AN ENHANCED JUNK EMAIL SPAM DETECTION USING MACHINE LEARNING BY SUPPORT VECTOR MACHINES OVER RANDOM FOREST.

C. Gnanendhra Reddy[1], S. Magesh Kumar[2*]

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

Abstract

Aim: The main aim of the research is to enhance Junk Email Spam Detection using Machine Learning by Novel Support vector machines over Random Forest.
Materials and Methods: Novel Support vector machine and Random Forest are implemented in this research work. Sample size is calculated using G power software and determined as 10 per group with pretest power 2, threshold 50 and Confidence Intervals 95%.
Results: Novel Support vector machine provides a higher of 93.52 % compared to the Random Forest algorithm with 91.41 % in email spam detection. There is a significant difference between two groups with significance value of p=0.019 (p<0.05).
Conclusion: Novel Support vector machine algorithm detects spam emails better than Random forest algorithm.

Keywords: Spam Filtering, Spam, Novel Support Vector Machine, Random Forest, Email Spam Detection, Machine Learning.

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

1. Introduction

The work is about Junk email spam detection using machine learning by Novel Support vector machines over Random Forest.Today, internet users are increasing Spam mail is the major problem and big challenges for researchers to reduce it (Mishra and Thakur 2013). Email users are increasing at a high rate and a huge number of people's privacy is getting risked by spam email and it also kills valuable time of people most often. Spam email can be malicious as well as it can be of commercial use as in for marketing which are not desirable to us. Hence, detecting and filtering spam emails from several emails is a must (Toma, Hassan, and Arifuzzaman 2021). Spam filtering emails are non-requested emails that sent unwanted messages to mass addresses by collecting email addresses from different sources such as chat rooms, websites, phone books and filled forms where user left their email ids. These spam mails are used for spreading viruses, advertising new products, and for sending phishing mails to make money (Vinitha and Renuka 2019). Even today the quantity of spam filtering SMS is quite lower than spam emails, but still there is enough quantity to create a miss-leading usage. In 2010-2012, it is reported that about 90% of emails are spam worldwide while this number is very low in terms of SMS, (Ali and Maqsood 2018). In Asia about 30% of total Messages were actually spam. Applications for email spam detection are extremely important for any organization. Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose.

Spam filtering is a process to detect unsolicited massage and prevent from entering into the user's inbox. Most of the anti-spam filtering methods have some inconsistency between false negatives and false positives which act as a barrier for most of the system to make successful anti spam filtering system. Therefore, an intelligent and effective spam-filtering system is the prime demand for web users (Vyas, Prajapati, and Gadhwal 2015).The task of spam filtering is to rule out deceptive messages automatically from a user's inboxes. These deceptive mails have already caused many problems such as filling mailboxes, overwhelming important personal mail, wasting on network, consuming users' time and energy to sort through it, not to mention all the other problems associated with spam filtering(More and Kulkarni 2013).Phishing emails are emails that pretend to be from a trusted company that target users to provide personal or financial information. Sometimes, they include links that may download malicious software on user's computers, when clicked (Kaddoura, Alfandi, and Dahmani 2020).Participants classified spam emails according to pairings of three stimulus features,presence or absence of awkward prose, abnormal message structure, and implausible premise. We examined dimensional interactions within general recognition theory . Classification accuracy was highest for categories containing either two non-normal dimension levels (Williams et al. 2019).Our team has extensive knowledge and research experience that has translated into high quality publications(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022)

The research gap identified from the existing system Random forest algorithm shows poor accuracy. The study is to improve the accuracy of Classification by incorporating Novel Support vector machine and comparing performance with Random forest. The proposed model improves classifiers to achieve more accuracy for email spam detection.

2. Materials and Methods

This study setting was done in the Soft Computing Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The number of required samples in research are two in which group 1 is a Novel Support vector machine compared with group 2 of Random forest algorithm (Conway and White 2011). The samples were taken from the device and iterated 10 times to get desired accuracy with G power 80%, threshold 0.05% and Confidence Intervals 95%. A dataset consisting of a collection of spam emails was downloaded from kaggle.

Novel Support vector machine

The Novel Support vector machine was utilized for classifying and differentiating input data types. This Novel Support vector machine is widely used in Machine Learning to make predictions(Baig 2021). Novel Support vector machine is often used in email spam filtering . It has a big effect on spam detection. So, the program detects spam mails.

Pseudocode for Novel Support vector machine

Step1: Import packages.
Step2: Create an input dataset.
Step3: Analyze the size of the taken input data.
Step4: Split the datasets for testing and training the dataset.
Step5: Apply Support Vector Machine algorithm.
Step6: Predict the results.

Random Forest Algorithm

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems(Mishra and Thakur 2013). It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It performs better results for classification problems

Pseudocode for Random Forest Algorithm

Step1: Import packages.
Step2: Create an input dataset.
Step3: Analyze the size of the taken input data.
Step4: Split the datasets for testing and training the dataset.
Step5: Apply the RF algorithm.
Step 6: Predict the results.

Recall that the testing setup includes both hardware and software configuration choices. The laptop has an Intel Core i5 5th generation CPU with 12GB of RAM, an x86-based processor, a 64-bit operating system, and a hard drive. Currently, the software runs on Windows 10 and is programmed in Python. Once the program is finished, the accuracy value will appear. Procedure: Wi-Fi laptop connected. Chrome to Google Collaboratory search Write the code in Python. Run the code. To save the file, upload it to the disc, and create a folder for it. Log in using the ID from the message. Run the code to output the accuracy and graph.

Statistical Analysis

SPSS is a software tool used for statistics analysis. The proposed system utilized 10 iterations for each group with predicted accuracy noted and analyzed. Independent samples t-test was done to obtain significance between two groups (Zhang, Zhu, and Yao 2004). Independent samples t-test was done to obtain significance between two groups. Dependent variable is no.of white list words and independent variable is no.of black list words.

3. Results

Table 1 shows the accuracy value of iteration of Novel Support vector machine and Random Forest. Table 2 represents the Group statistics results which depict a Novel Support vector machine with mean accuracy of 93.52%, and standard deviation is 1.77. Random Forest has a mean accuracy of 91.41% and standard deviation is 1.83. The Proposed Novel Support vector machine algorithm provides better performance compared to the Random Forest algorithm. Table 3 shows the independent samples T-test value for Novel Support vector machine and Random Forest with Mean difference as 8.1, std Error Difference as 0.80. The results achieved with $p=0.019$ ($p<0.05$) shows that two groups are statistically insignificant.

Figure 1 shows the bar graph comparison of mean of accuracy on Novel Support vector machine and Random Forest algorithm. Mean accuracy of Novel Support vector machine is 93.52% and Random Forest is 91.41%.

4. Discussion

In this study,Junk email spam detection using the Novel Support vector machine algorithm has significantly higher accuracy, approximately 93.52% in comparison to Random Forest 91.41%. Novel Support vector machine appears to produce more consistent results with minimal standard deviation.

These spam emails may cause serious threat to the user i.e,the email addresses used for any online registrations may be collected by the malignant third parties and they expose the genuine user to various kinds of attacks. Another method of spamming is by creating a temporary email register and receiving emails that can be terminated after some certain amount of time (Ali and Maqsood 2018). Spam email is very annoying for email account users to get relevant information. Detection of email spam has actually been applied to email services for the public with various methods. The server administrator must add a separate or modular spam detection feature so that e-mail accounts can be protected from spam email (Santoso 2019). Phishing is one of the major challenges faced by the world of e-commerce today. Thanks to phishing attacks, billions of dollars have been lost by many companies and individuals. In 2012, an online report put the loss due to the phishing attack at about $1.5 billion. This global impact of phishing attacks will continue to be on the increase and thus requires more efficient phishing detection techniques to curb the menace(Akinyelu and Adewumi 2014). Sending and receiving emails have continued to take the lead being the easiest and fastest way of e-communication despite the presence of other forms of e-communication such as social networking (Abdullahi et al. 2021). The rise in online transactions through email has globally contributed to the increasing rate of spam emails, which has been a major problem in the field of computing.

The limitation of this research is that it is not possible to consider all given feature variable parameters for training. If you have no occurrences of a class label and a certain attribute value together then the frequency-based probability estimation will be zero. A big data set is required for making reliable predictions of the probability of each class. Future scope of proposed work will be

junk email spam detection using class labels for lesser time complexity.

5. Conclusion

In this study Junk email spam detection using the Novel Support vector machine algorithm has significantly higher accuracy, approximately (93.52%) in comparison to Random Forest (91.41%). Novel Support vector machine appears to produce more consistent results with minimal standard deviation.

Declaration

Conflict of Interests

No conflict of interests in this manuscript

Authors Contribution

Author CGR was involved in data collection, data analysis, manuscript writing. Author SMK was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly 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. VK Technologies, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering

6. References

Abdullahi, Muhammad, Abdulmalik D. Mohammed, Sulaimon A. Bashir, and Opeyemi O. Abisoye. 2021. "A Review on Machine Learning Techniques for Image Based Spam Emails Detection." In *2020 IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)*. IEEE. https://doi.org/10.1109/cybernigeria51635.2021.9428826.

Akinyelu, Andronicus A., and Aderemi O. Adewumi. 2014. "Classification of Phishing Email Using Random Forest Machine Learning Technique." *Journal of Applied Mathematics* 2014: 1–6.

Ali, Syed Sarmad, and Junaid Maqsood. 2018. ".Net Library for SMS Spam Detection Using Machine Learning: A Cross Platform Solution." In *2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE. https://doi.org/10.1109/ibcast.2018.8312266.

Baig, Azhar. 2021. "Email Spam Detection Using SVM." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2021.36383.

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. "Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review." *Chemosphere*. https://doi.org/10.1016/j.chemosphere.2022.134596.

Conway, Drew, and John White. 2011. *Machine Learning for Email: Spam Filtering and Priority Inbox*. "O'Reilly Media, Inc."

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. "Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review." *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2022.127234.

Kaddoura, Sanaa, Omar Alfandi, and Nadia Dahmani. 2020. "A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach." In *2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. IEEE. https://doi.org/10.1109/wetice49692.2020.00045.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. "A Novel Face Recognition Model for Fighting against Human Trafficking in Surveillance Videos and Rescuing Victims." *Soft Computing*. https://doi.org/10.1007/s00500-022-06931-1.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. "Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation." *Journal of Thermal Analysis and Calorimetry*. https://doi.org/10.1007/s10973-022-11298-4.

Mishra, Rachana, and R. S. Thakur. 2013. "Analysis of Random Forest and Naïve Bayes for Spam Mail Using Feature Selection Catagorization." *International Journal of*

Computer Applications. https://doi.org/10.5120/13844-1670.

More, Sujeet, and S. A. Kulkarni. 2013. "Data Mining with Machine Learning Applied for Email Deception." In *2013 International Conference on Optical Imaging Sensor and Security (ICOSS)*. IEEE. https://doi.org/10.1109/icoiss.2013.6678403.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. "Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm." *Computational Intelligence and Neuroscience* 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha. 2022. "Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands." *Silicon*. https://doi.org/10.1007/s12633-022-01825-1.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. "A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102155.

Santoso, Budi. 2019. "An Analysis of Spam Email Detection Performance Assessment Using Machine Learning." *Jurnal Online Informatika*. https://doi.org/10.15575/join.v4i1.298.

Toma, Tasnia, Samia Hassan, and Mohammad Arifuzzaman. 2021. "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection." In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*. IEEE. https://doi.org/10.1109/acmi53878.2021.9528108.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan Subramani, and Jayashri N. Nair. 2022. "Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends." *Energy*. https://doi.org/10.1016/j.energy.2022.123806.

Vinitha, V. Sri, and D. Karthika Renuka. 2019. "Performance Analysis of E-Mail Spam Classification Using Different Machine Learning Techniques." In *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE. https://doi.org/10.1109/icacce46606.2019.9080000.

Vyas, Tarjani, Payal Prajapati, and Somil Gadhwal. 2015. "A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering." In *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE. https://doi.org/10.1109/icecct.2015.7226077.

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. "Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture." *Chemosphere* 301 (August): 134638.

Williams, Sarah E., Dawn M. Sarno, Joanna E. Lewis, Mindy K. Shoss, Mark B. Neider, and Corey J. Bohil. 2019. "The Psychological Interaction of Spam Email Features." *Ergonomics* 62 (8): 983–94.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review." *Chemosphere* 299 (July): 134390.

Zhang, Le, Jingbo Zhu, and Tianshun Yao. 2004. "An Evaluation of Statistical Spam Filtering Techniques." *ACM Transactions on Asian Language Information Processing* 3 (4): 243–69.

Tables and Figures

Table 1. This table contains Accuracy Values for Support Vector Machine(SVM) and Random Forest Algorithm (RFA)

S.NO	SVM	RFA
1	96.80	94.61

2	94.59	93.00
3	93.30	92.40
4	92.66	91.70
5	91.10	90.60
6	92.20	88.50
7	95.00	90.40
8	94.32	92.40
9	91.30	89.10
10	94.00	91.40

Table 2. Group Statistics Results-SVM has a mean (93.52%), std.deviation (1.77), whereas for RF has mean (91.41%), std.deviation (1.83).

Group Statistics					
	Groups	**N**	**Mean**	**Std deviation**	**Std. Error Mean**
Accuracy	SVM	10	93.5270	1.7725	0.5605
	RF	10	91.4110	1.8346	0.5801

Table 3. The significance value p=0.019 (p<0.05) shows that two groups are statistically significant.

		Independent Samples Test								
		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	**Equal variances assumed**	0.000	0.019	2.623	18.000	0.017	2.1160	0.8067	0.4211	3.8108
	Equal variances not assumed			2.623	17.979	0.017	2.1160	0.8067	0.4211	3.8110

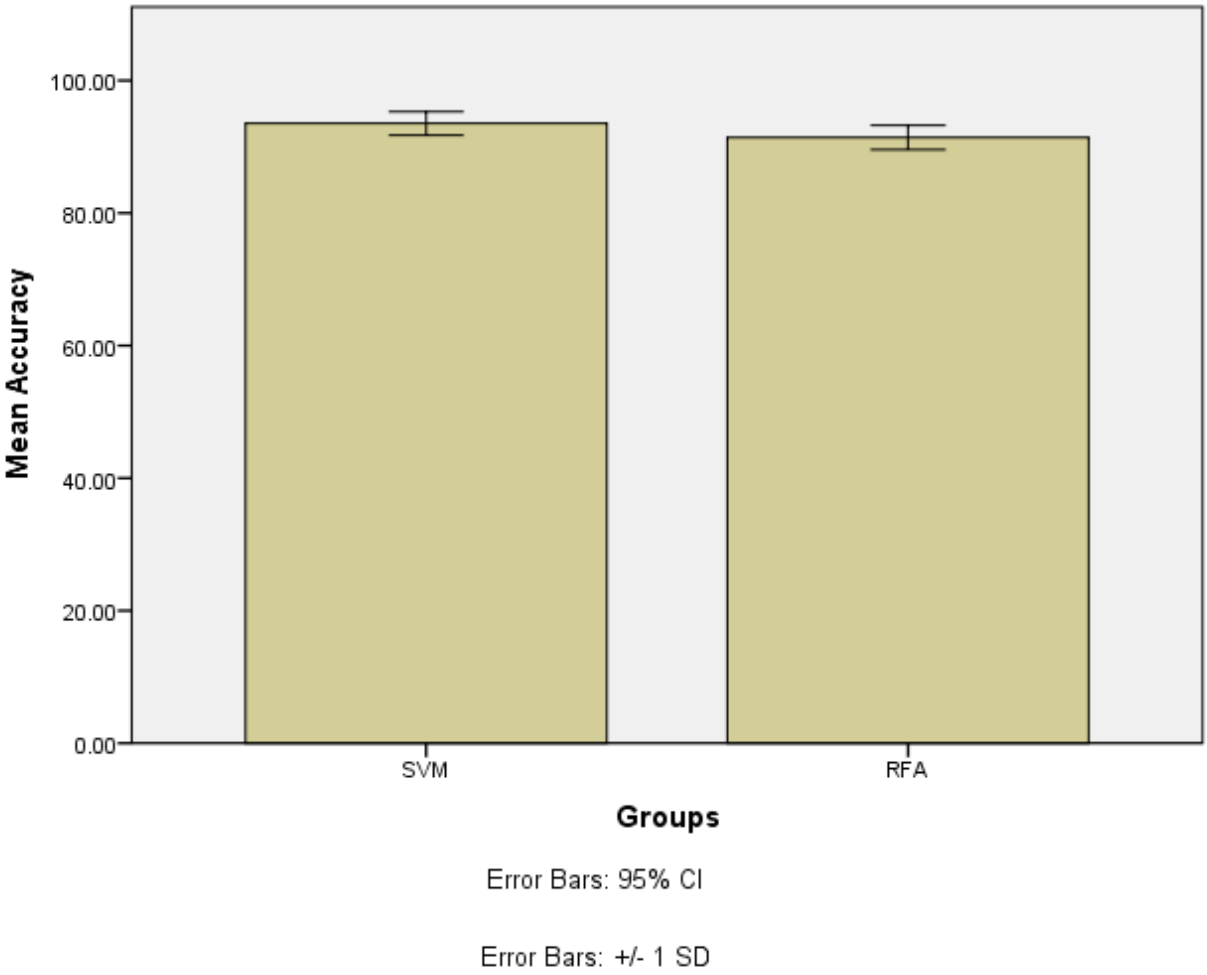

Fig. 1. Bar Graph Comparison on mean accuracy of Support vector machine (93.52%) and Random forest algorithm(91.4%). X-axis is having SVM, RFA, Y-axis is having Mean Accuracy with ± 1 SD.