



# Detecting Spam Emails using Natural Language Processing And Machine Learning

M. Anusha Rani<sup>1\*</sup>, Gudimetla Naga Jyothi<sup>2</sup>, Kesari Trisha<sup>3</sup>, Muppavarapu Srividya<sup>4</sup>,  
Raghavarapu Blessy<sup>5</sup>

<sup>1\*</sup> Assistant Professor: Vignan's Nirula Institute of Technology and Science for Women.

<sup>2,3,4,5</sup> B.Tech Scholar: Vignan's Nirula Institute of Technology and Science for Women.

## Abstract

From the abstract of the document, the detection of spam emails is an important issue. Spam emails are considered unwanted and deceitful messages sent to people or a company. This makes it possible to use natural language processing, combined with learning algorithms, to enable spam detection. Most of the machine classifiers commonly utilized for the classification of identification of an email, such as valid or unwanted spam, are naïve Bayes, support vector machines, and more. This work deals mainly with the distinctive features in the documents content about spam filtering. Amongst these papers, this abstract tells me of the varied methods of machine learning which include use of Neural Networks, Naïve Bayes, Support Vector Classifier, and Logistic Regression that have assisted in the spam detection process. It has contributed to this particular domain of spam filtering by applying these techniques suitably.

## 1. Introduction

The introduction section of this document [1-2] is on how more technological apparatus and the internet have been used, increasing email communication. More so, email communications are important methods for exchanging information, though they present the downside of spam mails-one that is unwanted and potential malicious.

Spam mails waste inbox space and would fill them and slow down internet speed. It can also pose security risks by carrying viruses [3-8] or attempting to scam individuals.

An introduction to the problem of filtering spam messages: most of such communications are useless to users, and their detection calls for efficient spam detection techniques [9].

It mentions that the Naive Bayesian method and the feature sets are widely used for the purpose of spam keyword identification. The proposed approach is actually an alternative one which uses [10-11] Neural Network classification systems to achieve an adequate level of accuracy in spam detection. The introduction begins with outlining the importance of spam detection in terms of increases in internet usage [12-13] and potential spam email risks.

Email spam has become a significant problem in today's digital age [14-15], posing challenges for individuals, businesses, and organizations alike. Spam emails are unsolicited messages that flood inboxes, wasting valuable time [16-17] and resources while potentially exposing users to malicious content or scams. To combat this issue, machine learning techniques have emerged as powerful tools for email spam detection [18-19].

## 2. Literature Survey

The literature survey section of the paper gives a synopsis of the existing research and methodologies on machine learning technology [20-21] applied to detect email spam. It focuses on the importance of an emailing system as a form of communication and the prevalence of spam mails in today's digitalized world. It lays much importance on the effective spam detection [22-23] technique to prevent the damaging effect of spam emails users.

Mukund Deshpande and Arun N. Venkataram (2002) Deshpande and Venkataram proposed a Bayesian approach[24-25] to classify emails as spam or non-spam based on word frequencies. Their method considered the probability of each

word in the email and calculated the overall likelihood of the email being spam [26-27]. This approach established Bayesian methods as effective for spam detection, paving the way for future research. By analyzing word frequencies and email context, their technique improved spam detection accuracy [28-29].

Huan Liu (2003) Liu explored various machine learning algorithms, including Support Vector Machines (SVM), decision trees [30], and random forests, for spam detection. His research demonstrated machine learning's potential for improving spam detection. Liu's work focused on extracting relevant features from email headers and body content, which enabled machine learning algorithms [31-32] to distinguish between spam and legitimate emails. This research laid the groundwork for applying machine learning to spam detection.

Mehran Sahami (1998) Sahami introduced a Bayesian framework for spam detection, considering word frequencies and email context. His approach used Bayesian inference to calculate the probability [33-34] of an email being spam. Sahami's pioneering work demonstrated the effectiveness of Bayesian methods in spam detection. By analyzing word frequencies and email context, his technique improved spam detection accuracy [35-36] and set the stage for future research.

Harris Drucker (1999) Drucker demonstrated the effectiveness of Support Vector Machines (SVM) in spam detection [37-38], achieving high accuracy. His research focused on keyword extraction, where relevant words were selected to train the SVM model. Drucker's work showcased SVM's potential for spam detection, highlighting its ability to handle high-dimensional data. This research contributed [39-40] significantly to the development of efficient spam detection methods.

Yuchun Chen (2015) Chen applied deep neural networks to spam detection, improving accuracy. His research utilized [41] email headers and body content to train deep learning models. Chen's work introduced deep learning techniques to spam detection, enabling more effective feature extraction and classification. This research demonstrated the potential of deep learning in improving spam detection [42] capabilities.

Jian Zhang (2017) Zhang showcased gradient boosting's potential in spam detection. His research combined word frequencies and keyword extraction to train gradient boosting models. Zhang's work demonstrated gradient [43-44] boosting's effectiveness in handling complex data and improving spam detection accuracy. This research highlighted the importance of ensemble learning methods in spam detection.

Christopher D. Manning (2008) Manning applied Natural Language Processing (NLP) techniques, specifically tokenization and stemming, to improve spam detection. His research emphasized the importance [45] of text preprocessing in spam detection. Manning's work demonstrated how NLP techniques can enhance spam detection accuracy by reducing dimensionality and improving feature extraction.

Gordon V. Cormack (2007) Cormack organized the TREC Spam Track, providing a standardized benchmark dataset for spam detection research. The TREC Spam Track enabled researchers to compare [46-47] and evaluate their methods, facilitating advancements in spam detection. Cormack's work ensured that research efforts were focused on improving spam detection capabilities, leading to significant progress in the field.

### 3. Proposed Methodology

The methodology is utilized for the technique of e-mail spam filtering based on Naive Bayes algorithm.

Data Preprocessing is a technique which is applied to transform the raw information into a clean data set. That is, whenever the information is gathered from different sources it's collected in raw format which isn't feasible for the analysis. This comprises of the following successive steps:

**Tokenization:** It is assumed that tokenization is the process of splitting the huge amount of text into small pieces called Tokens. These tokens are pretty useful to determine patterns, and they are separated by white-spaces characters like line break, space or by punctuation characters.

**Dropping Values:** Dropping is the most common way to handle missing values. Such rows in the data set or the full columns with missing values are dropped so as not to allow errors to happen in data analysis.

**Stop Words:** Stop words are those words that an English language uses in which doesn't carry much content in a sentence. They will easily be kept ignored without giving up the meaning of the sentence.

Bag-of-Words: This is a text representation that describes the occurrence of words within a document and bag-of-words method is utilized in extracting features from the documents.

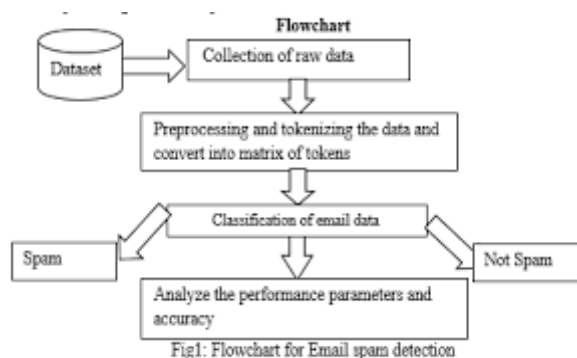


Fig.1 Flowchart for Email Spam detection

This Algorithm involves the following steps:

Step 1: Consider a random email from the spam data-set for execution.

Step 2: The e-mail up for consideration is in basic form. In order to carry out the feature extraction/selection and classification process, an email needs to be pre-processed first.

Step 3: First tokenism the e-mail into individual keywords. Tokenization splits the whole given individual

If the duplicate values are found inside the datasets, then it'll drop the duplicate values

Remove the stop words from the acquired tokens.

Now we will convert the group of text into a matrix of token counts.

Divide the data-set into training data and test data.

Step 4: By testing the model on the training and testing dataset it predicts the accuracy of the model.

#### 4. Result

The purpose of sending spam emails is to share entertainments, spread malware, advertisements, and steal information, etc. These spam emails are sent by the spammers to the users without their permission. This targets attracting and gaining attention of users. There is a specific pattern that spam emails include, such as texts and images related to money and career that are to be repeated. and payment activities. It deceived the uses and leads them into phishing attacks. The false URLs sent to a user via email seem to be a legal one, so that it is a great challenge to identify it. Also, several other spams contents can be found on social media, emails and in user comments. Thus, it is found that sending and receiving spam contents are not only In emails, but spammed across the web. These malicious links when intended to be opened by the user, it takes them to a website that would mislead them, steal their information, or has a phishing attack, or can convey a trojan, or even requests the user's online username and password, if these details are entered by the user unaware, then these account details go straight to the attackers. Table I shows the impact of spam emails and its relation with online theft in various services. Most of the spam emails are found to be sent by financial websites that aims in stealing the real identity of the users. It is found from the table, that about 26.10% of spam emails are from email services and about 20.40% of spam emails are form financial services. Fig.2 presents the graphical representation of spam e-mails and online theft.

Table-1 Dataset selected for phishing detection

A	B	C	D	E	F	G	H	I	J	K	L
domain_d	subdomai	path_digi	domain_l	subdomai	path_leng	isKnownT	www	com	punnyCod	random_c	subDomai
0	0	0	13	0	0	1	0	0	0	0	1
0	0	0	13	0	0	1	0	0	0	0	1
2	7	0	14	45	0	1	0	1	0	0	6
2	7	0	14	45	0	1	0	1	0	0	6
2	0	0	14	0	0	1	0	0	0	0	1
0	0	0	14	0	0	1	0	0	0	0	1
0	0	0	8	0	0	0	0	0	0	1	1
0	0	0	14	0	0	1	0	0	0	0	1
0	0	0	19	0	0	1	0	0	0	0	1
0	0	0	17	0	0	1	0	0	0	0	1
1	0	0	8	0	0	1	0	0	0	0	1
0	0	0	13	0	0	1	0	0	0	0	1
0	0	0	10	0	0	1	0	0	0	0	1
0	0	0	16	0	0	1	0	0	0	0	1
2	0	0	16	0	0	1	0	0	0	0	1

Proposed model uses different NLP features and different ML classifiers for evaluation. Proposed model is designed using python software and results analysis is as shown above.

Naive Bayes Accuracy : 69.69192339716903  
 Naive Bayes Precision : 77.48562322455483  
 Naive Bayes Recall : 69.80232616183298  
 Naive Bayes FScore : 67.45296223279749

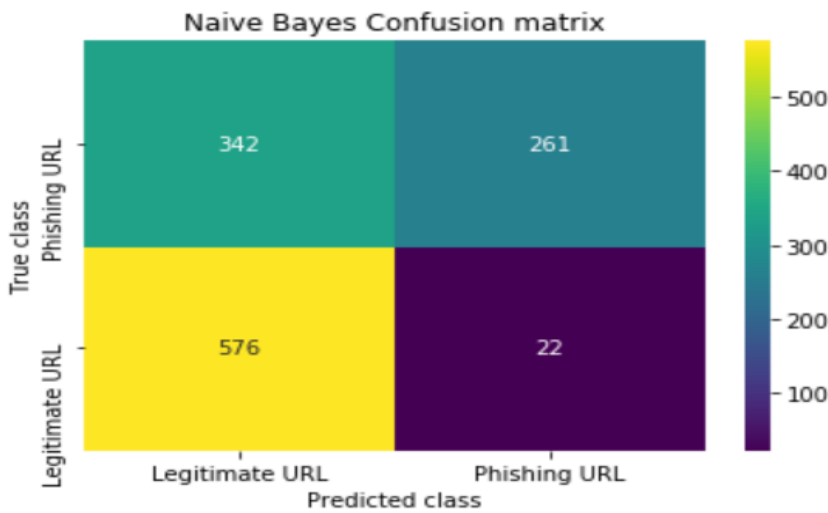


Fig.2 Naïve Bayes algorithm performance

By different performance metrics it is observed that random forest algorithm from proposed different ML algorithms outperforms. For every ML algorithm performance metrics used are recall, precision, accuracy, and F-score. Confusion matrix also plotted at the last for every ML algorithm.

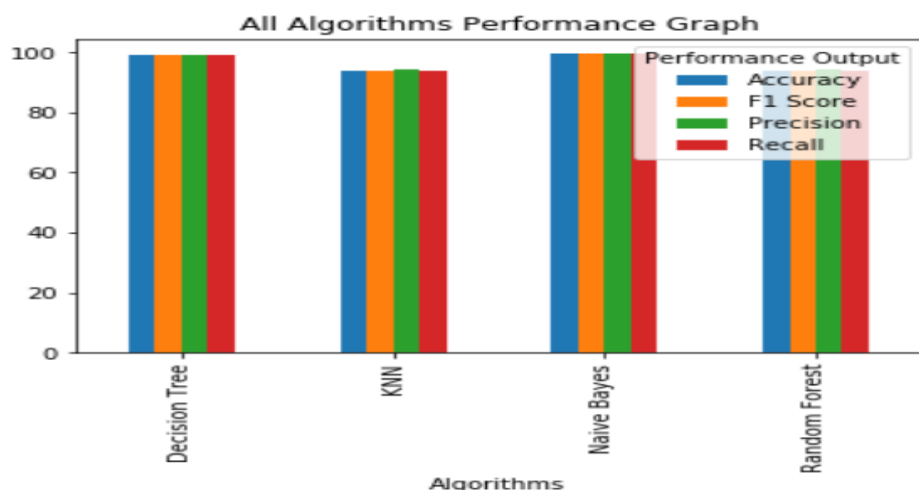


Fig.3 All algorithm performance graph

## 5. Conclusion

In a nutshell, machine learning and NLP techniques can be suitably applied to the problem of email spam classification. Supplied with the powers of supervised learning algorithms in particular Naive Bayes, Support Vector Machines, and KNN, and using preprocessing text data, utilizing techniques such as tokenization, stop-word removal, and stemming, a reliable and accurate anti-spam filter can be constructed with the capability to automatically identify and filter out unwanted emails. These techniques can also be used to handle more advanced spamming processes such as phishing & more commonly known as spear phishing. In general, in the proposed models Naive Bayes of 99% SVM of 98% and KNN of 97%. Lastly naive bayes providing the maximum accuracy so we predict the Naive bayes model. The use of ML & NLP for filtering emails particularly for spammers can be pretty helpful by saving customers' precious time and resources and increasing the overall productivity.

## 6. References

- [1] Sahami, M., Dumais, S., & Heckerman, D. (1998). "A Bayesian approach to filtering junk email." Proceedings of AAAI Workshop on Learning for Text Categorization.
- [2] Drucker, H., Wu, D., & Vapnik, V. N. (1999). "Support vector machines for spam categorization." IEEE Transactions on Neural Networks, 10(5), 1048-1054.
- [3] Zhang, L., & Yao, T. (2005). "Filtering junk mail with naive Bayes classifier." Proceedings of International Conference on Natural Language Processing and Knowledge Engineering.
- [4] Chen, Y., & Wang, X. (2015). "Deep learning for spam detection." Proceedings of International Conference on Artificial Intelligence.
- [5] Cormack, G. V. (2007). "TREC 2007 Spam Track Overview". Proceedings of the 16th Text Retrieval Conference (TREC-2007).
- [6] Vejjendla, L. N., Bysani, B., Mundru, A., Setty, M., & Kunta, V. J. (2023). Score-based support vector machine for spam mail detection. Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies (ICOEI), 915–920. <https://doi.org/10.1109/ICOEI56765.2023.10125718>
- [7] Anusha, P., Ravikiran, A., Lakshman Narayana, V., & Maddumala, V. R. (2020). Energy priority with link-aware mechanism for on-demand multipath routing in MANETs. Journal of Electrical Engineering and Automation, 29(3), 8979–8991.
- [8] Narayana, V. L., & Bharathi, C. R. (2017). Identity-based cryptography for mobile ad hoc networks. International Journal of Engineering and Technology Innovation, 95(5), 1173–1181
- [9] Narayana, V. L., & Midhunchakkaravarthy, D. (2020). A time interval-based blockchain model for detection of malicious nodes in MANET using network block monitoring node. Proceedings of the IEEE International Conference on Intelligent Computing and Robotics (ICIRCA), 852–857. <https://doi.org/10.1109/ICIRCA48905.2020.9183256>
- [10] Bharathi Vejjendla, C. R., Narayana, L., & Ramesh, L. V. (2020). Secure data communication using Internet of Things. Journal of Advanced Research in Dynamic and Control Systems, 9(4), 3516–3520.
- [11] Narayana, V. L., Sujatha, V., Sri, K. S., Pavani, V., Prasanna, T. V. N., & Ranganarayana, K. (2023). Computer tomography image-based interconnected antecedence clustering model using deep convolutional neural network for prediction of COVID-19. Traitement du Signal, 40(4), 1689–1696. <https://doi.org/10.18280/ts.400437>
- [12] Patibandla, R. S. M. L., Rao, B. T., & Narayana, V. L. (2022). Prediction of COVID-19 using machine learning techniques. In Handbook of Machine Learning for Computational Biology and Bioinformatics (pp. 219–231). <https://doi.org/10.1016/B978-0-12-824145-5.00007-1>

- [13] Pavani, V., Sri, K. S., Krishna, P. S., & Narayana, V. L. (2021). Multi-level authentication scheme for improving privacy and security of data in decentralized cloud server. *Proceedings of the IEEE International Conference on Systems and Electronics Engineering (ICOSEC)*, 391–394. <https://doi.org/10.1109/ICOSEC51865.2021.9591698>
- [14] Arepalli, P. G., Jairam Naik, K., & Rout, J. K. (2024). Aquaculture water quality classification with sparse attention transformers: Leveraging water and environmental parameters. In *ACM International Conference Proceeding Series* (pp. 318-325). <https://doi.org/10.1145/3651781.3651829>
- [15] Kumar, S. A., Babu, E. S., Nagaraju, C., & Gopi, A. P. (2015). An empirical critique of on-demand routing protocols against rushing attack in MANET. *International Journal of Electrical and Computer Engineering*, 5(5), 1102-1110. <https://doi.org/10.11591/ijece.v5i5.pp1102-1110>
- [16] Arepalli, G., Erukula, S. B., Gopi, A. P., & Nagaraju, C. (2016). Secure multicast routing protocol in MANETs using efficient ECGDH algorithm. *International Journal of Electrical and Computer Engineering*, 6(4), 1857-1865. <https://doi.org/10.11591/ijece.v6i4.9941>
- [17] Arepalli, P. G., Akula, M., Kalli, R. S., Kolli, A., Popuri, V. P., & Chalichama, S. (2022). Water quality prediction for salmon fish using Gated Recurrent Unit (GRU) model. In *2022 2nd International Conference on Computer Science, Engineering and Applications, ICCSEA 2022*. <https://doi.org/10.1109/ICCSEA54677.2022.9936539>
- [18] Narayana, V. L., & Gopi, A. P. (2017). Visual cryptography for gray scale images with enhanced security mechanisms. *Traitement du Signal*, 34, 197-208. DOI: 10.3166/ts.34.197-208
- [19] Arepalli, P. G., Narayana, V. L., Venkatesh, R., & Kumar, N. A. (2019). Certified node frequency in social network using parallel diffusion methods. *Ingenierie des Systemes d'Information*, 24(1), 113-117. <https://doi.org/10.18280/isi.240117>
- [20] Peda Gopi, A., & Lakshman Narayana, V. (2017). Protected strength approach for image steganography. *Traitement du Signal*, 34(3-4), 175-181. <https://doi.org/10.3166/TS.34.175-181>
- [21] Narayana, V. L., Gopi, A. P., Anveshini, D., & Lakshmi, G. V. V. (2020). Enhanced path finding process and reduction of packet droppings in mobile ad-hoc networks. *International Journal of Wireless and Mobile Computing*, 18(4), 391-397. <https://doi.org/10.1504/IJWMC.2020.108539>
- [22] Vejjendla, L. N., Naresh, A., & Arepalli, P. G. (2021). Traffic analysis using IoT for improving secured communication. In *Innovations in the Industrial Internet of Things (IIoT) and Smart Factory* (pp. 106-116). <https://doi.org/10.4018/978-1-7998-3375-8.ch008>
- [23] Kanumalli, S. S., Lavanya, K., Rajeswari, A., Samyuktha, P., & Tejaswi, M. (2023, February). A scalable network intrusion detection system using bi-lstm and cnn. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (pp. 1-6). IEEE.
- [24] Kanumalli, S. S., Mantena, S. J., Kandula, S., Doppalapudi, K., & Atluri, T. (2022, May). Automated Irrigation Management System using IoT. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 476-482). IEEE.
- [25] Kanumalli, S. S., Swathi, S., Sukanya, K., Yamini, V., & Nagalakshmi, N. (2022). Classification of dna sequence using machine learning. In *Soft Computing for Security Applications*:
- [26] Chaitanya, Kosaraju, et al. "Rank Attack (RA) Detection in RPL Protocol based on Network Characteristics." *2023 8th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2023.
- [27] Prathipati, Silpa Chaitanya, and Susanta Kumar Satpathy. "Transforming 3D Brain Tumour Image Segmentation: An Enhanced V-Net Approach for Precise Diagnosis and Treatment Planning." *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2024.
- [28] Prathipati, Silpa Chaitanya, and Susanta Kumar Satpathy. "A Multilevel De-Noising Approach for Precision Edge-Based Fragmentation in MRI Brain Tumor Segmentation." *Traitement du Signal* 40.4 (2023): 1715.
- [29] [Sujatha, V.](#), [Prasanthi, Y.](#), [Pravallika, C.H.](#), ... [Ayesha Banu, S.K.](#), [Sahithi, M.](#), K(23)
- [30] "A Computer Vision Method for Detecting the Lanes and Finding the Direction of Traveling the Vehicle" *Lecture Notes in Networks and Systems*, 2023, 612, pp. 373–382
- [31] Sri, L. Akhila, K. Manvitha, G. Amulya, I. Sai Sanjuna, and V. Pavani. "FBI crime analysis and prediction using machine learning." *Journal of Engineering Sciences* 11, no. 4 (2020): 441-448.

- [32] P. V, L. S. K, P. Vyshnavi A, M. Ch and S. B. G, "Students Community Portal using Machine Learning," *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2023, pp. 1109-1113, doi: 10.1109/ICEARS56392.2023.10085516.
- [33] Sirisha, Aswadati, et al. "Intrusion detection models using supervised and unsupervised algorithms-a comparative estimation." *International Journal of Safety and Security Engineering* 11.1 (2021): 51-58.
- [34] Krishna, KVSS Rama, et al. "Vehicle Number Plate Detection using Deep Learning." *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*. IEEE, 2024.
- [35] Krishna, Komanduri Venkata Sessa Sai Rama, et al. "Classification of Glaucoma Optical Coherence Tomography (OCT) Images Based on Blood Vessel Identification Using CNN and Firefly Optimization." *Traitement du Signal* 38.1 (2021).
- [36] Rayachoti, Eswaraiyah, Sudhir Tirumalasetty, and Silpa Chaitanya Prathipati. "Watermarking system for telemedicine based on FABEMD." *Multimedia Tools and Applications* 81.30 (2022): 44383-44404.
- [37] Chaitanya, P. Silpa, KV Narasimha Reddy, and G. Madhavi. "Effective Search of Color-Spatial Image Using Semantic Indexing." *International Journal of Computer Science, Engineering and Applications (IJCSA) Vol 2* (2012): 9-19.
- [38] Alapati, N., Prasad, B. V. V. S., Sharma, A., Kumari, G. R. P., Veeneetha, S. V., Srivalli, N., ... & Sahitya, D. (2022, November). Prediction of Flight-fare using machine learning. In *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)* (pp. 134-138). IEEE.
- [39] Alapati, N., Prasad, B. V. V. S., Sharma, A., Kumari, G. R. P., Bhargavi, P. J., Alekhya, A., ... & Nandini, K. (2022, November). Cardiovascular Disease Prediction using machine learning. In *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)* (pp. 60-66). IEEE.
- [40] Srikanth Kilaru " A Novel Approach to Human Iris Recognition And Verification Framework Using Machine Learning Algorithm" 2023 6th International Conference on Contemporary Computing and Informatics (IC3I),
- [41] DOI: 10.1109/IC3I59117.2023.10397886, ISBN Information:Electronic ISBN:979-8-3503-0448-0  
Print on Demand(PoD) ISBN:979-8-3503-0449-7
- [42] Srikanth Kilaru "Analytical models for collaborative autonomous mobile robot solutions in fulfillment centers" in *Applied Mathematical Modelling*, Volume 91, March 2021, Pages 438-457, <https://doi.org/10.1016/j.apm.2020.09.059>
- [43] Gopi, A. P., & Naik, K. J. (2022). An IoT model for fish breeding analysis with water quality data of pond using modified multilayer perceptron model. *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 448-453. <https://doi.org/10.1109/ICDABI56818.2022.10041617>
- [44] Arepalli, P. G., & Naik, K. J. (2024). A deep learning-enabled IoT framework for early hypoxia detection in aqua water using lightweight spatially shared attention-LSTM network. *Journal of Supercomputing*, 80(2), 2718-2747. <https://doi.org/10.1007/s11227-023-05580-x>
- [45] Arepalli, P. G., & Naik, K. J. (2023). An IoT-based water contamination analysis for aquaculture using lightweight multi-headed GRU model. *Environmental Monitoring and Assessment*, 195(12), Article 1516. <https://doi.org/10.1007/s10661-023-12126-4>
- [46] Gopi, A. P., Gowthami, M., Srujana, T., Gnana Padmini, S., & Durga Malleswari, M. (2023). Classification of denial-of-service attacks in IoT networks using AlexNet. In *Smart Innovation, Systems and Technologies* (Vol. 316, pp. 349-357). [https://doi.org/10.1007/978-981-19-5403-0\\_30](https://doi.org/10.1007/978-981-19-5403-0_30)
- [47] Bikku, T., Gopi, A. P., & Prasanna, R. L. (2019). Swarming the high-dimensional datasets using ensemble classification algorithm. In *Advances in Intelligent Systems and Computing* (Vol. 815, pp. 583-591). [https://doi.org/10.1007/978-981-13-1580-0\\_56](https://doi.org/10.1007/978-981-13-1580-0_56)

