A Comparative Study of Machine Learning and Quantum Models for Spam Email Detection

Cameron Williams
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA
email: cwilliams1936@tuskegee.edu

Taieba Tasnim
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA
email: ttasnim6386@tuskegee.edu

Berkeley Wu
Auburn City School
Auburn, Alabama, USA
email: tulipfan002@hotmail.com

Mohammad Rahman
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA
email: mrahman@tuskegee.edu

Fan Wu
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA
email: fwu@tuskegee.edu

***Abstract*—This research focused on evaluating the performance of seven different machine learning algorithms including Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Feedforward Neural Network (FNN), Convolutional Neural Network (CNN), and Quantum Convolutional Neural Network (QCNN) using a single labeled email dataset. Each algorithm was applied to the same set of data and tested for its ability to detect spam and classify various types of abnormal behavior patterns. The study aimed to benchmark the accuracy of each model in a consistent environment to understand how well they handled real-world classification challenges. After processing and training the models, their outputs were compared based on accuracy, with results compiled into a bar chart for clear comparison. The findings highlight the strengths and limitations of each approach, providing insight into which models are better suited for tasks, such as spam detection, anomaly detection, and pattern recognition in email-based data.**

Keywords-KNN; FNN; CNN; SVM; QCNN; Machine Learning; Deep Learning; Quantum Computing.

I. Introduction

In today's digital communication ecosystem, spam emails continue to pose significant security and productivity challenges. Beyond mere nuisance, spam messages are frequently used as vectors for phishing, malware distribution, and social engineering attacks. As these threats evolve in complexity, traditional rule-based filtering systems are no longer sufficient, prompting a growing reliance on Machine Learning (ML) models for automated, adaptive detection.

Machine learning offers the ability to extract patterns and anomalies from large volumes of textual data, enabling more accurate and scalable spam filtering. While various algorithms have been employed in this domain including probabilistic models, distance-based classifiers, and deep neural networks, yet comparative studies under consistent experimental conditions remain limited. Furthermore, emerging paradigms such as quantum inspired learning have not been thoroughly benchmarked against classical approaches in real-world spam detection tasks.

This study addresses this gap by evaluating and comparing the performance of seven classification algorithms: Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Convolutional Neural Network (CNN), Feedforward Neural Network (FNN), Support Vector Machines (SVM), and Quantum Convolutional Neural Network (QCNN) on a standardized email dataset. Each model is tested using identical preprocessing, training, and evaluation pipelines to ensure fair comparison.

In this research, our main contributions are outlined as follows:

- Developed a standardized evaluation pipeline to compare the performance of traditional machine learning, deep learning, and quantum learning models using a single, preprocessed spam email dataset.
- Implemented and benchmarked seven classification algorithms, Naive Bayes, KNN, Logistic Regression, CNN, Neural Network, SVM, and QCNN under consistent conditions to assess their effectiveness in spam detection.
- Provided critical analysis of model performance, revealing the strengths of classical and deep learning methods, and highlighting the limitations of emerging quantum models like QCNN in handling text-based classification tasks.

Our findings aim to inform researchers and practitioners of the comparative efficacy of different machine learning approaches in email-based classification tasks, especially as interest grows in hybrid and quantum inspired cybersecurity solutions.

The remainder of this paper is organized as follows. Section II reviews related work on classical and quantum-inspired models, with emphasis on CNN and QCNN advancements. Section III outlines the methodology, including data acquisition, preprocessing, and model implementation. Section IV defines the evaluation metrics used to assess performance. Section V presents experimental results and a comparative analysis of all models. Section VI discusses key findings and model behaviors. Finally, Section VII concludes the paper and

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-306-4

highlights directions for future research in quantum machine learning.

II. LITERATURE REVIEW

Spam email detection has long been a central focus in cybersecurity, with the Naive Bayes classifier recognized for its simplicity and effectiveness. As shown by Zaragoza et al., it performs well on high-dimensional text data by applying the Bayes' Theorem with the independence assumptions of features [1]. Enhancements such as Laplace smoothing and hybrid models have further improved its accuracy, particularly on imbalanced datasets.

KNN is another widely used technique, valued for its intuitive, non-parametric structure. In spam filtering, KNN classifies emails based on their similarity to labeled examples. However, as noted by Eskin et al. [2], its computational cost on large datasets has led to the adoption of dimensionality reduction techniques such as Principal Component Analysis (PCA) to improve scalability.

Logistic Regression remains a popular method for binary classification due to its interpretability and scalability. As discussed by Bolton and Hand [3], it effectively models relationships between input features and class labels, making it particularly suited for text-based spam detection where features like word frequency and presence of specific terms can be strong predictors. Its transparent coefficients offer insight into the importance of features, which is valuable in both research and regulatory settings.

SVMs are widely used in spam filtering due to their ability to model non-linear boundaries through kernel functions. Compared to traditional techniques like blacklists and whitelists, SVMs offer superior generalization on high-dimensional email data. However, their performance heavily depends on kernel selection. Singh et al. [4] evaluated linear and Gaussian kernels using the SpamAssassin dataset and found that kernel choice significantly affects accuracy. Their results, validated on Gmail data, highlight SVM's effectiveness and adaptability in real-world spam detection tasks.

In recent years, deep learning models like CNN have been adapted for spam detection. Although originally designed for image recognition, CNN can classify text by learning local feature patterns. Jeong et al. [5] showed that CNNs with Spatial Pyramid Average Pooling (SPAP) effectively detect malware in document byte streams, demonstrating their versatility across data types.

FNN have proven effective in spam detection, particularly when optimized using metaheuristic algorithms. Jantan et al. [6] applied an Enhanced Bat Algorithm (EBAT) to train FNN, achieving strong performance on SPAMBASE and UK-2011 datasets. Similarly, Alsudani et al. [7] combined FNN with Crow Search Optimization and LSTM, reaching 99.1% testing accuracy, underscoring the benefits of hybrid approaches.

QCNN has recently gained attention as a novel framework for high-dimensional data classification. Using quantum principles such as entanglement and superposition, QCNN enable efficient representation and manipulation of complex data structures [8]. Cong et al. demonstrated their potential for exponential speedups in structured classification problems [9]. Empirical benchmarks comparing QCNN and CNN show that, under classical simulation and comparable settings, classical CNNs remain stronger on binary image classification [10]. Although current implementations remain constrained by hardware limitations, QCNN has shown promise in cybersecurity applications such as pattern recognition and intrusion detection, positioning them as a forward-looking candidate for future email security systems. Adversarial attacks occur in text, audio, and graph data. Published studies show textual adversarial examples and defenses, multi-targeted audio perturbations that mislead speech recognizers, and attacks on graph neural networks [11] [12] [13]. This means spam filters should be tested for robustness, not only accuracy.

In summary, while numerous models have been explored for spam classification, few studies have benchmarked classical, deep learning, and quantum-inspired approaches under consistent conditions. This research addresses that gap through a unified comparative analysis using a standardized dataset and evaluation framework.

III. METHODOLOGY

A. Data Acquisition

This study used a labeled email dataset obtained from Kaggle, a widely recognized platform for open source machine learning resources [14]. The data set contained approximately 5,700 email samples, each labeled spam (1) or non-spam (0), and was downloaded in Excel format. Each entry included raw email text and a corresponding binary label.

To prepare the data for model training, several preprocessing steps were performed: duplicate removal, conversion to lowercase, punctuation removal, and stop-word filtering. These steps ensured text uniformity and improved model learning efficiency. The data set showed moderate class imbalance about 745 spam emails versus 4,955 non-spam. This imbalance was taken into account during the model evaluation to avoid biased results.

We used a single, moderately sized Kaggle dataset, which provides a controlled, unified benchmark across the seven models but also limits generalizability. Figure 1 depicts an email along with the possible factors contributing to its classification as spam.

B. Tools, Training Environment, and Hardware Integration

Model development and testing were conducted on Google Colab, which provided sufficient computing resources, including GPU support for deep learning tasks [15] [16]. The integration of the platform with Google Drive allowed for easy storage and access to datasets and scripts. The preprocessing steps, the model configurations and the evaluation procedures were kept consistent between experiments to ensure fair benchmarking and reproducibility.

The entire workflow for this research is illustrated in Figure 2. It begins with data collection and preprocessing,

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

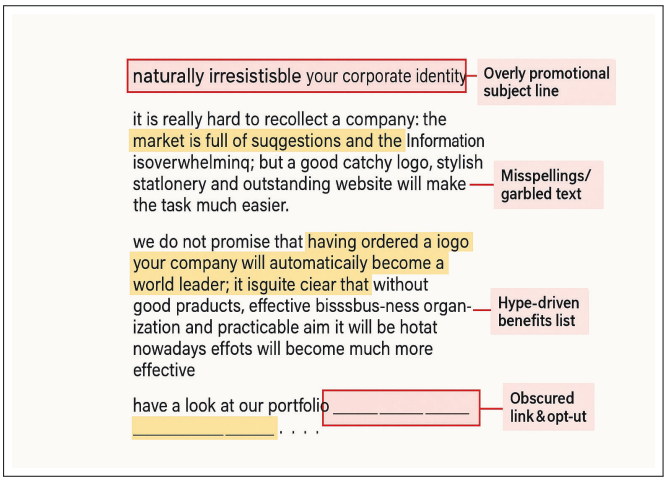

Figure 1. Sample of a potential spam email.

followed by text vectorization using either Term Frequency-Inverse Document Frequency (TF-IDF) or Count Vectorizer methods. After vectorization, several models were selected and trained such as Naive Bayes, KNN, Logistic Regression, SVM, FNN, CNN, and QCNN. Model performance was evaluated based on metrics such as accuracy and precision, and the results were visualized for comparison.

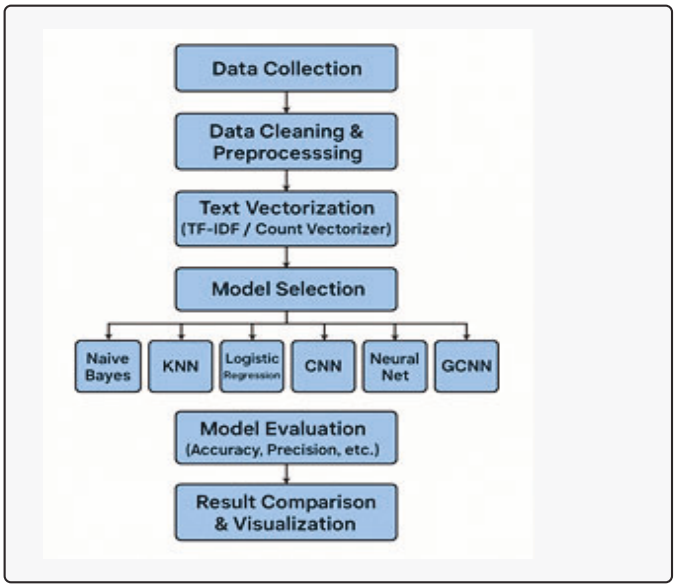

Figure 2. Spam Detection Workflow.

IV. EVALUATION METRICS

To evaluate how well our models performed on new data, we used a set of metrics commonly applied to classification problems. These metrics helped us measure how accurately each model could separate one class from another, especially in binary scenarios. Two of the main metrics we focused on were the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR shows how often the model correctly identifies positive cases, while FPR reveals how often it incorrectly labels negative cases as positive. Together, these metrics provided a clearer picture of each model's strengths and potential weaknesses when applied to real-world data.

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{2}$$

The True Positive Rate (TPR) shows how well the model correctly identifies positive cases out of all actual positives, while the False Positive Rate (FPR) measures how often negative cases are mistakenly labeled as positive. A True Positive (TP) is a case where the model correctly predicts a positive result, and a True Negative (TN) is when it correctly identifies a negative one. False Positives (FP) occur when negative cases are wrongly marked as positive, and False Negatives (FN) occur when the model misses a positive case and marks it as negative instead. These metrics are especially important when dealing with imbalanced datasets, as they help reveal how well the model can tell the difference between the two classes.

Along with TPR and FPR, we also measured Precision and Recall to better understand how the models handled positive predictions. Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

and shows the percentage of correct positive predictions out of all the positive results the model gave. Recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

and tells us how many of the actual positive cases were successfully identified by the model. These two metrics helps evaluate the trade-off between being accurate and being thorough in catching all positive cases.

To capture a balance between Precision and Recall, we used the F1 Score, which is the harmonic mean of the two. It is calculated as:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The F1 Score gives a single number that reflects both correctness and coverage of positive predictions, ranging from 0 to 1, where 1 means perfect performance. We also looked at Accuracy, which is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

This metric shows the percentage of total predictions the model got right, including both positive and negative outcomes. While accuracy is easy to understand, it can be misleading when classes are imbalanced, so we used it alongside the other metrics for a fuller picture.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

V. EXPERIMENTAL WORK

This study focused on a labeled dataset comprising spam and non-spam emails, aiming to evaluate and compare the performance of a range of classification algorithms under consistent experimental conditions. Standard preprocessing steps were applied, including the removal of special characters, stop words, and irrelevant symbols. We applied duplicate removal, lowercasing, punctuation and special-character removal, stop-word filtering, tokenization, and TF–IDF vectorization with unigrams and bigrams; no stemming or lemmatization was used [17] [18]. The cleaned text was then converted into numerical format using TF-IDF vectorization, ensuring standardized input across all models and enabling a fair and reproducible evaluation.

Seven classification algorithms were selected to assess their effectiveness in spam detection: Naive Bayes, KNN, Logistic Regression, CNN, FNN, SVM, and QCNN. These models represent a diverse spectrum of methodologies, ranging from classical statistical approaches and distance-based learning to deep learning and experimental quantum-inspired techniques.

Naive Bayes was chosen as the baseline model due to its long-standing success in text classification tasks. Its probabilistic framework, simplicity, and computational efficiency make it particularly suitable for high-dimensional textual data. KNN, an instance-based learner, was included to model similarity-based classification by evaluating the distance between new samples and labeled training instances. While KNN can be effective in small to medium datasets, it becomes computationally intensive as dataset size increases.

Logistic Regression was included for its interpretability and strong binary performance; its feature weights make it a solid benchmark. To assess deep learning, we added CNN and FNN. The CNN reshaped emails into matrices to enable convolutions that capture local patterns, while the FNN used stacked dense layers to model non-linear interactions. Both worked as expected but delivered only modest gains, likely due to the small dataset and limited hyperparameter tuning.

SVM was incorporated for its robust performance in handling overlapping and non-linearly separable classes. By using kernel functions, SVM effectively maps input features to higher-dimensional spaces to identify optimal separating hyperplanes. Its strong generalization made it one of the more competitive models in the study.

As an exploratory addition, a QCNN was implemented using quantum-inspired simulation on classical hardware. QCNN uses quantum entanglement and superposition principles to potentially encode and process high-dimensional data more efficiently. However, the QCNN in this experiment underperformed significantly relative to other models [19]. This could be attributed to limitations in current hardware simulation, immature software frameworks, or the mismatch between the model's structure and the nature of text data.

Model accuracies are shown in a comparative bar chart. Classical methods performed strongly, with SVM highest and Logistic Regression and Naive Bayes close behind. KNN served as an additional classical reference. QCNN was included as a future oriented, quantum inspired baseline.

Overall, this study provided a fair evaluation of multiple classification models on a shared spam email dataset. Traditional algorithms such as Naive Bayes and Logistic Regression outperformed others in terms of accuracy and efficiency. While deep learning models like CNN and FNN showed potential, they underperformed due to data limitations and minimal tuning. The QCNN, though promising in theory, delivered the lowest performance, highlighting the current gap between quantum-inspired approaches and practical text classification tasks.

VI. RESULTS AND DISCUSSIONS

The classification results from seven models applied to the spam email dataset are summarized across multiple performance metrics: accuracy, precision, recall, and F1-score. As shown in Figure 3, most classical machine learning models demonstrated strong overall performance, with accuracy values exceeding 95%, indicating their reliability for binary classification in structured text data.

The Naive Bayes classifier achieved an accuracy of 98.67%, precision of 98.66%, recall of 98.68%, and an F1-score of 98.67% (Figures 3–6). Its strength lies in the simplicity of its probabilistic model and independence assumptions, which work well for bag-of-words representations. The model's high performance despite minimal computational complexity makes it well-suited for real-time spam detection on low-resource devices.

The KNN model, while still achieving a respectable accuracy of 95.20%, showed slightly lower scores across all metrics (precision, recall, and F1-score: 95.20%). This model relies on distance based similarity, which can be affected by noisy or high-dimensional data. In practical settings, KNN can be effective in behavior based filtering but may struggle in large-scale or high-noise environments. As shown in Figure 4, precision follows the same pattern with the classical models performing strongly, KNN slightly lower, and QCNN clearly behind.

The CNN, adapted from its typical use in image processing to structured email data, performed exceptionally well. With 99.39% accuracy (Figure 3) and balanced scores across precision (99.40%), recall (99.38%), and F1-score (99.39%), CNN demonstrated its ability to extract useful local patterns from structured inputs. This suggests CNN's adaptability for non-image classification tasks when data is appropriately reshaped. As shown in Figure 5, recall is highest for SVM and the FNN, with CNN, Naive Bayes, and Logistic Regression also performing strongly, while KNN is slightly lower and QCNN is substantially weaker.

Logistic Regression, a linear model traditionally used for binary classification, also performed well, achieving 98.67% in precision, precision and F1 score, and slightly higher recall at 98.68%. Its interpretability and simplicity make it an excellent baseline model, particularly in environments that require explainable AI such as healthcare or financial domains.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

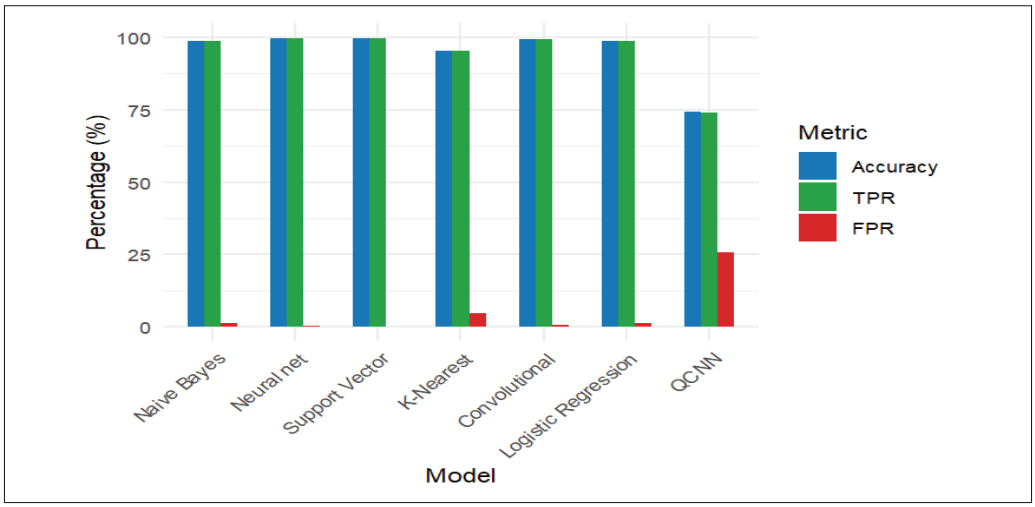

Figure 3. Model performance comparison by accuracy.

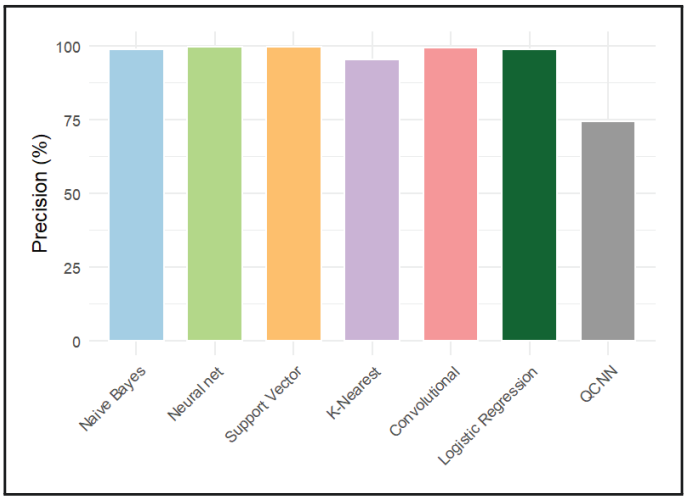

Figure 4. Precision comparison of classification models.

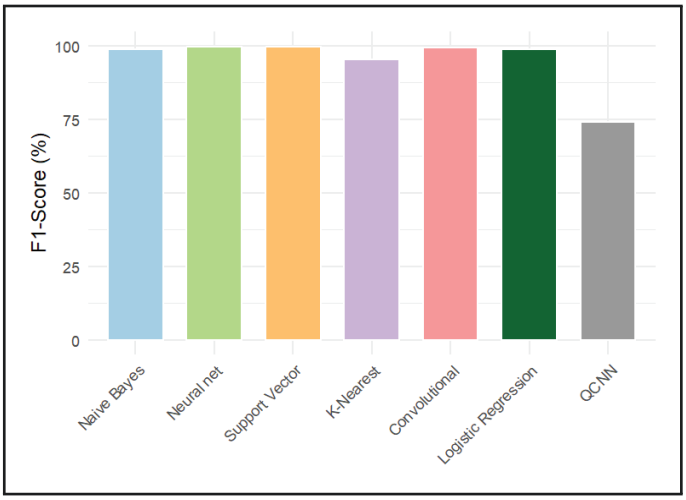

Figure 6. F1-score comparison of classification models.

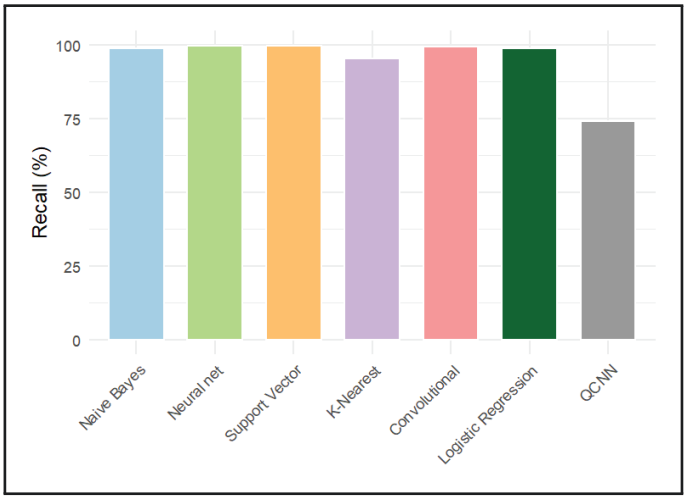

Figure 5. Recall comparison of classification models.

Among all models, the Neural Networ and SVM delivered the highest performance. The FNN achieved 99.65% precision with consistent metrics in precision (99.66%), recall (99.64%), and F1 score (99.65%). Its layered architecture captured complex, nonlinear relationships, contributing to its robustness. The SVM outperformed all others with an accuracy of 99.70%, supported by a precision of 99.70%, recall of 99.68%, and F1-score of 99.69%. Its strength lies in identifying optimal decision boundaries in high-dimensional feature spaces, making it highly suitable for separating subtle class differences.

The QCNN, tested here as an experimental model, achieved noticeably lower results: 74.17% accuracy, 74.22% precision, 74.16% recall, and 74.19% F1-score. These outcomes reflect the current limitations of quantum-inspired models when implemented on classical simulation hardware. While theoretically promising, QCNN still requires further algorithmic development and hardware support to compete with classical models on real-world datasets like spam classification. On the QCNN baseline, we include a QCNN to show where quantum-

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

inspired models for text stand today within one consistent pipeline. Given classical simulation limits and simple text encodings, the QCNN performs below strong classical baselines. This is a useful starting point for future work on native quantum hardware and richer encodings rather than a claim of current superiority.

The comparative analysis presents in Figure 4 and Figure 6 reflects their overall balance. Although the QCNN underperformed compared to classical models, its inclusion serves as an early benchmark for integrating quantum-enhanced techniques into cybersecurity. With future advancements in quantum hardware and optimization strategies, such models may offer significant potential.

VII. Conclusion and Future Work

This study evaluated the effectiveness of seven classification algorithms: Naive Bayes, KNN, CNN, Logistic Regression, FNN, SVM, and QCNN on a labeled dataset of spam and non-spam emails. Each model demonstrated unique strengths, with traditional machine learning techniques consistently achieving high accuracy and computational efficiency.

Naive Bayes reached 98.67% accuracy and is fast and simple, a good fit for lightweight spam filters. Logistic Regression matched 98.67% and is easy to interpret, useful where transparency matters. SVM led with 99.69%, handling non-linear and obfuscated patterns well. CNN and FNN performed solidly but showed limited gains at this data scale, likely due to dataset size and modest tuning. QCNN underperformed under classical simulation, reflecting current limits for text. Overall, SVM and FNN offer the best balance of precision and recall. For tight compute budgets choose Logistic Regression or Naive Bayes, and prefer Logistic Regression when interpretability is required. QCNN serves as a forward-looking baseline rather than a competitive option today.

Future work will focus on evaluating these models on larger, real-world datasets and exploring advanced feature engineering, ensemble methods, and native quantum hardware implementations to further enhance spam detection performance.

ACKNOWLEDGMENT

The work is supported in part by the National Science Foundation under NSF grants #2417608, and #2234911, #2209637, #2131228, and #2100134. Any opinions, findings and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] H. Zaragoza, P. Gallinari, and M. Rajman, "Machine learning and textual information access", in *Workshop at the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)*, Lyon, France, 2000, pp. 1–13.

[2] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection", in *Applications of Data Mining in Computer Security*, Springer, 2002, pp. 77–101.

[3] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review", *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002. DOI: 10.1214/SS/1042727940.

[4] M. Singh, R. Pamula, and S. K. Shekhar, "Email spam classification by support vector machine", in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 2018, pp. 878–882.

[5] Y. S. Jeong, J. Woo, S. Lee, and A. R. Kang, "Malware detection of hangul word processor files using spatial pyramid average pooling", *Sensors (Basel)*, vol. 20, no. 18, p. 5265, Sep. 2020.

[6] A. B. Jantan, W. A. H. M. Ghanem, and S. A. A. Ghaleb, "Using modified bat algorithm to train neural networks for spam detection", *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 24, pp. 6788–6799, 2017.

[7] S. W. A. Alsudani, H. A. M. Nasrawi, M. H. Shattawi, and A. Ghazikhani, "Enhancing spam detection: A crow-optimized ffnn with lstm for email security", *WJCM Science*, 2024, Available online: 01 April 2024. DOI: 10.31185/wjcms.199.

[8] T. Tasnim, M. Rahman, and F. Wu, "Comparison of CNN and QCNN performance in binary classification of breast cancer histopathological images", in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 3770–3777.

[9] I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks", *Nature Physics*, vol. 15, pp. 1273–1278, 2019. DOI: 10.1038/s41567-019-0648-8.

[10] T. Tasnim, A. Saha, M. Rahman, and F. Wu, "Quantum vs classical: Performance benchmarking of CNN and QCNN in binary image classification", in *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, Jan. 2025, pp. 203–208. DOI: 10.1109/CCWC62904.2025.10903816.

[11] H. Kwon and S. Lee, "Detecting textual adversarial examples through text modification on text classification systems", *Applied Intelligence*, vol. 53, no. 16, pp. 19 161–19 185, 2023. DOI: 10.1007/s10489-022-03313-w.

[12] K. Ko, S. Kim, and H. Kwon, "Multi-targeted audio adversarial example for use against speech recognition systems", *Computers & Security*, vol. 128, p. 103 168, 2023. DOI: 10.1016/j.cose.2023.103168.

[13] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data", in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, UK: ACM, 2018, pp. 2847–2856. DOI: 10.1145/3219819.3220078.

[14] J. Csie, *Spam email dataset*, https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset [retrieved: June, 2025], 2021.

[15] *Google colaboratory*, https://colab.research.google.com [retrieved: June, 2025], 2024.

[16] T. Tasnim, M. Rahman, and F. Wu, "A comparative analysis of cpu and gpu-based cloud platforms for cnn binary classification", in *The 2024 IARIA Annual Congress on Frontiers in Science, Technology, Services, and Applications*, Porto, Portugal, 2024, pp. 198–201.

[17] W. A. Awad and S. M. Elseuofi, "Machine learning methods for spam e-mail classification", *International Journal of Computer Science & Information Technology*, vol. 3, no. 1, pp. 173–184, 2011.

[18] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: Review and approaches", *Artificial Intelligence Review*, vol. 53, pp. 5019–5081, 2020.

[19] S. Oh, J. Choi, J. Kim, and J. Kim, "Quantum convolutional neural network for resource-efficient image classification: A quantum random access memory (qram) approach", in *2021 International Conference on Information Networking (ICOIN)*, 2021, pp. 50–52.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org