Comparison of Novel Recurrent Neural Network Over Artificial Neural network in Predicting Email spammers with improved accuracy

Chillakuru.Neeharika[1,*] , S.Kalaiarasi[1]

[1]*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

ABSTRACT: The main aim is to compare Novel Recurrent Neural Network over Artificial Neural Network in predicting Email spammers with improved accuracy. Material and Methods : This research study contains two groups namely Novel Recurrent Neural Network and Artificial Neural Network. Each group consists of a sample size of 10 and the study parameters are calculated using clincalc with preset parameters as alpha 0.8, beta 0.2 and CI as 90%. Results and Discussion : The Novel Recurrent Neural Network has the highest accuracy 97.96% when compared to Artificial Neural Network it has 93.79% accuracy in Electronic Mail spam prediction with significance value p=0.000($p<0.05$) that is significantly better. The G-power value is 80%. When used as a spam predictor for electronic mail, the Novel Recurrent Neural Network performance analysis outperforms the best results than the Artificial Neural Network performance.

Keywords: Artificial Neural Network, Electronic Mail, Machine Learning, Novel Recurrent Neural Network, Spam, Unsupervised approach, Vulnerability.

INTRODUCTION

E-mail is a message that may contain text, files, images, or other attachments and is sent over a network to a specific individual or group of individuals (Lai et al. 2008). Despite the availability of numerous chat apps such as WhatsApp, Facebook Messenger, and Snapchat,Email has remained an important part of daily digital life (Parasuraman 1982). The number of global e-mail users is expected to increase to 4.48 billion in 2024, up from 3.8 billion in 2018 (Yadav and Srivastava 2020). In terms of the most popular email clients, Apple and Google are constantly fighting for first place. Email is the cheapest source for communication and it creates vulnerabilities as well. A business Electronic mail system is

[*]Corresponding Authour:Chillakurussvcsv@gmail.com

one of the types that is highly vulnerable if it is not used. Electronic mail spam is a part of spam security that can be demonstrated by the spam vulnerability. To defend your network from the various potential threats, such as malware, phishing attacks,compromised web links, and other dangerous information, it is critical to eliminate as much spam as possible (Rost, Sander, and Schneider 1994). Many of us who have websites might have received an email stating that your website has vulnerabilities. A great way to limit the no.of vulnerabilities is to actively follow the best cyber security and also we can use SMTP authentication for protecting the Email box. Spam filters also safeguard your servers from being swamped with non-essential emails, as well as from being infected with spam software, which may transform them into spam servers themselves (Szyman and Barbucha 2022). One of the biggest issues with today's internet is email spam. Rule-based techniques, white and blacklists, collaborative spam filtering, challenge-response systems, and other ways have been suggested to automatically categorize communications as spam or valid (Aydogan and Karci 2018).Vulnerability is the risk of malicious emails that are designed to bypass spam filters and deliver malware or phishing attacks to the user's system. In this innovation the work used two unsupervised approaches and compared two different neural networks to find which is giving the best accuracy for email spam prediction.
Several reference papers have utilized the same algorithm to improve the prediction of Email spammers. Over the past five years, a significant number of research articles have been published on email spam prediction using Machine learning, with 231 articles available on Google Scholar, 20 journal papers on IEEE Xplore, and 155 articles on ScienceDirect. Various algorithms have been proposed for predicting email spam, including the Novel Recurrent Neural Network and Artificial Neural Network, which is an unsupervised approach technique (AndoohginShahri et al. 2018). It is worth noting that the Novel Recurrent Neural Network is also an unsupervised approach for email spam prediction (Hulten, Goodman, and Rounthwaite 2004). It has a variety of functions like storing the information when the input is being read step by step and it generalizes services of the model to input arrangements. The other algorithm is Artificial Neural network, this is also an unsupervised approach.The artificial neural network has the greatest detection rate of whether a file is spam or ham (Montgomery 2011).

The disadvantage of the current method is that as the quantity of the dataset grows, the performance of supervised machine learning algorithms declines (Bredenberg 2010). So it wants to improve the accuracy of the existing proposed system using machine learning. The main moto of this project is to improve the accuracy of Email spam prediction using Novel Recurrent Neural Network i.e; unsupervised approach in comparison to Artificial neural network ie; unsupervised approach to improve the accuracy (Takahashi, Sakai, and Sakurai 2010).

MATERIALS AND METHODS

The Saveetha School of Engineering at Saveetha Institute of Medical and Technical Sciences is where the study is being done. In order to complete the study, two study groups were considered for this study, where Novel Recurrent Neural Network and Artificial Neural Network. The sample size were calculated as 10 for each group using GPower, by preseting the parameter values of power as 80%, and threshold of 0.001%, and a confidence interval of 95%. An inventory stocks dataset was obtained from the Kaggle repository.
The free source websites Kaggle were used to get the data. The Novel Recurrent Neural Network and Artificial neural network were used to estimate software effort on the dataset (spam.cvv), which comprises 4602 columns and 58 rows. The Visual Evaluation Tool is

used to assess this research project for display (Volkov and State University of Telecommunications 2022).

Novel Recurrent Neural Network

An example of a novel recurrent neural network is one in which the output from the previous step is used as the input for the current step. Traditional neural networks' inputs and outputs are independent, but in order to predict the subsequent word in a phrase, the preceding words must be remembered. This led to the development of a Novel Recurrent Neural Network that, with the aid of a Hidden Layer, resolved the issue. The Hidden state, which retains some information about a sequence, is the Novel Recurrent Neural Network's primary and most significant characteristics (Schrödel 2011).

Novel Recurrent Neural Network Algorithm

Step 1:Data preprocessing: Use word embeddings, padding, and tokenization to transform the email messages into numerical sequences. Create training and test sets from the data.
Step 2:RNN architecture definition: A lengthy Short-Term Memory (LSTM) network, a sort of RNN that can process lengthy sequences of data and recall information from earlier inputs, is a typical design for email spam prediction. The input, output, and forget gates of the LSTM network regulate the information flow across the network.
Step 3:RNN model training: To minimize the loss function, backpropagation and gradient descent are used to train the model on the training set.
Step 4:Review the model: Calculate the accuracy for the trained model on testing the testing set to assess how well it performed.

Artificial Neural Network

Artificial Neural Network is a computational model that mimics how nerve cells function in the human brain. Artificial neural networks employ learning algorithms that can independently make adjustments or learn as new input is received. As a result, they are an extremely effective tool for non-linear statistical data modeling. An artificial neural network is made up of three or more interconnected layers. The first layer is made up of input neurons (Isik et al. 2020).

Those neurons send data to the deeper layers, which send the final output data to the final output layer. The inner layers are all hidden and formed by units that change the information received from layer to layer adaptively through a series of transformations (Goldberg 2017). Each layer serves as both an input and output layer,allowing the Artificial Neural network to comprehend more complex objects. These inner layers are referred to collectively as the neural layer (Takesue 2010). Backpropagation sends information backward each time the output is labeled as an error during the supervised training phase. Each weight is adjusted in proportion to how much it contributed to the error (Karim, Salleh, and Khan 2016).

Algorithm for Artificial Neural Network

Input: Training and Testing data

Output: Accuracy Score

Step 1 : Import the libraries and datasets of eye images.
Step 2 : Load the dataset and add the dense layer with activation function.
Step 3 : Split and compile the model's dataset into training and testing.
Step 4 : Train and evaluate the model's performance and feature of the dataset.
Step 5 : Make predictions on the performance of the model accuracy score using the training graph with matplotlib and improve the model.
Step 6 : The result of the improved model will be analyzed for accuracy.

Statistical Analysis

To find the effectiveness of these machine learning algorithms we have conducted a statistical analysis of the proposed model using the SPSS (IBM,2021) and the Google colab is used to implement the algorithms. The independent variables are Word frequency, Word frequency parts, Word frequency conference and Frequency meeting. The Accuracy, Precession are the dependent variables. The mean, median, standard deviation,and standard mean errors were determined using an independent sample T-test (Rajalingam 2020).

RESULTS

All of the variables are pertinent to the statistics group's findings. It is clear from comparing the improved Novel Recurrent Neural Network's accuracy and standard deviation results to those of the Artificial Neural Network Independent Sample T-test. The relevance of equality of variance, which states that the results in the study work are meaningful and associated with one another, is what causes the accuracy difference between the Novel Recurrent Neural Network and Artificial Neural Network.
Table 1 shows the accuracy of Novel Recurrent Neural Network accuracy is 97.96%, whereas Artificial Neural Network accuracy is 93.79%. Novel Recurrent Neural Network generates more than Artificial Neural Network due to its effective categorization feature based on airline fare. According to Table 2, the standard deviation for artificial neural networks is 1.96833 and for novel recurrent neural networks is 1.54071. The independent samples T-test results for the novel recurrent neural network and the artificial neural network show a mean difference of 4.17000 and a standard deviation error difference of.79045. The significance value is $p=0.000$ ($p<0.05$) that is significantly better. The comparison of accuracy of the Novel Recurrent Neural Networks with Artificial Neural Networks is shown in figure 1.

Table 1. Novel Recurrent Neural Network and Artificial Neural Network Iteration Values are as follows

Iterations	Novel Novel Recurrent Neural Network	Artificial Neural network
1	96.20	92.10
2	96.20	92.50
3	96.20	92.10
4	96.10	92.10
5	99.30	92.60
6	98.90	93.30

7	99.10	93.90
8	99.10	95.40
9	99.20	96.50
10	99.30	97.40

Table 2. Analysis of Novel Recurrent Neural Network and Artificial Neural Network methods using group statistics. Novel Recurrent Neural Network and Artificial Neural Network have respective means of 97.9600 and 93.7900, as well as standard deviations of 1.54071 and 1.96833, and standard errors of.48721 and.62244.

	Group Name	**N**	**Mean**	**Standard Deviation**	**Standard Error Mean**
Accuracy	Novel Recurrent Neural Network	10	97.9600	1.54071	0.48721
	Artificial Neural Network	10	93.7900	1.96833	0.62244

Table 3. For Novel Recurrent Neural Network and Artificial Neural Network algorithm, Independent Samples of T test reveals statistical significance.Statistical significance is 0.000 ($p<0.05$).

Independent Sample 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 Differences**	**95% Confidence Interval of the Difference**	
								Lower	**Upper**

Accuracy	**Equal Variances assumed**	0.292	0.596	5.275	18	0.000	4.17000	0.79045	2.50933	5.83067
	Equal Variances not assumed			5.275	17.018	0.000	4.17000	0.79045	2.50933	5.83757

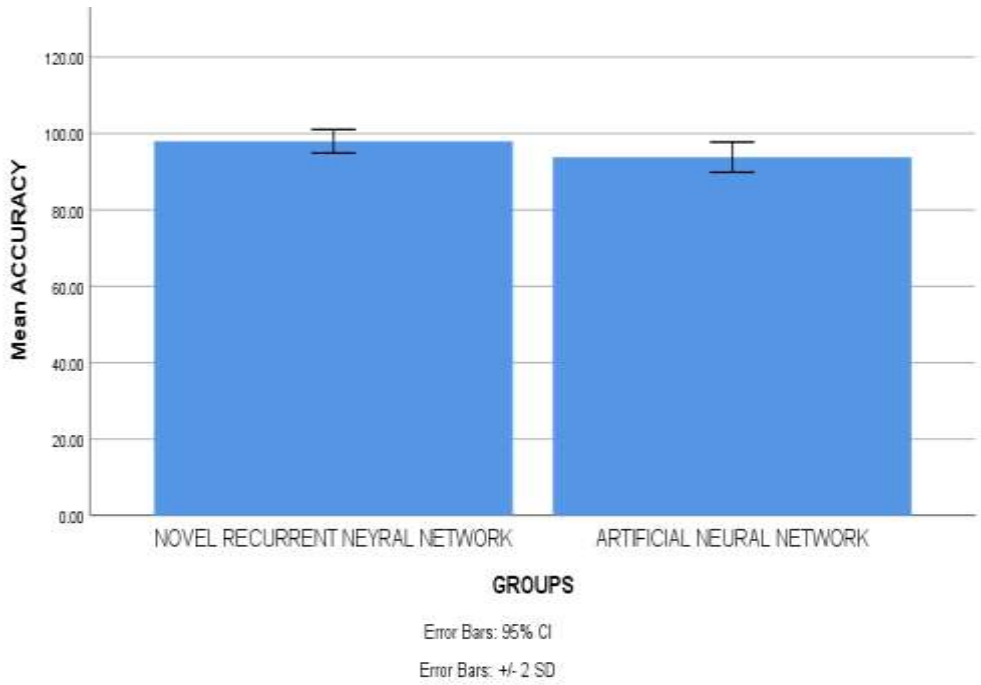

Fig. 1. Groups such as Novel Recurrent Neural Network and Artificial Neural Network are represented on the X-axis.The mean accuracy of the novel recurrent neural network was found to be better than artificial neural network as shown on the Y-axis - Mean Efficiency of detection is ±2 SD.

DISCUSSION

When it came to predicting email spam in this study, the Novel Recurrent Neural Network greatly outperformed the Artificial Neural Network (97% vs 94% accuracy). Results from

novel recurrent neural networks often exhibit greater consistency and smaller standard deviation (Baktir and Atay 2022) .
The corresponding results from the research showed that a novel recurrent neural network for email spam prediction has an accuracy of 97%. The stated Artificial Neural Network's suggested work has 94% accuracy and is used to forecast email spam (Yang et al. 2012). The Novel Recurrent Neural Network and Artificial Neural network are the parameters used to predict Email spam (Yu 2020). My research indicates that Novel Recurrent Neural Networks have the best accuracy 97% and Artificial Neural Networks have the lowest accuracy 94% when compared to other machine learning techniques (Lagrana 2016). The dataset's value tends to increase as the required precision is attained (Egan 2004). Using novel recurrent neural networks with additional deep learning and machine learning methods improves performance.

An Artificial Neural network internal workings might be difficult to analyze and comprehend, making it impossible to determine how it arrived at a certain outcome or decision. Artificial Neural networks are prone to overfitting to training data, resulting in poor performance on new data. The planned next work will predict Email spam using supervised machine learning approaches. Our future innovation will concentrate on improving accuracy for predicting Email spam (Dong and Zhou 2018).

CONCLUSION

In this study, email spam prediction using the Novel Recurrent Neural Network has the accuracy of 97.96% and the Artificial Neural Network has an accuracy of 93.79%. By comparing both the algorithms the Novel Recurrent Neural Network has more accuracy. The discussion of the research paper also proves that the Novel Recurrent Neural Network provides better accuracy than Artificial Neural Network, when finding the email spam.

DECLARATIONS

Conflict of Interests

The authors of this paper declare no conflict of interest.

Authors Contribution

CNH contributed to data collecting, data analysis and manuscript writing. Author SKA was involved in conceptualization, guidance and critical review of manuscript.

Acknowledgements

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. FinuraBioteks,Chennai.
2. Saveetha School of Engineering.
3. Saveetha University.
4. Saveetha Institute of Medical and Technical Sciences.

REFERENCES

1. AndoohginShahri, Mona, Mohammad DavarpanahJazi, Glenn Borchardt, and Mehdi Dadkhah. 2018. "Detecting Hijacked Journals by Using Classification Algorithms." *Science and Engineering Ethics* 24 (2): 655–68.
2. Aydogan, Murat, and Ali Karci. 2018. "Spam Mail Detection Using Naive Bayes Method with Apache Spark." *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. https://doi.org/10.1109/idap.2018.8620737.
3. Baktir, Nuriye, and Yılmaz Atay. 2022. "Comparative Analysis of Machine Learning Approaches in the Spam-Mail Classification Problem." *Bilişim Teknolojileri Dergisi*. https://doi.org/10.17671/gazibtd.1014764.
4. Bredenberg, Al. 2010. "E-Mail: Spam." *Encyclopedia of Information Assurance*. https://doi.org/10.1081/e-eia-120046363.
5. Dong, Cailing, and Bin Zhou. 2018. "Spam Detection: E-mail/Social Network." *Encyclopedia of Social Network Analysis and Mining*. https://doi.org/10.1007/978-1-4939-7131-2_294.
6. Egan, Michael. 2004. *Email Etiquette*. New Line Publishing.
7. Goldberg, Yoav. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.
8. Hulten, G., J. Goodman, and R. Rounthwaite. 2004. "Filtering Spam E-Mail on a Global Scale." *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters - WWW Alt. '04*. https://doi.org/10.1145/1013367.1013478.
9. Isik, Sahin, Zuhal Kurt, Yildiray Anagun, and Kemal Ozkan. 2020. "Spam E-Mail Classification Recurrent Neural Networks for Spam E-Mail Classification on an Agglutinative Language." *International Journal of Intelligent Systems and Applications in Engineering*. https://doi.org/10.18201/ijisae.2020466316.
10. Karim, Ahmad, Rosli Salleh, and Muhammad Khurram Khan. 2016. "SMARTbot: A Behavioral Analysis Framework Augmented with Machine Learning to Identify Mobile Botnet Applications." *PloS One* 11 (3): e0150077.
11. Lagrana, Fernando. 2016. *E-Mail and Behavioral Changes: Uses and Misuses of Electronic Communications*. John Wiley & Sons.
12. Lai, Wei-Ru, Che-Hui Liao, Chang-Ching Lu, and Ming-Kuan Liao. 2008. "A Prediction Mechanism of Mail Retrieval Based on the User Behavior Analysis for Electronic Mail Database System." *2008 Eighth International Conference on Intelligent Systems Design and Applications*. https://doi.org/10.1109/isda.2008.348.
13. Montgomery, Mike. 2011. "Spam Score in the Mail Mascot." *SciVee*. https://doi.org/10.4016/28543.01.
14. Parasuraman, A. 1982. "More on the Prediction of Mail Survey Response Rates." *Journal of Marketing Research*. https://doi.org/10.2307/3151626.
15. Rajalingam, Mallikka. 2020. *Text Segmentation and Recognition for Enhanced Image Spam Detection: An Integrated Approach*. Springer Nature.
16. Rost, Burkhard, Chris Sander, and Reinhard Schneider. 1994. "PHD-an Automatic Mail Server for Protein Secondary Structure Prediction." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/10.1.53.
17. Schrödel, Tobias. 2011. "E-Mail Und SPAM." *Hacking Für Manager*. https://doi.org/10.1007/978-3-8349-6475-5_7.
18. Szyman, Paweł, and Dariusz Barbucha. 2022. "Link Prediction in Organizational Social Network Based on E-Mail Communication." *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2022.09.463.

19. Takahashi, Kenichi, Akihiro Sakai, and Kouichi Sakurai. 2010. “Spam Mail Blocking in Mailing Lists.” *Multimedia*. https://doi.org/10.5772/7628.
20. Takesue, Masaru. 2010. “Cascaded Simple Filters for Accurate and Lightweight Email-Spam Detection.” *2010 Fourth International Conference on Emerging Security Information, Systems and Technologies*. https://doi.org/10.1109/securware.2010.34.
21. Volkov, M. V., and State University of Telecommunications. 2022. “Application of Bayesian E-Mail Filtering to Detect Spam.” *Modern Information Security*. https://doi.org/10.31673/2409-7292.2022.034854.
22. Yadav, Nikhil, and Kajal Srivastava. 2020. “Student Performance Prediction from E-Mail Assessments Using Tiny Neural Networks.” *2020 IEEE Integrated STEM Education Conference (ISEC)*. https://doi.org/10.1109/isec49744.2020.9397817.
23. Yang, Zhen, L. A. I. Ying-Xu, Li-Juan Duan, and L. I. Yu-Jian. 2012. “Spam Collaborative Filtering in Enron E-Mail Network.” *Acta Automatica Sinica*. https://doi.org/10.3724/sp.j.1004.2012.00399.
24. Yu, Szde. 2020. “Crime Hidden in Email Spam.” *Encyclopedia of Criminal Activities and the Deep Web*. https://doi.org/10.4018/978-1-5225-9715-5.ch057.