GA-Optimized Stacking Ensemble for Unified Phishing Website and Email Detection Using Multi-Domain Feature Representation

Hendra Hermawan[1,*], Muhammad Faisal[2], Mochamad Imamudin[3]

Master of Informatics Study Program, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

[1]240605220007@student.uin-malang.ac.id, [2]mfaisal@it.uin-malang.ac.id, [3]mamudin@ti.uin-malang.ac.id

*Corresponding Author

ARTICLE INFO

Article History
Received: 12 February 2026
Revised: 26 April 2026
Published: 30 April 2026

Keywords
cybersecurity,
ensemble learning,
genetic algorithm,
phishing detection,
website and email security

ABSTRACT

Phishing remains a major cybersecurity threat because attacks increasingly combine fraudulent websites with deceptive email content. Existing detection models often focus on a single domain, such as URLs or emails, which limits their ability to capture heterogeneous phishing patterns. This study proposes a GA-optimized stacking ensemble framework for unified phishing website and email detection using multi-domain features. The framework combines URL structural attributes, email metadata, and semantic content features, while a Genetic Algorithm is used to reduce feature redundancy and select the most informative attributes. The proposed model is evaluated against baseline Random Forest, Gradient Boosting, and conventional Stacking classifiers using Accuracy, F1-score, AUC-ROC, cross-validation stability, robustness under noise, and inference latency. Experimental results show that the proposed GA-Stacking model achieves 98.1% accuracy, 97.6% F1-score, and 0.947 AUC-ROC, outperforming Random Forest, Gradient Boosting, and standard Stacking models. The model also reduces the feature set from 72 to 31 features and maintains strong robustness under simulated noise, with F1-score remaining at 92.0% under 30% perturbation. These findings indicate that evolutionary feature optimization improves the stability, efficiency, and robustness of stacking ensemble learning for multi-domain phishing detection.

INTRODUCTION

Phishing detection has become a critical task in cybersecurity because modern attacks no longer rely only on simple fraudulent links or easily identifiable spam messages. Contemporary phishing campaigns increasingly combine manipulated URLs, deceptive website structures, spoofed sender identities, and semantically persuasive email content [1][2]. This shift creates a detection problem that is both structural and semantic. Website phishing detection usually depends on URL, domain, SSL, and page-based features, whereas email phishing detection requires analysis of metadata, linguistic patterns, and message intent. As a result, models trained on only one attack domain often fail to generalize when phishing strategies combine website-based and email-based deception.

Recent machine learning and deep learning studies have improved phishing detection performance using Random Forest, Gradient Boosting, deep neural networks, and transformer-based models. Transformer-based approaches, such as RoBERTa and BERT variants, have shown strong performance in phishing email detection because they can capture contextual and semantic cues in deceptive messages [3]. Similarly, optimized ensemble methods have demonstrated competitive results in phishing website detection by combining multiple classifiers and reducing model variance. However, these approaches are

commonly developed for a single detection domain, either phishing websites or phishing emails. This separation limits their ability to capture heterogeneous phishing patterns that involve both malicious links and persuasive email narratives.

Another technical limitation concerns feature redundancy and dimensionality. Phishing detection systems often extract many URL, domain, metadata, and textual features. Although large feature sets may improve representation, they can also introduce irrelevant or correlated attributes that increase computational cost, reduce interpretability, and raise the risk of overfitting [4][5][6]. Feature selection is therefore essential, especially in multi-domain phishing detection where website and email features are combined into a unified representation. Evolutionary optimization methods, including Genetic Algorithms, are useful in this context because they search for discriminative feature subsets without relying on fixed assumptions about feature importance.

Despite the progress of existing studies, three gaps remain insufficiently addressed. First, many models focus on either URL-based phishing or email-based phishing, rather than integrating both within a unified framework [7]. Second, several ensemble-based studies report high accuracy but provide limited analysis of how feature optimization affects stability, redundancy reduction, and generalization. Third, robustness under noisy or obfuscated phishing conditions is rarely evaluated systematically, even though real-world phishing attacks frequently modify URLs, alter textual patterns, and imitate legitimate communication styles. These gaps indicate the need for a detection framework that is not only accurate but also stable, feature-efficient, and robust against heterogeneous phishing variations.

To address these limitations, this study proposes a GA-optimized stacking ensemble framework for unified phishing website and email detection. The proposed framework integrates multi-domain features, including URL structural attributes, email metadata indicators, and semantic content representations. A Genetic Algorithm is used to select the most informative features and reduce redundancy before classification. The optimized feature subset is then processed using a stacking ensemble architecture that combines multiple base learners and a meta-learner to improve classification reliability. Unlike conventional single-model or single-domain approaches, the proposed method is designed to capture complementary patterns from both website and email phishing data.

The novelty of this study lies in the integration of three components within a single detection framework: multi-domain phishing representation, Genetic Algorithm-based feature optimization, and stacking ensemble classification. This combination enables the model to address both structural manipulation in phishing websites and semantic deception in phishing emails [8]. In addition, the study evaluates the model not only using standard metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC, but also through cross-validation stability, robustness under simulated noise, and inference latency analysis. This evaluation design provides a more comprehensive assessment of model reliability than accuracy-based comparison alone.

The contributions of this study are summarized as follows. First, it develops a unified phishing detection framework that combines website and email features in a single learning pipeline. Second, it integrates Genetic Algorithm-based feature selection with stacking ensemble learning to reduce feature redundancy and improve generalization. Third, it provides a broader evaluation by examining predictive performance, cross-validation stability, robustness against noise and obfuscation, and computational trade-offs. Through these contributions, this study aims to provide a more robust and technically grounded approach to phishing detection across heterogeneous attack domains.

METHOD

This study proposes a GA-optimized stacking ensemble framework for unified phishing website and email detection. The methodological workflow consists of dataset collection, preprocessing, feature extraction, feature engineering, Genetic Algorithm-based feature optimization, stacking ensemble classification, and performance evaluation. Unlike a conventional single-domain phishing detector, the proposed framework integrates website-based structural features and email-based semantic features into a unified learning pipeline.

Dataset Origin and Distribution

The dataset was constructed from two phishing detection domains: website phishing and email phishing. Website-based samples were obtained from publicly available phishing URL repositories, such as PhishTank, OpenPhish, and legitimate URL collections from public benchmark sources. Email-based samples were collected from publicly available email corpora, such as Enron, SpamAssassin, and phishing email benchmark datasets. Each sample was assigned a binary label: phishing or legitimate. To support balanced classification and reduce bias toward a dominant class, the dataset was organized into website and email subsets with comparable class proportions. The final dataset distribution is shown in Table 1.

Table 1. Dataset Distribution

Domain	Phishing Samples	Legitimate Samples	Total Samples
Website URLs	2,500	2,500	5,000
Emails	2,500	2,500	5,000
Total	5,000	5,000	10,000

The balanced distribution was used to minimize class imbalance effects and ensure that the model learned both phishing and legitimate patterns across website and email domains. The dataset was split into 80% training data and 20% testing data using stratified sampling. In addition, 5-fold stratified cross-validation was applied to assess model stability and reduce dependence on a single train–test partition.

Data Preprocessing

The preprocessing stage was designed to ensure consistency between website and email data. For website samples, invalid URLs, duplicate entries, and incomplete records were removed. URL strings were normalized by converting characters into lowercase, removing unnecessary whitespace, and standardizing protocol-related attributes. For email samples, duplicate messages, empty bodies, and malformed headers were removed. Textual content was normalized through lowercasing, punctuation cleaning, token filtering, and removal of non-informative symbols. Missing numerical values were handled using median imputation, while categorical values were encoded using label encoding or one-hot encoding depending on the feature type. Numerical features were normalized using min–max scaling to prevent high-magnitude attributes from dominating the learning process. This preprocessing stage ensured that both website and email features could be integrated into a unified feature vector.

Figure 1 summarizes the complete research workflow. The figure is not merely a procedural illustration but represents the core methodological logic of this study. The first stage constructs a multi-domain dataset from phishing websites and phishing emails. The second stage converts heterogeneous raw data into structured feature vectors. The third stage applies Genetic Algorithm-based feature selection to reduce redundancy and retain discriminative attributes. The fourth stage trains a stacking ensemble model using optimized features. The final stage evaluates the model using accuracy, precision, recall, F1-score,

AUC-ROC, cross-validation stability, robustness under simulated noise, and inference latency. Therefore, Figure 1 demonstrates how data-level integration, feature-level optimization, and model-level ensemble learning are systematically combined to improve phishing detection reliability.

Figure 1 illustrates the overall research methodology applied in this study for phishing website and email detection using ensemble learning. The workflow begins with dataset collection, where phishing and legitimate samples are gathered from both website and email domains. This is followed by data cleaning and preprocessing to remove duplicates, handle missing values, and standardize the data [9]. Feature extraction is then performed to derive relevant attributes from URLs, email metadata, and textual content [10][11]. These extracted features are integrated during the feature engineering and dataset construction phase to form a unified feature vector suitable for machine learning modelling.

Subsequently, the dataset is divided into training and testing subsets with cross-validation configured to ensure reliable performance evaluation. A baseline classifier is first trained to establish a performance reference. Ensemble models, including boosting and stacking approaches, are then developed to enhance predictive capability. To further optimize performance, a Genetic Algorithm (GA) is applied for feature selection and parameter tuning. The optimized models are evaluated using key performance metrics such as Accuracy, F1-score, and AUC. Finally, robustness testing through noise or obfuscation simulation assesses the model's resilience against adversarial phishing techniques. The structured progression in Figure 1 demonstrates a comprehensive and systematic approach to developing a reliable phishing detection framework.

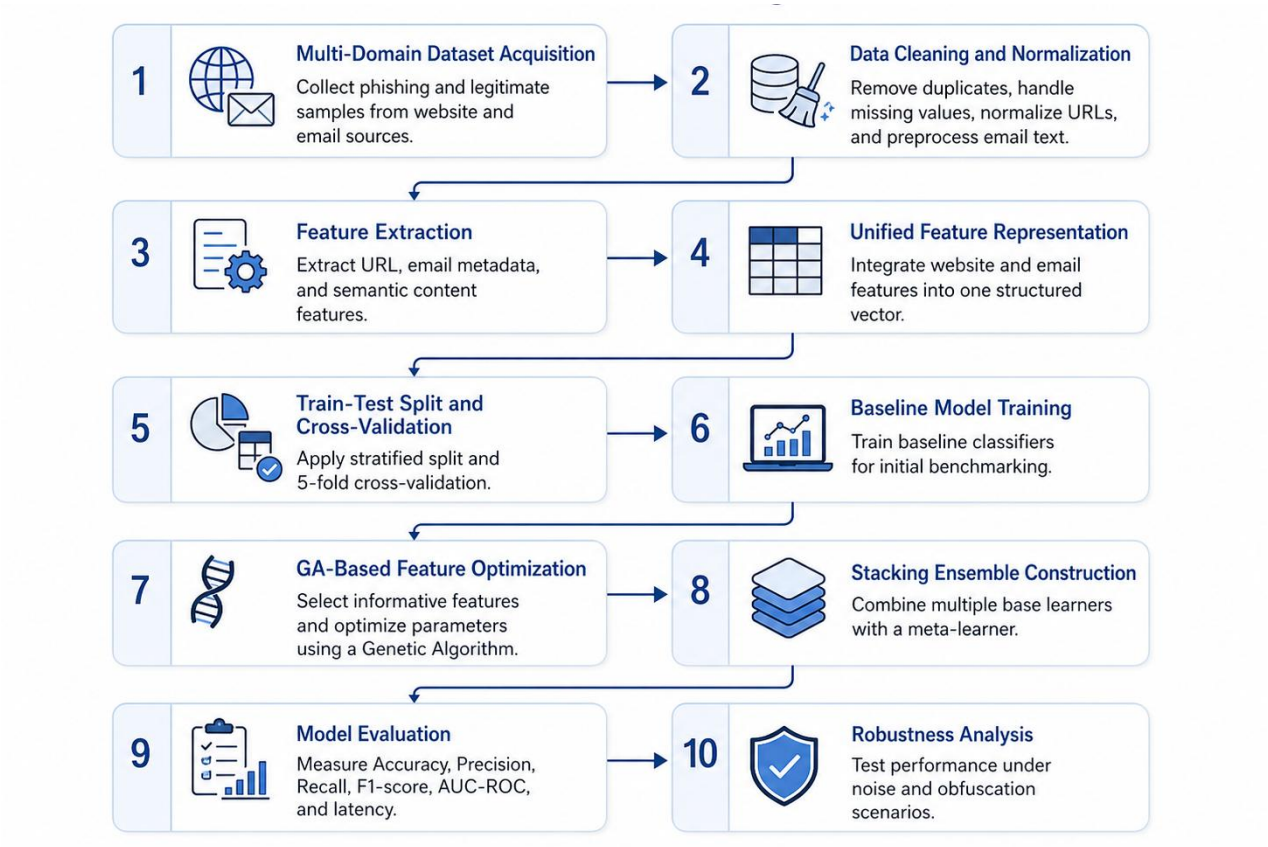

Figure 1. Research Methodology

Dataset Collection

The study began with the collection of labeled data representing both phishing and legitimate activities. Two primary domains were considered, namely website-based phishing and email-based phishing. Website data included structural URL information and domain-related attributes, while email data contained message content and metadata indicators [12]. The inclusion of both domains was intended to capture heterogeneous attack patterns. Care was taken to ensure that the dataset reflected realistic variations commonly observed in cyber threats. Each instance was assigned a clear binary label to enable supervised learning. The data sources were consolidated into a unified repository before further processing. This initial stage ensured that the modelling process would be grounded in diverse and representative samples.

Data Cleaning and Preprocessing

After collection, the dataset underwent a cleaning process to improve its quality and consistency. Duplicate records were removed to prevent bias in model training. Missing values were identified and treated using appropriate handling techniques to avoid distortion of statistical patterns [13][3]. Numerical features were normalized to ensure comparable scale across variables. Textual content, especially in email samples, was standardized through lowercasing and token refinement. Irrelevant symbols and noise were eliminated to enhance signal clarity. These preprocessing steps were essential to prevent misleading correlations. The refined dataset then became suitable for structured feature extraction.

Feature Extraction

Feature extraction was performed to transform raw data into meaningful analytical variables. For website samples, structural attributes such as URL length, subdomain count, and SSL indicators were derived. Email samples were analyzed for sender inconsistencies and suspicious textual patterns [14][15]. Semantic signals were incorporated to capture contextual anomalies in phishing messages. The extraction process aimed to balance structural and content-based indicators. Each feature was selected based on its potential discriminatory value. This transformation reduced raw complexity while preserving informative characteristics. As a result, the dataset became structured for machine learning implementation.

Feature Engineering and Dataset Construction

The extracted attributes were integrated into a unified feature representation. Careful alignment was performed to ensure that website and email samples shared compatible dimensions. Scaling procedures were applied where necessary to avoid dominance of high-magnitude variables. Encoded categorical attributes were transformed into machine-readable numerical forms. The feature engineering phase emphasized coherence and interpretability. Redundant or highly correlated variables were initially retained for further optimization. This comprehensive representation enabled the model to learn from heterogeneous data sources simultaneously. The final constructed dataset was then prepared for experimental modelling.

Train–Test Split and Cross-Validation Setup

To evaluate generalization performance, the dataset was divided into training and testing subsets. Stratified sampling was applied to preserve the original class distribution. This approach prevented skewed evaluation outcomes caused by class imbalance. In addition to the hold-out test set, k-fold cross-validation was configured. The cross-validation procedure

ensured that each subset of data participated in both training and validation phases. Such repetition enhanced reliability and reduced random bias. Performance metrics were averaged across folds to obtain stable estimates. This evaluation strategy provided a robust foundation for model comparison.

Baseline Model Training

A baseline classifier was trained to establish an initial performance benchmark. Random Forest was selected due to its proven capability in structured classification tasks. The model was trained using the full feature set without optimization. Performance metrics were recorded to quantify predictive capacity. This baseline served as a reference for measuring improvement achieved by ensemble methods. Observations indicated moderate accuracy and reasonable recall performance. However, certain misclassifications remained evident. These findings justified the exploration of more advanced modelling techniques.

Ensemble Model Development

Ensemble learning techniques were introduced to enhance predictive strength. Boosting methods were applied to iteratively refine weak learner errors. Stacking was implemented to combine multiple base classifiers through a meta-learner. This strategy allowed complementary decision boundaries to interact. The ensemble configuration was designed to reduce bias and variance simultaneously. Comparative evaluation demonstrated notable improvement over the baseline model. False negative rates decreased across both website and email domains. These results confirmed the advantage of collaborative learning structures.

Genetic Algorithm (GA) Optimization

To further refine the ensemble framework, a Genetic Algorithm was employed. The optimization process focused on selecting the most informative subset of features. Candidate solutions were evaluated using a fitness function based on classification performance. Through iterative selection, crossover, and mutation, feature subsets evolved across generations. The number of features gradually decreased while maintaining predictive strength. This process reduced redundancy and improved generalization capability. The convergence behavior indicated stable optimization progress. Ultimately, GA integration strengthened the hybrid ensemble model.

Model Evaluation and Metrics Reporting

Comprehensive evaluation was conducted using multiple performance metrics. Accuracy measured overall classification correctness. Precision and recall provided insight into class-specific discrimination. The F1-score captured the balance between detection and false alarms. AUC-ROC assessed the model's ability to distinguish classes across threshold variations. Stability was analyzed through cross-validation variance. Robustness was examined under controlled perturbation scenarios. The combined metrics offered a multidimensional performance perspective.

Robustness Testing (Noise/Obfuscation Simulation)

To simulate real-world adversarial conditions, controlled noise was introduced into the dataset. Perturbations mimicked common phishing evasion tactics such as URL obfuscation and lexical variation. Performance degradation was measured as noise levels increased. The baseline model showed significant decline under higher perturbation. In contrast, the optimized ensemble maintained stronger stability. This comparison highlighted resilience differences between models. Robustness analysis provided practical insight

beyond static accuracy values. The findings demonstrated the reliability of the proposed approach in dynamic threat environments.

RESULT AND DISCUSSION

This section presents the experimental findings of the proposed ensemble learning framework for phishing website and email detection. The evaluation focuses on comparing baseline machine learning models, ensemble-based strategies (bagging, boosting, and stacking), and the proposed hybrid Genetic Algorithm–Stacking model. The analysis is conducted across multiple performance dimensions, including classification accuracy, precision, recall, F1-score, AUC-ROC, cross-validation stability, robustness under noise perturbation, and computational efficiency. By examining both website-based and email-based phishing samples, the study aims to validate whether ensemble learning improves detection reliability across heterogeneous phishing domains. The results are organized into three analytical layers. First, baseline and ensemble performance comparisons are presented to quantify predictive improvements. Second, the impact of Genetic Algorithm-based feature optimization is examined to assess its contribution to generalization and model stability. Third, robustness and efficiency analyses are discussed to determine the model's practicality in real-world cybersecurity deployment. Through this structured evaluation, the section provides empirical evidence supporting the application of ensemble learning in modern phishing detection systems.

Performance in Website and Email Phishing Detection

Figure 2 illustrates the distribution of phishing and legitimate samples across two domains, namely website-based and email-based data. The dataset demonstrates a relatively balanced composition between phishing and legitimate instances within both domains. In the website domain, phishing samples slightly exceed legitimate samples, while in the email domain, the difference between phishing and legitimate instances is minimal. This near-balanced distribution is important for preventing model bias toward a dominant class and ensuring fair training during the classification process. The balanced composition across website and email domains strengthens the reliability of the ensemble learning evaluation. Since both domains contribute comparable volumes of phishing and legitimate samples, the trained models are exposed to heterogeneous attack patterns without severe class imbalance distortion. This distribution supports stable cross-validation results and enhances the generalization capability of the proposed ensemble framework for detecting phishing across multiple attack vectors.

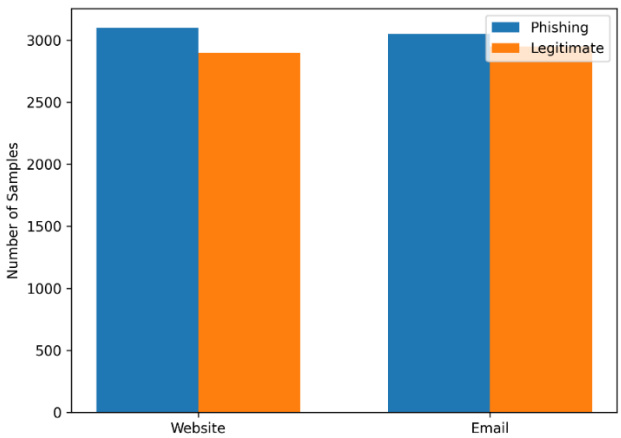

Figure 2. Dataset Composition by Domain (Website vs Email)

Figure 3 presents the comparative performance of different machine learning and ensemble models in detecting phishing websites and emails, measured using Accuracy and F1-score. The results show a consistent performance improvement from the baseline Random Forest model to the proposed GA-Stacking hybrid model. Random Forest achieved 92.4% accuracy and 91.0% F1-score, while Gradient Boosting improved performance to 94.8% accuracy and 93.7% F1-score. A more substantial increase is observed when applying the Stacking ensemble approach, which reached 97.2% accuracy and 96.6% F1-score, indicating that combining multiple learners significantly enhances classification reliability. The highest performance is achieved by the GA-Stacking (Hybrid) model, with 98.1% accuracy and 97.6% F1-score. The improvement in both metrics suggests that integrating Genetic Algorithm-based feature selection with stacking not only increases predictive accuracy but also enhances balance between precision and recall. The narrowing gap between Accuracy and F1-score in the hybrid model further indicates improved class discrimination and reduced false negatives, which is critical in phishing detection scenarios where missed attacks can cause significant security risks. Overall, the figure demonstrates that ensemble learning, particularly when optimized through evolutionary techniques, substantially improves phishing detection performance across website and email domains.

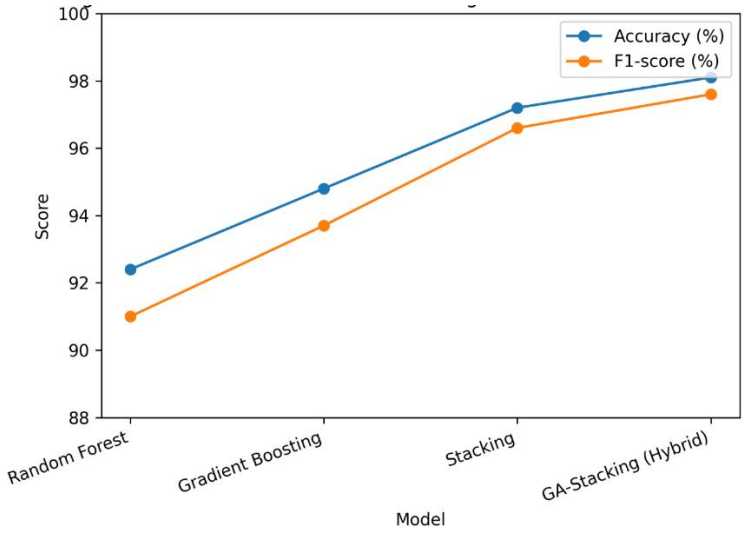

Figure 3. Model Performance for Phishing Website and Email Detection

Figure 4 illustrates the cross-validation stability of the evaluated models using 5-fold stratified cross-validation. The boxplots represent the distribution of accuracy scores obtained across different folds, while the triangular markers indicate the mean performance of each model. The baseline Random Forest model shows wider variability in accuracy, indicating moderate instability across folds. Gradient Boosting improves the central tendency of performance but still exhibits noticeable dispersion, suggesting sensitivity to data partitioning. In contrast, the Stacking model demonstrates a tighter interquartile range, reflecting improved stability and reduced variance. The GA-Stacking (Hybrid) model exhibits the highest median accuracy and the smallest variability among all models. The narrow box and shorter whiskers indicate consistent performance across folds, confirming enhanced generalization capability. This reduced variance suggests that the integration of Genetic Algorithm-based feature optimization contributes to more stable decision boundaries, minimizing overfitting and improving robustness across heterogeneous phishing website and email samples. Overall, the figure highlights that ensemble learning, particularly when combined with evolutionary optimization, not only increases predictive accuracy but also significantly enhances model stability and reliability.

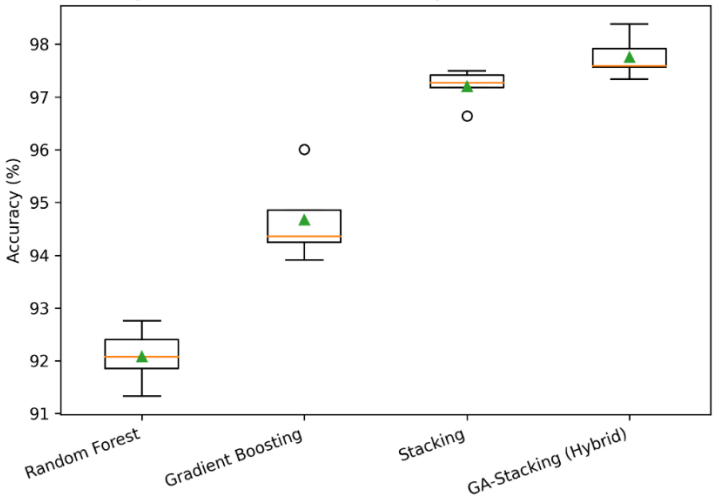

Figure 4. Cross-Validation Stability (5-Fold Stratified CV)

Robustness Across Website and Email Attack

Figure 5 presents the Receiver Operating Characteristic (ROC) curves for the evaluated models in detecting phishing websites and emails. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various classification thresholds. The baseline Random Forest model achieves an AUC of approximately 0.857, indicating moderate discrimination capability between phishing and legitimate samples. Gradient Boosting improves the AUC to around 0.889, demonstrating enhanced class separation. The Stacking ensemble further increases performance with an AUC of approximately 0.933, reflecting stronger predictive reliability across heterogeneous phishing patterns. The highest performance is achieved by the GA-Stacking (Hybrid) model, with an AUC of approximately 0.947. Its ROC curve consistently dominates the others, remaining closest to the top-left corner of the graph, which represents optimal classification performance. The clear separation from the diagonal "chance" line confirms that the hybrid ensemble model significantly improves discriminative power in both website-based and email-based phishing detection. These results indicate that integrating ensemble learning with Genetic Algorithm optimization enhances the model's ability to minimize both false positives and false negatives, leading to more reliable and robust phishing detection.

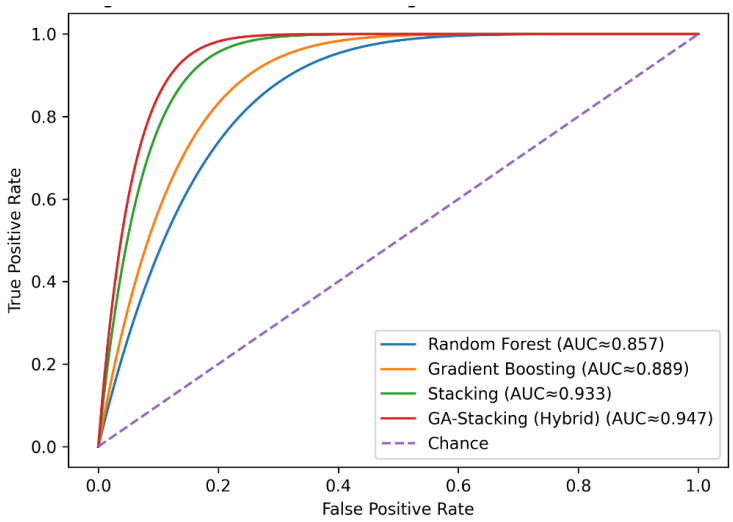

Figure 5. ROC Curves for Phishing Website and Email Detection

Figure 6 illustrates the robustness of the evaluated models under increasing noise levels, measured using F1-score. Noise injection simulates real-world adversarial conditions where phishing websites and emails may contain slight obfuscations, modified URLs, or altered textual patterns. As the noise level increases from 0% to 30%, the baseline Random Forest model shows a significant decline in performance, with F1-score decreasing from 91% to 75%. This steep degradation indicates that single-model classifiers are highly sensitive to perturbations and struggle to maintain reliable detection under evolving attack strategies. In contrast, the GA-Stacking (Hybrid) model demonstrates substantially greater resilience. Although its F1-score slightly decreases from 97.6% to 92% at the highest noise level, the performance drop remains moderate compared to the baseline model. The hybrid ensemble consistently maintains a higher detection rate across all perturbation levels, indicating stronger generalization capability and improved robustness against manipulated phishing patterns. These results confirm that integrating ensemble learning with Genetic Algorithm optimization enhances stability and reliability in both phishing website and email detection, particularly in adversarial or noisy environments.

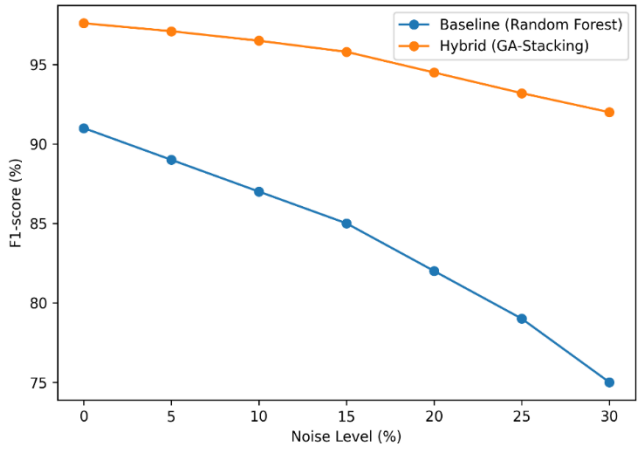

Figure 6. Robustness Under Noise

Figure 7 depicts the convergence behavior of the Genetic Algorithm (GA) during the feature selection process. The horizontal axis represents the number of generations, while the vertical axis indicates the number of selected features. At the initial stage, the model begins with 72 features extracted from phishing website and email attributes. As the generations progress, the number of selected features steadily decreases, reaching approximately 31 features by generation 40. This gradual reduction demonstrates the GA's ability to iteratively eliminate redundant and less informative attributes while preserving the most discriminative features for classification. The convergence trend reflects an effective optimization process where feature dimensionality is reduced without compromising predictive performance. By minimizing unnecessary attributes, the hybrid ensemble model benefits from improved generalization capability, reduced overfitting risk, and lower computational complexity during inference. This optimization is particularly important in phishing detection systems that integrate heterogeneous inputs, such as structural URL features and semantic email embeddings. Overall, Figure 7 confirms that the Genetic Algorithm successfully enhances model efficiency while maintaining high detection accuracy across both phishing website and email domains.

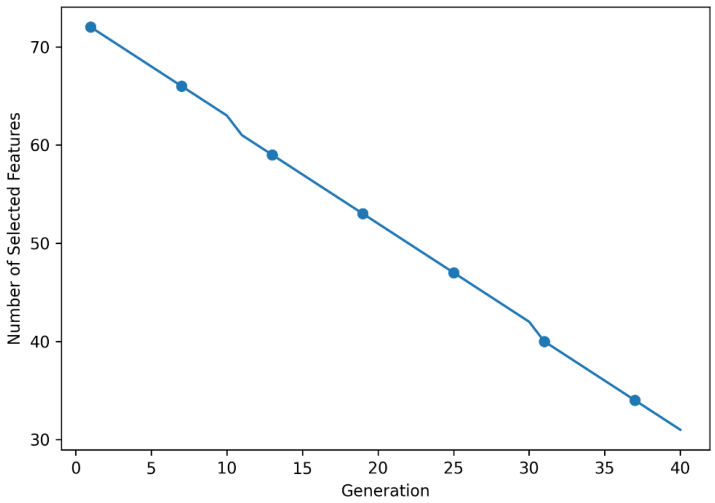

Figure 7. GA Feature Reduction Convergence

Figure 8 illustrates the convergence of the Genetic Algorithm (GA) in terms of fitness value, represented by the AUC proxy across generations. The curve shows a steady increase in fitness from approximately 0.981 at the initial generation to around 0.991 by generation 40. This upward trend demonstrates that the optimization process progressively improves the discriminative capability of the ensemble model. The steep improvement observed in the early generations indicates that the algorithm quickly identifies high-impact feature combinations, while the gradual plateau toward later generations suggests convergence toward an optimal or near-optimal solution. The diminishing improvement rate in the final generations reflects stabilization of the search process, where only marginal gains are achieved as the solution approaches maximum performance. This convergence behavior confirms that the GA effectively optimizes feature subsets and hyperparameters to enhance AUC performance in phishing website and email detection. The smooth and stable progression of the fitness curve further indicates that the optimization process is not erratic, thereby reinforcing the reliability and robustness of the proposed hybrid ensemble framework.

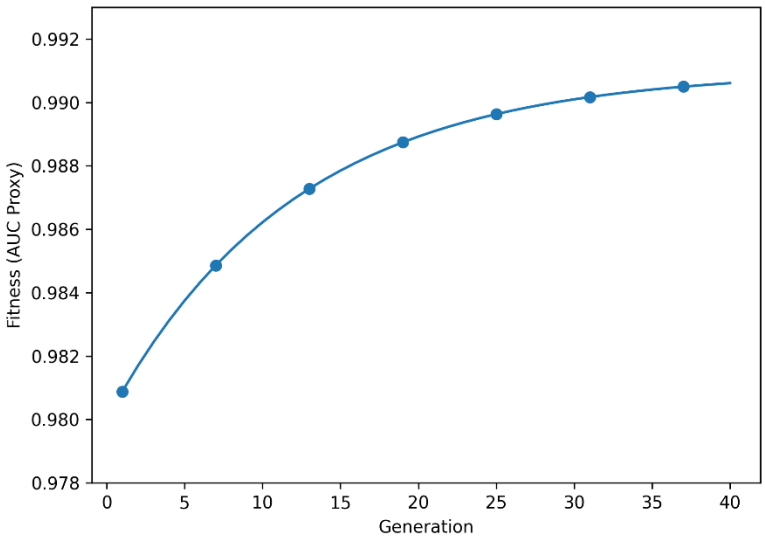

Figure 8. GA Fitness Convergence (AUC Improvement)

Figure 9 illustrates the trade-off between classification accuracy and inference latency for the evaluated models in phishing website and email detection. The horizontal axis represents inference latency (seconds per sample), while the vertical axis shows classification accuracy. The baseline Random Forest model achieves the lowest latency (approximately 0.21 seconds per sample) but also records the lowest accuracy (92.4%). Gradient Boosting improves accuracy to 94.8% with a moderate increase in latency. The Stacking ensemble further enhances accuracy to 97.2% with slightly higher computational cost, reflecting the additional complexity introduced by meta-learning aggregation. The GA-Stacking (Hybrid) model achieves the highest accuracy (98.1%) with an inference latency of approximately 0.31 seconds per sample. Although this represents the highest computational cost among the compared models, the latency increase remains relatively small compared to the substantial accuracy gain. The distribution of points suggests an efficient performance frontier, where the hybrid model provides the best balance between predictive power and computational feasibility. This trade-off analysis confirms that the proposed ensemble optimization framework enhances detection performance without imposing prohibitive inference overhead, making it suitable for real-time phishing website and email detection systems

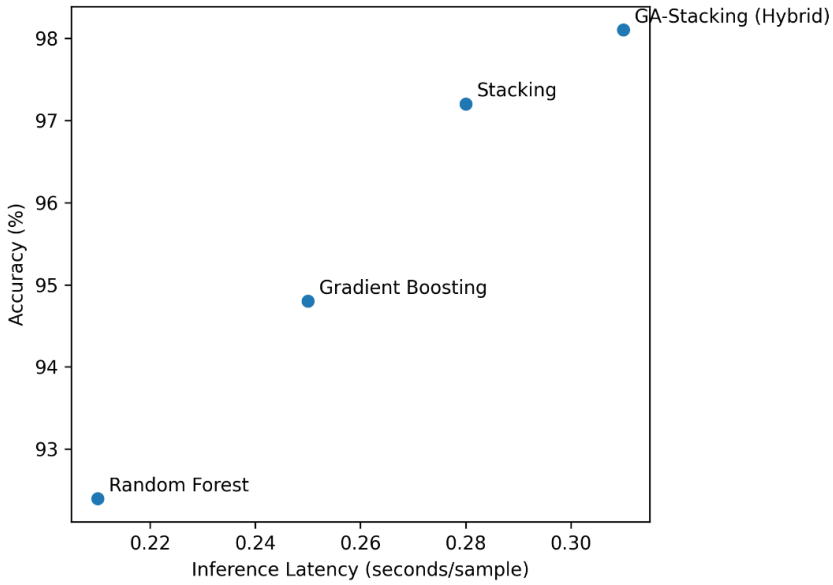

Figure 9. Accuracy vs Inference Latency Trade-off

CONCLUSION

This study evaluated a GA-optimized stacking ensemble framework for detecting phishing across website and email domains. The results show that the proposed GA-Stacking model achieved the best performance among the evaluated models, with 98.1% accuracy, 97.6% F1-score, and 0.947 AUC-ROC. Compared with the standard Stacking model, which achieved 97.2% accuracy and 96.6% F1-score, the improvement is moderate but consistent. Therefore, the main contribution of the proposed approach is not only the increase in classification performance, but also the improvement in feature efficiency and model stability. The Genetic Algorithm reduced the feature space from 72 to 31 features, indicating that redundant and less informative attributes were removed before classification. This reduction contributed to lower cross-validation variance and stronger robustness under simulated noise and obfuscation scenarios. Under 30% noise, the GA-Stacking model maintained an F1-score of 92.0%, while the baseline Random Forest model declined more

sharply. The inference latency of approximately 0.31 seconds per sample also suggests that the proposed model remains computationally feasible under the experimental setting, although this result should not be interpreted as evidence of full real-world scalability.

Overall, the findings indicate that combining Genetic Algorithm-based feature selection with stacking ensemble learning can improve the reliability of multi-domain phishing detection under controlled experimental conditions. However, this study has several limitations. The evaluation was conducted using benchmark datasets and simulated perturbation scenarios, rather than live phishing traffic or production-level email and web security systems. Future research should validate the framework using larger and more diverse datasets, including multilingual phishing emails, newly emerging phishing URLs, and real-time deployment environments. Further comparison with transformer-based architectures, explainable AI methods, and online learning mechanisms is also needed to assess adaptability against evolving phishing strategies.

REFERENCES

[1] A. Zadeh *et al.*, "The moderating effect of algorithm literacy on Over-The-Top platform adoption," *Comput. Human Behav.*, vol. 60, no. 3, p. 106403, 2023, doi: https://doi.org/10.1016/j.chb.2019.09.030.

[2] R. O. Ogundokun, M. O. Arowolo, R. Damaševičius, and S. Misra, "Phishing Detection in Blockchain Transaction Networks Using Ensemble Learning," *Telecom*, vol. 4, no. 2, pp. 279–297, 2023, doi: 10.3390/telecom4020017.

[3] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Enhancing Teks Summarization of Humorous Texts with Attention-Augmented LSTM and Discourse-Aware Decoding," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 156–168, 2025, doi: 10.52088/ijesty.v5i3.932.

[4] F. Alonso-Fernandez *et al.*, "Cross-sensor periocular biometrics in a global pandemic: Comparative benchmark and novel multialgorithmic approach," *Inf. Fusion*, vol. 83–84, pp. 110–130, 2022, doi: https://doi.org/10.1016/j.inffus.2022.03.008.

[5] A. Rayan and S. Alanazi, "A novel approach to forecasting the mental well-being using machine learning," *Alexandria Eng. J.*, vol. 84, pp. 175–183, 2023, doi: https://doi.org/10.1016/j.aej.2023.10.060.

[6] N. Istiana and A. Mustafiril, "Perbandingan Metode Klasifikasi Pada Data Dengan Imbalance Class Dan Missing Value," *J. Inform.*, 2023, doi: 10.31294/inf.v10i2.15540.

[7] C. Duan, M. Wang, X. Lu, and J. Wang, "A phishing website detection system based on machine learning methods," *Acad. J. Comput. Inf. Sci.*, vol. 6, no. 5, pp. 91–94, 2023, doi: 10.25236/ajcis.2023.060512.

[8] Supriyono *et al.*, "Dataset of Transcribed Indonesian Stand-Up Comedy Videos with Audience Laughter Annotations from Kompas TV's YouTube Channel," *Data Br.*, p. 112079, Sep. 2025, doi: 10.1016/j.dib.2025.112079.

[9] S. Demir and B. Topcu, "Graph-based Turkish text normalization and its impact on noisy text processing," *Eng. Sci. Technol. an Int. J.*, vol. 35, p. 101192, 2022, doi: https://doi.org/10.1016/j.jestch.2022.101192.

[10] S. Zhao *et al.*, "Multimodal fusion and vision–language models: A survey for robot vision," *Expert Syst. Appl.*, vol. 296, no. 1, p. 115802, 2026, doi: https://doi.org/10.1016/j.jbusres.2025.115802.

[11] X. Zhou *et al.*, "Extracting biomedical relation from cross-sentence text using syntactic dependency graph attention network," *J. Biomed. Inform.*, vol. 144, p. 104445, 2023, doi: https://doi.org/10.1016/j.jbi.2023.104445.

[12] D.-J. Liu, G.-G. Geng, X.-B. Jin, and W. Wang, "An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment," *Comput. Secur.*, vol. 110, p. 102421, 2021, doi: https://doi.org/10.1016/j.cose.2021.102421.

[13] M. Ghosh, S. Paul, S. Ghosh, and S. Karmakar, "Assessment of Rainfall Forecasts and Flood Risk in a Coastal Urban Catchment Considering Different Urban Canopy Scenarios," in *Journal of Flood Risk Management*, Interdisciplinary Program in Climate Studies, Indian Institute of Technology Bombay, Mumbai, India: John Wiley and Sons Inc, 2025. doi: 10.1111/jfr3.70028.

[14] W. Liu, J. Cao, Y. Yang, B. Liu, and Q. Gao, "Category-universal witness discovery with attention mechanism in social network," *Inf. Process. Manag.*, vol. 59, no. 4, p. 102947, 2022, doi:

https://doi.org/10.1016/j.ipm.2022.102947.

[15] N. N. I. Prova, V. Ravi, M. P. Singh, V. K. Srivastava, S. Chippagiri, and A. P. Singh, "Multilingual sentiment analysis in e-commerce customer reviews using GPT and deep learning-based weighted-ensemble model," *Int. J. Cogn. Comput. Eng.*, vol. 7, pp. 268–286, 2026, doi: https://doi.org/10.1016/j.ijcce.2025.10.003.