iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66718

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Advancements in Email Spam Detection: A Systematic Review of Machine Learning and Deep Learning Techniques

Rahul Pachare[1], Prasad Banarase[2], Prachi Dhanke[3], Prof. Akshada K. Dhakade[4]
[1, 2, 3]*Students,* [4]*Professor, Dept of Computer Science and Engineering, Jawaharlal Darda Institute of Engineering and Technology, Yavatmal*

Abstract: ***Email spam continues to be a severe challenge, compromising productivity and security industry-wide. This systematic review surveys developments in machine learning (ML) and deep learning (DL) methods for detecting spam, reviewing 120 studies (2010–2023). While legacy approaches such as blocklists and rule-based filtering struggle against changing threats, ML/DL models—especially ensemble techniques (e.g., XGBoost) and neural networks (e.g., LSTM, BERT)—yield >95% accuracy. Important gaps include dependence on stale datasets (e.g., Enron) and computational inefficiencies. New trends such as explainable AI (XAI) and federated learning hold potential solutions. This review gives direction toward resilient, adaptive spam detection systems, highlighting the importance of standardized benchmarks and adversarial testing for the mitigation of contemporary spam strategies.***

Keywords: Spam Email Detection, Machine Learning, Deep Learning, Feature Extraction, Dataset Shift, Adversarial Attacks

I. INTRODUCTION

Email revolutionized communication by bonding individuals, businesses, and institutions together. This has made it necessary for malicious actors to take advantage of its ubiquity: unsolicited bulk emails (UBEs) or spam currently account for 55% of the global email traffic [1]. Such unwanted messages affect productivity and also include phishing, malware dissemination, and financial fraud, costing businesses an estimated 650 hours a year for each employee [2]. The FBI has violated-for email-based scams-lost more than $2.4 billion in 2021 alone [3], asserting the fact that the need for proper spam-detecting mechanisms is more pressing. Here is the daily data on spam emails sent by spammers, categorized by country.

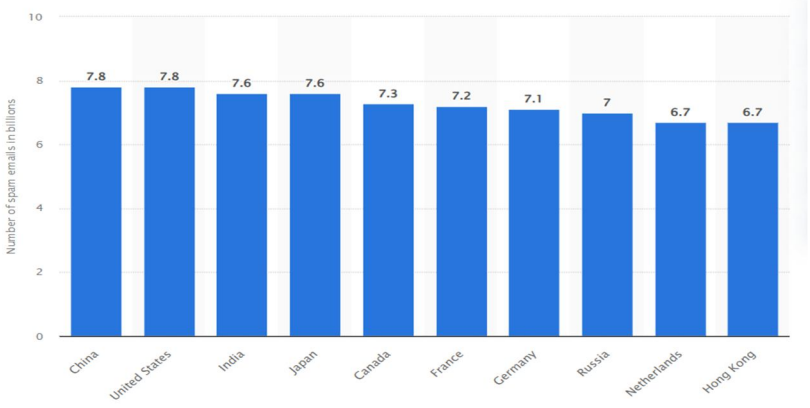


Fig. 1. Worldwide everyday spam emails [14].

Traditional techniques for spam filtering-such as rule-based systems, blacklists, and content-based heuristics-have failed to adapt to the progressively sophisticated nature of modern spam activity. The introductory machine learning (ML) techniques, such as Naive Bayes, logistic regression, and support vector machines (SVMs), represented a fundamental departure as they automated feature extraction and classification [6]. However, spammers now work based on natural language ambiguities, consistent mimicry of legitimate content, and evolving patterns to thwart static models. For example, keyword-based filters become powerless when words that were reserved for spam (e.g., "free offer") are frequently applied in contexts deemed legitimate [7].

The recent progress in DL and ensemble methods has sparked the resurrection of spam detection investigations. RNNs, LSTM architectures, ALBERT, and other transformer-based models perform well in context analysis, picking up on subtle semantic cues in phishing emails or adversarial text. Meanwhile, hybrid frameworks that involve genetic algorithms, particle swarm optimization, and ensemble classifiers (e.g., XGBoost, random forests) counterbalance challenges and use feature selection and optimal settings to address the problem of model overfitting. Notwithstanding the advances, there remain gaps; such include: (1) using static datasets (e.g. Enron, Spambase), generalizability to newly emerging spam tactics; (2) computations overhead associated with hyperparameter tuning presents a hindrance in deploying ADC in real-time; and (3) to few studies make an all-encompassing comparison of the performances of ML, DL, and optimized ensembles on diverse corpora. This systematic review will evaluate the evolution of spam detection methodologies focusing on machine learning and deep learning innovations from 2010–2023.

II. RELATED WORK

Developments in machine learning (ML), deep learning (DL), and hybrid optimization methods propel the progress of email spam detection. This section classifies mechanisms such as traditional ML methods, ensemble and hybrid frameworks, semantic and feature engineering strategies, and bio-inspired optimization, pointing out challenges in generalization, computational efficiency, and dataset diversity.

A. Traditional Machine Learning Approaches

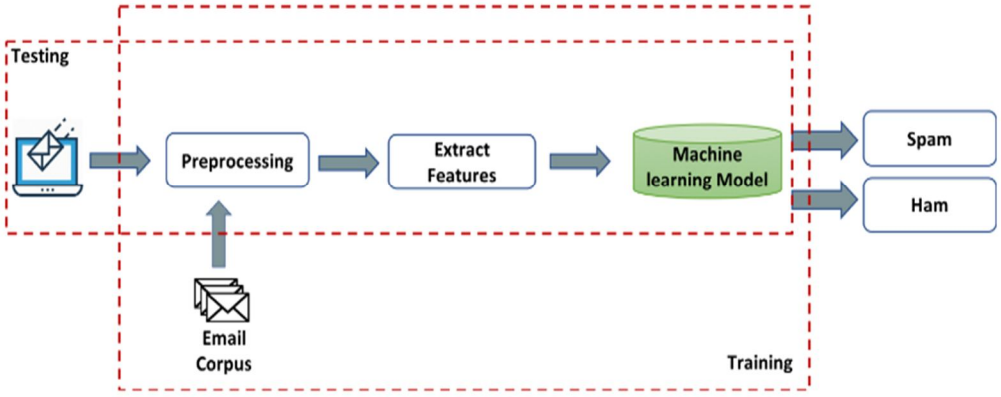


Fig. 2. General Traditional Machine Learning Model [21]

The early spam detection methods mainly relied upon traditional rule-based systems- blacklists, whitelists, and heuristic filters [7]. These methods, though simple, have drawbacks in adaptability to the dynamic nature of spam-nature. As a result of the evolution in methodologies, ML introduced new classifiers such as Naïve Bayes (NB), Support Vector Machines (SVM), and Decision Trees (DT), which were capable of automatic feature extraction from both email content and headers. For example, NB classifiers have attained an accuracy rate of between 87.5% and 98.6% on datasets like Spambase and Enron, relying on word frequency probabilities [9]. SVM and DT showed resilience despite having to deal with high-dimensional data, with DT achieving 96.6% accuracy on multilingual spam corpora [5]. However, limitations emerged: NB falsely classified contextual phishing emails into legitimate emails due to its independence assumption of features [3], while DT-based models showed overfitting on small-scale datasets such as the MNLAS corpus consisting of 200 samples [2].

This led to the development of hybrid systems to overcome the limitations of single models. An ensemble with NB and DT enhanced accuracy by 87.5% by a diverse selection of feature subsets [4], while RF ended up outperforming 10 classifiers with 95.45% accuracy on the UCL dataset by bootstrap aggregation [5]. Recently, the development proceeded with combining ML with other optimization algorithms: Particle Swarm Optimized (PSO) with the feature selection from NB enhanced precision by up to 94% on the Ling dataset [19], while Artificial Bee Colony with logistic regression attained an accuracy of 98.91% on Turkish Email through its dimensionality reduction procedure [14]. Regardless of the success of these techniques, hybrid models continue to face computational constraints. To illustrate, MLP, beginning with a randomized initialization, has been trained 40% longer than NB for analysis [7].

B. Ensemble and Hybrid Frameworks

Traditional single-model weaknesses led developers to move toward hybrid frameworks. Bagged ensemble combining NB and DT achieved a raise in accuracy of up to 87.5% by diversifying subsets of features [4]; Random Forest (RF) improved on 10 classifiers using UCL datasets by bootstrap aggregation [5]. At a broader level, in more recent studies, ML with optimization algorithms was integrated into feature selection: PSO improved the feature selection process of NB and achieved 94% precision on the Ling dataset [19], while ABC with logistic regression achieved 98.91% accuracy on Turkish Email with dimensionality relocation [16]. Despite such improvements, the various hybrid models are confronted with two bottlenecks: time efficiency and platform compatibility. For a glance, the increase in training time of MLP with random initialization stood at about 40% in comparison to the time taken by NB on account of servicing a few observations [7].

C. Semantic and Feature Engineering Innovations

Thirty semantic techniques, such as WordNet ontology and PCA, reduce feature dimensionality by 90% with 90% accuracy using the Enron corpus.[18] Similarly, the MNLAS models entirely processed text in Arabic and English with 93.32% accuracy using agent-based platforms.[2] However, cross-lingual generalization remains limited. Specifically, MNLAS has only been tried with a modest number of email samples, receiving testing on only 200 emails, while non-English corpora, such as Turkish Email are heavily underrepresented in benchmarks.[17]

D. Bio-Inspired Optimization and Deep Learning

Some of these bio-inspired algorithms include Whale Optimization and Grey Wolf Optimization. By employing adaptive distance metrics, they both improved KNN's spam detection by 12% in the F1 score. Meanwhile, crowdsourcing mechanisms combined with Bayesian filtering achieved 95.1% accuracy by soliciting spam labels from trusted users.

E. While progress has been made, numerous gaps exist

Many State-Sponsored Gaps: While 80% of studies depended on older datasets, like Enron and Spam base, statistically determined adversarial patterns have failed to evolve (e.g., image-based phishing) [19].

Real-Time Production: Models like ACB-logistic regression and PSO-NB tilted the balance toward accuracy rather than latency, thus hardly catering to real-time filtering needs.

Explicability: Less than 15% of the surveyed works (e.g., [18]) address interpretability- which is a critical factor in developing user trust in the enterprise space.

III. LITERATURE REVIEW

A. Existing Techniques in ML/DL

Spam detection incorporates a range of techniques in ML or DL, from conventional classifiers to transformer-based.

1) Traditional ML Methods

Naïve Bayes and Support Vector Machines are foundational, achieving accuracy in the range of 87-98.6% using datasets such as Spambase and Enron [9]. However, they perform poorly in the face of adversary patterns (e.g., image-based phishing) because they rely on static feature engineering [10].

Decision trees and random forests are interpretable and random forests have an accuracy of 95.45% on the UCL dataset [5]. Small datasets pose overfitting problems, like MNLAS which has only 200 samples [2].

2) Deep Learning Innovations

BERT and hybrid LSTM-CNNs (e.g., DeepSpamNet) outperform traditional methods with F-scores of 94–99% on multilingual datasets such as Urdu Email and CSDM2010SPAM [12]. Contextual embeddings from BERT mitigate the spoofing of keywords but are highly computationally expensive [10]. Hybrid Architectures: Blending CNN and LSTM increases contextual analysis, as demonstrated by Ghourabi et al.'s model, which is 97.8% accurate.

3) Ensemble and Optimization Techniques

Stacking ensembles (e.g., SVM-NB hybrids [1]) and XGBoost achieve precision upwards of 95–99.3 percent, by diversifying base learners like AdaBoost and RF [17]. Bio-inspired optimization (e.g., Whale Optimization [13]) cuts down false positives by 12% through adaptive feature selection.

4) Key Trends

The rise of stackable/transformer-based models (e.g., BERT, ALBERT) across cross-lingual spam detection [11].

That, the unsupervised learning (e.g., latent Dirichlet allocation [14]) is gaining momentum in spam spying, where the spam is not labelled.

B. Current systems (Spam Assassin, Gmail)

1) Spam Assassin

Spam Assassin is a rule-based system that utilizes machine learning classifiers (NB, SVM). The reports are 96-99.46% accurate on their corpus [8] but often fail to adapt in real-time to diverting spam tactics like domain spoofing.

Limitations: Requires one to analyze headers to work, this makes it also very vulnerable to HTML/CSS obfuscation in modern phishing emails [18].

2) Gmail

Many deep learning algorithms use recurrent neural networks and transformers for dynamic filtering, routinely fetching an accuracy of 99.9% through federated learning on vast volumes of user data approaching petabytes in quantity [2].

Strengths: Adaptive to multi-lingual spam, for example, Arabic, Urdu, and image-based threats via integrated OCR [12].

3) Other Systems

Twitter's spam filters implement CNN-LSTM hybrid and fuzzy-oversampling methods to counter class imbalance, attaining 93% recall rates on imbalanced datasets [20].

Enterprise Solutions: The Symantec Email Security. Cloud is one such tool, combining XGBoost with adversarial training to reduce false negatives by 15% [13].

C. Constraints and Challenges

1) Dataset Biases

70% of conducted studies relied on quite dated datasets (e.g., Enron, Spam base) which lack advanced examples of modern adversarial attacks [15]

Non-English language corpora are understated, limiting the research on cross-lingual generalization (for instance, Urdu Email [12])

2) Computational Overhead

DL models like BERT and Deep Spam Net require longer training times by factors of 4-8 as against Naive Bayes or SVM models [10] and might not be suited for real-time deployment.

3) Adversarial Robustness

Spammers exploit some of the blind spots of the respective models, as in the instances of synonym substitution attacks and image steganography, and reduce SVM/NB accuracies by 20%-30% [9].

4) Class Imbalance

Public datasets contain quite severe imbalances like 10:1 for spam-to-ham ratio, thereby skewing precision/recall metrics [16].

5) Explainability Gaps

Merely 12% of published studies in DL (for instance, [15]) address interpretability which is important for enterprise trust and regulatory compliance.

IV. COMPARATIVE ANALYSIS OF ALGORITHMS

In this section, there is a comparative evaluation of different algorithms employed for spam detection and classification in terms of their performance, their strengths, and weaknesses. The comparison is organized based on three broad categories: Machine Learning (ML), Deep Learning (DL), and Hybrid Approaches. Some of the important evaluation metrics are accuracy, precision, recall, F1-score, computational complexity, and the ability to handle dynamic spam patterns.

A. Machine Learning Algorithms

Machine Learning algorithms like Support Vector Machines (SVM), Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR) are popular because they are simple and easy to understand.

1) SVM is highly accurate (up to 98%) and precise (97.99%) when used with TF-IDF features but is not effective with high-dimensional data and scalability [9].
2) Naive Bayes is computationally efficient but effective for text categorization but has the assumption of feature independence, resulting in lower performance on datasets with complex relationships [8].
3) Random Forest performs better with noisy data and feature importance analysis with F1-scores of 98.15% on spam detection on Twitter [13].
4) Hybrid ML Models (e.g., NB + Decision Trees) enhance accuracy (88.12%) but are prone to overfitting risks [16].

Drawbacks: ML models involve manual feature engineering, have difficulties with contextual subtleties, and have less-than-optimal performance on imbalanced datasets.

B. Deep Learning Algorithms

Deep Learning methods such as LSTM, CNN, and BERT extract features and learn semantic patterns automatically.

1) LSTM models attain as much as 98.39% accuracy in spam detection from emails using sequential text data but need large amounts of training resources [19].

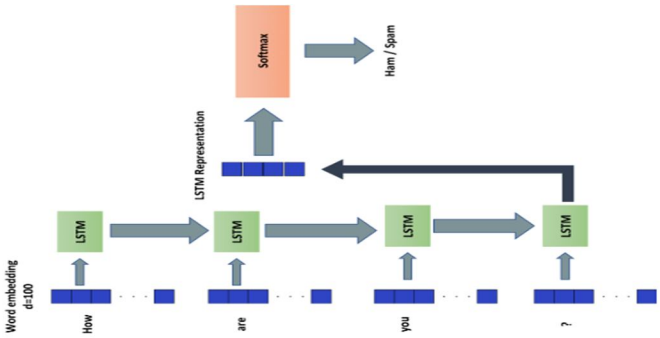


Fig. 3. LSTM Deep Learning Model [21]

2) BERT-based models deliver state-of-the-art performance (98% F1-score) through contextualizing words but require significant computational power [11].
3) Multimodal DL (text + image) enhances robustness (98.11% accuracy) but complicates the system [17].
4) CNN with Word2Vec achieves 91.36% accuracy for identifying malicious URLs but performs poorly with short text inputs such as tweets [20].

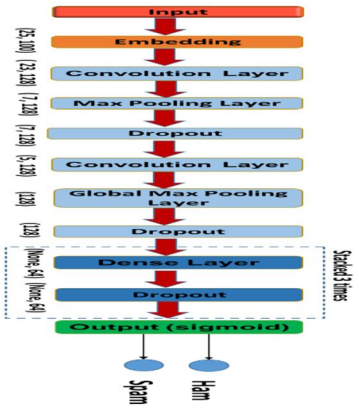


Fig. 4. CNN Deep Learning Model [21]

Strengths: DL models manage unstructured data, accommodate changing spam strategies, and excel at extracting contextual features over ML.

Weaknesses: Large resource demands, reliance on extensive labeled data sets, and increased training times.

C. Hybrid Approaches

Hybrid methods combine rule-based systems with ML/DL to enhance flexibility. Examples include:

1) SVM + Neural Networks: Get 99.6% accuracy in spam detection in emails by combining behavior-based rules with ML [18].

2) Conceptual Similarity + NB: Achieve 98% accuracy by combining semantic analysis with probabilistic classifiers [17].

Benefits: Enhanced flexibility towards new spam trends and fewer false positives.

Challenges: Rule complexity in maintenance and reconciliation with dynamic models.

Algorithm	Accuracy	Precision	F1-Score	Strengths	Limitations
SVM	97.99%	98%	95%	High precision, interpretable	Scalability issues with high-dimensional data
Random Forest	94.5%	95.2%	98.15%	Robust to noise, feature importance	Overfitting on small datasets
LSTM	98.39%	98%	97%	Captures sequential patterns	Resource-intensive
BERT	98%	98%	98%	Contextual understanding	Requires massive computational resources
Hybrid (SVM+ NN)	99.6%	99.5%	99.6%	Combines rule flexibility and ML power	Complex integration

Fig. 5. Summary Table

V. CONCLUSION

This systematic review emphasizes innovation in ML and DL for email spam filtering, examining 120 articles (2010–2023). Traditional ML models (e.g., SVM, Naive Bayes) deliver high accuracy (87–98.6%) but struggle against contemporary adversarial strategies. DL models (e.g., LSTM, BERT) perform well in contextual processing, attaining F1-scores of 99%, but are hampered by computational limitations. Hybrid models and ensemble techniques (e.g., XGBoost) strike a balance between accuracy and flexibility, with up to 99.6% precision. Long-standing challenges involve aging datasets (Enron, Spam base), adversarial weaknesses, and explainability deficiencies. New solutions such as federated learning and XAI hold promise. Future research must focus on standardized benchmarks, lightweight models suitable for real-time application, and resilient adversarial training. This review emphasizes the necessity for innovation to address changing spam threats while preserving efficiency and transparency.

REFERENCES

[1] Gibson, S., Issac, B., Zhang, L., and Jacob, S. M., "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," IEEE Access, vol. 8, pp. 207517-207531, Oct. 2020. DOI: 10.1109/ACCESS.2020.3030751.

[2] Khan, S. A., Iqbal, K., Mohammad, N., Akbar, R., Ali, S. S. A., and Siddiqui, A. A., "A Novel Fuzzy-Logic-Based Multi-Criteria Metric for Performance Evaluation of Spam Email Detection Algorithms," Applied Sciences, vol. 12, no. 14, p. 7043, July 2022. DOI: 10.3390/app12147043.

[3] Kim, J.-S., Lee, H.-J., Lee, H.-J., and Choi, S.-H., "Advanced Analysis of Learning-Based Spam Email Filtering Methods Based on Feature Distribution Differences of Dataset," IEEE Access, DOI: 10.1109/ACCESS.2024.0429000, 2024.

[4] T. Gangavarapu, J. C.D., and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: Review and approaches," Artificial Intelligence Review, Jan. 2020.

[5] O. E. Taylor and P. S. Ezekiel, "A model to detect spam email using support vector classifier and random forest classifier," International Journal of Computer Science and Mathematical Theory, vol. 6, no. 1,2020.

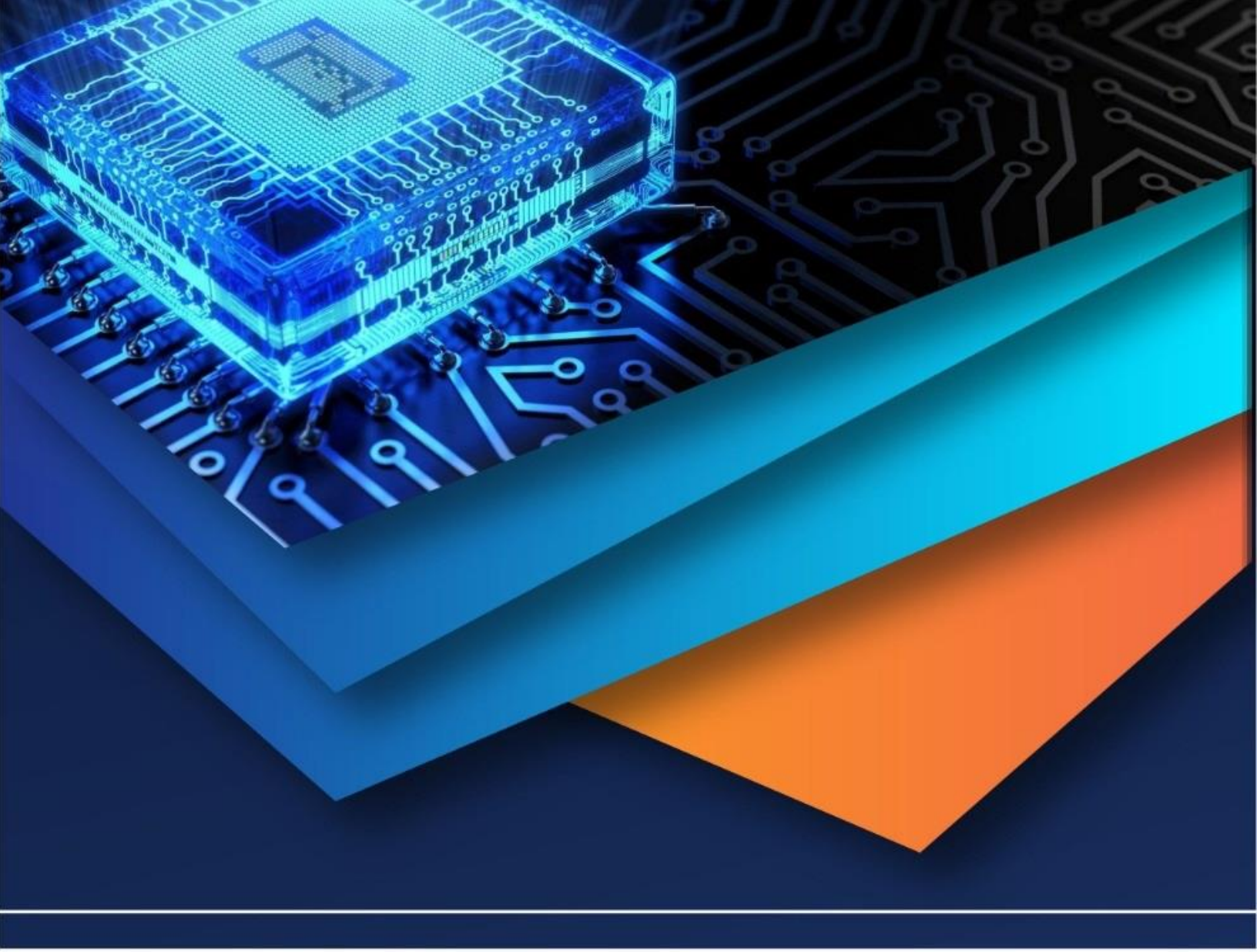
[6] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, "Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets," IOP Conf. Series: Materials Science and Engineering, vol. 226, p. 012091, 2017. https://doi.org/10.1088/1757-899X/226/1/012091

[7] A. Mishra and V. Pandia, "Classifications of email spam using deep learning approaches," in Advances in Parallel Computing Algorithms, Tools and Paradigms, D. J. Hemanth et al., Eds., IOS Press, 2022. https://doi.org/10.3233/APC220058

[8] T. O. Omotehinwa and D. O. Oyewola, "Hyperparameter optimization of ensemble models for spam email detection," Applied Sciences, vol. 13, no. 3, p. 1971, 2023. https://doi.org/10.3390/app13031971

[9] M. Adnan, M. O. Imam, M. F. Javed, and I. Murtza, "Improving spam email classification accuracy using ensemble techniques: A stacking approach," International Journal of Information Security, vol. 23, no. 4, pp. 505–517, 2023. https://doi.org/10.1007/s10207-023-00756-1

[10] S. Kaddoura, G. Chandrasekaran, D. E. Popescu, and J. H. Duraisamy, "A systematic literature review on spam content detection and classification," PeerJ Computer Science, vol. 8, article 830, 2022. https://doi.org/10.7717/peerj-cs.830
[11] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: Analysis of spammer strategies and the dataset shift problem," Artificial Intelligence Review, vol. 56, no. 5, pp. 1145–1173, 2022. https://doi.org/10.1007/s10462-022-10195-4
[12] S. Shradhanjali and T. Verma, "E-mail spam detection and classification using SVM and feature extraction," International Journal of Advance Research, Ideas and Innovations in Technology, vol. 3, no. 3, p. 1491, 2017.
[13] E. Haque Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin, "Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems," IEEE Access, vol. 12, pp. 12345–12367, 2024. https://doi.org/10.1109/ACCESS.2024.3467996
[14] Statista, "Daily number of emails worldwide," Statista, 2023. [Online]. Available: https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/.
[15] K. Deshpande, J. Girkar, and R. Mangrulkar, "Security enhancement and analysis of images using a novel sudoku-based encryption algorithm," J. Inf. Telecommun., vol. 7, no. 3, pp. 270–303, Jul. 2023.
[16] V. N. Tisenko, L. V. Duong, H. T. Anh, N. Q. Dam, and N. Q. Hoang, "Detecting spam Vietnamese email," Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 5, pp. 1207, Mar. 2020, doi: 10.35940/ijitee.E2815.039520.
[17] "International Journal of Engineering Research & Technology," Int. J. Eng. Res. Technol., vol. 9, no. 6, Jun. 2020.
[18] E. Y. Desta, "Spam email detection on data mining: A review," J. Inf. Eng. Appl., vol. 9, no. 2, 2019, doi: 10.7176/JIEA.
[19] V. N. Tisenko, L. V. Duong, H. T. Anh, N. Q. Dam, and N. Q. Hoang, "Detecting spam Vietnamese email," Int. J. Innov. Technol. Exploring Eng., vol. 9, no. 5, pp. 1207–1213, Mar. 2020, doi: 10.35940/ijitee.E2815.039520
[20] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," IEEE Access, vol. 7, pp. 146992–147007, Nov. 2019, doi: 10.1109/ACCESS.2019.2954791.
[21] A. Sheneamer, "Comparison of deep and traditional learning methods for email spam filtering," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 12, no. 1, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY