FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 12, December (Special Issue), 2025, pp 746 – 749
DOI: https://doi.org/10.33003/fjs-2025-0912-4387

OPTIMIZING SVM FOR PHISHING DETECTION: A COMPARATIVE PERFORMANCE STUDY

***Ismaila Idris, Aji L. Igba, Sikiru O. Subairu, Moses D. Noel and Sulemam Ahmed**

Department of Cyber Security Science, Federal University of Technology, Minna, Niger State, Nigeria.

*Corresponding authors' email: ismi.idris@futminna.edu.ng

ABSTRACT

Phishing attacks continue to pose significant cybersecurity threats, particularly through deceptive emails designed to steal user credentials or deliver malware. Existing detection systems rely heavily on machine learning models such as Support Vector Machines (SVM), whose performance is sensitive to hyperparameter optimization. This study investigates how different optimization techniques improve the performance of an SVM-based phishing email detection model. Four approaches were compared using a labelled phishing Kaggle dataset: a baseline linear SVM, a Grid Search–optimized SVM, a Stochastic Gradient Descent (SGD)–based SVM, and a Bayesian-optimized SVM with an RBF kernel. Each model was evaluated using standard performance metrics. The baseline SVM achieved 92.87% accuracy; Grid Search reached 96.00%; SGD achieved 92.00%; and Bayesian Optimization achieved the highest accuracy of 96.67%. Hyperparameter optimization significantly enhances SVM performance in phishing detection, with Bayesian Optimization offering the most efficient and accurate configuration.

Keywords: Phishing Detection, SVM Optimization, Hyper Parameter Tuning, Cybersecurity

INTRODUCTION

Phishing email attacks have become one of the most pervasive global cyber threats, affecting individuals, businesses, and government institutions. By exploiting human vulnerabilities and evading traditional security filters, phishing campaigns continue to cause substantial financial and data losses. To strengthen email-security mechanisms, researchers have increasingly applied machine learning techniques, with Support Vector Machines (SVMs) remaining among the most effective classifiers for text-based detection tasks. A comprehensive review in the study (Said Saloum et al., 2022), highlights that most existing phishing email studies rely heavily on NLP-driven feature extraction and SVM classification, typically using Term Frequency–Inverse Document Frequency (TF-IDF) or embedding-based representations. Their survey of 100 publications shows that the Nazario corpus and Python dominate experimental setups and reveals a significant research gap in non-English phishing datasets, particularly Arabic. Building on this, (Chinta et al., 2025), evaluated CNN, XGBoost, RNN, and SVM models, reporting that although SVM performed second best among traditional classifiers, its precision, recall, and F1-score remained notably lower than those of the proposed hybrid architectures. Their BERT-LSTM system achieved an exceptional F1-score of 99.24%, recall of 99.55%, precision of 99.61%, and accuracy of 99.55%, with minimal overfitting, illustrating the growing effectiveness of deep hybrid models.

Other studies reinforce the role of optimization in improving traditional ML classifiers as demonstrated in the research work (Fatima et al,. 2024), compared optimized models across three datasets and found that Stochastic Gradient Descent (SGD) achieved the highest overall performance, while algorithms such as Extra Trees, Random Forest, Logistic Regression, Linear SVM, and MLP also benefited substantially from hyper parameter tuning. Similarly, (Qiqieh et al., 2025), demonstrated that optimized SVMs, particularly those enhanced with nature-inspired algorithms such as Harris Hawks Optimization, achieve superior accuracy for various cyber-threats, including fake news, IoT attacks, malicious URLs, and spam, consistently outperforming untuned ML models.

Research on malware detection echoes these findings, most especially in this recent study (Akinshola-Awe et al., 2025), which showed that classifiers such as KNN, Decision Trees, and SVM can detect malware, but require proper hyper parameter tuning to reach high accuracy. Their Grid Search–based optimization and balanced datasets yielded significant improvements in SVM performance, reducing false positives and outperforming conventional antimalware techniques. Additional experiments using cross-validated Grid Search reported strong gains in overall precision, recall, and error reduction.

Optimization-enhanced hybrid models have also shown promising results, like the study (Chen ,2025), applied the Mayfly Optimization Algorithm to SGD and Naïve Bayes for phishing-website detection, with the SGD+MO hybrid achieving the highest accuracy (0.921). Strong precision across classes indicated that hybrid optimization continues to improve model reliability for real-world phishing scenarios. Meanwhile,(Fatima et al., 2024), emphasized that Spam emails tend to be phishing and continue to threaten user security, motivating the use of optimized ML methods for reliable detection. Using Count Vectorizer and TF-IDF across the Ling Spam, UCI SMS Spam, and a proposed dataset, this work evaluates several classifiers with multiple hyper parameter tuning strategies. SGD emerges as the top performer on all datasets, while other ensembles, linear, and neural models also achieve strong overall metrics.

Beyond these methods, probabilistic approaches have received renewed interest as the work (Zhang, Pang, and Yin 2022), highlight the effectiveness of Bayesian algorithms due to their ability to model uncertainty, adapt to evolving phishing behaviors, and remain robust under noisy or limited data conditions. Their studies show that improved Bayesian filters surpass traditional methods, achieving higher transmission consistency and data-security levels, such as 92% email-flow consistency and over 95% security, demonstrating why Bayesian models are increasingly preferred for phishing email filtering.

Collectively, the literature shows that while SVM remains a strong and widely used classifier, its effectiveness is highly dependent on the careful selection of hyper parameters such as kernel functions, regularization strengths, and learning

rates. Manual tuning typically yields suboptimal results, which has led to a strong research shift toward systematic optimization techniques. Motivated by these findings, this study conducts a rigorous comparison of four SVM optimization strategies to determine which approach best enhances SVM performance for phishing email detection. The contributions include (i) a structured evaluation of optimization methods, (ii) demonstrable performance gains through Bayesian Optimization, and (iii) a reproducible methodology for optimized SVM-based phishing detection.

MATERIALS AND METHODS

This study employs comparison and evaluates machine learning phishing email detection. The methodology is organized into three phases: problem analysis, system design, implementation using Python codes, and comparative evaluation to systematically identify the best performance in phishing detection. As illustrated in Figure 1, the framework for comparative process is shown below in the research work**.**

Dataset Description

The phishing email dataset used in the baseline research work (Fares et al., 2024), consists of Email_Text (email content) and Email_Type (identifying) labelled emails categorized as either phishing or legitimate. Each email was preprocessed through tokenization, stop-word removal, TF-IDF vectorization, and normalization. Table 1 below shows the result of the preprocessed phishing email dataset used in this research work.

Table 1: Dataset Before And AfterCleaning

Dataset Characteristic	Before Cleaning	After Cleaning	Reduction/Change
Total Instances	10,000 emails	9,872 emails	128 removed (.28%)
Feature (Raw)	1,250 Features	500 Features	750 removed (60%)
Text Length (Avg)	1,245 Characters	872 Characters	30% reduction
HTML Tag Present	68% of emails	0% of emails	Complete Remove
URL Present	4.2 avg per email	0 avg per email	Complete Remove
Email Addresses	3.1 avg per email	0 avg per email	Complete Remove
Special Characters	142 avg per email	12 avg per email	92% reduction
Duplicate Emails	312 detected	0 remaining	Complete Remove
Null Values	427 fields	0 fields	Complete Remove

Model Framework

The framework begins with data preprocessing and splitting, followed by developing four SVM variants: baseline, Grid Search, SGD-based, and Bayesian-optimized models. Each variant is trained and evaluated, after which its test accuracy is compared. The highest-performing model is then identified, forming the basis for the study's final insights and recommendations.

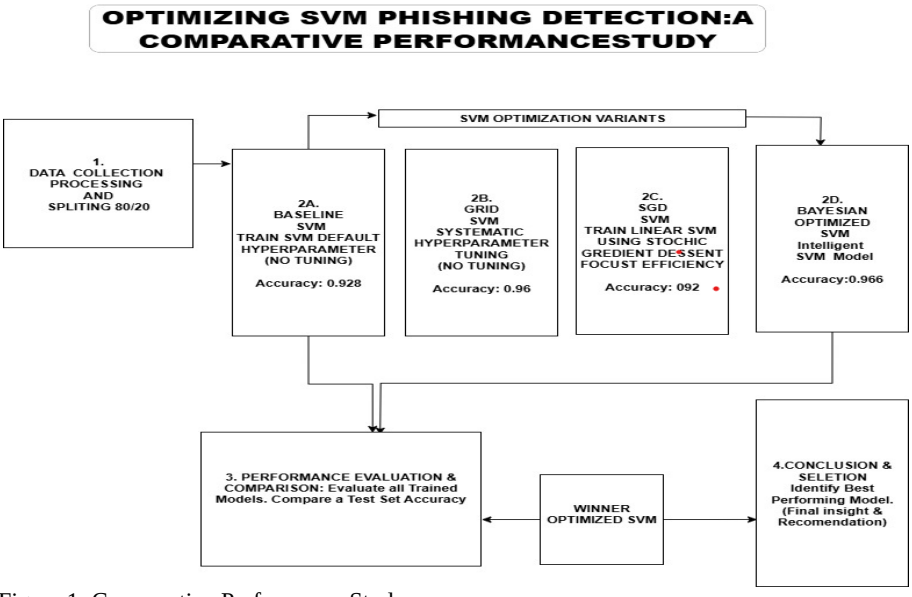

Figure 1: Comparative Performance Study

Baseline SVM Model

The baseline model, trained with default hyperparameters, provides a reference for comparison. It achieves moderate accuracy and demonstrates the need for tuning to improve detection performance.

Grid Search Optimization

Systematic hyperparameter tuning using GridSearchCV significantly improves performance. By exploring combinations of *C*, *gamma*, and *kernel* options, this model achieves higher accuracy and stronger classification metrics than the baseline.

SGD Optimization

The SGD Classifier using hinge loss (a linear SVM equivalent) prioritizes computational efficiency. While faster to train, its performance is typically lower than both the Grid

Search and Bayesian-optimized models, reflecting the trade-off between speed and accuracy.

Bayesian Optimization

Bayesian Optimization achieved the best performance by using probabilistic modelling to efficiently select optimal hyperparameters. Using the search space kernel = rbf, C ∈ [0.1, 50], and gamma ∈ [1e-4, 1], it produced the highest accuracy and most balanced precision, recall, and F1-scores. These results confirm that Bayesian-tuned SVMs outperform other optimization approaches for phishing detection.

RESULTS AND DISCUSSION

Results show the effectiveness of each optimization technique, as summarized below:

Model Performance Comparison

The baseline SVM provided a strong starting point, but its accuracy left room for enhancement. When Grid Search tuning was applied, the model's performance improved noticeably, demonstrating the value of systematic parameter exploration. The SGD-based SVM, however, delivered only comparable results to the baseline and did not offer meaningful gains. In contrast, the Bayesian-optimized SVM, particularly with the RBF kernel, emerged as the most effective approach, achieving the highest level of accuracy among all models. Overall, the progression from the baseline to the optimized variants clearly shows that hyperparameter tuning significantly strengthens the classification capability of the SVM, with Bayesian optimization offering the most impactful improvement. Table 2 below represents a quantitative analysis of the models involved.

Table 2: Accuracy Comparison of SVM Variants

Model	Accuracy (%)	Accuracy gain
Baseline SVM	92.87	-
Grid Search SVM	96	3.13
SGD SVM	92	-0.87
Bayesian SVM	96.67	3.8

The displayed figure 3.2 below presents the performance outcomes of the Bayesian-optimized Support Vector Machine (SVM) model, alongside the optimal hyperparameters identified through the Bayesian search process. It summarizes the classifier's precision, recall, and F1-scores across both classes, as well as the macro and weighted averages. This figure serves as a consolidated view of how Bayesian optimization enhances the SVM's decision-making capability by systematically selecting the most effective combination of parameters.

```
Best Parameters (Bayesian): OrderedDict([('C', 8.713), ('gamma', 0.005), ('kernel', 'rbf'), ('de
gree', 3)])

Bayesian Optimized SVM Performance:
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       742
           1       0.97      0.96      0.97       758

    accuracy                           0.97      1500
   macro avg       0.97      0.97      0.97      1500
weighted avg       0.97      0.97      0.97      1500

Accuracy: 0.9666666666666667
```

Figure 2: Bayesian Optimized SVM

Discussion

This study examines four variants of the Support Vector Machine (SVM) classifier to determine the most effective optimization approach for detecting phishing emails. The comparative analysis shows that Bayesian Optimization consistently delivers superior performance, achieving the highest detection accuracy among all tested classifiers, where all experiments adopted the Kaggle dataset of the baseline study. In contrast, the Stochastic Gradient Descent (SGD)–based SVM offers only marginal gains over the baseline model, indicating limited suitability for this task. A comparative assessment of model accuracy reveals distinct differences in how each optimization strategy affects detection capability. The summarized results illustrate both absolute accuracy levels and the relative improvements achieved through optimization. The Bayesian-optimized SVM emerged as the strongest performer, reaching a peak accuracy of 0.9667 percent. Its advantage lies in its ability to intelligently explore the hyperparameter space, allowing it to uncover optimal configurations that are often overlooked by manual tuning or exhaustive search methods.

Grid Search optimization also led to a noticeable improvement, increasing accuracy to 0.96 percent. This confirms the importance of systematic hyperparameter

optimization in phishing detection tasks. However, its higher computational cost and slightly lower precision make it less efficient than other classifiers. The SGD-based SVM demonstrated only a modest accuracy increase, reaching 0.92 percent. Although SGD is well-suited for very large-scale datasets, its performance appears less stable with the feature representations used in this study. As a result, it delivers weaker gains compared to more robust optimization strategies.

CONCLUSION

Based on the experimental results, Bayesian Optimization is the most effective approach for tuning SVM models in phishing detection applications. It achieves the highest accuracy improvement of 0.38 percent while maintaining strong precision. For scenarios where computational resources or time constraints are a concern, Grid Search remains a practical alternative, though it should be considered secondary to Bayesian methods for optimal performance. Future work will incorporate ensemble learning, while deep neural architectures, or explainability frameworks such as SHAP and LIME, also remain as options.

REFERENCES

Akinshola-Awe, F. J., A. A. Obiniyi, Gilbert Aimufua, Kene Anyachebelu, and Binyamin Adeniyi Ajayi. 2025. "Detection and Classification of Malware Using Grid Search Optimization Technique." *Science World Journal* 20(2):740–47. https://doi.org/10.4314/swj.v20i2.40

Chen, Xiao. 2025. "Hybrid Phishing Detection Using Stochastic Gradient Descent and Naïve Bayes Optimized with the Mayfly Algorithm." *Informatica (Slovenia)* 49(21):121–38. https://doi.org/10.31449/inf. v49i21.8056.

Chinta, Purna Chandra Rao, Chethan Sriharsha Moore, Laxmana Murthy Karaka, Manikanth Sakuru, Varun Bodepudi, and Srinivasa Rao Maka. 2025. "Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering." *European Journal of Applied Science, Engineering and Technology* 3(2):41–54. https://doi.org/10.59324/ejaset.2025.3(2).04.

Fares, Hajar, Jihad Kilani, Fatima Ezzahra Fagroud, Hicham Toumi, Fatima Lakrami, Youssef Baddi, and Noura Aknin. 2024. "Machine Learning Approach for Email Phishing Detection." Pp. 746–51 in *Procedia Computer Science*. Vol. 251. Elsevier B.V.

Fatima, Rubab, Mian Muhammad Sadiq Fareed, Saleem Ullah, Gulnaz Ahmad, and Saqib Mahmood. 2024. "An Optimized Approach for Detection and Classification of Spam Email's Using Ensemble Methods." *Wireless Personal Communications* 139(1):347–73. https://doi.org/10.1007/s11277-024-11628-9.

Qiqieh, Issa, Omar Alzubi, Jafar Alzubi, K. C. Sreedhar, and Ala' M. Al-Zoubi. 2025. "An Intelligent Cyber Threat Detection: A Swarm-Optimized Machine Learning Approach." *Alexandria Engineering Journal* 115:553–63. https://doi.org/10.1016/j.aej.2024.12.039.

SAID SALOUM, TAREK GABER, SUNIL VADERA, and KHALED SHAALAN. 2022. "A Sytematic Litereture Review on Phishing Email Detection Using Natural Language Processing Techniques." *IEEE Access*.

Zhang, Yahao, Jin Pang, and Hongshan Yin. 2022. "The Optimization Analysis of Phishing Email Filtering in Network Fraud Based on Improved Bayesian Algorithm." *International Journal of Circuits, Systems and Signal Processing* 16:504–8. https://doi.org/10.46300/9106.2022.16.62.