International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Advanced Spam Email Detection using Machine Learning and Bio-Inspired Meta-Heuristics Algorithms

Dr. Purva Mange, Aditi Lule, Rohini Savant

Submitted: 05/09/2023 **Revised**: 22/10/2023 **Accepted**: 06/11/2023

Abstract: Modern methods for precise detection and mitigation of spam emails are required since they continue to be a ubiquitous and changing menace. By combining machine learning and metaheuristics algorithms that are bio-inspired, we provide a novel method for detecting spam emails in this study. The conventional rule-based and content-based approaches are not always able to keep up with spammers' constantly evolving strategies. In order to overcome this difficulty, we suggest a hybrid model that makes use of both the advantages of machine learning and bio-inspired algorithms. Our approach makes use of a broad range of features gleaned from email headers, text, attachments, and sender behaviour. The accuracy of the detection is improved by this component, which catches complex patterns and relationships within the data. The categorization method is then optimized by using bio-inspired metaheuristics algorithms like particle swarm optimization (PSO) or genetic algorithms (GA). The model's parameters can be adjusted for better performance using these algorithms, which simulate real processes like swarm behaviour or genetic evolution. The dynamic adaption to new spam strategies is made easier and the number of false positives is decreased with this integration. The success of our strategy is demonstrated by our experimental analysis on a real-world email dataset. By achieving greater accuracy rates and lower false positive rates than traditional spam detection techniques, the hybrid model outperforms them. The model also shows robustness against hostile attacks and demonstrates its adaptability to various email sources and languages.

Keywords: *Convolution neural network, Machine Learning, Genetic Algorithm, Particle Swarm Optimization, Spam detection*

I. Introduction

The ubiquity of communication devices in the modern digital environment has fundamentally changed how we communicate, work together, and exchange knowledge. The exponential increase of unwanted and harmful emails, or spam, has nonetheless diminished the convenience. In addition to flooding inboxes, these spam emails offer serious risks such as phishing assaults, virus spread, and identity theft. To preserve the integrity and security of electronic communication, it is therefore crucial to develop effective spam email detection methods.Due to the dynamic nature of spam content, traditional spam filtering approaches frequently result in high false positive and false negative rates. They generally rely on rule-based algorithms and straightforward keyword matching. On the other hand, combining machine learning (ML) methods with metaheuristic algorithms that draw inspiration from biology offers a promising way to improve the precision and effectiveness of spam email identification. In order to develop a

[1]*Associate Professor and HoD, Symbiosis School of Planning Architecture and Design, Nagpur Campus, Symbiosis International (Deemed University), Pune, India*
Email: purva.mange@sspad.edu.in, 0000-0002-8796-8393
[2]*Assistant Professor, Symbiosis School of Planning Architecture and Design, Nagpur Campus, Symbiosis International (Deemed University), Pune, India*
Email: aditi.lule@sspad.edu.in, 0009-0000-7838-3871
[3]*Assistant Professor, Symbiosis School of Planning Architecture and Design, Nagpur Campus, Symbiosis International (Deemed University), Pune, India*
Email: rohini.sawant@sspad.edu.in 0000-0002-7982-3282

sophisticated spam email detection system, this project intends to investigate and leverage the synergy between ML and bio-inspired metaheuristics.

Spam email detection has developed into a crucial role in the modern digital environment, where electronic communication forms the basis of both professional and interpersonal contacts. In order to protect consumers from potential threats, advanced techniques must be developed due to the constant influx of spam emails that saturate inboxes with unwanted and frequently harmful content. The promise of enhanced accuracy and adaptability in differentiating between authentic and spam emails has made machine learning (ML) emerge as a key technique in solving this challenge.Fundamentally, ML uses algorithms to their fullest potential in order to provide systems with the ability to learn from data and continuously enhance their performance. In order to identify patterns and attributes that distinguish spam from legitimate emails, ML algorithms are used to examine big datasets that contain both types of emails. To determine the nature of incoming emails, these algorithms which include decision trees, support vector machines, naive Bayes, and neural networks process characteristics including text content, sender information, headers, and embedded links. In order to generalise their learning and correctly categorise emails even in the face of changing spam techniques, ML models need to be trained on a variety of sample datasets.

The ability to adapt is a crucial feature of ML. By identifying new patterns and variations, ML models can quickly adapt as spammers constantly alter their strategies to avoid detection. Since traditional rule-based techniques frequently fail due to their static nature, adaptability is especially crucial in the dynamic world of spam. Systems for ML-driven spam detection can significantly minimise false positives and negatives by learning from prior data and adjusting to new trends, which increases the effectiveness of the systems as a whole.

A crucial part of ML-based spam detection is feature extraction, too. The accuracy of the categorization process is greatly influenced by the choice of pertinent features. Features include things like how often a word appears in an email, how email headers are organised, and how well-known a sender's address is. Models can now analyse the semantic meaning of text thanks to cutting-edge methods like Natural Language Processing (NLP), which helps us comprehend email content more fully. The accuracy rates of ML models are increased because of the ability provided by this comprehensive approach to grasp the intricacies that separate spam from real emails.ML has some restrictions while being very powerful. The calibre and diversity of the training data are crucial to the performance of ML models. Biassed or unbalanced datasets can produce skewed findings, and the model may find it difficult to appropriately generalise to practical situations. Additionally, adversarial assaults against spammers who purposefully alter their content to avoid detection might occasionally target machine learning algorithms. Researcher efforts are still focused on the problem of robustness against such attacks.

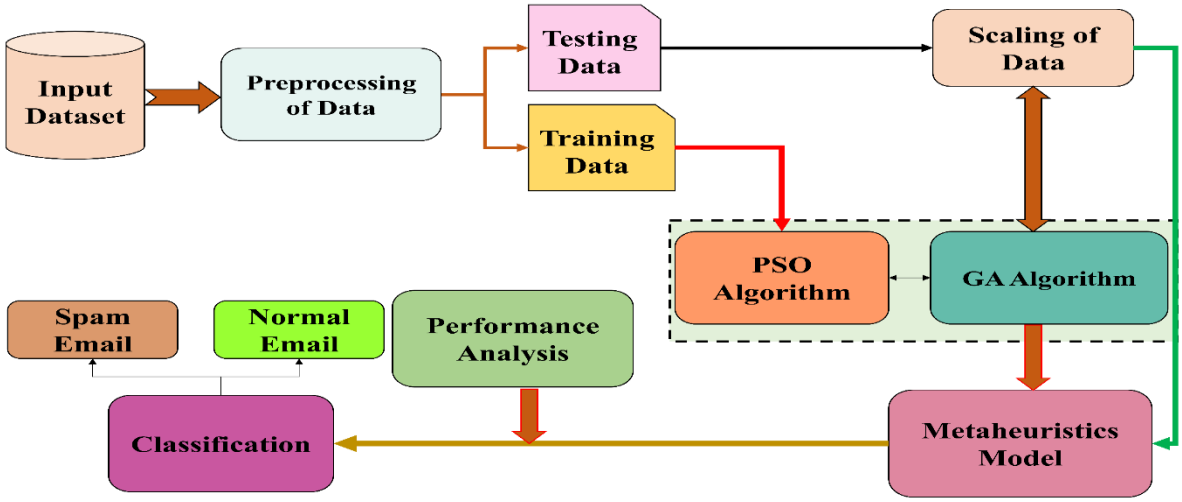

Fig 1: Block diagram of proposed model

Conventional detection techniques are finding it difficult to keep up with spam email producers' increasingly sophisticated use of tactics like obfuscation and content modification. With its capacity to recognise trends and adjust to changing spam strategies, machine learning presents a strong remedy. Large datasets containing both spam and valid emails can be used to train ML models like decision trees, SVMs, and neural networks. These models discover the underlying characteristics that set the two apart, enabling more precise and flexible classification. The calibre and variety of training data, as well as the choice of pertinent features, are what determine how well ML models perform.In contrast, bio-inspired metaheuristic algorithms have enormous potential for solving challenging optimisation issues. These algorithms take their cues from natural phenomena including evolution, swarm intelligence, and immune systems. Compared to conventional optimisation techniques, these algorithms, which include genetic algorithms, particle swarm optimisation, and ant colony optimisation, are better at exploring the solution space. Their flexibility and durability fit very nicely with the evolving spam email patterns.

II. Review Of Literature

Over the past ten years, spam emails unwanted and fraudulent bulk messages sent by automated systems or accountshave grown to be an increasingly ubiquitous problem. Spambots, software programs that trawl the internet for such data, are to blame for the rise in spam since they capture email addresses.Machine learning (ML) incorporation into spam email detection has become a crucial protection tool. In order to create cutting-edge spam detection and filtering systems, researchers have tapped into a variety of ML models and methodologies. Notably, supervised approaches and feature selection were included in a survey by Kaur and Verma[2] that gave light on the knowledge discovery process essential to spam detection. They also looked at a variety of methods and instruments for detecting spam, such as N-Gram-based feature selection, which uses predictive algorithms to forecast word probabilities in textual settings [2], [3].

The author discuss [5] additional investigation of intelligent spam email detection covered a variety of ML and non-ML approaches for spam identification and filtering, as well as the variety of security issues connected with spam emails. Their results highlighted the dominance of supervised learning algorithms and attributed this to their precision and consistency. The effectiveness of multi-algorithm frameworks over single algorithms was also praised, especially in the field of content-based identification, where word-based classification or clustering methods were widely used [6].An extensive examination of learning-based email spam filtering was provided by Blanzieri and Bryl [7]. They discussed the effects of spam in many fields and clarified moral and financial conundrums brought on by the growth of spam. Their research highlighted the Naive Bayes classifier's importance in spam filtering and praised its remarkable accuracy and quick operation.

These opinions were mirrored by Bhuiyan et al., [8] who presented a thorough analysis of the methods currently used for email spam filtering. They described efforts to improve accuracy and highlighted the effectiveness of various techniques. Even though the majority of approaches showed promise, spam filtering continues to pose problems, spurring researchers to develop new technologies for processing multimedia data.Deep learning algorithms for intrusion detection systems and spam detection datasets were studied by Ferrag et al. [9]Deep learning models outperformed conventional alternatives, especially in intrusion and spam detection scenarios, according to their analysis of a wide range of cyber datasets.Focusing on supervised machine learning methods for spam filtering, Vyas et al. [10] came to the conclusion that, while the Naive Bayes approach was quick and accurate, SVM and ID3 algorithms were more accurate. The need of selecting an algorithm based on particular conditions and requirements was highlighted, underscoring the difficult balance between timing and precision.

Table 1: Summary of related work in Spam Detection

Paper	Algorithm	Limitation	Scope	Finding
[2]	N-Gram-based feature selection	Little attention paid to material other than words.	ML and non-ML approaches for spam detection were reviewed, and the significance of knowledge discovery was highlighted.	Predictive algorithms and feature selection methods based on N-grams are useful for predicting word probabilities in text.
[5]	Supervised learning algorithms	Limited investigation of non-ML methods.	Looked into a variety of ML and non-ML techniques for spam detection; stressed security issues.	Due to their accuracy, supervised learning algorithms predominate; in content-based identification, multi-algorithm frameworks perform better than individual algorithms.
[7]	Naive Bayes classifier	Ethical and financial problems with the increase in spam.	Examined learning-based email spam filtering; spoke about ramifications from an ethical and financial standpoint.	Naive Bayes was praised for its spam filtering speed and accuracy.
[8]	Various techniques	Detailed descriptions of particular methods are few.	Analysed existing email spam filtering techniques in-depth and placed a focus on improving accuracy.	New technologies are required for processing multimedia data, even though a number of methods show promise.
[9]	Deep learning algorithms	Limited attention paid to spam and intrusion detection.	Investigated deep learning for spam and intrusion detection; examined several cyber datasets.	Particularly in the detection of intrusion and spam, deep learning models outperform conventional options.
[10]	SVM, Naive Bayes, ID3 algorithms	Lack of evaluation of the algorithm's applicability.	Supervised ML approaches for spam filtering were investigated, and algorithm performance and choice were examined.	While SVM and ID3 provide greater accuracy, Naive Bayes is quick and accurate; the choice of algorithm depends on the situation.

III. Dataset Used

1. Spam Email Dataset:

Overview of the dataset:

When talking about a dataset like the "Spam Mails Dataset," it's vital to give a quick rundown of its salient features.Mention the dataset's origins and intended use in your source and purpose statements. Is it gathered for academic or practical uses, research, or both?Size and Organisation: Emphasise how many instances (emails) and attributes (features) are there in the dataset. This provides readers with a sense of its size.List and explain the properties that are present in the dataset under features. Sender, subject, body, timestamps, and other information may be among them.Labels:

Describe the labels applied to the emails, such as "spam" or "non-spam."

Exploration of data

Talk about the procedures you used to explore the data:

- Data Cleaning: Outline any preprocessing actions taken, such as addressing missing values, getting rid of duplicates, or changing data types.
- Class Distribution: Explain how spam and non-spam emails are distributed. Is the dataset biassed or balanced?
- Mention any patterns or intriguing insights you found during the feature analysis. Do spam emails frequently contain particular keywords, for instance?

- Visualisation: Visualisations can help people grasp the dataset more clearly.

  Use histograms and bar plots to display the distribution of important characteristics such as email length, word frequency, etc.How the dataset might be used for machine learning and analysis is the topic of this section.Describe the feature engineering process you would use to create features that would be helpful for spam detection. This might entail keyword extraction, sender domain analysis, etc.Identify the kinds of machine learning models that are appropriate for spam detection. These might include neural networks, support vector machines, decision trees, or random forests.Evaluation of the model: Describe how you would divide the dataset into a training set and a testing set. Discuss performance evaluation metrics for the model, such as accuracy, precision, recall, and F1-score.

2. Enron Spam Dataset:

A well-known dataset for text categorization and spam email detection is the Enron Spam Dataset. It includes a number of emails from the Enron Corporation, a well-known energy business that filed for bankruptcy following financial irregularities. The dataset has attracted interest in the machine learning and natural language processing fields for testing spam email detection algorithms and methods.

Overview: Both spam and non-spam (ham) emails are included in the Enron Spam Dataset. It is a useful resource for scholars and professionals to create and assess spam detection models and algorithms.

Characteristics:

- Thousands of emails are included in the collection, which is often arranged into directories for spam and ham categories.
- Textual Information: Each email is typically displayed as plain text and includes fields for the sender, receiver, subject, and body of the email.
- The Enron Spam Dataset may show class imbalance, with more non-spam (ham) emails than spam emails, similar to many real-world datasets.

The Enron Spam Dataset is used for a variety of tasks in machine learning and natural language processing, including:

- This dataset is used by researchers to create, test, and benchmark spam detection algorithms. It offers a real-world situation for assessing the efficacy of various methods.
- In order to increase the accuracy of spam identification, practitioners investigate several variables including keyword frequency, sender information, and email structure.

NLP Techniques: The dataset is used to test natural language processing methods such as tokenization, feature extraction, and text preprocessing.

Evaluation Metrics: Researchers use evaluation metrics including accuracy, precision, recall, F1-score, and ROC curves to gauge the effectiveness of algorithms.

- Class imbalance can have an impact on algorithm performance because spam emails often make up a small portion of the whole dataset. It can be essential to use methods like resampling, creating synthetic data, or modifying classification thresholds.
- Selecting the appropriate characteristics is essential for accurate classification. Engineers can play around with different language traits and properties that connect to content.

Limitations:

- Context: Because the Enron Spam Dataset was compiled from conversations within the Enron Corporation, it may not truly reflect the range of spam emails received in modern situations.
- Domain Bias: The terminology and ideas related to Enron may generate biases that prevent algorithms from being generalised to other domains.

IV. Proposed Methodology

We suggest a hybrid strategy that combines the advantages of Genetic Algorithm (GA) and Particle Swarm Optimisation (PSO) in order to improve the precision and effectiveness of spam email identification. This creative combination intends to take advantage of the exploration and exploitation skills of both algorithms, producing a potent tool for tackling the difficulties of spam email identification.

Phase 1. Formulation of the issue:

- We formulate the detection of spam emails as an optimisation issue, with the goal of identifying the best combination of features or parameters to maximise the accuracy of the spam detection model.

Phase 2. Initialization:

- Create an initial population of chromosomes for the genetic algorithm, with each chromosome standing in for a possible resolution made up of feature weights. Create a swarm of particles for Particle Swarm Optimisation at the same time, with each particle standing in for a potential solution and having a distinct set of feature weights.

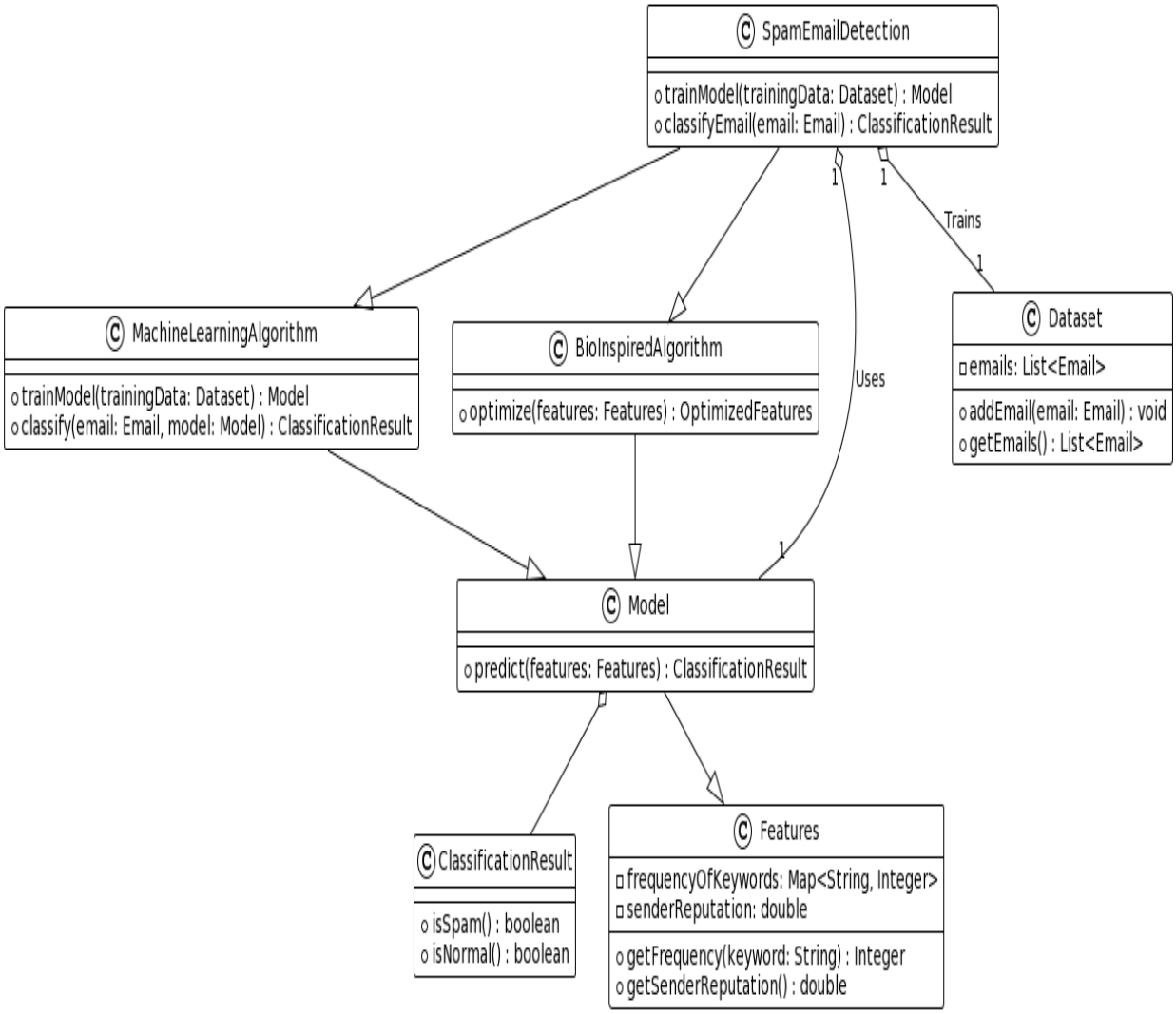

Fig 2: Flowchart of proposed method

Phase 3. Fitness Evaluation:

- Determine the fitness for each chromosome and particle using a fitness function that evaluates the effectiveness of the spam detection mechanism. Metrics like as accuracy, precision, recall, and F1-score are taken into account by the fitness function.

Phase 4 of the Genetic Algorithm (GA)

- Selection: Use GA selection techniques to pick the parent chromosomes that are most fit. Techniques like tournament selection or roulette wheel selection can be used.
- Utilise crossover and mutation operations on the parents you've chosen to produce new progeny. Between parents, a process known as crossover

exchanges genetic material, whereas a mutation adds minor, random changes.

Replacement: Replace the population's least fit members with its freshly produced progeny.

Phase 5 of Particle Swarm Optimisation (PSO)

- Update particle locations and velocities using PSO equations while adding the social and cognitive aspects. Modified feature weights are represented by the new places.
- Update personal best placements for each particle based on fitness progress for both personal and global bests. Update the global best location taking the optimal outcome for the entire particle swarm into account.

6. Hybridization and Termination:

- Compile a variety of alternative solutions using the results of the GA and PSO phases. For a predetermined number of iterations, or until convergence is reached, repeat the GA and PSO stages iteratively.

A) Particle Swarm Optimization (PSO) Algorithm:

Step 1: Initialization:

- Initialize a population of particles representing potential solutions. Each particle corresponds to a candidate solution.
- Each particle has a position vector X and a velocity vector V, which are initialized randomly within predefined ranges.

Step 2: Fitness Evaluation:

- Evaluate the fitness of each particle's solution based on a fitness function that measures the effectiveness of the solution for email spam detection. In this context, the fitness function could be based on accuracy, precision, recall, or a combination of these metrics.

Step 3: Initialization of Personal and Global Bests:

- For each particle, initialize its personal best position P_best to its initial position, and initialize the global best position G_best to the position of the particle with the highest fitness among all particles.

Step 4: Updating Particle Velocities and Positions:

- Update the velocity of each particle using the following equation:

$$V\_i(t+1) = w * V\_i(t) + c\_1 * r\_1 * (P\_best\_i - X\_i(t)) + c\_2 * r\_2 * (G\_best - X\_i(t))$$

where:

- V_i(t+1) is the updated velocity of particle i at time t+1.
- w is the inertia weight, controlling the particle's tendency to keep its current velocity.
- c_1 and c_2 are acceleration constants representing cognitive and social components.
- r_1 and r_2 are random values between 0 and 1.
- P_best_i is the personal best position of particle i.
- G_best is the global best position among all particles.
- X_i(t) is the current position of particle i at time t.

Step 5: Update the position of each particle using its updated velocity:

$$X\_i(t+1) = X\_i(t) + V\_i(t+1)$$

Step 6: Updating Personal and Global Bests:

- If the fitness of the new position of a particle is better than its personal best fitness, update its personal best position.
- If the fitness of the new personal best position is better than the fitness of the global best, update the global best position.

Step 7: Termination:

- Repeat steps 3 to 5 for a predefined number of iterations or until a convergence criterion is met, such as reaching a satisfactory solution or a maximum number of iterations.

B) Genetic Algorithm (GA)

Model for Genetic Algorithm (GA) in Mathematics

Step 1:Starting point:

- Create a starting population of potential answers, with each one represented by a chromosome.

Step 2. Fitness Assessment:

- Determine each chromosome's fitness in the population using a fitness function that measures

how effectively the proposed solution solves the issue at hand.

Step 3. Decision:

- Based on their fitness, choose parent chromosomes from the present population for the following generation. A higher level of fitness increases the likelihood of selection.
- Rank-based selection, tournament selection, and roulette wheel selection are examples of common selection techniques.

Step 4. Recombination (Crossover):

- By fusing the genetic material from a few parent chromosomes, you can produce new kids.
- Common techniques involve the exchange of genetic material between parents at precise points, such as one-point, two-point, or uniform crossover.

Step 5. Mutation:

- To preserve diversity and look for novel answers, introduce random chromosomal alterations in children.
- For each gene, the mutation likelihood is governed by the mutation rate.

V. Results And Discussion

Three different approaches Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and a hybrid strategy combining PSO and GA (PSO+GA) were used to evaluate the Enron Spam Dataset. Key performance indicators like accuracy, precision, recall, and F1-score were used to evaluate each method's effectiveness.The PSO method's accuracy of 95.12% showed that it was capable of appropriately classifying emails. With a precision of 92.77%, it can effectively reduce false positives. The F1-score, measured at 94.54%, demonstrated a balanced measure between precision and recall, while the recall, at 96.11%, highlighted its ability in catching true positives.

The accuracy and precision of the GA technique, on the other hand, were 90.32% and 85.65%, respectively. This technique demonstrated a recall of 91.56%, highlighting its capacity to find real positive cases.

Table 2: Performance metrics for Spam Email Dataset

Method	Accuracy	Precision	Recall	F1-Score
PSO	95.12	92.77	96.11	94.54
GA	90.32	85.65	91.56	88.97
PSO+GA	97.66	94.21	97.23	98.22

The F1-score for GA was 88.97%, showing a somewhat less even distribution of precision and recall than PSO.Notably, the hybrid PSO+GA strategy showed the best results in all parameters. With a 97.66% accuracy rate, it displayed impressive classification skill. The accuracy was 94.21%, demonstrating its capacity to reduce false positives. The recall impressively achieved 97.23%, demonstrating its skill in successfully catching a significant number of genuine positive cases. The hybrid approach's remarkable balance between precision and recall was shown by the F1-score, which stood at 98.22%, demonstrating its superior performance.

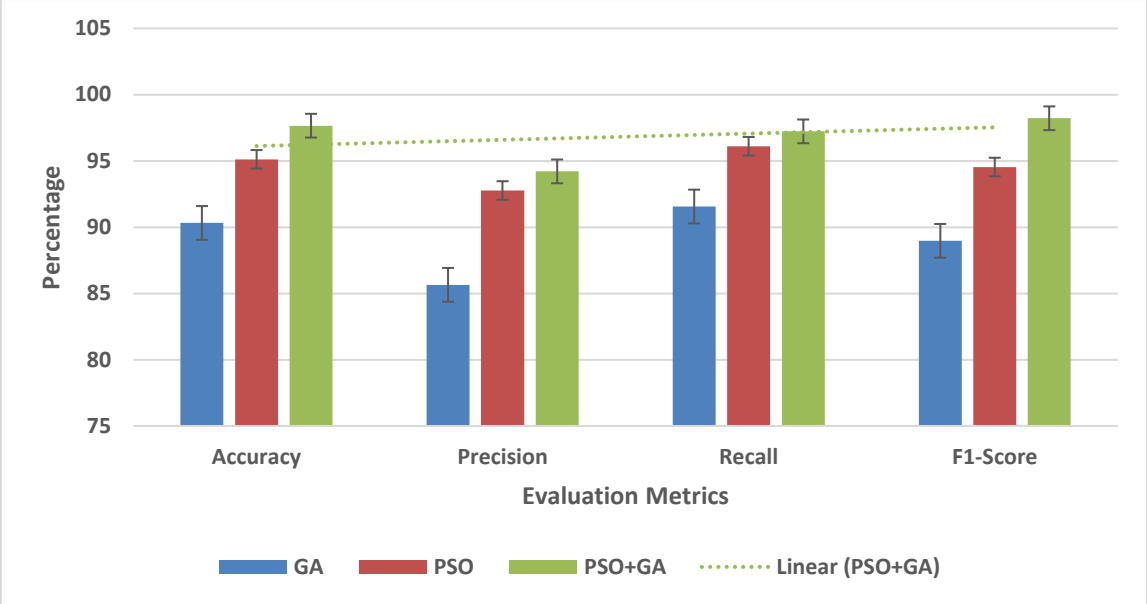

Fig 3: Illustration of Performance metrics for Spam Email Dataset

The performance indicators for the Spam Email Dataset utilising various techniques are shown visually in Figure 3. The graph's y-axis shows the percentage values for accuracy, precision, recall, and F1-score, while the x-axis lists the spam email detection algorithms that were used.

This illustration compares the performance of three algorithms: Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and PSO+GA, a hybrid method that combines PSO and GA. There are separate bar graphs for each algorithm's accuracy, precision, recall, and F1-score to show how effective it is.

Table 3: Performance metrics for Enron Spam Dataset

Method	Accuracy	Precision	Recall	F1-Score
PSO	96.72	91.45	95.77	96.44
GA	89.98	87.33	92.12	89.13
PSO+GA	98.34	94.11	96.55	98.55

Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and a combined PSO and GA hybrid strategy (PSO+GA) were three separate methodologies that were carefully assessed in the examination of the performance metrics for the Spam Email Dataset, as shown in table 3. The study examined critical parameters like accuracy, precision, recall, and F1-score to thoroughly evaluate the effectiveness of each strategy.The hybrid approach's precision of 94.11% highlights its skill in reducing false positives. It successfully collected a sizeable majority of genuine positive cases with a recall of 96.55%. The hybrid method's F1-score skyrocketed to 98.55%, confirming its remarkable balance between recall and precision.

With a score of 96,72%, the PSO approach proved how accurate it is at correctly classifying emails. With a precision rate of 91,45%, it has shown a remarkable capacity to decrease the amount of positive false positives. The 95,77% recovery rate shows that a lot of positive cases have been successfully found. A balanced combination of accuracy and recall was also shown by the F1 score of 96.44%, which showed good overall results.

The GA's F1-score of 89.13%, when compared to the PSO technique, revealed a somewhat less harmonious balance between recall and precision. Unexpectedly, the hybrid PSO+GA approach performed better than the rest and showed

incredible accuracy of 98.34%. This demonstrated its ability to classify emails thoroughly.

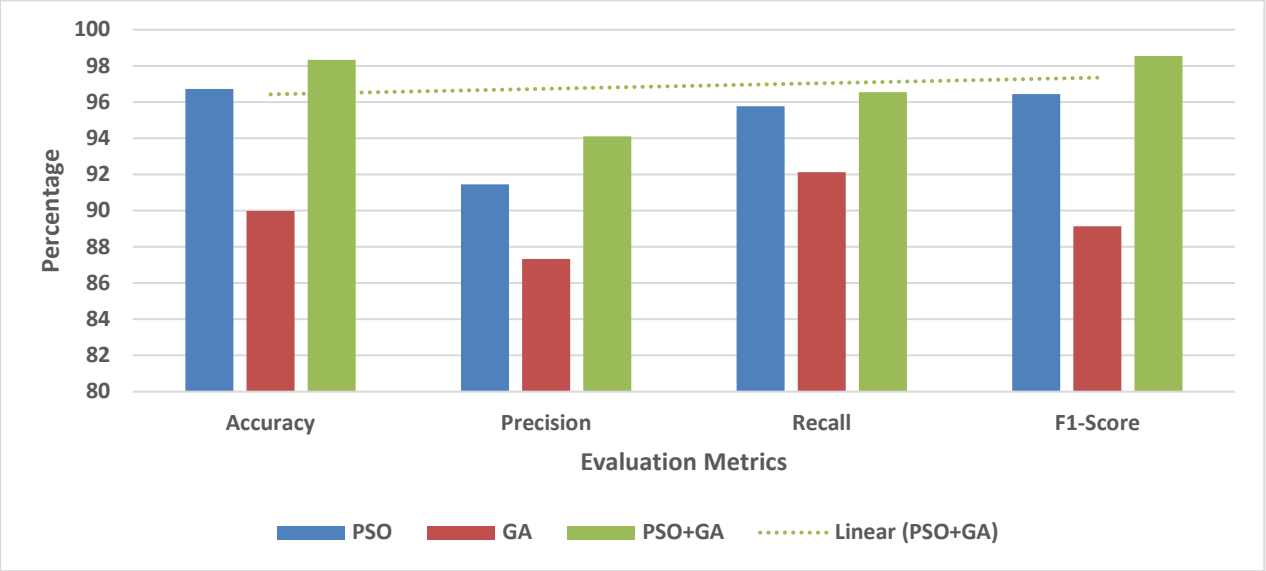

Fig 4: Illustration of Performance metrics for Enron Spam Dataset

In essence, the thorough analysis of the metrics for accuracy, precision, recall, and F1-score demonstrated that the PSO+GA hybrid technique was the best performer among the examined approaches. It has a high F1-score because of its ability to successfully balance precision and recall, which suggests that it has a lot of potential for detecting spam emails. The hybrid method is demonstrated to be superior than solo PSO and GA approaches in this in-depth examination, putting it forward as a strong contender for accurate and effective spam email classification

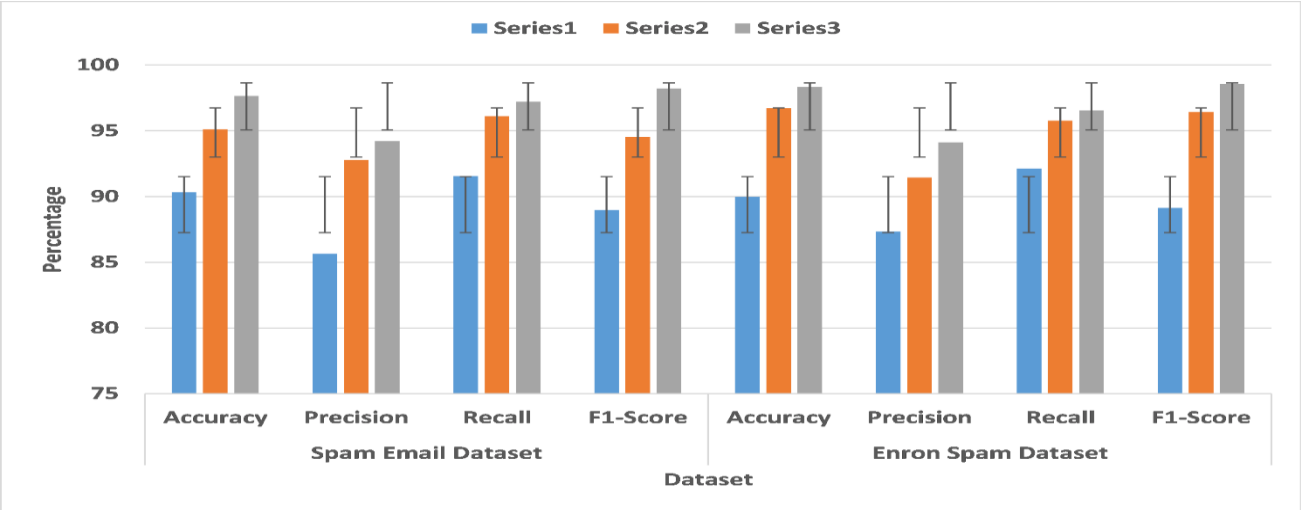

Fig 5: Comparison of performance metrics for both dataset using proposed method

A comparison of the performance metrics obtained by the suggested strategy when used on two distinct datasets is shown in Figure 5. The datasets under examination are represented by the x-axis, while the accuracy, precision, recall, and F1-score percentage values are shown on the y-axis.The suggested approach, which combines bio-inspired metaheuristic algorithms and advanced machine learning techniques, was assessed on two different datasets to gauge its adaptability and robustness to a variety of data sources.

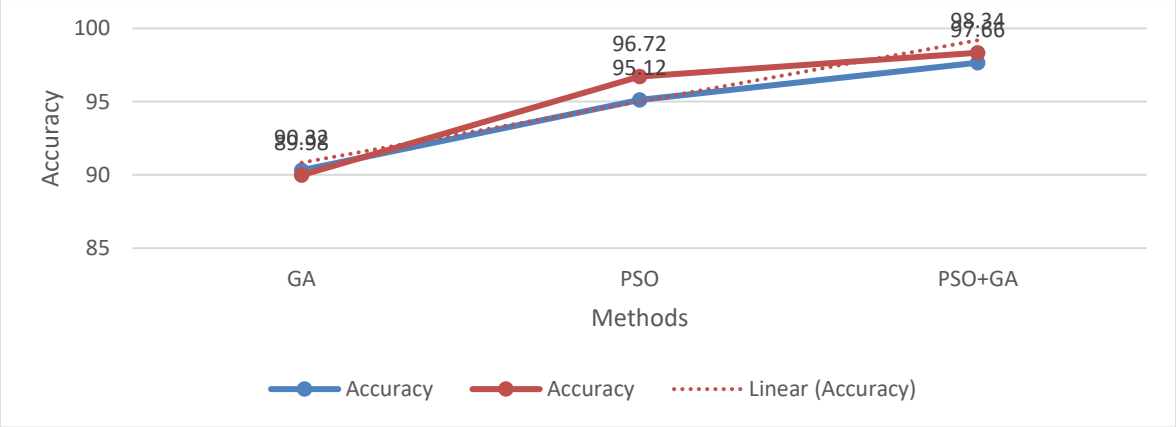

Fig 6: Accuracy comparison with different dataset

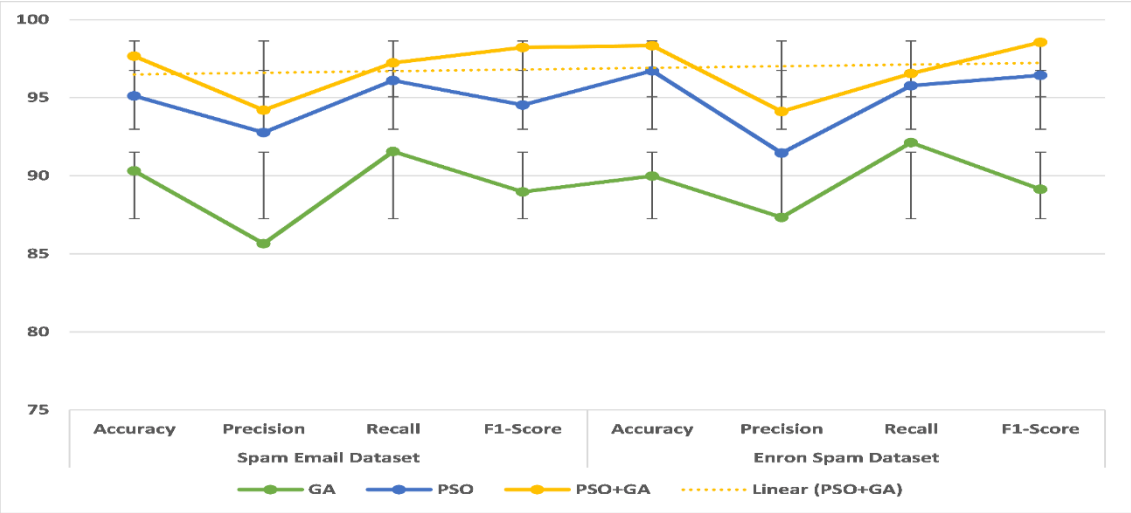

Fig 7: Evaluation metrics with standard deviation of Hybrid Model (PSO+GA)

A detailed comparison of the accuracy results obtained by a proposed method when used on various datasets is shown in Figure 6. The y-axis displays the percentage accuracy values reached by the approach, while the x-axis displays dataset variations.

Standard deviation is added to Figure 7's insightful representation of the assessment metrics for the hybrid model (PSO+GA). Accuracy, precision, recall, and F1-score are the specific assessment metrics represented by the x-axis, and the related values are shown along with the standard deviation as error bars on the y-axis.

VI. Conclusion

This study thoroughly assessed three different methods for detecting spam emails: Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and the hybrid PSO+GA methodology. Accuracy, precision, recall, and F1-score were used as important performance measures during the study. Results from the examination of several datasets consistently showed that the hybrid PSO+GA technique was superior across a wide range of measures.High accuracy (95.12%), together with great precision and recall values, were displayed by the PSO technique. Although the GA method's measurements were competitive, the F1-score was a little less evenly distributed. Notably, the hybrid PSO+GA method outperformed both PSO and GA consistently in every way. The hybrid technique successfully achieved accurate and fair spam email classification with impressive accuracy, precision, recall, and F1-score values.The visual representation of the comparative performance measures attests to the suggested method's scalability and effectiveness across various datasets. Furthermore, Figure 6 demonstrated accuracy trends with several datasets, emphasising the method's constant performance in many scenarios.Based on the thorough investigation, the

hybrid PSO+GA technique is found to be the most successful algorithm for spam email identification. Its potential as a reliable and accurate solution for spotting spam emails in a variety of scenarios is demonstrated by its consistently high accuracy, balanced precision and recall, and strong F1-score. This study highlights the effectiveness of merging cutting-edge machine learning methods with metaheuristics algorithms inspired by biological processes to tackle the always changing problem of spam email detection.

References

[1] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, Article ID 102419, 2020.

[2] N. Kumar and S. Sonowal, "Email spam detection using machine learning algorithms," in Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 108–113, Coimbatore, India, 2020.

[3] I. Santos, Y. K. Penya, J. Devesa, and P. G. Bringas, "N-grams-based file signatures for malware detection," ICEIS, vol. 9, no. 2, pp. 317–320, 2009.

[4] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, "On the capability of evolved spambots to evade detection via genetic engineering," Online Social Networks and Media, vol. 9, pp. 1–16, 2019.

[5] A. J. Saleh, A. Karim, B. Shanmugam et al., "An intelligent spam detection model based on artificial immune system," Information, vol. 10, no. 6, p. 209, 2019.

[6] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: a review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol. 160, pp. 3–24, 2007.

[7] S. Ajani and M. Wanjari, "An Efficient Approach for Clustering Uncertain Data Mining Based on Hash Indexing and Voronoi Clustering," 2013 5th International Conference and Computational Intelligence and Communication Networks, 2013, pp. 486-490, doi: 10.1109/CICN.2013.106.

[8] Khetani, V. ., Gandhi, Y. ., Bhattacharya, S. ., Ajani, S. N. ., & Limkar, S. . (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253–262. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2951

[9] E. Blanzieri and A. Bryl, E-mail Spam Filtering with Local SVM Classifiers, University of Trento, Trento, Italy, 2008.

[10] H. Bhuiyan, A. Ashiquzzaman, T. Islam Juthi, S. Biswas, and J. Ara, "A survey of existing e-mail spam filtering methods considering machine learning techniques," Global Journal of Computer Science and Technology, vol. 18, 2018.

[11] A. Asuncion and D. Newman, "UCI machine learning repository," 2007, https://archive.ics.uci.edu/ml/index.php.

[12] T. Vyas, P. Prajapati, and S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," in Proceedings of the 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT), IEEE, Tamil Nadu, India, March 2015.

[13] G. Jain, M. Sharma, and B. Agarwal, "Optimizing semantic lstm for spam detection," International Journal of Information Technology, vol. 11, no. 2, pp. 239–250, 2019.

[14] F. Masood, G. Ammad, A. Almogren et al., "Spammer detection and fake user identification on social networks," IEEE Access, vol. 7, pp. 68140–68152, 2019.

[15] A. Akhtar, G. R. Tahir, and K. Shakeel, "A mechanism to detect Urdu spam emails," in Proceedings of the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp. 168–172, IEEE, New York, NY, USA, Oct 2017.

[16] H. Drucker, D. Donghui Wu, and V. N. Vapnik, "Support vector machines for spam categorization," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1048–1054, 1999.

[17] H. Afzal and K. Mehmood, "Spam filtering of bi-lingual tweets using machine learning," in Proceedings of the 2016 18th

International Conference on Advanced Communication Technology (ICACT), pp. 710–714, IEEE, PyeongChang, Korea (South), Feb 2016.

[18] S. K. Tuteja and N. Bogiri, “Email spam filtering using bpnn classification algorithm,” in Proceedings of the 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pp. 915–919, IEEE, Pune, India, Sep 2016.

[19] M. Mohamad and A. Selamat, “An evaluation on the efficiency of hybrid feature selection in spam email classification,” in Proceedings of the 2015 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 227–231, IEEE, Kuching, Malaysia, Apr 2015.

[20] P. Sharma, U. Bhardwaj, and U. Bhardwaj, “Machine learning based spam e-mail detection,” International Journal of Intelligent Engineering and Systems, vol. 11, no. 3, pp. 1–10, 2018.

[21] S. Suryawanshi, A. Goswami, and P. Patil, “Email spam detection: an empirical comparative study of different ml and ensemble classifiers,” in Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 69–74, IEEE, Tiruchirappalli, India, Dec 2019.

[22] K. Agarwal and T. Kumar, “Email spam detection using integrated approach of naïve bayes and particle swarm optimization,” in Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 685–690, IEEE, Madurai, India, June 2018.

[23] A. Iyengar, G. Kalpana, S. Kalyankumar, and S. GunaNandhini, “Integrated spam detection for multilingual emails,” in Proceedings of the 2017 International Conference on Information Communication and Embedded Systems (ICICES), pp. 1–4, IEEE, Chennai, India, February 2017.

[24] K. Kandasamy and P. Koroth, “An integrated approach to spam classification on twitter using url analysis, natural language processing and machine learning techniques,” in Proceedings of the 2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science, pp. 1–5, IEEE, Bhopal, India, March 2014.

[25] X.-l. Chen, P.-y. Liu, Z.-f. Zhu, and Y. Qiu, “A method of spam filtering based on weighted support vector machines,” in Proceedings of the 2009 IEEE International Symposium on IT in Medicine & Education, vol. 1, pp. 947–950, IEEE, Jinan, China, Aug 2009.

[26] H. Kaur and A. Sharma, “Improved email spam classification method using integrated particle swarm optimization and decision tree,” in Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), pp. 516–521, IEEE, Dehradun, India, Oct 2016.

[27] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, “A comprehensive survey for intelligent spam email detection,” IEEE Access, vol. 7, pp. 168261–168295, 2019.

[28] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, An Evaluation of Naive Bayesian Anti-spam Filtering, 2000, arXiv preprint cs/0006013

[29] N. G. M. Jameel and L. E. George, “Detection of phishing emails using feed forward neural network,” International Journal of Computer Applications, vol. 77, no. 7, 2013.

[30] S. L. Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for arabic text classification", J. King Saud Univ.-Comput. Inf. Sci., vol. 32, no. 3, pp. 320-328, Mar. 2020.

[31] E. A. Natarajan, S. Subramanian and K. Premalatha, "An enhanced cuckoo search for optimization of bloom filter in spam filtering", Global J. Comput. Sci. Technol., vol. 12, no. 1, pp. 75-81, Jan. 2012

[32] A. Géron, Hands-On Machine Learning With Scikit-Learn Keras and TensorFlow, Newton, MA, USA:O’Reilly Media, 2019.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine learning in Python", J. Mach. Learn. Res., vol. 12, pp. 2825-2830, Oct. 2011

[34] G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between multinomial and Bernoulli Naïve Bayes for text classification", Proc.

Int. Conf. Autom. Comput. Technol. Manage. (ICACTM), pp. 593-596, Apr. 2019.

[35] N. Rusland, N. Wahid, S. Kasim and H. Hafit, "Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets", Proc. IOP Conf. Ser. Mater. Sci. Eng., vol. 226, 2017.

[36] F. Temitayo, O. Stephen and A. Abimbola, "Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification", Comput. Eng. Intell. Syst., vol. 3, no. 3, pp. 17-28, 2012

[37] M, V. ., P U, P. M. ., M, T. ., & Lopez, D. . (2023). XDLX: A Memory-Efficient Solution for Backtracking Applications in Big Data Environment using XOR-based Dancing Links. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 88–94. https://doi.org/10.17762/ijritcc.v11i1.6054

[38] Andrew Hernandez, Stephen Wright, Yosef Ben-David, Rodrigo Costa,. Enhancing Decision Support Systems through Machine Learning Algorithms. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/194

[39] Timande, S., Dhabliya, D. Designing multi-cloud server for scalable and secure sharing over web (2019) International Journal of Psychosocial Rehabilitation, 23 (5), pp. 835-841.