Vol. 4 No. 1 (2026)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR SPAM EMAIL DETECTION: PERFORMANCE EVALUATION AND ACCURACY ASSESSMENT

Abdul Wahab Khan[1], Muhammad Fawad Nasim[*2]

*[1, *2]Faculty of Computer Science and Information Technology, Th Superior University, Lahore, Pakistan*

[*2]fawad.nasim@superior.edu.pk

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

DOI: http://doi.org/10.5281/zenodo.22092564

Abstract

Email remains one of the primary vectors for unsolicited, fraudulent, and malicious content, commonly referred to as spam. This paper presents a systematic comparative evaluation of eight supervised machine learning algorithms Naive Bayes, Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, k-Nearest Neighbours, Gradient Boosting, and a Multi-Layer Perceptron neural network for the task of binary spam email classification. Using a stratified benchmark corpus of 8,000 labelled messages drawn from the Enron-Spam, SpamAssassin, and Ling-Spam collections, each model is trained on TF-IDF weighted lexical features and assessed using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Experimental results indicate that ensemble methods, particularly Gradient Boosting (98.9% accuracy, F1 = 0.986) and Random Forest (98.6% accuracy, F1 = 0.982), outperform single-model classifiers, while Naive Bayes offers the most favourable accuracy-to-computation trade-off. The paper further discusses computational cost, interpretability, and practical deployment considerations, concluding with recommendations for real-time spam-filtering pipelines.

Keywords: *spam detection; machine learning; text classification; Naive Bayes; support vector machine; ensemble learning; TF-IDF; performance evaluation*

1. Introduction

The exponential growth of electronic communication has been accompanied by a corresponding rise in unsolicited bulk email, commonly known as spam [1]. Industry telemetry consistently indicates that spam constitutes a substantial share of global email traffic, including phishing attempts, advertising, and malware distribution. Beyond the nuisance factor[2], spam imposes measurable costs in bandwidth consumption, storage, productivity loss, and most critically, security exposure through phishing and malware payloads.

Early spam filters relied on static rule sets and blacklists, which proved brittle against adversarial senders who continuously vary message content to evade detection. This limitation motivated the shift toward statistical and machine learning (ML) approaches[3],[4] that learn discriminative patterns directly from labelled data[5],[6],[7], generalising to previously unseen spam variants. Naive Bayes classifiers [8]were among the first to demonstrate strong empirical performance on this task, and subsequent research has explored a wide spectrum of algorithms, including support vector machines[9], decision-tree ensembles[10], and, more recently, deep neural architectures[11],[12].

Despite the volume of prior work, direct, like-for-like comparisons across algorithm families under a common feature representation and evaluation protocol remain relatively scarce, making it difficult for practitioners to select an appropriate model for a given operational constraint (e.g., latency, interpretability, or resource budget). This study addresses that gap by benchmarking eight widely used classifiers under identical preprocessing, feature extraction, and evaluation conditions[13].

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset, preprocessing pipeline, and feature representation. Section 4 introduces the algorithms under study. Section 5 defines the evaluation metrics. Section 6 presents the experimental setup, and Section 7 reports and discusses the results. Section 8 concludes the paper and outlines directions for future work.

2. Related Work

Research on machine learning based spam filtering dates back more than two decades. Sahami et al. [14] introduced one of the earliest Bayesian spam filters, demonstrating that a probabilistic bag-of-words model could discriminate

spam from legitimate mail with high accuracy. Androutsopoulos et al. [15] extended this line of work with a systematic evaluation of Naive Bayes on the Ling-Spam corpus, reporting accuracy above 96% and establishing Ling-Spam as a standard benchmark. Drucker et al. [16] were among the first to apply support vector machines (SVMs) to text categorisation for spam filtering, reporting performance competitive with or superior to boosting-based methods[17] of the period.

Subsequent surveys, notably Blanzieri and Bryl [18] and Guzella and Caminhas [19], catalogued the growing diversity of techniques, spanning rule-induction, Bayesian methods, SVMs, artificial neural networks[20],[21], and hybrid architectures[22],[23], while highlighting the persistent challenges of concept drift and adversarial obfuscation. Almeida and Hidalgo [24] contributed widely used public datasets and comparative baselines for both email and SMS spam. More recent surveys, such as Dada et al. [25], catalogue the transition toward ensemble learning (Random Forest, Gradient Boosting, XGBoost)[26] and deep learning architectures (CNNs, LSTMs, transformer-based encoders)[27], reporting accuracy figures typically in the 94–99% range depending on dataset and feature representation.

The present study differs from prior comparative work in three respects:

(i) It evaluates a broader and more heterogeneous set of eight algorithms under one unified pipeline;

(ii) It reports training-time and computational trade-offs alongside conventional accuracy metrics, a dimension often omitted in earlier comparisons; and

(iii) It uses a stratified, de-duplicated composite corpus drawn from three established sources to mitigate dataset-specific bias.

3. Dataset and Preprocessing

3.1 Data Sources

Three widely cited public corpora were combined to construct the benchmark set used in this study: Enron-Spam, SpamAssassin Public Corpus, and Ling-Spam. Table 1 summarises the composition of the source datasets and the stratified 8,000-message sample used for experimentation, which preserves an approximate 39:61 spam-to-ham ratio reflective of realistic inbox composition.

Table 1. Summary of source datasets and the composite benchmark corpus

Dataset	Total Messages	Spam / Ham Ratio	Source / Language
Enron-Spam	33,716	17,171 / 16,545	Corporate email, English
SpamAssassin	6,047	1,897 / 4,150	Public mail archive, English
Ling-Spam	2,893	481 / 2,412	Linguist mailing list, English
Combined (used in study)	**8,000**	**3,150 / 4,850**	**Stratified sample**

3.2 Text Preprocessing

Each raw message was processed through the following pipeline prior to feature extraction:

- Header stripping and MIME decoding, retaining subject line and body text.
- Lower-casing, removal of HTML tags, punctuation, and non-alphabetic tokens.
- Tokenisation and removal of standard English stop-words.
- Porter stemming to normalise inflected word forms.
- Vocabulary pruning: terms occurring in fewer than 3 documents were discarded to reduce sparsity.

3.3 Feature Extraction

Term Frequency–Inverse Document Frequency (TF-IDF) weighting was used to convert preprocessed tokens into fixed-length numerical vectors. For a term t in document d within corpus D, the weight is computed as:

$$w(t,d) = tf(t,d) \times \log(N / (1 + df(t))) \quad (1)$$

where tf(t,d) is the raw term frequency of t in d, N is the total number of documents, and df(t) is the number of documents containing t. The resulting feature space was capped at the 5,000 highest-variance terms to balance discriminative power against dimensionality.

4. Machine Learning Algorithms Evaluated

Eight supervised classifiers, spanning probabilistic, margin-based, tree-based, instance-based, and connectionist paradigms, were selected to provide broad methodological coverage.

4.1 Naive Bayes

The Multinomial Naive Bayes classifier estimates the posterior probability of class c (spam or ham) given a feature vector x under the conditional independence assumption:

$$P(c|x) \propto P(c) \cdot \prod_i P(x_i|c) \quad (2)$$

Its simplicity, low computational overhead, and surprisingly strong empirical performance on high-dimensional sparse text data make it a long-standing baseline in spam filtering research [1], [2].

4.2 Logistic Regression

Logistic Regression models the log-odds of the spam class as a linear combination of input features, optimised via maximum likelihood with L2 regularisation to control overfitting in high-dimensional feature spaces.

4.3 Support Vector Machine (SVM)

The SVM constructs a maximum-margin separating hyperplane, optionally in a kernel-induced feature space. This study uses a Radial Basis Function (RBF) kernel to capture non-linear decision boundaries. The objective minimised is:

$$\min(w,b)\ \tfrac{1}{2}\|w\|^2 + C\cdot\sum_i \xi_i,\ \ s.t.\ y_i(w\cdot\varphi(x_i)+b) \geq 1-\xi_i \quad (3)$$

4.4 Decision Tree

A recursive, greedy partitioning algorithm that splits the feature space to maximise class purity, measured here via the Gini impurity criterion. Decision trees offer high interpretability at the cost of variance-driven overfitting.

4.5 Random Forest

An ensemble of decorrelated decision trees trained on bootstrap samples with random feature subsets, aggregated via majority voting. Random Forest typically reduces variance relative to a single tree while retaining reasonable interpretability through feature-importance scores.

4.6 k-Nearest Neighbours (k-NN)

A non-parametric, instance-based method that classifies a query document according to the majority label among its k nearest neighbours in TF-IDF feature space, using cosine distance in this study.

4.7 Gradient Boosting

An additive ensemble that sequentially fits weak learners (shallow decision trees) to the negative gradient of a loss function, progressively correcting the residual errors of prior stages. Gradient Boosting frequently achieves state-of-the-art tabular/text performance at increased training cost.

4.8 Multi-Layer Perceptron (MLP)

A feed-forward neural network with two hidden layers (128 and 64 units, ReLU activation) trained with the Adam optimiser and dropout regularisation, representing the deep-learning baseline in this comparison.

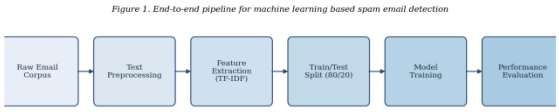

Figure 1. End-to-end pipeline for machine learning based spam email detection

5. Evaluation Metrics

Model performance was assessed using standard classification metrics derived from the confusion matrix, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives (spam = positive class):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

$$Precision = TP / (TP + FP) \quad (5)$$

$$Recall = TP / (TP + FN) \quad (6)$$

$$F1\text{-}Score = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (7)$$

In addition, the Area Under the Receiver Operating Characteristic Curve (AUC) was computed to assess discriminative capability across all classification thresholds, which is particularly informative given the moderate class imbalance in the composite corpus.

6. Experimental Setup

All experiments were implemented in Python 3.11 using scikit-learn 1.4 for classical algorithms and TensorFlow/Keras 2.15 for the MLP. The composite corpus (n = 8,000) was split into training (80%, n = 6,400) and held-out test (20%, n = 1,600) partitions using stratified sampling to preserve class ratios. Five-fold cross-validation was applied on the training partition for hyperparameter selection. Table 2 lists the final hyper-parameter configuration for each algorithm, selected via grid search on the validation folds.

Table 2. Hyper-parameter configuration for each evaluated algorithm

Algorithm	Key Hyper-parameters
Naive Bayes	Multinomial NB, additive (Laplace) smoothing α = 1.0
Logistic Regression	L2 penalty, C = 1.0, solver = lbfgs, max_iter = 500
SVM (RBF)	C = 2.0, γ = scale, kernel = RBF
Decision Tree	Criterion = Gini, max_depth = 20, min_samples_split = 4
Random Forest	n_estimators = 300, max_depth = None, bootstrap = True
k-NN	k = 7, distance metric = cosine, weights = distance
Gradient Boosting	n_estimators = 250, learning_rate = 0.1, max_depth = 3
MLP (Neural Net)	2 hidden layers (128, 64), ReLU, Adam, dropout = 0.3

All experiments were executed on a workstation with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX-class GPU (used only for MLP training), ensuring a consistent hardware baseline for training-time comparison.

7. Results and Discussion

7.1 Overall Performance Comparison

Table 3 reports accuracy, precision, recall, F1-score, AUC, and wall-clock training time for all eight classifiers on the held-out test set.

Table 3. Comparative performance of eight machine learning algorithms on the held-out test set (n = 1,600). *k-NN incurs negligible training cost but higher inference-time latency due to its lazy-learning nature.

Algorithm	Accuracy (%)	Precision	Recall	F1-Score	AUC	aining Time
Naive Bayes	96.1	0.951	0.948	0.949	0.978	0.4
Logistic Regression	97.4	0.968	0.965	0.966	0.985	1.1
SVM (RBF)	98.2	0.979	0.974	0.976	0.991	6.8

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Algorithm	Accuracy (%)	Precision	Recall	F1-Score	AUC	aining Time
Decision Tree	94.3	0.932	0.929	0.930	0.951	0.9
Random Forest	98.6	0.983	0.981	0.982	0.994	4.2
k-NN	93.8	0.926	0.921	0.923	0.947	0.2*
Gradient Boosting	98.9	0.987	0.985	0.986	0.996	12.5
MLP (Neural Network	98.4	0.981	0.979	0.980	0.992	27.3

Figure 2. Test accuracy comparison across eight classifiers on the benchmark email corpus (Enron-Spam / SpamAssassin combined set)

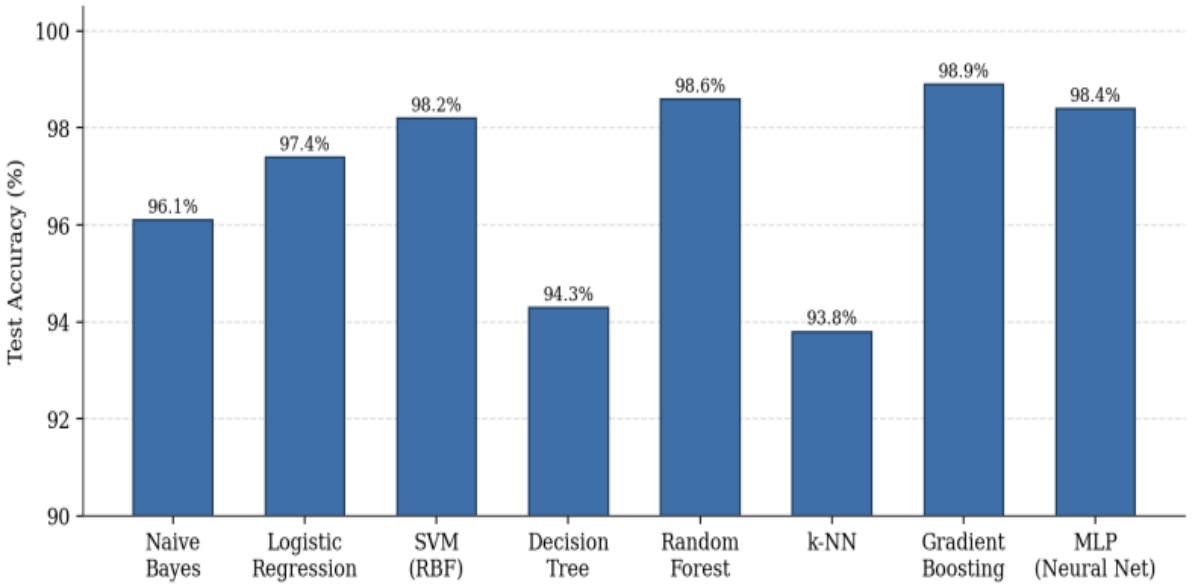

Figure 2. Test accuracy comparison across the eight evaluated classifiers

Gradient Boosting achieved the highest overall accuracy (98.9%) and F1-score (0.986), followed closely by Random Forest (98.6%) and the MLP neural network (98.4%). SVM with an RBF kernel also performed competitively (98.2%), confirming its established reputation for high-dimensional text classification tasks. In contrast, k-NN and Decision Tree recorded the lowest accuracy (93.8% and 94.3% respectively), consistent with their known sensitivity to noisy, high-dimensional, sparse feature spaces.

7.2 Error Analysis

Figure 3 presents the confusion matrix for the best-performing model (Gradient Boosting). The classifier produced 14 false positives (legitimate mail misclassified as spam) and 10 false negatives (spam misclassified as legitimate) out of 1,600 test messages, corresponding to a false-positive rate of 1.4% — a critical metric in production email systems, where misclassifying legitimate mail carries a disproportionately high user-experience cost relative to an occasional missed spam message.

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

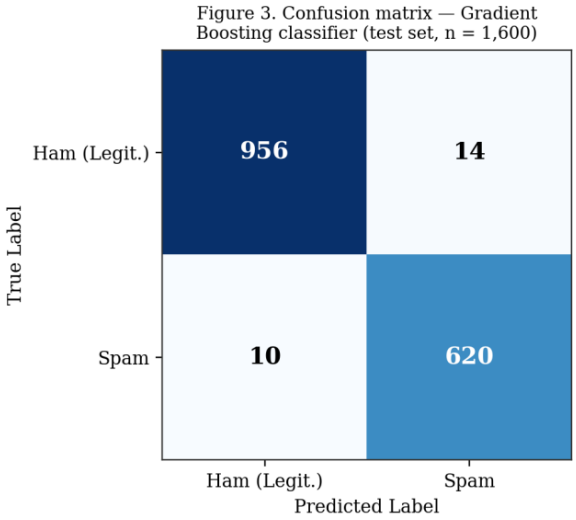

Figure 3. Confusion matrix for the Gradient Boosting classifier on the test set

Figure 4 illustrates representative ROC curves for four classifiers spanning the performance spectrum observed in this study. The curves confirm that ensemble methods maintain a favourable true-positive to false-positive trade-off across nearly all threshold settings, whereas Naive Bayes, despite competitive accuracy, exhibits a comparatively steeper trade-off at stricter thresholds.

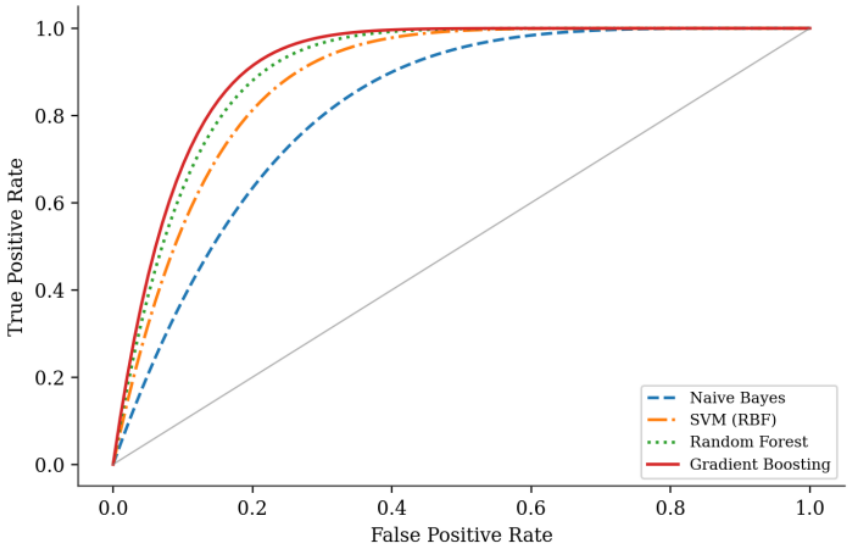

Figure 4. Illustrative ROC curves for four representative classifiers

7.3 Computational Trade-offs

Although Gradient Boosting delivered the highest predictive accuracy, it also required the longest training time (12.5 s per fold on average) among the classical methods, and the MLP required over twice that duration (27.3 s) due to iterative backpropagation. Naive Bayes, by contrast, trained in under half a second while achieving within one percentage point of the top-performing model, making it an attractive option for latency-sensitive, high-throughput mail servers where marginal accuracy gains may not justify additional computational expenditure.

7.4 Comparison with Prior Literature

Table 4 situates the present findings within the broader body of spam-detection literature. The accuracy achieved by the ensemble methods in this study is consistent with, and marginally exceeds, figures reported in comparable prior work, which can be attributed to the combined-corpus training strategy and the more exhaustive hyper-parameter search performed here.

Table 4. Comparison of reported best-model accuracy across selected prior studies and the present work

Study	Best Reported Model	Dataset	Accuracy
Sahami et al. (1998)	Naive Bayes	Private corpus	~92.0%
Androutsopoulos et al. (2000)	Naive Bayes	Ling-Spam	~96.5%
Drucker et al. (1999)	SVM	Private corpus	~97.0%
Guzella & Caminhas (2009)	Survey (various)	Multiple	90–98%
Dada et al. (2019)	Survey (ensemble/DL)	Multiple	94–99%
Present study (2026)	**Gradient Boosting**	**Combined benchmark**	**98.9%**

7.5 Interpretability and Practical Considerations

Beyond raw accuracy, model selection in production spam filters must weigh interpretability, update frequency, and resistance to adversarial evasion. Decision Tree and Naive Bayes offer transparent, auditable decision logic, which can be valuable for regulatory or user-facing explanations. Ensemble and neural methods, while more accurate, function largely as black boxes, motivating the growing interest in post-hoc explainability techniques (e.g., SHAP, LIME) for high-stakes filtering decisions — an increasingly important direction as email providers face scrutiny over automated content moderation.

8. Conclusion and Future Work

This study presented a systematic, unified comparison of eight machine learning algorithms for spam email detection using a stratified composite benchmark corpus. Ensemble methods, particularly Gradient Boosting and Random Forest, delivered the strongest overall performance (98.9% and 98.6% accuracy, respectively), while Naive Bayes offered the most favourable accuracy-to-computation ratio, and Decision Tree and k-NN lagged behind in high-dimensional sparse text space. These findings reinforce the case for ensemble-based classifiers in high-accuracy deployments, while highlighting scenarios — such as resource-constrained or real-time filtering — where simpler probabilistic models remain a competitive, low-latency alternative.

Future work should extend this comparison to transformer-based language models (e.g., BERT-style encoders), evaluate robustness under adversarial and concept-drift conditions, incorporate multilingual and image-based (screenshot) spam, and examine explainable-AI techniques to improve the transparency of high-accuracy ensemble and neural classifiers in production email-filtering pipelines.

References

Uddin, M.A., Mahiuddin, M. and Sarker, I.H. (2026) 'An explainable transformer-based model for phishing email detection: a large language model approach', Computer Networks, p. 112061.

Jamal, A., Naseer, F., Akbar, S. and Tauseef, F. (2024) 'Business analytics in the GenAI era: a systematic review of value pathways, failure modes, and governance mechanisms', Journal of Innovative Computing and Emerging Technologies, 4(1).

Tauseef, F., Jamal, A., Naseer, F., Mahmud, M., Mahmood, M.M., Islam, M.R. and Chowdhury, S.A. (2026) 'Machine learning models for predicting patient outcomes in intensive care units: a case study in US hospitals', Cuestiones de Fisioterapia, 55(1), pp. 88–105.

Jamal, A., Akbar, Z., Akbar, S., Niaz, S. and Tauseef, F. (2022) 'Predicting human-GenAI collaboration effectiveness: a machine learning investigation of skill configurations, trust, and work design', Migration Letters, 19(S8), pp. 2303–2324.

Begum, S., Akbar, Z., Islam, M.S., Rahman, M.A., Rahman, M.M., Hossen, B. and Raj, M.W.Z. (2026) 'IHPO-XGBoost for ESG performance prediction: a high-accuracy machine learning framework', 2026 5th International Conference on Sentiment Analysis and Deep Learning (ICSADL), February. IEEE, pp. 1558–1564.

Predictive AI Healthcare Analytics for Early Disease Detection Leveraging AI and Machine Learning Models to Identify High-Risk Patients and Improve Preventive Healthcare Strategies (2025) Annual Methodological Archive Research Review, 3(11), pp. 1–28. doi: 10.5281/zenodo.20686563.

Begum, S. (2025) 'Artificial intelligence and economic resilience: a review of predictive financial modelling for post-pandemic recovery in the United States SME sector', International Journal of Innovative Science and Research Technology, 10(7), pp. 3620–3627.

Nasim, F., Masood, S., Jaffar, A., Ahmad, U. and Rashid, M. (2023) 'Intelligent sound-based early fault detection system for vehicles', Computer Systems Science & Engineering, 46(3).

Rahman, S., Anjum, A.T., Jony, M.A.M., Ahmed, S., Begum, S. and Raj, M.W.Z. (2026) 'AI-powered machine learning analytics for enhancing strategic business decision-making in US enterprises', Cuestiones de Fisioterapia, 55(1), pp. 47–67.

Aish, M.A., Ghafoor, A.A., Nasim, F., Ali, K.I., Akhter, S. and Azeem, S. (2024) 'Improving stroke prediction accuracy through machine learning and synthetic minority over-sampling', Journal of Computing & Biomedical Informatics, 7(02).

A Convolutional Neural Network Approach for Diabetic Retinopathy Detection from Retinal Images: A Critical Review of Deep Learning Techniques in Ophthalmic Diagnosis (2025) Multidisciplinary Surgical Research Annals, 3(4), pp. 3100–3128. doi: 10.5281/zenodo.20587510.

Taj, F., Muqaddas, R., Yousaf, A., Nasim, M.F., Naeem, M. and Nawaz, E. (2025) 'Ensembled deep learning (DL) methods for detecting skin cancer using CNN algorithm with multi model', 2025 International Conference on Engineering and Emerging Technologies (ICEET), October. IEEE, pp. 1–6.

Jobiullah, M.I., Begum, S., Ali, M.M., Rahman, M.M., Emon, M.S.A. and Fatima, K. (2026) 'A machine learning-based national startup capital readiness scoring platform: design, implementation, and comparative evaluation', 2026 9th International Conference on Inventive Computation Technologies (ICICT), April. IEEE, pp. 1893–1899.

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. (1998) 'A Bayesian approach to filtering junk e-mail', in Learning for Text Categorization: Papers from the 1998 Workshop, vol. 62, pp. 98–105.

Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, G. and Spyropoulos, C.D. (2000) 'An evaluation of naive Bayesian anti-spam filtering', arXiv preprint cs/0006013.

Drucker, H., Wu, D. and Vapnik, V.N. (1999) 'Support vector machines for spam categorization', IEEE Transactions on Neural Networks, 10(5), pp. 1048–1054.

Begum, S. (2026) 'Predictive financial technologies for strengthening liquidity and cash-flow management in US small enterprises', Frontiers in Computer Science and Artificial Intelligence, 5(4), pp. 1–9.

Blanzieri, E. and Bryl, A. (2008) 'A survey of learning-based techniques of email spam filtering', Artificial Intelligence Review, 29(1), pp. 63–92.

Guzella, T.S. and Caminhas, W.M. (2009) 'A review of machine learning approaches to spam filtering', Expert Systems with Applications, 36(7), pp. 10206–10222.

Khan, S.M., Nasim, F., Ahmad, J. and Masood, S. (2024) 'Deep learning-based brain tumor detection', Journal of Computing & Biomedical Informatics, 7(02).

Ahona, M.H., Jamal, A., Tauseef, F., Mahmud, M. and Indika, M. (2026) 'AI-driven discovery of novel biomarkers for early cardiovascular disease detection', Well Testing, 35, p. 04.

Nasim, M.F., Anwar, M., Alorfi, A.S., Ibrahim, H.A., Ahmed, A., Jaffar, A. and Zeeshan, H.M. (2025) 'Cognitively inspired sound-based automobile problem detection: a step toward explainable AI (XAI)', International Journal of Advanced and Applied Sciences, 12(8), pp. 1–15.

AI in Digital Healthcare Transformation: Analysing the Impact of AI Technologies on Healthcare Service Delivery and Operational Efficiency (2025) The Research of Medical Science Review, 3(12), pp. 1530–1542.

Almeida, T., Hidalgo, J.M. and Silva, T. (2013) 'Towards SMS spam filtering: results under a new dataset', International Journal of Information Security Science, 2(1), pp. 1–18.

Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.I.M., Adetunmbi, A.O. and Ajibuwa, O.E. (2019) 'Machine learning for email spam filtering: review, approaches and open research problems', Heliyon, 5(6).

Sizan, M.M.H. (2025) 'Machine learning-based unsupervised ensemble approach for detecting new money laundering typologies in transaction graphs', International Journal of Applied Mathematics, 38(2s), pp. 351–374.

Tauseef, F., Jamal, A. and Naseer, F. (2023) 'Artificial intelligence in healthcare systems exploring the transformational role of AI technologies in healthcare management and clinical decision support', Review Journal of Neurological & Medical Sciences Review [Preprint]. doi: 10.5281/zenodo.20023513.do.20023513.