Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

Rank based RELIEFG Method for Spam Mail Detection and Classification

B. Aruna Kumari[1], C. Nagaraju[2]
[1]Research Scholar, YSR Engineering College of Yogi Vemana University, Computer Science & Engineering, Proddatur, arunakumarib1421@gmail.com
[2]Professor, YSR Engineering College of Yogi Vemana University, Computer Science & Engineering, Proddatur, nagaraju.c@yvu.edu.in

Article History:
Received: 12-01-2025
Revised: 15-02-2025
Accepted: 01-03-2025

Abstract:

Email plays a crucial role in day-to-day communication. An increase in spam emails poses a significant threat, leading to financial and resource losses. It's not easy to differentiate spam emails from legitimate emails due to large no.of features in datasets. Developing effective spam detection method is crucial for email security with missing values, outliers and noise. In the literature many methods have been existed among those rank-based Relief method is suitable for detection and classification. However, the relief method considers one hit and one miss. Due to this reason, it fails for multiclass classification. Due to the importance of Relief method, researchers have made many extensions to Relief. Among those ReliefF is the one of the latest variants of Relief method which improves accuracy by reducing the computational cost in the presence of missing values and noise for multiclass classifications. However, it fails in presentce of outliers. In this paper, a new variant ReliefG is proposed. This method reduces the computational cost by finding the highest rank of attributes and generating crisp set values by integrating efficient fuzzy membership function into a sigmoid function to improve the accuracy in the presence of outliers.

Keywords: Crisp set, Rank-based methods, Relief, Iterative Relief, ReliefF.

1. Introduction

Communication has undergone significant evolution in recent years, presenting various methods for individuals to share ideas and thoughts. From traditional in-person conversations, phone calls, and letters to modern online communication, methods have become more diverse and cost – effective. The widespread use of mobile phones has also drawn the attention of cyber-criminals, who misuse technology for malicious intent [1]. An often-used strategy involves sending unsolicited electronic messages, commonly known as spam. Attackers can use these messages to deceive recipients and lead them to respond and potentially face additional charges or simply promote offers and products. Spam messages are sent as text, images, and multimedia through mobile text messages, emails, and social media. Email or electronic mail, is an essential and widely used communication tool on the internet. Harness the power to exchange messages with anyone possessing an email address, no matter their location across the globe. Email relies on various protocols within the TCP/IP suite and has long been a vital medium of communication, offering near instant global connectivity. Email spam is a major issue in today's digital age, creating challenges for organizations, businesses and individuals [2]. Spam emails are unwanted messages that fill up inboxes, wasting precious time and resources, all while risking the exposure of users to harmful content or scams. Machine learning techniques have become

powerful tools for detecting email spam, addressing this issue. Email spam detection aims to classify incoming emails as legitimate or spam. Traditional rule-based methods are less effective due to evolving spam. Machine learning offers a more adaptable approach using patterns and features from large email datasets.

2. Literature Survey

Author, Year	Aim	Method	Results	Future Scope
Verónica Bolón-Canedo, et.al., [3], 2012	To assess feature selection methods with increasing irrelevant features, data noise, and attribute interactions.	Correlation-based Feature Selection, INTERACT algorithm, Consistency-based Filter, ReliefF, Information gain. SVM – RFE, FS-P Embedded methods and WrapperSubsetEval Methods are used and compared.	ReliefF filters offer independence from the induction algorithm and outpace embedded and wrapper methods in terms of speed.	In future work, this procedure could be further developed to tackle regression problems or be utilized for image analysis.
Girish Chandrashekar, Ferat Sahin, [4], 2014	To offer a general introduction to variable elimination that can be used for a wide range of machine learning problems.	Filter methods, Wrapper methods, Embedded methods are used and compared.	Feature selection provides benefits such as gaining insight into the data, improving classifier models, enhancing generalization, and identifying irrelevant variables.	Future research must prioritize hybrid methods that combine multiple feature selection methods to maximize accuracy, computational efficiency, and stability. This approach is essential for overcoming individual weaknesses through the combination of algorithms.
Newton Spolaor, et.al., [5], 2012.	By evaluating each feature separately, we can gain a comprehensive understanding of their	RF and IG-ML are used	The two feature selection methods were assessed across ten benchmark datasets, demonstrating	In the future, it will be possible to expand the concepts of ReliefF for multi-label data.

	individual impact.		highly promising results.	
Hidayet Tacki, Fatema Nusrat [6], 2023.	To introduce a machine learning – based model that will filters spam emails.	C4.5, ID3, RndTree, C-SVC, Naïve Bayes algorithms are used. Various feature selection algorithms like Forward selection, ReliefF and Fisher Filtering was used and compared.	In the comparison of studies for accuracy, the C-SVC algorithm proved to be the most successful. It achieved the highest correct recognition rate of 93.13%, outperforming the RndTree algorithm. In experiments involving feature selection, the forward selection method yielded superior results. Additionally, binary transformation demonstrated significantly better success in data transformation.	Forward selection produced good results; however, future research should explore advanced feature selection methods like RFE, L1 regularization (Lasso), or genetic algorithms to optimize processing time and accuracy, especially with larger datasets.
Panem Charanarur, et.al., [7], 2023.	The focal point of this project is to significantly enhance the efficiency of	KNN, Naïve Bayes, ETC, Random Forest, SVC, AdaBoost, xgboost, LR, GBDT, BgC,	The KN and Naive Bayes algorithms outperform than other algorithms, by	Consider deep learning models like RNNs or Transformers, especially pre-trained ones like BERT, for

	spam email detection and management.	Decision Tree are used and compared.	achieving the highest accuracy and precision scores.	more accurate email spam detection.
Dr. Purva Mange, Aditi Lule, Rohini Savant, [8], 2023.	The aim is to detect and mitigate spam emails precisely.	PSO (Particle Swarm Optimisation), Genetic algorithm and the hybrid Genetic algorithm + PSO methods was used and compared.	The hybrid model produced lower false positive rates and superior accuracy compared to conventional spam detection techniques.	In the future, this hybrid model could be expanded to cover additional domains such as phishing attempts and fraudulent transactions.

3. Relief

The original Relief algorithm is based on the principles of instance-based learning, making it a powerful and effective approach. Relief assesses each feature by calculating a proxy statistic, making it an effective method for individual feature evaluation and selection. This statistic is crucial for determining the feature's 'quality' or 'relevance' to the target concept [9]. Remember, these statistics represent feature weights or feature scores, which range from worst to best i.e., from -1 to +1. The Relief algorithm could only handle binary classification problems and did not account for missing data. Due to this, it was limited in addressing real-world datasets with missing values or multiclass classification problems. Updating the weight vector by using following formula:

$$W_i = W_i - (x_i - nH_i)^2 + (x_i - nM_i)^2 \quad \text{Eq (1)}$$

In the above equation, i represents the components, nH denotes nearHit, and nM specifies nearMiss

3.1 Variants of Relief

The extended version, Relief – A, averages the contribution of k nearest hits/misses for the random instance. So that noise and redundancy in the data are reduced, leading to improved accuracy. The Relief – B method is an extension to Relief -A, it can handle incomplete data, specifically missing values. This method is limited to handling binary class problems. It is designed to manage data with missing values for a particular attribute of instances. When calculating the diff function, it does not take into account instances with missing values. The Relief-C method is same as Relief-B. It can handle data with missing values for a given attribute of instances by ignoring the instances with missing values for evaluation. This method can also handle two-class problems. The Relief – D is an extension to Relief – C method and it evaluates neighbours only if their attribute values differ from those of a random instance. Three innovative strategies proposed to tackle incomplete data in Relief such as

ReliefB, ReliefC, and ReliefD. When missing value is encountered, Relief – D sets the "diff" function to the class-conditional probability of different feature values between two instances, but these three strategies are limited to binary class problems. To overcome this limitation, the Relief – E method was proposed to handle multi-class endpoints. It employs membership generalization over multi-class neighbours and randomly selects 'k' hits and 'k' misses for each instance. Recently two more powerful methods iterative relief and ReliefF extended and which are described below.

3.2 Iterative relief

An iterative approach for Relief, denoted as Iterative Relief. It is an extension to relief and it's variant algorithms. At the outset, all features are assigned equal weights. Any negative feature weights are treated as outliers and are rounded up to zero. The algorithm iterates until the weights converge. Iterative Relief requires multiple iterations for convergence, resulting in significantly higher computational costs are increased when dealing with large datasets. So to overcome this drawback than ReliefF method is derived.

3.3 ReliefF

ReliefF is a significant improvement of the original Relief and iterative relief algorithms. ReliefF selects multiple neighbors to enhance robustness and identifies discriminative features between spam and non-spam emails [10]. ReliefF seeks the k nearest hits and misses, averaging their impact on feature weights. The value of k can be adjusted to suit each specific problem. The weight measure of a feature 'A' is represented as W[A] as follows:

$$\mathrm{w[A]} = \mathrm{w[A]} - \frac{1}{\mathrm{m-k}}\sum_{j=1}^{k}(\,\mathrm{diff}\,(\mathrm{A}, \mathrm{x_i}\,, \mathrm{Hit_j}\,)\,) + \frac{1}{\mathrm{m-k}}\sum_{j=1}^{k}\left(\frac{P(C) - diff\left(A, \mathrm{x_i}, \mathrm{Miss_j}, c\right)}{1-P(C)}\right) \qquad \text{Eq (2)}$$

In Eq(2), diff (A, x, $\mathrm{Hit_j}$) denotes distance function calculating the difference between the values of feature A with hit or miss operations. The j-th nearest hit is represented by $\mathrm{Hit_j}$, j-th nearest miss from class C is denoted by $\mathrm{Miss}_{j,\,c}$ and prior probability of class C is shown by P(C).

ReliefF's effectiveness is compromised when dealing with highly correlated or redundant features. ReliefF may not perform well with outliers. ReliefF method handles multiclass classification, multi label classification very effectively, when the number of attributes is very low (for example, just one or two), the algorithm may struggle to identify meaningful patterns because it lacks sufficient discriminative power. On the other hand, when the number of attributes is very high, ReliefF can become computationally expensive and less reliable due to the curse of dimensionality, which may lead to inaccurate results. So, in this paper, we are proposing a new variant i.e., ReliefG which will work effectively with large datasets as well as outliers.

4. Proposed Method

The ReliefG is proposed. The no. of hits and no.of misses are automatically selected by using fuzzy membership function. The crisp set values are evaluated by integrating trapezoidal membership function into sigmoid function such that we can eliminate the outliers. Because the sigmoid function can work by enhancing lower values and reducing the highest value to lower value. Simultaneously it reduces the computational costs. It gives the highest accurate results evem in presence of more attributes and it automatically selects the k value.

4.1 Sigmoid function and its characteristics

The sigmoid membership function is an essential tool for smoothly transitioning between different degrees of membership. Its versatile nature makes it a valuable asset in various applications.

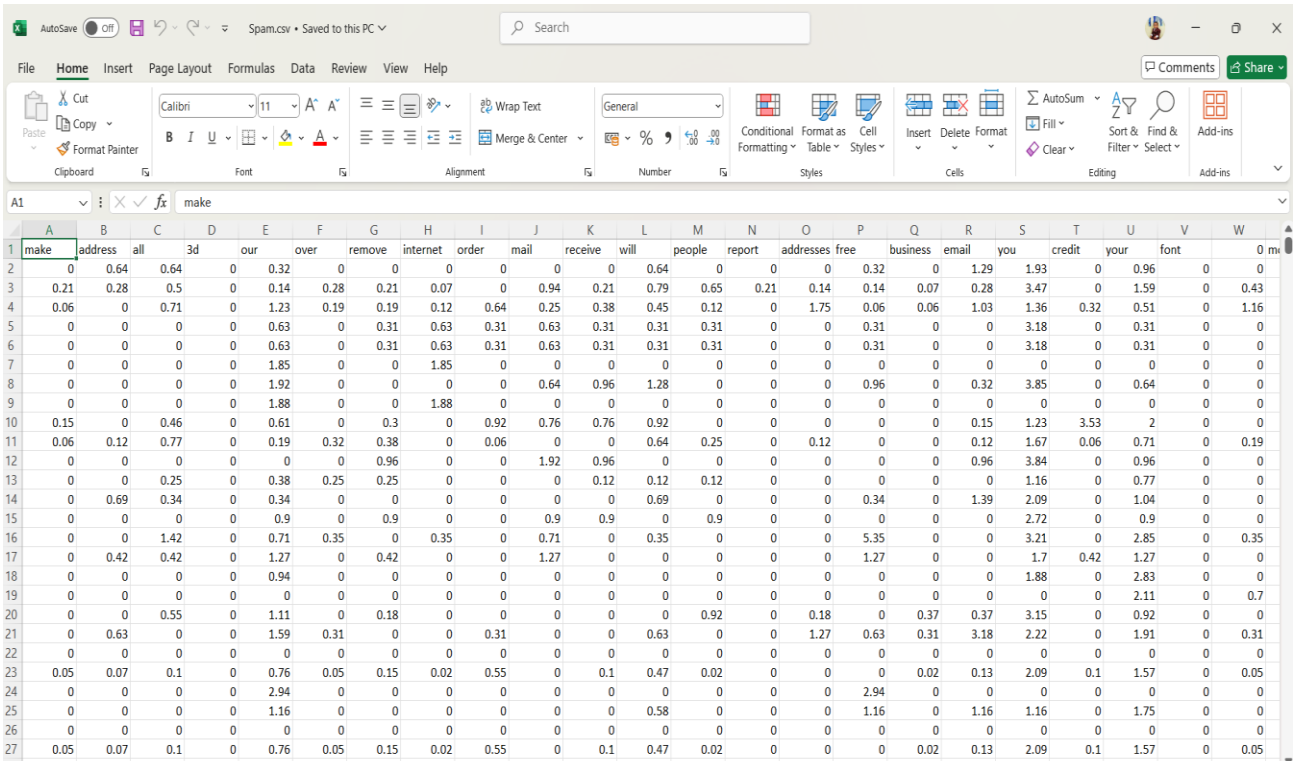
$$\mu(x) = \frac{1}{1+e^{\wedge}(-a(x-c))} \qquad \text{Eq (3)}$$

In Eq (3), the $\mu(x)$ represents the membership degree of x in the fuzzy set, ranging from 0 to 1. Here, x is the input variable, 'a' controls the slope of the function, and c is the threshold value.

The sigmoid function creates an 'S' curve that moves from 0 to 1. The slope of the curve is controlled by the parameter "a". When a > 0, the function increases from 0 to 1 as x increases; when a < 0, the function decreases from 1 to 0 as x increases. The sigmoid function's non-linear nature empowers it to adeptly navigate intricate relationships between variables within fuzzy systems. To effectively eliminate outliers, we are integrating the power of the sigmoid function with the trapezoidal function. This innovative approach will ensure a more robust and accurate analysis.

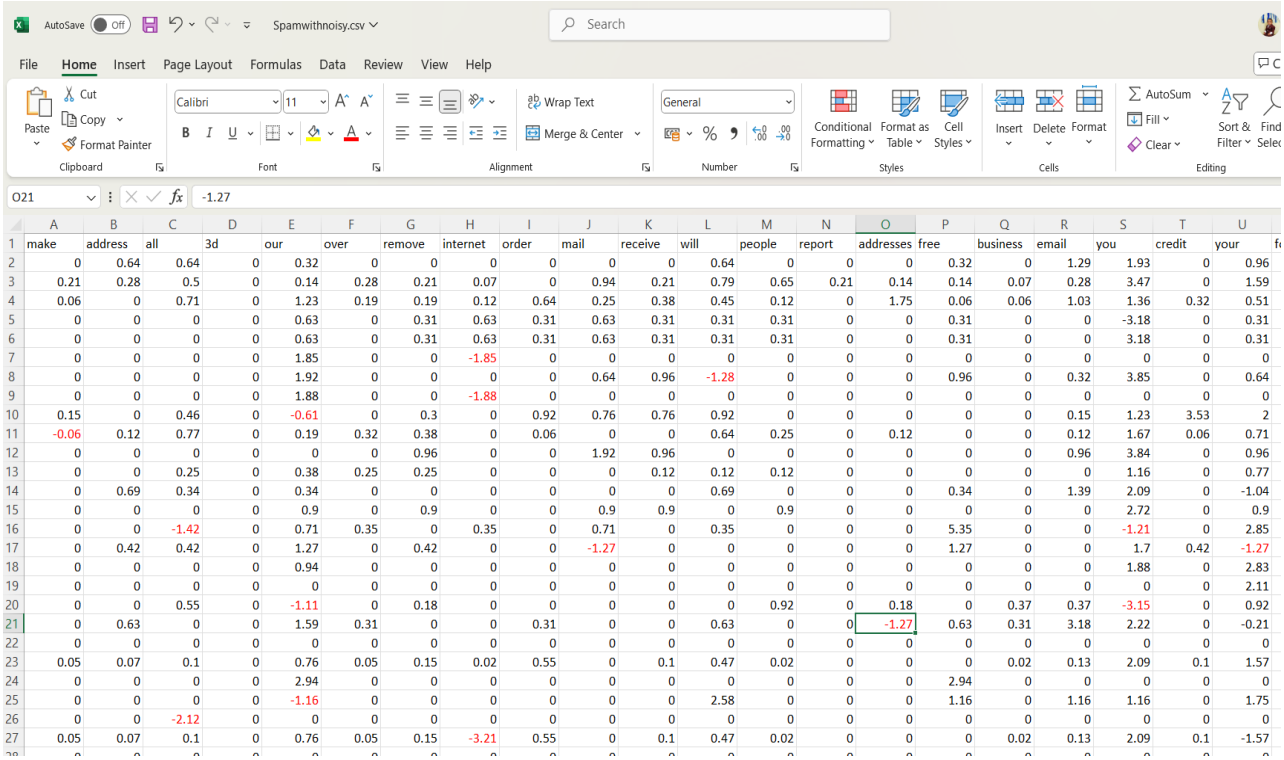
5. Dataset Description

The SPAM dataset consists of 1813 spam mails and 2788 ham emails. Among 4601 instances, there are 57 attributes, consisting of frequently occurring words or characters. The first 48 attributes are continuous real numbers, ranging from 0 to 100, labeled as "WORD." This indicates the percentage of words present in the email. The next six attributes are continuous real numbers, ranging from 0 to 100, labeled as "char_freq_CHAR," indicating the percentage of characters in the email. The remaining attributes include a mix of continuous real numbers and integers. The final column indicates whether the email is spam (denoted as "1") or ham (denoted as "0").



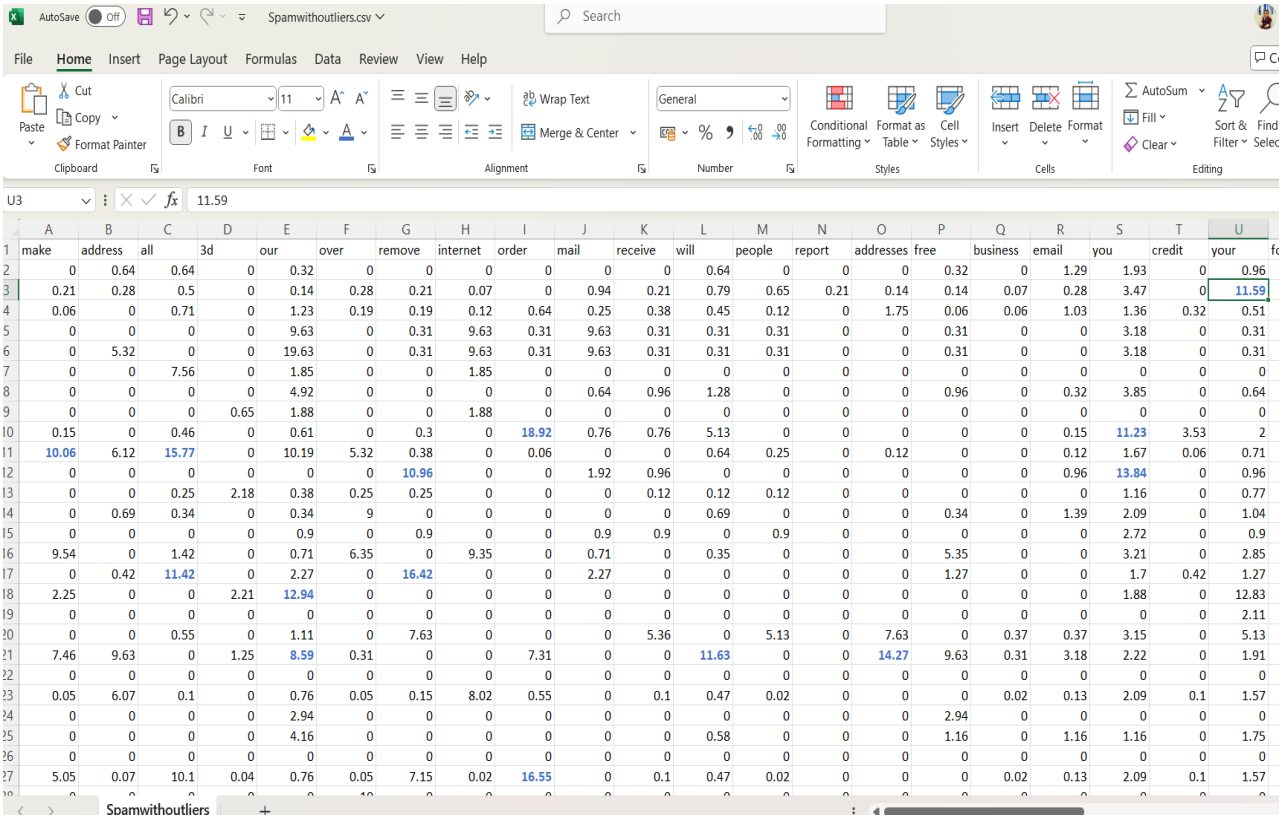
make	address	all	3d	our	over	remove	internet	order	mail	receive	will	people	report	addresses	free	business	email	you	credit	your	font	0
0	0.64	0.64	0	0.32	0	0	0	0	0	0	0.64	0	0	0	0.32	0	1.29	1.93	0	0.96	0	0
0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21	0.79	0.65	0.21	0.14	0.14	0.07	0.28	3.47	0	1.59	0	0.43
0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0	1.75	0.06	0.06	1.03	1.36	0.32	0.51	0	1.16
0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	0.31	0	0	3.18	0	0.31	0	0
0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	0.31	0	0	3.18	0	0.31	0	0
0	0	0	0	1.85	0	0	1.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1.92	0	0	0	0	0.64	0.96	1.28	0	0	0	0.96	0	0.32	3.85	0	0.64	0	0
0	0	0	0	1.88	0	0	1.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.15	0	0.46	0	0.61	0	0.3	0	0.92	0.76	0.76	0.92	0	0	0	0	0	0.15	1.23	3.53	2	0	0
0.06	0.12	0.77	0	0.19	0.32	0.38	0	0.06	0	0	0.64	0.25	0	0.12	0	0	0.12	1.67	0.06	0.71	0	0.19
0	0	0	0	0	0	0.96	0	0	1.92	0.96	0	0	0	0	0	0	0.96	3.84	0	0.96	0	0
0	0	0.25	0	0.38	0.25	0.25	0	0	0	0.12	0.12	0.12	0	0	0	0	0	1.16	0	0.77	0	0
0	0.69	0.34	0	0.34	0	0	0	0	0	0	0.69	0	0	0	0.34	0	1.39	2.09	0	1.04	0	0
0	0	0	0	0.9	0	0.9	0	0	0.9	0.9	0	0.9	0	0	0	0	0	2.72	0	0.9	0	0
0	0	1.42	0	0.71	0.35	0	0.35	0	0.71	0	0.35	0	0	0	5.35	0	0	3.21	0	2.85	0	0.35
0	0.42	0.42	0	1.27	0	0.42	0	0	1.27	0	0	0	0	0	1.27	0	0	1.7	0.42	1.27	0	0
0	0	0	0	0.94	0	0	0	0	0	0	0	0	0	0	0	0	0	1.88	0	2.83	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.11	0	0.7
0	0	0.55	0	1.11	0	0.18	0	0	0	0	0	0.92	0	0.18	0	0.37	0.37	3.15	0	0.92	0	0
0	0.63	0	0	1.59	0.31	0	0	0.31	0	0	0.63	0	0	1.27	0.63	0.31	3.18	2.22	0	1.91	0	0.31
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.05	0.07	0.1	0	0.76	0.05	0.15	0.02	0.55	0	0.1	0.47	0.02	0	0	0	0.02	0.13	2.09	0.1	1.57	0	0.05
0	0	0	0	2.94	0	0	0	0	0	0	0	0	0	0	2.94	0	0	0	0	0	0	0
0	0	0	0	1.16	0	0	0	0	0	0	0.58	0	0	0	1.16	0	1.16	1.16	0	1.75	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.05	0.07	0.1	0	0.76	0.05	0.15	0.02	0.55	0	0.1	0.47	0.02	0	0	0	0.02	0.13	2.09	0.1	1.57	0	0.05

Fig 1. Actual SPAM dataset



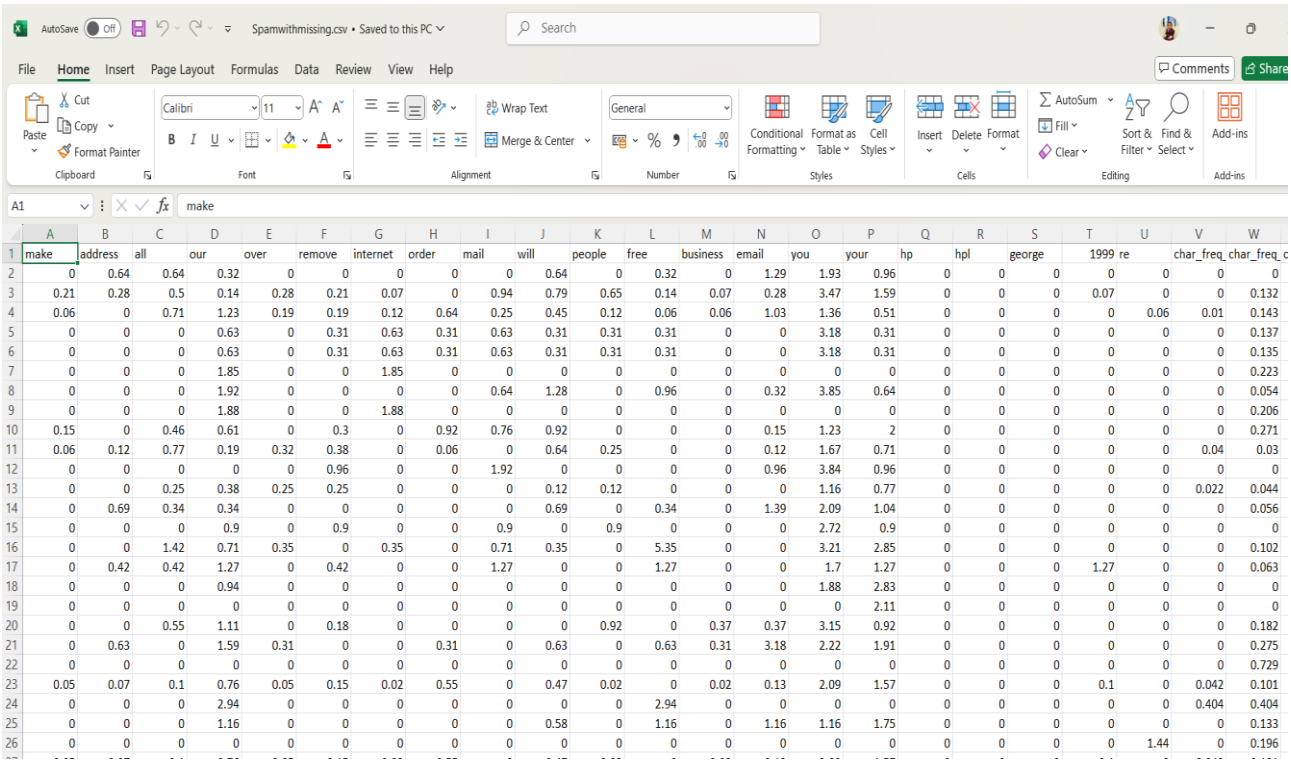
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	make	address	all	3d	our	over	remove	internet	order	mail	receive	will	people	report	addresses	free	business	email	you	credit	your
2	0	0.64	0.64	0	0.32	0	0	0	0	0	0	0.64	0	0	0	0.32	0	1.29	1.93	0	0.96
3	0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21	0.79	0.65	0.21	0.14	0.14	0.07	0.28	3.47	0	1.59
4	0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0	1.75	0.06	0.06	1.03	1.36	0.32	0.51
5	0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	0.31	0	0	-3.18	0	0.31
6	0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	0.31	0	0	3.18	0	0.31
7	0	0	0	0	1.85	0	0	-1.85	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1.92	0	0	0	0	0.64	0.96	-1.28	0	0	0	0.96	0	0.32	3.85	0	0.64
9	0	0	0	0	1.88	0	0	-1.88	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0.15	0	0.46	0	-0.61	0	0.3	0	0.92	0.76	0.76	0.92	0	0	0	0	0	0.15	1.23	3.53	2
11	-0.06	0.12	0.77	0	0.19	0.32	0.38	0	0.06	0	0	0.64	0.25	0	0.12	0	0	0.12	1.67	0.06	0.71
12	0	0	0	0	0	0	0.96	0	0	1.92	0.96	0	0	0	0	0	0	0.96	3.84	0	0.96
13	0	0	0.25	0	0.38	0.25	0.25	0	0	0	0.12	0.12	0.12	0	0	0	0	0	1.16	0	0.77
14	0	0.69	0.34	0	0.34	0	0	0	0	0	0	0.69	0	0	0	0.34	0	1.39	2.09	0	-1.04
15	0	0	0	0	0.9	0	0.9	0	0	0.9	0.9	0	0.9	0	0	0	0	0	2.72	0	0.9
16	0	0	-1.42	0	0.71	0.35	0	0.35	0	0.71	0	0.35	0	0	0	5.35	0	0	-1.21	0	2.85
17	0	0.42	0.42	0	1.27	0	0.42	0	0	-1.27	0	0	0	0	0	1.27	0	0	1.7	0.42	-1.27
18	0	0	0	0	0.94	0	0	0	0	0	0	0	0	0	0	0	0	0	1.88	0	2.83
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.11
20	0	0	0.55	0	-1.11	0	0.18	0	0	0	0	0	0.92	0	0.18	0	0.37	0.37	-3.15	0	0.92
21	0	0.63	0	0	1.59	0.31	0	0	0.31	0	0	0.63	0	0	-1.27	0.63	0.31	3.18	2.22	0	-0.21
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0.05	0.07	0.1	0	0.76	0.05	0.15	0.02	0.55	0	0.1	0.47	0.02	0	0	0	0.02	0.13	2.09	0.1	1.57
24	0	0	0	0	2.94	0	0	0	0	0	0	0	0	0	0	2.94	0	0	0	0	0
25	0	0	0	0	-1.16	0	0	0	0	0	0	2.58	0	0	0	1.16	0	1.16	1.16	0	1.75
26	0	0	-2.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0.05	0.07	0.1	0	0.76	0.05	0.15	-3.21	0.55	0	0.1	0.47	0.02	0	0	0	0.02	0.13	2.09	0.1	-1.57

Fig2. SPAM dataset with noisy data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	make	address	all	3d	our	over	remove	internet	order	mail	receive	will	people	report	addresses	free	business	email	you	credit	your
2	0	0.64	0.64	0	0.32	0	0	0	0	0	0	0.64	0	0	0	0.32	0	1.29	1.93	0	0.96
3	0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21	0.79	0.65	0.21	0.14	0.14	0.07	0.28	3.47	0	11.59
4	0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0	1.75	0.06	0.06	1.03	1.36	0.32	0.51
5	0	0	0	0	9.63	0	0.31	9.63	0.31	9.63	0.31	0.31	0.31	0	0	0.31	0	0	3.18	0	0.31
6	0	5.32	0	0	19.63	0	0.31	9.63	0.31	9.63	0.31	0.31	0.31	0	0	0.31	0	0	3.18	0	0.31
7	0	0	7.56	0	1.85	0	0	1.85	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	4.92	0	0	0	0	0.64	0.96	1.28	0	0	0	0.96	0	0.32	3.85	0	0.64
9	0	0	0	0.65	1.88	0	0	1.88	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0.15	0	0.46	0	0.61	0	0.3	0	18.92	0.76	0.76	5.13	0	0	0	0	0	0.15	11.23	3.53	2
11	10.06	6.12	15.77	0	10.19	5.32	0.38	0	0.06	0	0	0.64	0.25	0	0.12	0	0	0.12	1.67	0.06	0.71
12	0	0	0	0	0	0	10.96	0	0	1.92	0.96	0	0	0	0	0	0	0.96	13.84	0	0.96
13	0	0	0.25	2.18	0.38	0.25	0.25	0	0	0	0.12	0.12	0.12	0	0	0	0	0	1.16	0	0.77
14	0	0.69	0.34	0	0.34	9	0	0	0	0	0	0.69	0	0	0	0.34	0	1.39	2.09	0	1.04
15	0	0	0	0	0.9	0	0.9	0	0	0.9	0.9	0	0.9	0	0	0	0	0	2.72	0	0.9
16	9.54	0	1.42	0	0.71	6.35	0	9.35	0	0.71	0	0.35	0	0	0	5.35	0	0	3.21	0	2.85
17	0	0.42	11.42	0	2.27	0	16.42	0	0	2.27	0	0	0	0	0	1.27	0	0	1.7	0.42	1.27
18	2.25	0	0	2.21	12.94	0	0	0	0	0	0	0	0	0	0	0	0	0	1.88	0	12.83
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.11
20	0	0	0.55	0	1.11	0	7.63	0	0	0	5.36	0	5.13	0	7.63	0	0.37	0.37	3.15	0	5.13
21	7.46	9.63	0	1.25	8.59	0.31	0	0	7.31	0	0	11.63	0	0	14.27	9.63	0.31	3.18	2.22	0	1.91
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0.05	6.07	0.1	0	0.76	0.05	0.15	8.02	0.55	0	0.1	0.47	0.02	0	0	0	0.02	0.13	2.09	0.1	1.57
24	0	0	0	0	2.94	0	0	0	0	0	0	0	0	0	0	2.94	0	0	0	0	0
25	0	0	0	0	4.16	0	0	0	0	0	0	0.58	0	0	0	1.16	0	1.16	1.16	0	1.75
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	5.05	0.07	10.1	0.04	0.76	0.05	7.15	0.02	16.55	0	0.1	0.47	0.02	0	0	0	0.02	0.13	2.09	0.1	1.57

Fig 3. SPAM dataset with outliers



make	address	all	our	over	remove	internet	order	mail	will	people	free	business	email	you	your	hp	hpl	george	1999	re	char_freq	char_freq_c
0	0.64	0.64	0.32	0	0	0	0	0	0.64	0	0.32	0	1.29	1.93	0.96	0	0	0	0	0	0	0
0.21	0.28	0.5	0.14	0.28	0.21	0.07	0	0.94	0.79	0.65	0.14	0.07	0.28	3.47	1.59	0	0	0	0.07	0	0	0.132
0.06	0	0.71	1.23	0.19	0.19	0.12	0.64	0.25	0.45	0.12	0.06	0.06	1.03	1.36	0.51	0	0	0	0	0.06	0.01	0.143
0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	3.18	0.31	0	0	0	0	0	0	0.137
0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	3.18	0.31	0	0	0	0	0	0	0.135
0	0	0	1.85	0	0	1.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.223
0	0	0	1.92	0	0	0	0	0.64	1.28	0	0.96	0	0.32	3.85	0.64	0	0	0	0	0	0	0.054
0	0	0	1.88	0	0	1.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.206
0.15	0	0.46	0.61	0	0.3	0	0.92	0.76	0.92	0	0	0	0.15	1.23	2	0	0	0	0	0	0	0.271
0.06	0.12	0.77	0.19	0.32	0.38	0	0.06	0	0.64	0.25	0	0	0.12	1.67	0.71	0	0	0	0	0	0.04	0.03
0	0	0	0	0	0.96	0	0	1.92	0	0	0	0	0.96	3.84	0.96	0	0	0	0	0	0	0
0	0	0.25	0.38	0.25	0.25	0	0	0	0.12	0.12	0	0	0	1.16	0.77	0	0	0	0	0	0.022	0.044
0	0.69	0.34	0.34	0	0	0	0	0	0.69	0	0.34	0	1.39	2.09	1.04	0	0	0	0	0	0	0.056
0	0	0	0.9	0	0.9	0	0	0.9	0	0.9	0	0	0	2.72	0.9	0	0	0	0	0	0	0
0	0	1.42	0.71	0.35	0	0.35	0	0.71	0.35	0	5.35	0	0	3.21	2.85	0	0	0	0	0	0	0.102
0	0.42	0.42	1.27	0	0.42	0	0	1.27	0	0	1.27	0	0	1.7	1.27	0	0	0	1.27	0	0	0.063
0	0	0	0.94	0	0	0	0	0	0	0	0	0	0	1.88	2.83	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.11	0	0	0	0	0	0	0
0	0	0.55	1.11	0	0.18	0	0	0	0	0.92	0	0.37	0.37	3.15	0.92	0	0	0	0	0	0	0.182
0	0.63	0	1.59	0.31	0	0	0.31	0	0.63	0	0.63	0.31	3.18	2.22	1.91	0	0	0	0	0	0	0.275
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.729
0.05	0.07	0.1	0.76	0.05	0.15	0.02	0.55	0	0.47	0.02	0	0.02	0.13	2.09	1.57	0	0	0	0.1	0	0.042	0.101
0	0	0	2.94	0	0	0	0	0	0	0	2.94	0	0	0	0	0	0	0	0	0	0.404	0.404
0	0	0	1.16	0	0	0	0	0	0.58	0	1.16	0	1.16	1.16	1.75	0	0	0	0	0	0	0.133
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.44	0	0.196

Fig 4. SPAM dataset with data preprocessing

For experimentation, here we considered SPAM dataset as shown in Fig 1, SPAM dataset with noisy data is represented in Fig 2. The SPAM dataset with outliers is denoted in Fig 3. Finally, the SPAM dataset with data preprocessing is shown in Fig 4.

6. Experimental Results

The proposed method has been implemented and tested on the SPAM dataset. The database contains 57 features utilized to differentiate between two types of emails, comprising specific words or characters that frequently appear in an email. In this paper, a confusion matrix is used to evaluate the method's performance. The matrix consists of four outcomes: TN, TP, FP, and FN [12] Statistical parameters such as precision, recall, accuracy, and f1-score are computed from the confusion matrix to analyze performance using tables and graphs.

- **Accuracy:** Accuracy specifies no.of accurate predictions were made regarding the total number of values.

$$\text{Accuracy} = \frac{\text{truepositive}+\text{truenegative}}{(\text{truepositive}+\text{falsepositive}+\text{truenegative}+\text{falsenegative})} \quad \text{Eq (4)}$$

- **Precision:** Precision is the key to accurately anticipating true positive outcomes.

$$\text{Precision} = \frac{\text{truepositive}}{(\text{truepositive}+\text{falsepositive})} \quad \text{Eq (5)}$$

- **Recall:** Utilized for obtaining accurate negative values.

$$\text{Recall} = \frac{\text{truenegative}}{(\text{truepositive}+\text{falsenegative})} \quad \text{Eq (6)}$$

- **F1 Score:** It is used for classifying dataset values as positive or negative.

$$F1 - score = \frac{2*Precision*Recall}{Precision* recall} \quad \text{Eq (7)}$$

Table 1. Accuracy

Datasets	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	94.17897	94.52649	94.72649	95.04778
data filling with mean	94.61337	94.09209	94.17897	94.9609
noisy data	93.83145	94.30026	94.71833	94.87402
data with outliers	94.0609	94.23961	94.46585	94.52649
data with missing values elimination	95.13466	95.12218	95.13466	95.44778

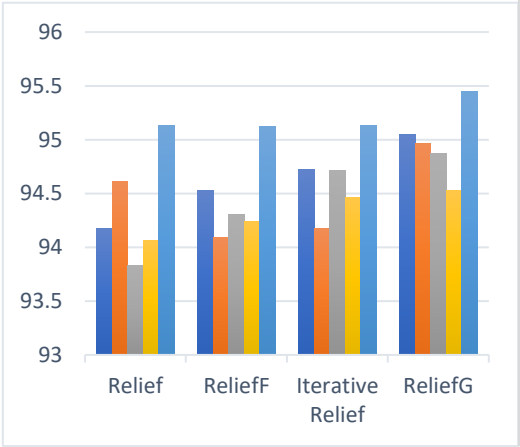


Fig 5. Accuracy

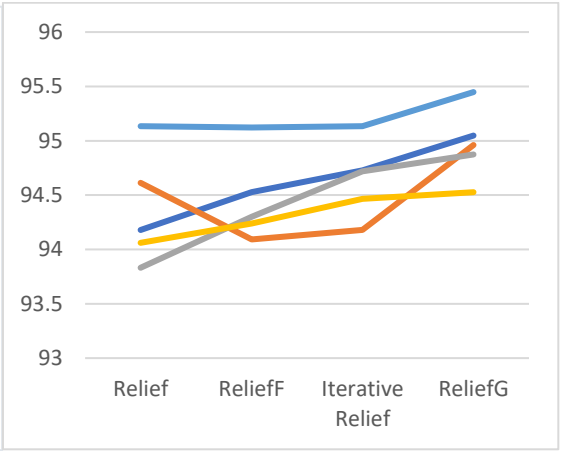


Fig 6. Accuracy

In above Fig 5 and Fig 6, thick blue bar & line represents accuracy of quality data for all variants of Relief. Orange colored bar & line denotes accuracy of data after preprocessing for all relief methods. Grey colored results shown that the accuracy of noisy data for relief and it's variants. Accuracy of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the accuracy of dataset with missing values elimination.

Table 2. Precision

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	94.54148	94.7826	94.88874	96.44444
data with Preprocessing	94.9891	94.14316	94.54148	96.43652
noisy data	94.69026	94.4206	93.92624	96.42857

data with outliers	94.85393	94.77124	94.55337	96.18834
data with missing values elimination	95.84245	95.87852	95.84245	95.95478

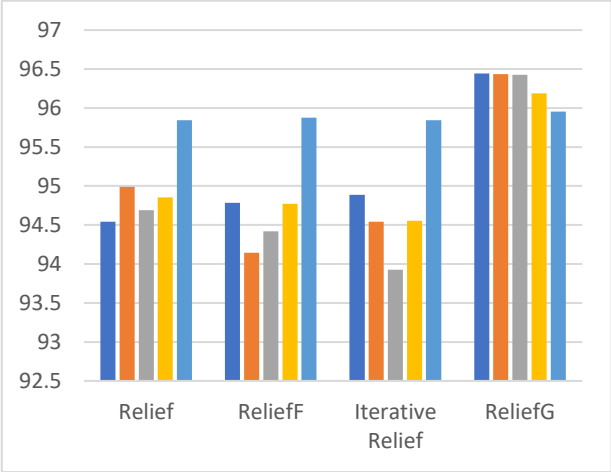


Fig 7. Precision

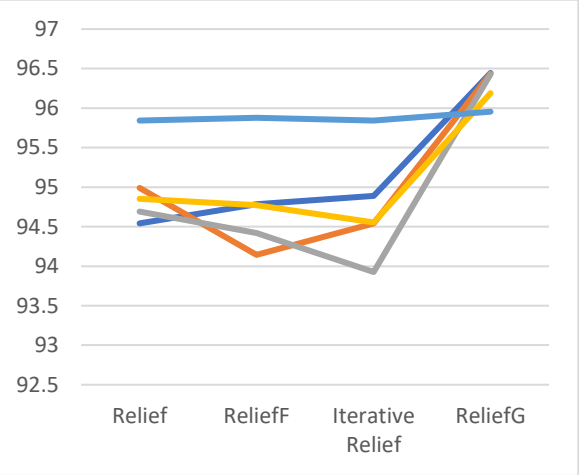


Fig 8. Precision

In above Fig 7 and Fig 8, thick blue bar & line represents precision values of quality data for all variants of Relief. Orange colored bar & line denotes precision values of data after preprocessing for all relief methods. Grey colored results shown that the precision values of noisy data for relief and it's variants. precision values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the precision values of dataset with missing values elimination.

Table 3. Recall

Type of data	**Relief**	**ReliefF**	**Iterative Relief**	**ReliefG**
Quality data	91.15789	91.78947	92	92.36842
data with Preprocessing	91.08947	91.36842	91.15789	92.15789
noisy data	90.10526	91.03157	91.15789	91.94736
data with outliers	90.73684	91.57894	91.36842	91.61578
data with missing values elimination	90.21052	91.05263	92.21052	92.42105

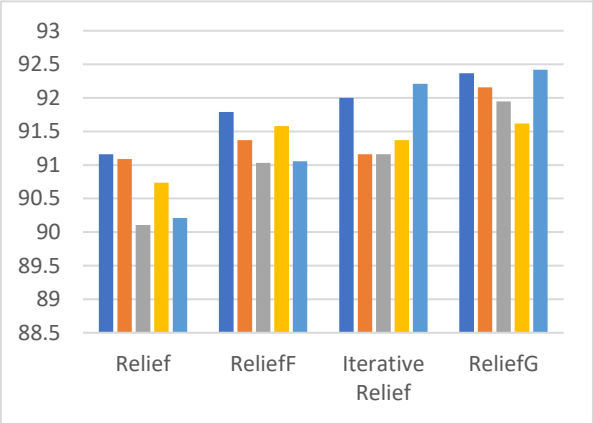


Fig 9. Recall

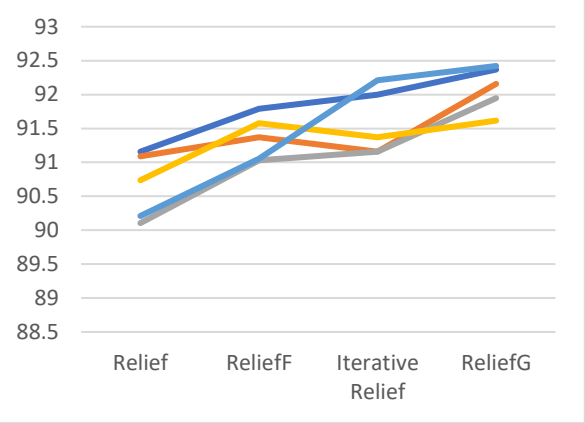


Fig 10. Recall

In above Fig 9 and Fig 10, thick blue bar & line represents recall values of quality data for all variants of Relief. Orange colored bar & line denotes recall values of data after preprocessing for all relief methods. Grey colored results shown that the recall values of noisy data for relief and it's variants. recall values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the recall values of dataset with missing values elimination.

Table 4. F1-Score

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	92.81886	93.26203	93.27641	93.83783
data with Preprocessing	93.36188	92.73504	92.81886	93.72294
noisy data	92.34088	93.51753	92.52136	93.6078
data with outliers	93.69565	93.14775	92.93361	93.1596
data with missing values elimination	93.99141	94.44444	93.99141	94.90374

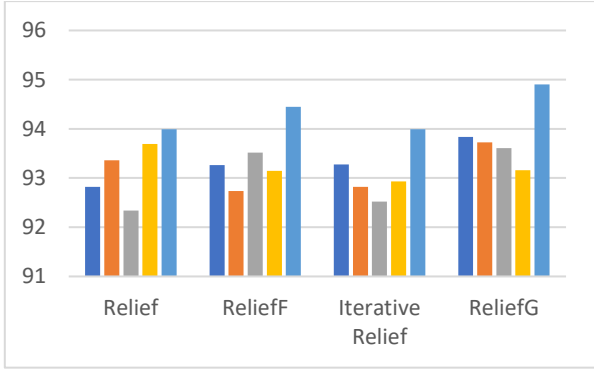


Fig 11. F1-Score

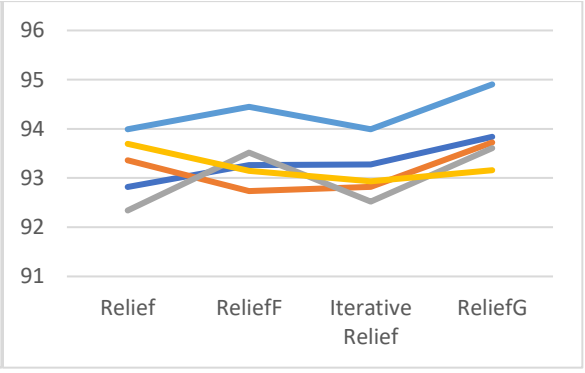


Fig 12. F1-Score

In above Fig 11 and Fig 12, thick blue bar & line represents F1-Score of quality data for all variants of Relief. Orange colored bar & line denotes F1-Score of data after preprocessing for all relief methods. Grey colored results shown that the F1-Score of noisy data for relief and it's variants. F1-Score of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the F1-Score of dataset with missing values elimination.

Table 5. Mean Absolute Error

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	0.058	0.054	0.054	0.049
data with Preprocessing	0.053	0.059	0.058	0.05
noisy data	0.061	0.052	0.06	0.051
data with outliers	0.05	0.055	0.057	0.054
data with missing values elimination	0.048	0.045	0.048	0.049

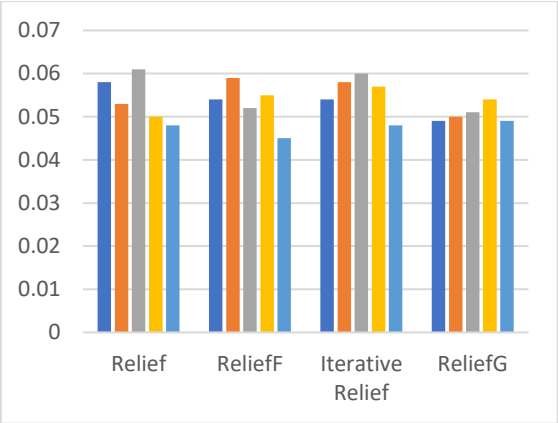


Fig 13. Mean Absolute Error

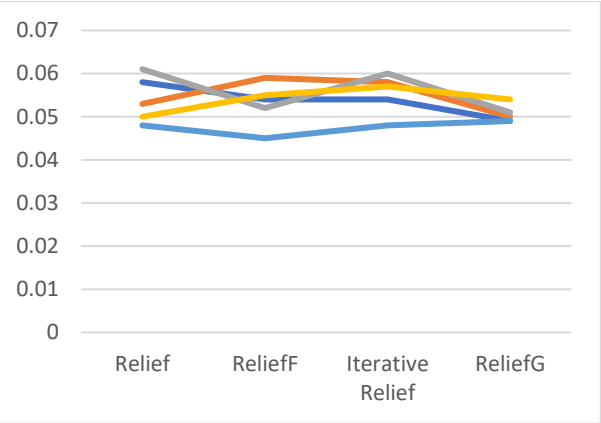


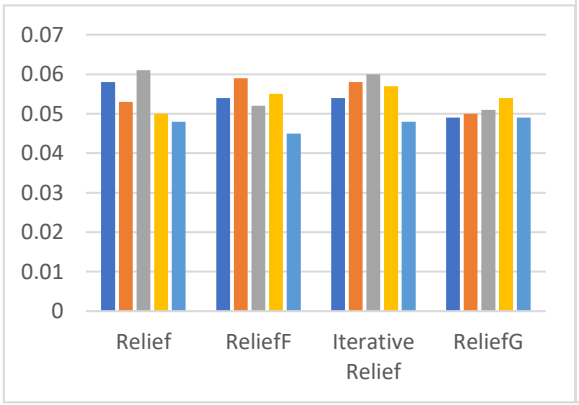
Fig 14. Mean Absolute Error

In above Fig 13 and Fig 14, thick blue bar & line represents MAE values of quality data for all variants of Relief. Orange colored bar & line denotes MAE values of data after preprocessing for all relief methods. Grey colored results shown that the MAE values of noisy data for relief and it's variants. MAE values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the MAE values of dataset with missing values elimination.

Table 6. Mean Squared Error

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	0.058	0.054	0.054	0.049

data with Preprocessing	0.053	0.059	0.058	0.05
noisy data	0.061	0.052	0.06	0.051
data with outliers	0.05	0.055	0.057	0.054
data with missing values elimination	0.048	0.045	0.048	0.049



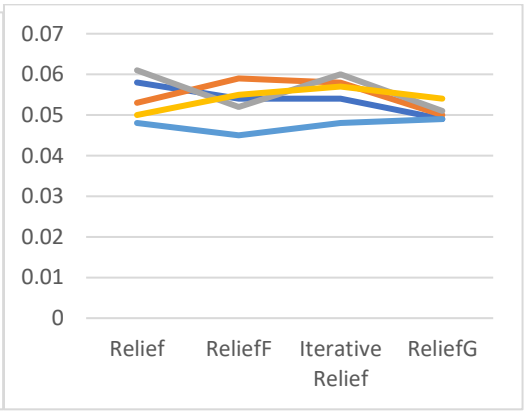
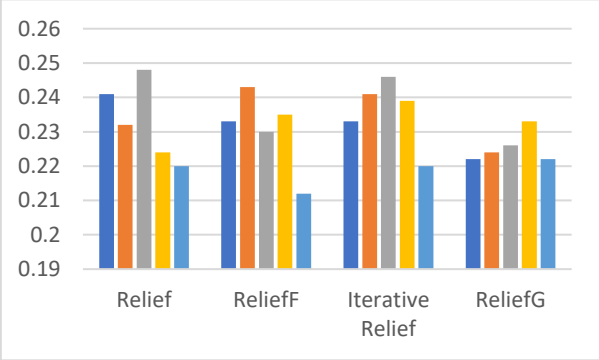


Fig 15. Mean Squared Error **Fig 16. Mean Squared Error**

In above Fig 15 and Fig 16, thick blue bar & line represents MSE values of quality data for all variants of Relief. Orange colored bar & line denotes MSE values of data after preprocessing for all relief methods. Grey colored results shown that the MSE values of noisy data for relief and it's variants. MSE values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the MSE values of dataset with missing values elimination.

Table 7. Root Mean Squared Error

Type of data	**Relief**	**ReliefF**	**Iterative Relief**	**ReliefG**
Quality data	0.241	0.233	0.233	0.222
data with Preprocessing	0.232	0.243	0.241	0.224
noisy data	0.248	0.23	0.246	0.226
data with outliers	0.224	0.235	0.239	0.233
data with missing values elimination	0.22	0.212	0.22	0.222



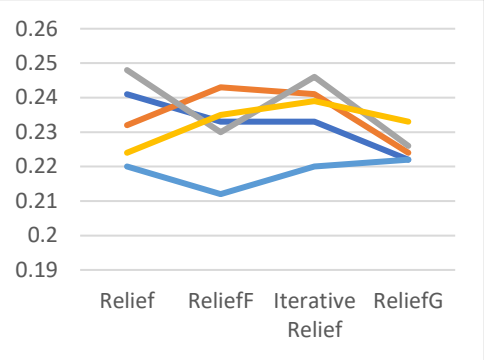
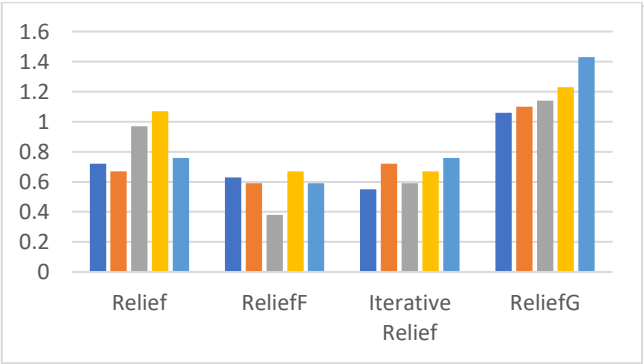


Fig 17. Root Mean Squared Error **Fig 18. Root Mean Squared Error**

In above Fig 17 and Fig 18, thick blue bar & line represents RMSE values of quality data for all variants of Relief. Orange colored bar & line denotes RMSE values of data after preprocessing for all relief methods. Grey colored results shown that the RMSE values of noisy data for relief and it's variants. RMSE values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the RMSE values of dataset with missing values elimination.

Table 8. t-test values

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	0.72	0.63	0.55	1.06
data with Preprocessing	0.67	0.59	0.72	1.1
noisy data	0.97	0.38	0.59	1.14
data with outliers	1.07	0.67	0.67	1.23
data with missing values elimination	0.76	0.59	0.76	1.43



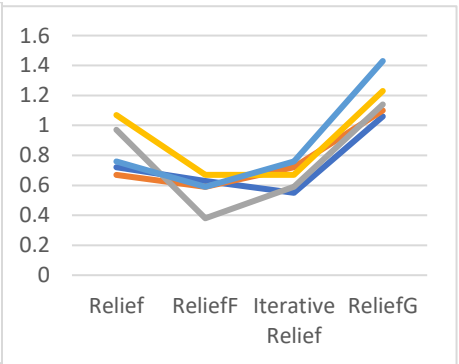


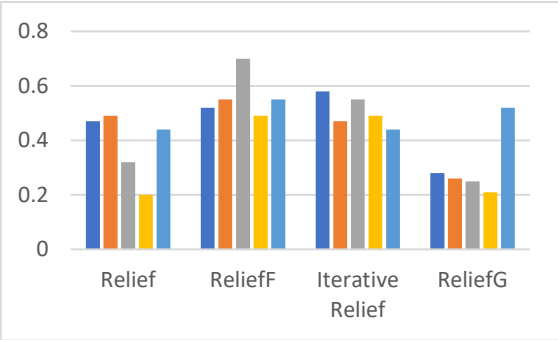
Fig 19. t-test values **Fig 20. t-test values**

In above Fig 19 and Fig 20, thick blue bar & line represents t-test values of quality data for all variants of Relief. Orange colored bar & line denotes t-test values of data after preprocessing for all relief methods. Grey colored results shown that the t-test values of noisy data for relief and it's variants. t-

test values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the t-test values of dataset with missing values elimination.

Table 9. p-test values

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	0.47	0.52	0.58	0.28
data with Preprocessing	0.49	0.55	0.47	0.26
noisy data	0.32	0.7	0.55	0.25
data with outliers	0.2	0.49	0.49	0.21
data with missing values elimination	0.44	0.55	0.44	0.52



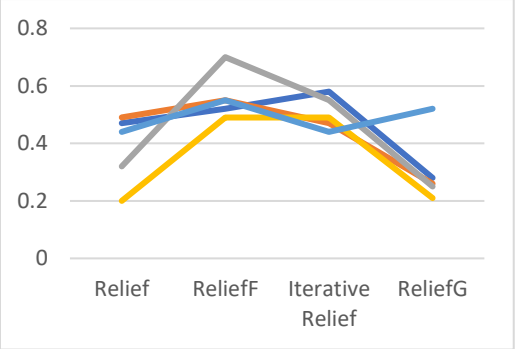


Fig 21. p-test values

Fig 22. p-test values

In above Fig 21 and Fig 22, thick blue bar & line represents p-test of quality data for all variants of Relief. Orange colored bar & line denotes p-test values of data after preprocessing for all relief methods. Grey colored results shown that the p-test values of noisy data for relief and it's variants. p-test values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the p-test values of dataset with missing values elimination.

Table 10. f-test values

Type of data	Relief	ReliefF	Iterative Relief	ReliefG
Quality data	0.98	0.98	0.99	0.92
data with Preprocessing	0.98	0.99	0.98	0.94
noisy data	0.98	0.99	0.99	0.97
data with outliers	0.97	0.98	0.98	0.97
data with missing values elimination	0.98	0.99	0.98	0.98

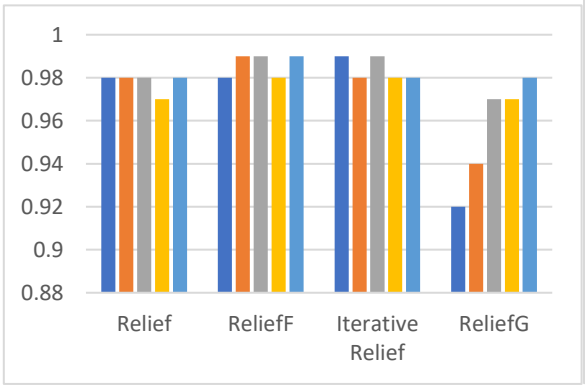


Fig 23. f-test values

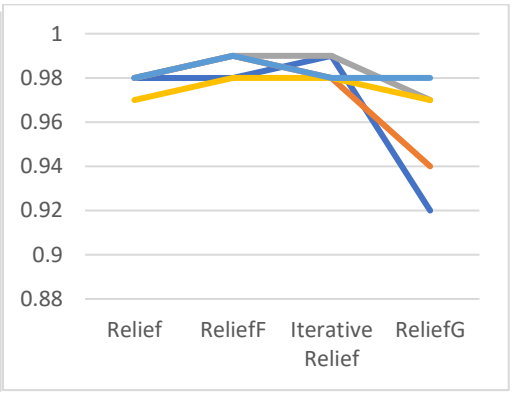


Fig 24. f-test values

In above Fig 23 and Fig 24, thick blue bar & line represents f-test values of quality data for all variants of Relief. Orange colored bar & line denotes f-test values of data after preprocessing for all relief methods. Grey colored results shown that the f-test values of noisy data for relief and it's variants. F-test values of dataset with outliers is represented by yellow colored bar and line. Finally the light blue colored bar, line represents the f-test values of dataset with missing values elimination.

7. Conclusions

This paper suggests an innovative approach for improving spam detection by utilizing sophisticated feature selection method ReliefG. This method builds on the principles of original Relief algorithm, enhancing its ability to handle various challenges like noise, missing values and outliers. The subsequent phase involves utilizing Random Forest (RF) for classification, leveraging the chosen features. The ReliefG method produced better accuracy over Relief, ReliefF and Iterative relief methods. However, the method reduces computational cost by identifying key attributes and generating crisp set values. This is done by integrating an efficient fuzzy membership function into a sigmoid function to improve accuracy, especially in the presence of outliers.

References

[1] Ahmed AI – Ajeli, Raaid Alubady, Eman S. AI – Shamery (2020), Improving Spam Email Detection using Hybrid Feature Selection and Sequential Minimal Optimisation, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 19, No. 1, PP: 535-542.

[2] Mrs. Anitha Reddy, et.al., (2023), Email Spam Detection Using Machine Learning, Journal of Survey in Fisheries Sciences, Vol. 10, No. 1, PP: 2658-2664.

[3] Veronica Bolon-Canedo, et.al., (2012), A Review of Feature Selection Methods on Synthetic data, Knowledge and Information Systems, Springer.

[4] Girish Chandrashekar, Ferat Sahin, (2014), A Survey on Feature Selection Methods, Computers and Electrical Engineering, Vol. 40, PP: 16-28.

[5] Newton Spolaor, et.al., (2012), Filter Approach Feature Selection Methods to Support Multi-label Learning Based on ReliefF and Information Gain, Springer, PP: 72-81.

[6] Hidayet Tacki, Fatema Nusrat (2023), Highly Accurate Spam Detection with the Help of Feature Selection and Data Transformation, The International Arab Journal of Information Technology, Vol. 20, No.1, PP: 29-37.

[7] Panem Charanarur, et.al., (2023), Machine – Learning – Based Spam Mail Detector, SN Computer Science, https://doi.org/10.1007/s42979-023-02330-x.

[8] Dr. Purva Mange, Aditi Lule, Rohini Savant, (2024), Advanced Spam Email Detection using Machine Learning and Bio-Inspired Meta-Heuristics Algorithms, International Journal of Intelligent Systems and Applications in Engineering, Vol. 12, No. 4, PP: 122-135.

[9] Ryan J. Urbanowicz, et.al., (2018), Relief – Based Feature Selection: Introduction and Reviews, Journal of Biomedical Informatics.

[10] B. Aruna Kumari, et.al., (2024), MultiSURF: Optimal Feature Selection Technique for Spam Mail Detection and Classification, Nanotechnology Perceptions, Vol. 20, No. S8, PP: 452-461.

[11] Bruce A. Draper, et.al., (2003), Iterative Relief, CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Explore, DOI: 10.1109/CVPRW.2003.10065.

[12] C. Nagaraju, et.al., (2024), A Generalized Two-Level Ensemble Method for Spam Mail Detection, Journal of Electrical Systems, Vol. 20, No. 2, PP: 1570-1579.