanced identification of overlapping cells. This may be avoided in the future by combining high-accuracy approaches with a Support Vector Machine.

5. Conclusion

In this paper a 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 Support Vector Machine which implements a Novel Kernel-based Technique outperforms KNN in terms of Accuracy. The Proposed Support Vector Machine proved with better accuracy (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 collection, 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.

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

Cruz, Dela, C. Jennifer, Valiente, Leonardo C. Castor, Celine Margaret T. Mendoza, B. Arvin Jay, L. Song Cherry Jane, and P. Torres Bailey Brian. 2017. "Determination of Blood Components (WBCs, RBCs, and Platelets) Count in Microscopic Images Using Image Processing and Analysis." 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and

Management (HNICEM). https://doi.org/10.1109/hnicem.2017.8269515.

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.

Hackeling, Gavin. 2017. Mastering Machine Learning with Scikit-Learn. Packt Publishing Ltd.

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.

Kigerl, Alex C. 2018. "Email Spam Origins: Does the CAN SPAM Act Shift Spam beyond United States Jurisdiction?" Trends in Organized Crime. https://doi.org/10.1007/s12117-016-9289-9.

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.

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.

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.

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.
"SMS Spam Detection Using Machine Learning and Deep Learning Techniques." n.d. Accessed February 7, 2022. https://ieeexplore.ieee.org/abstract/document/9441783.
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 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%
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%

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

GROUP	N	Mean	Std. Deviation	Std.Error Mean

Accuracy	Support Vector Machine	5	97.9850	0.21793	.06892
	KNN	5	91.6000	1.07497	.33993

Table 3. Comparison of the significance level for Support Vector Machine and kNN algorithms with value p = 0.001. Both Support Vector Machine and KNN have a significance level less than 0.001 in terms of accuracy with a 95% confidence interval.

	Levene's Test for Equality of Variances		T-test for Equality of means						
								95% confidence interval of the Difference	
	F	Sig.	t	df	Sig(2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Accuracy	16.314	.001	18.409	18	.000	6.38500	.34685	5.65629	7.11371
			18.409	9.739	.000	6.38500	.34685	5.60935	7.16065

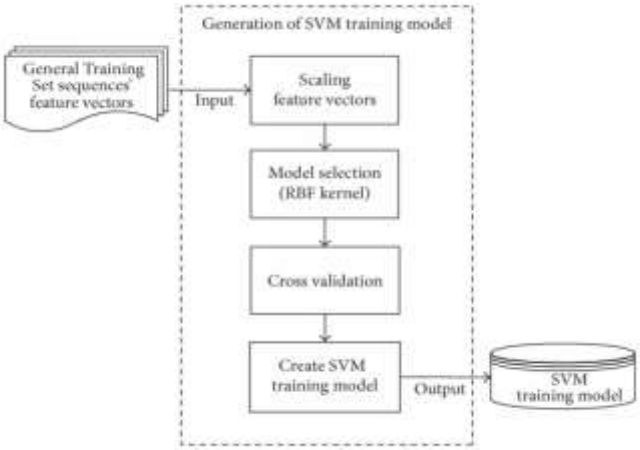

Fig. 1. Flowchart for Support Vector Machine (SVM) Algorithm

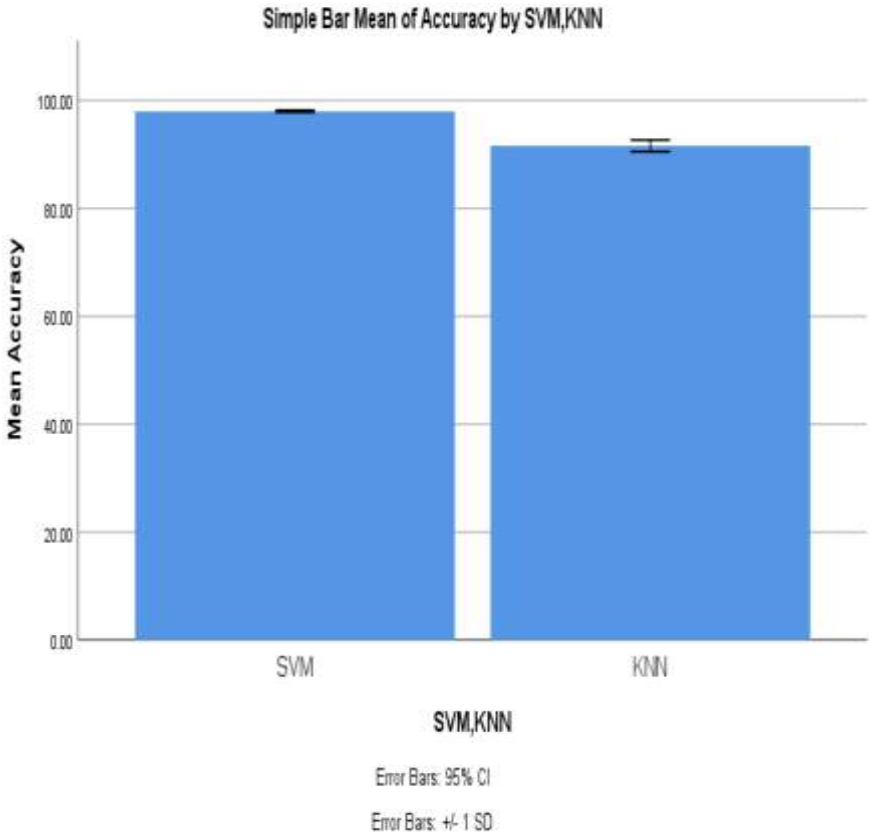

Fig. 2. Comparison of mean accuracy of KNN and Support Vector Machine algorithms. The standard errors appear to be less in Support Vector Machine compared to KNN. Support Vector Machine appears to produce more consistent results with higher accuracy. X-Axis: KNN vs Support Vector Machine Algorithm. Y-Axis: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