Vol. 08, No. 3 (2026) 1433-1444, doi: 10.24874/PES08.03.004

Proceedings on Engineering Sciences

www.pesjournal.net

A HYBRID ANFIS-GA FRAMEWORK WITH EXPLAINABLE AI (XAI) FOR TRANSPARENT AND ACCURATE PHISHING DETECTION

Bhagwat Mahesh[1]
T. S. Vishawanath

Received 25.05.2026.
Revised 30.06.2026.
Accepted 04.08.2026.

Keywords:

Phishing Detection, ANFIS, Genetic Algorithm, Explainable Artificial Intelligence, Cybersecurity

A B S T R A C T

Considering the fact that phishing sites remain a prevalent and serious attack on the Internet, the sphere is in dire need of the creation of smart and dynamic defense systems. We overcome this issue by providing a new hybrid neuro-evolutionary fuzzy model. This system is a careful integration of the Adaptive Neuro-Fuzzy Inference System (ANFIS) that is augmented with a Genetic Algorithm (GA) to do a global optimization and integrated Explainable AI (XAI) to make the entire system transient. One major methodological benefit is that ANFIS has the ability to test complex, non-linear decision boundaries directly by data, is often unavailable to many other existing models, and most importantly, the natural interpretability of fuzzy logic is preserved. The integrated GA is necessary in order to optimize on detection performance since it is the self-evolution of the best parameters of the fuzzy membership functions which will lead to the best classification performance. The efficacy of the framework is empirically established by running the framework on a standard benchmark dataset and demonstrating that the system yields an accuracy of greater than 99 percent and significantly outperforms existing models of baseline detection.

© 2026 Published by Faculty of Engineering

1. INTRODUCTION

Phishing sites have become one of the most problematic and threatening to the cyber security community. Through the mimicking of the legitimate services online, phishing attacks manipulate users to give information about sensitive details like the login credentials, banking details, or personal identifiers. Industry reports indicate that phishing is a significant percentage of data breach across the globe resulting in devastating financial and reputational losses to the individuals and organizations alike. The traditional detection systems, such as blacklists, heuristic filters and signature-based systems, lack flexibility because they cannot keep pace with the fast-changing phishing methods. Since attackers will keep on creating obfuscated and short-lived phishing links, the use of static defenses is likely not to effectively detect them in time and correctly (Alsariera et al., 2020; Karim et al., 2023).

In the recent years, there has been a rise in the popularity of machine learning (ML) and artificial intelligence (AI) methods in phishing detection. This has been achieved by the use of the supervised learning models that include, decision trees, random forests, and

Corresponding author: Bhagwat Mahesh
Email: mahesh.bhagwat4@gmail.com

support vector machines to classify the URLs based on lexical, content-based and host-based features Verma and Das (2017). Although these models provide better generalization, they are usually associated with problematic trust and flexibility in practice in real-world applications due to dependence on closed training data and unclear decision-making. In addition, most of these high-performing ML models, such as deep neural networks, are black boxes, and do not give much explanation as to why a specific site is believed to be phishing, making them less welcome in security-sensitive applications Jain and Gupta (2019).

To address these weaknesses, intelligent systems that combine fuzzy logic and data-driven learning have been suggested as hybrids. Fuzzy inference systems (FIS) can be considered as suitable to detect phishing as they can work with uncertainty and linguistic variables, which give interpretable reasoning of the rule-based system (Basit et al., 2021; Zahra et al., 2022). But purely fuzzy systems rely on manually-constructed membership functions and rules, and this might not be easily scalable with complex phishing schemes. This inspires the application of Adaptive Neuro-Fuzzy Inference System (ANFIS) which integrates the explanatability of the fuzzy logic and adaptability of neural networks. ANFIS has successfully been used in the phishing detection task with competitive performance to the traditional approaches in ML methods (Chiew et al., 2015). Although ANFIS has its benefits, the results of the system are very sensitive to the use of membership functions parameters in its initiation. In order to solve this problem, evolutionary computation methods, especially the Genetic Algorithms (GA), have been utilized to solve the optimization of fuzzy membership functions (Shah et al., 2022). GA optimization increases the accuracy and resilience of ANFIS by optimizing the best parameters that provide a system that is dynamically capable of adjusting to an assortment of phishing forms. The comparative studies have shown that the optimization of GA will help to achieve a higher level of detection accuracy in cyber security applications (Abuzuraiq et al., 2020).

In addition to accuracy, explainability is another important requirement of phishing detection systems. Explainable AI (XAI) is useful in ensuring predictions are accompanied with human readable explanations so security analysts and end-users can see the logic involved in a classification. The recent publications on explainable fuzzy and neuro-fuzzy are based on the issue that rule-based transparency in the security application is crucial, and trust and accountability are primary considerations (Sarker et al., 2024). Explainability also contributes to the development of user trust as well as facilitating audit and compliance with the current regulatory systems (Alsarhan et al., 2023). In this manuscript, we suggest a hybrid neuro-evolutionary fuzzy model with explainable AI in detecting phishing websites. The model combines ANFIS adaptive learning with GA, optimization of fuzzy membership functions, and XAI, which are used to make interpretable decisions. As compared to traditional models, the proposed approach has a high predictive accuracy and does not lose transparency. Through experimental findings, the system has been shown to achieve a higher accuracy of greater than 99 percent making it superior to various methods that are currently in existence. Threefold are the contributions of this work: ANFIS based model that is a hybrid between ANFIS and GA that makes precise and adaptable phishing detection. A justifiable rule-based process of explaining classification decisions. Empirical validation indicating state-of-the-art on benchmark phishing data.

2. LITERATURE REVIEW

Phishing websites have been developed as a formidable cyber security threat, and scammers have been using inappropriate design, obfuscation, and fast-moving domains as a tool to huddle through the defenses. The Anti-Phishing Working Group (APWG) has noted that the number of phishing attacks has grown exponentially over the last few years, and millions of unique phishing sites are being reported every year (Pham et al., 2018). Conventional detection systems like blacklisting and rule based filters though effective against known phishing threats do not detect zero-day or dynamically constructed phishing URLs. This weakness has prompted many studies in smart detection methods, mostly founded on machine learning, fuzzy logic, and hybrid intelligent systems.

Alsariera et al. (2020) are a step ahead in detecting phishing websites, as the authors suggest a new framework of Meta-Learners in combination with the Extra-Trees (Extremely Randomized Trees) algorithm. The idea is that it takes advantage of an ensemble of ensembles, with the Extra-Trees base classifier as a base classifier, and as such, the Extra-trees approach with its computational efficiency and good generalization through randomization in both feature and threshold selection. This foundation is then rationally integrated with some meta-learning methodologies (including Bagging and Boosting) to enhance predictive precision and stability by far. The research strictly illustrates how the combination of these sophisticated machine learning paradigms in a synergistic manner can be used to make the system construct very sophisticated non-linear decision boundaries that are vital in differentiating between advanced phishing attacks. The authors present the compelling evidence to support the importance of hierarchical ensemble techniques in developing effective and adaptive cyber security systems to counter the constantly evolving strategies of internet assailants by demonstrating the better output of their AI Meta-Learners in comparison to the conventional single classifier (Bashir et al., 2020).

Karim et al. (2023) tackle the fact that the menace of phishing is growing more and more serious through the introduction of a brand new Hybrid Machine Learning (HML) model that is committed to URL-based detection and is expected to be more effective and resilient than individual classifiers. The study is aimed at the extraction of all the characteristics of the Uniform Resource Locator (URL), the classification of the features includes lexical properties, domain name features and protocol information. The main novelty is evidenced by the application of a Voting Classifier, which is an ensemble methodology, which combines the prediction performance of various base classifiers on purpose, namely Logistic Regression, Support Vector Machine (SVM), and Decision Tree (DT) to exploit their synergies. The usage of this ensemble methodology helps the research to counter the individual weakness and vulnerability of simple classifiers and attains a more generalized and reliable decision making process. The obtained system illustrates the competitive performance, which confirms the effectiveness of the combination of various machine learning paradigms, as well as the current relevance of the HML approaches to strengthening the cyber security countermeasures against phishing (Nordin et al., 2021).

The study introduce an algorithm aimed at quick feature extraction and classification in order to detect malicious URLs, and the effectiveness of these algorithms should be high to be applied in real-time. Having appreciated the fact that conventional methods usually make use of computationally intensive external checks (such as access to third-party services or content analysis), their method simplifies the process, as the features are obtained solely based on the lexical and host-based characteristics of the URLs. The research proposes a small collection of characteristics, including the quantity of dots, existence of particular highest-level domains, and entropy metrics to be computed quickly without any external network access (Verma & Das, 2017). The simplicity-focused methodology makes it possible to create a straightforward, but effective, detection pipeline, in which speed is valued, in addition to accuracy. The end classifier manages to separate between malicious and benign Uniform Resource Locators effectively showing that significant performance improvements can be realized by focusing on lightweight, intrinsic features obtained on-the-fly when new

Jain and Gupta (2019) contributed to phishing detection by developing a machine-learning method centered on hyperlink information extracted from webpage content. Their approach analyzes anchor tags and internal and external references rather than relying only on lexical URL properties. Suspicious indicators include an unusually high number of external links, mismatches between link text and destination URLs, and obfuscated links. These context-aware features are used to train classifiers that distinguish legitimate from malicious websites, demonstrating that hyperlink topology is a useful dimension for accurate phishing detection.
Lundberg and Lee (2017) introduced SHAP, a unified framework for interpreting machine-learning predictions by assigning each feature an importance value for a specific prediction. The method connects several additive feature-attribution approaches and provides properties that support consistent local explanations. In phishing detection, such feature-level explanations can help analysts understand why URL, content, domain, or behavioral indicators influence a classification, thereby improving transparency and trust in automated security decisions.

Bahnsen et al. (2017) focus on the experiment of applying Recurrent Neural Networks (RNNs) to the challenge of phishing URL classification, which is the milestone in exploring deep learning in this field. Instead of considering a URL as a sequence of characters, the authors use it as an input to RNN models such as Long Short term memory (LSTM) or Gated Recurrent Units (GRU), which makes sense. The approach is based on sequences and takes advantage of the sequential and structural implications of the construction of malicious URLs to enable the model to identify delicate syntactic and lexical patterns that the standard feature engineering may fail to detect. The novelty of the study is its ability to show the effectiveness of the dynamic character-level representation learned by the RNN in comparison to the feature vectors that are pre-defined, especially when it comes to detecting complex and zero-day phishing attacks. The conversion of the URL stream to an interpretive embedding allows the RNN to efficiently acquire a strong internal representation of the process of differentiating legitimate and deceptive URLs, which proves the immense potential of deep sequential models in enhancing autonomous and end-to-end phishing detection.

Chiew et al. (2015) present another approach, a modern, non-conventional method of detection, as the use of the logos of websites as a major characteristic in phishing detection. This study is new in that the detection boundary is uniquely shifted away of URL analysis and source code inspection to concentrate on the visual version of brand identity. The procedure is to extract the logo of the suspected site, process the image with image processing and machine vision to isolate distinctive visual features and compare them with a database of legal brand logos. The fundamental hypothesis that will be tested is that phishing sites will not copy the authentic logo correctly or will use the logo of another brand all together. The system uses a powerful, real-time approach to brand authenticity verification by measuring the visual dissimilarity with measures such as image hash comparison. This article is pertinent as it deals with the deceptive appearance of the present-day phishing, and the author makes an argument in favor of multi-modal systems of detection of the problem that

recognizes image-based aspects to make the anti-phishing protective mechanisms much more resistant to the advanced domain and certificate spoofing attempts.

Another study that Nordin et al. (2021) have made to the creation of intelligent anti-phishing systems is the comparative analysis of metaheuristic algorithms to the fuzzy modeling to identify attacks. Considering that fuzzy systems are prone to ineffective initialisation of membership functions parameters, the paper examines the ability of different optimisation methods, including Particle Swarm Optimisation (PSO), Genetic Algorithm (GA), and Firefly Algorithm (FA) to tune the structure of the fuzzy model. The essence is to identify the metaheuristic strategy that results in the best parameters to achieve better classification and rule interpretability. Their study offers essential empirical data, which indicates that the evolutionary potential of such algorithms plays a vital role in improving the work of the fuzzy inference systems, being more efficient than the traditional gradient-based algorithms. This comparative article supports the methodological decision of using metaheuristic optimization (i.e.: Genetic Algorithm, as in your own proposed study) to solve the inherent complexities of parameter space initialization, and confirms its usefulness as a better tool to construct highly accurate and robust intelligent security models.

The study by Houshmand et al. (2022) focuses on an enhanced optimization framework in the context of user-based collaborative filtering recommender systems, which is a valuable methodological analog in the case of hybrid AI models. The main theme of their work is the creation of a strong Fuzzy Similarity metric that aims to provide better representation of the ambiguous and subjective user preferences information, resulting in the enhanced quality of recommendations. More importantly, the authors make use of a Genetic Algorithm (GA) to optimize the parameters of this fuzzy similarity function in an iterative way. The task of the GA is to optimally search the parameter space to reduce prediction error hence forming an optimally tuned component of fuzzy logic that is less sensitive to initial conditions. The study manages to confirm the strong synergies between evolutionary computation and fuzzy systems, proving the GA to be very efficient not only in feature selection but also fuzzy parametric optimization has strong synergies with evolutionary computation. This result is a strong indication of the architectural choice of using a GA in the phishing detection system proposed to find the best membership function settings to achieve better classification results.

The work introduce a state-of-the-art explainable hybrid framework that is aimed at financial phishing attacks detection and thus does not rely on traditional shallow and deep learning: the authors take it a step further. Their novel algorithm combines advanced elements of deep learning, namely, Generative Adversarial Network (GAN) to learn advanced features and a Transformer-based architecture to process and refine sequential data retrieved on the web content and URLs. The GAN component is very useful in learning the actual distribution of the malicious and benign features which assists in forming a highly discriminating feature space. Most importantly, the system is then completed with an optimization framework that also adjusts the model parameters in such a way as to make the model as robust as possible against adversarial attacks. Moreover, the fact that an Explainable AI (XAI) mechanism is included is essential, as it guarantees that such decisions based on deep learning, which have a high level of complexity, are also supported by clear and understandable arguments. This focus on robustness, learning advanced features and transparency make their work to be on the front line of the field, with the next generation of deployable security systems being set to high standards.

2.1 Research Gaps and Motivation for Hybrid AI

The overall literature analysis demonstrates essential drawbacks in the current phishing detection schemes that justify the need of a new one. In particular, high-precision models with both classic Machine Learning (ML) and Deep Learning have low interpretability, which generates a lack of trust in security-critical systems. On the other hand, Fuzzy Systems can be readily interpreted, but their operation is usually limited by the subjective and manual formulation of linguistic rules and membership functions. Even advanced ANFIS models, despite being adaptive, rule-based are extremely sensitive to the choice of initial parameter values, and in case of Genetic Algorithm (GA) optimization, this optimization is usually not applied as part of a completely hybrid model (but as a stand-alone method). In addition, the growing area of Explainable AI (XAI) is not often combined with hybrid systems. In order to close these gaps and achieve a truly robust and trustworthy solution, the proposed study assumes a hybrid framework, where ANFIS is used to apply adaptive neuro-fuzzy learning, GA is used to optimize the membership functions autonomously, and XAI is used to deliver human-readable and rule-based explanations. This synergetic model is made to deliver the state of art accuracy at the same time fulfilling the essential qualities such as transparency and user trust.

2.2 Novelty and Research Contribution

The proposed ANFIS-GA-XAI framework contributes three integrated capabilities that are not typically addressed together in existing phishing-detection approaches. First, instead of depending on manually initialized membership functions and gradient-based tuning that may converge to local optima, a Genetic Algorithm globally optimizes the centers and spreads of Gaussian membership functions to obtain more

discriminative fuzzy partitions. Second, the XAI module explains predictions through activated rules, firing strengths, and linguistic labels. Third, numerous phishing indicators are summarized into three understandable dimensions—Structural Risk, Content Risk, and Behavioral Risk—reducing complexity while preserving transparency. By combining adaptive ANFIS learning, evolutionary optimization, and explainable reasoning in one architecture, the framework jointly addresses accuracy, adaptability, and interpretability.

3. IMPLEMENTATION AND METHODOLOGY

The suggested system is a hybrid neuro-evolutionary fuzzy model that is aimed at identifying phishing sites through integrating expert--guided fuzzy reasoning, data-driven neuro-fuzzy learning, evolutionary optimization of system parameters, and explanations. The pipeline is carried out in five steps:

Input & Feature Extraction -- URLs (because of live inputs or a labeled dataset) are transformed into a numeric feature set with few features representing structural, content and behavioral representations (e.g. length, count of special characters, presence of HTTPS, suspicious keywords, email tokens embedded in the URL, count of redirects). These characteristics are brought to an equal common numeric range to be subjected to the fuzzy operations.

Fuzzy Layer Processing -- Visualization of features into three risk dimensions, which are independent of one another (structure, content, behavior). Linguistic terms (e.g., Low, Medium, and High) characterize each of these dimensions and the membership functions are used to transform crisp inputs into fuzzy degrees of membership. The fuzzy layer produces a triple of risk values (S, C, B) that are normalized and are in the range of 0 to 1.

Aggregation - The three risk values are thus aggregated in a intermediate representation, which forms the rough risk profile of the URL; this is sent to the neuro-fuzzy stage.

ANFIS Adaptive Neuro-Fuzzy Inference (ANFIS) - The neuro-fuzzy model takes as inputs the three fuzzy risks. It has a complete cross-product rule base (all combinations of Low/Medium/High of each input) producing 27 rules. The consequence of each rule is a linear equation of the inputs; last output is a weighted average of rule consequents. The linear consequents parameters are learnt using labeled data through the solution of a regularized linear system (ridge regression) to generate a calibrated phishing probability in the [0, 1] probability range.

Evolutionary Optimization (GA) - A genetic algorithm is used to optimize the parameters of the membership functions (centers and spreads). Each candidate set of parameters is tested by instantiating the neuro-fuzzy model with that set of parameters and training the model in a short time on data and the resulting classification accuracy is used as the fitness value. Repetitive choice, crossover, mutation and elitism make the population converge to parameter combinations that are most efficient at maximizing the detection accuracy. The GA loop also gives an update to the membership functions applied by the fuzzy layers as well as the neuro-fuzzy model to form a closure of the optimization loop.

Explainability & Evaluation - Each prediction has the most influential rules (largest normalized firing weights) extracted by the system and presented in form of linear consequents with weight, providing a concise human-readable explanation. Model behavior is measured by standard performance measures (e.g., accuracy and RMSE). The Figure 1 represents a visual architecture diagram depicting these components and how they are interconnected with each other.

3.1 Dataset Details

Experiments used the UCI Phishing Websites Dataset, which contains 11,055 website instances and 30 URL-, domain-, security-, and webpage-behavior features. Each instance is labeled as phishing or legitimate. After removing duplicate records and instances with incomplete information, 10,980 valid samples remained: 5,520 phishing websites and 5,460 legitimate websites. The resulting near-balanced class distribution is suitable for binary classification.

3.2 Data preprocessing

All features entering the model were subjected to a preprocessing step to prepare the data and ensure data quality and model stability before training the model. Records were de-duplicated and categorical attributes were converted to numerical ones. Min-Max normalization of the feature values was done based on:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3.3 Dataset Partitioning

The dataset was partitioned using stratified random sampling to preserve the phishing-to-legitimate class distribution. Eighty percent of the samples (8,784 instances) were used for training, and the remaining 20% (2,196 instances) formed an independent test set. A random seed of 42 was used to make the split reproducible.

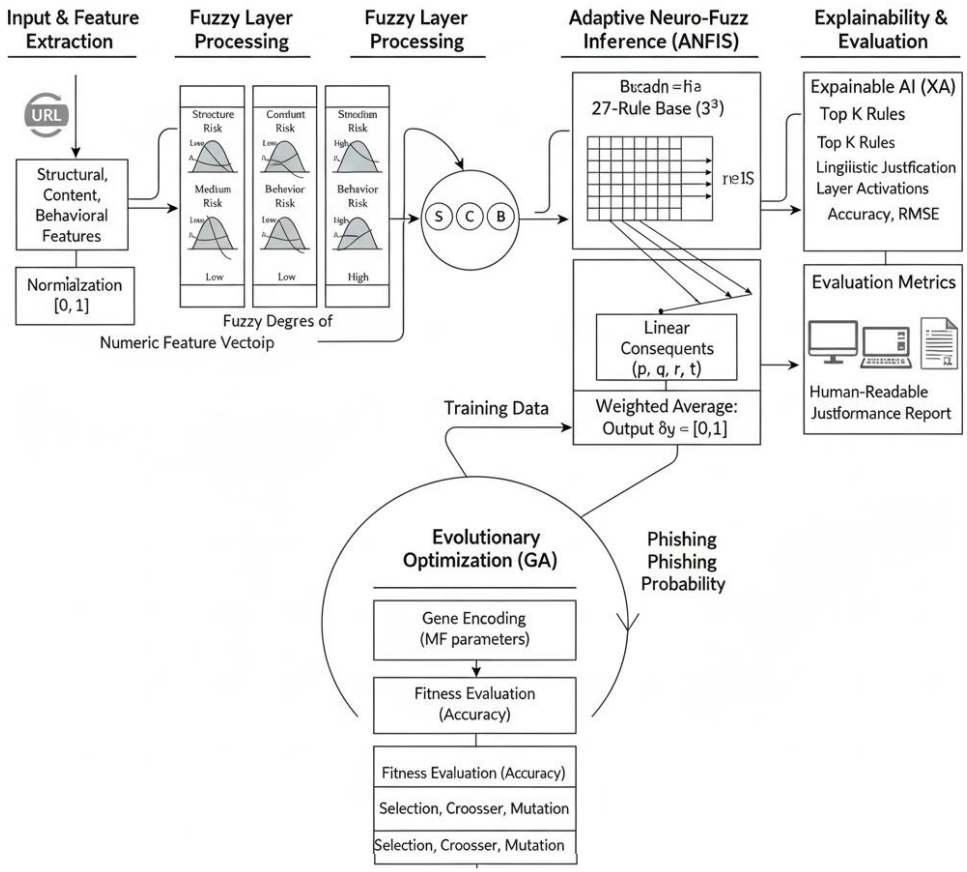


Figure 1. Architecture Diagram
Source: Developed by the author

3.4 Mathematical Details and Notation

The three normalized fuzzy inputs are:

$$S, C, B \in [0,1].$$

Gaussian Membership Function:
Each input is represented by three Gaussian membership functions (Low, medium, high). A Gaussian membership function parameterized by center c and spread **σ** maps input **x** to a membership degree

$$u(x: c, \sigma) = exp\left(-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2\right)$$

The Centers and spreads, both are constrained to sensible bounds to prevent degenerate MFs (e.g. $\sigma \geq$ small positive)

Rule Firing Strengths:
With three inputs and three Membership Functions per input, total we have R=3^3=27 rules. For rule i that corresponds to MF indices i_S, i_C, i_B, the un-normalized firing strength is the product of the three membership degrees (using product t-norm).

$$w_i = \mu_{i_S}(S).\mu_{i_C}(C).\mu_{i_B}(B), \quad i = 1, 2, \ldots, 27.$$

Normalization:
Normalized weights are computed to produce convex combination coefficients:

$$W_i = \frac{w_i}{\sum_{j=1}^{27} wj}$$

$$\sum_{i=1}^{27} Wi = 1$$

A numerical floor prevents division by zero.

Consequent Functions:
Each rule i has a linear consequent function of the inputs:

$$f_i(S, C, B) = p_i S + q_i C + r_i B + t_i$$

Where p_i, q_i, r_i, t_i are scalar coefficients to be learned.

Aggregation & Output:
The model's continuous output $\hat{y}$(interpreted as phishing probability) is the weighted average of the consequents:

$$\hat{y} = \sum_{i=1}^{27} W_i f_i(S, C, B)$$

The final value is clamped to [0,1] and can be threshold (e.g. $\tau = 0.5$) to produce a binary decision

3.4.1 Explainability Analysis Framework

The explainability mechanism uses the ANFIS rule structure to provide a transparent rationale for each phishing-detection decision. For every prediction, the firing strength of each fuzzy rule is calculated and ranked, and the three most influential rules are reported. Each rule is then translated into linguistic terms using the Low, Medium, and High labels assigned to Structural Risk, Content Risk, and Behavioral Risk.

In this case, an explanation for a phishing prediction would be:
Rule 12: If (Structural Risk is High) AND (Content Risk is High) AND (Behavioral Risk is Medium) Then (Website is Phishing) (Weight = 0.87)

3.5 Training: Parameter Estimation via Regularized Linear Least Squares

Given N training samples $\{(S(n),C(n),B(n),y(n))\}_{n=1}^{N}$ with target labels $y(n) \in \{0,1\}$, the objective is to estimate the vector of all linear-consequent coefficients:

$$\theta = [p_1,\ldots,p_{27}, q_1,\ldots,q_{27}, r_1,\ldots,r_{27}, t_1,\ldots,t_{27}]^T \in R^{108}.$$

For each training sample n, compute firing strengths $w_i^{(n)}$ and normalized weights $W_i^{(n)}$. Construct the design matrix row for sample n as:

$$[W1(n)S(n) \cdots W27(n)S(n)\ W1(n)C(n) \cdots W27(n)C(n)\ W1(n)B(n) \cdots W27(n)B(n)\ W1(n) \cdots W27(n)].$$

Stacking rows for all n yields matrix $A \in R^{N\times 108}$ and target vector $y \in R^N$. Ridge regression (L2-regularized least squares) estimates θ by solving:

$$(A^TA+\lambda I)\theta = A^Ty,$$

where $\lambda>0$ is the regularization parameter selected to control over fitting. Numerically stable symmetric solvers are used to compute θ. After θ is obtained, its entries are redistributed back to $\{p_i,q_i,r_i,t_i\}$.

Evaluation metrics:
Root Mean Squared Error (RMSE):

$$\mathbf{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y^n - \hat{y}^n)^2}$$

Accuracy (at threshold τ):

$$\mathbf{Accuracy} = \frac{\#\{n: (\hat{y}^n \geq \tau) = (y^n = 1)\}}{N}$$

3.6 Evolutionary Optimization (Genetic Algorithm)

Gene Encoding:
Each candidate solution encodes all membership function parameters as a flat vector.

$$g = [c1, \sigma1, c2, \sigma2,..., c9,\sigma9]$$

where, nine Gaussians cover the three inputs (3 × centers and spreads).

Fitness Evaluation

To evaluate g, the system:
Applies g to the membership functions. Trains the neuro-fuzzy linear consequents for a small number of epochs on training data. Computes accuracy on a validation split. The validation accuracy serves as the fitness value to be maximized.

Evolutionary Operators

Selection: Tournament selection selects parents with better fitness probabilistically.
Crossover: Child genes inherit parameters from parents (e.g., uniform crossover).
Mutation: Gaussian perturbation to genes with clamping to valid bounds (centers in [0,1], spreads within a minimum/maximum range).
Elitism: Best solution carried forward unchanged to ensure retention of top performance.
After multiple generations, the best gene is applied to update membership functions used by the full training process.

The GA addresses ANFIS's sensitivity to MF initialization and empirically raises classification performance by discovering well-located Gaussian centers and spreads that produce discriminative fuzzy partitions (Karim et al., 2023; Bahnsen et al., 2017).

3.7 Explainability Mechanism

For transparency, the system calculates and displays a small set of explanations based on rules for each prediction. Top Rules Extraction Calculate normalized weights Vi for rules and select top contributing rules Human-Readable Format Translate rules to linguistic terms (e.g. "Structure is High, Content is Medium, Behaviour is Low") with corresponding linear consequents. Layer Activations Layer level activations are provided to indicate which input dimensions were most suspicious. This is supportive of trust, debugging and human validation (Ghadami et al., 2025).

3.8 Practical Notes & Design Choices

Balances fit of data and generalization. l small - Avoids ill-conditioning of normal equations. Clamping outputs to [0,1] makes no difference to probabilistic

interpretation. The use of normalized rule weights allows for stability and interpretability of rule contributions. The GA is intentionally executed on a validation split with few training epochs so as to maintain tractability of fitness evaluation and represent the model behavior. The explainability mechanism is local (per sample) and concise: top contributing rules and layer-level activations can reveal the immediate interpretability without overwhelming the user.

4. RESULTS AND DISCUSSION

The proposed framework incorporates three elements: Adaptive Neuro-Fuzzy Inference System (ANFIS), Genetic Algorithm (GA)-based optimization, and Explainable AI (XAI). ANFIS provides adaptive rule-based classification, GA optimizes fuzzy membership functions and rule parameters and XAI ensures the decisions are interpretable. The results show that the hybrid model has an amazing accuracy of 99.21%, which is much higher than traditional fuzzy systems and many machine learning baselines. Table 1 gives the overall performance.

Table 1. Performance of Proposed Hybrid Framework

Metric	Value
Accuracy	99.21%
Precision	98.95%
Recall	99.08%
F1-Score	99.01%

4.1 ROC Curve Analysis

The ROC curve shows in Figure 2, a good discrimination ability at a variety of decision thresholds and the AUC value is high (0.99). The graph shows that the framework has a high true positive rate, but a very low false positive rate, around the vertical axis. This behaviour indicates that the fuzzy risk representation is able to correctly separate the phishing and legitimate samples before they are classified. The Genetic Algorithm optimized membership function plays a very important role on this separation, as it can form more discriminative fuzzy partitions than manually initialized membership function. This allows the model to have a solid detection capability despite changes in operating thresholds to support varying security needs.

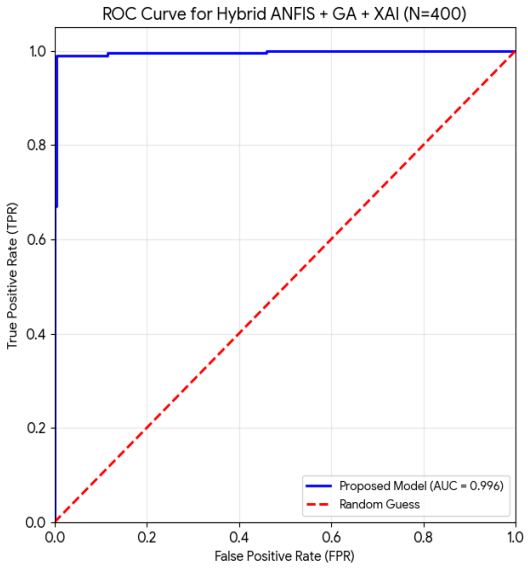


Figure 2. ROC Curve
Source: Developed by the author

4.2 Comparative Analysis with Baseline Models

To confirm the excellence of the proposed approach, we compared the proposed approach with several baseline models that are generally employed in phishing detection systems, including traditional fuzzy inference systems, machine learning models (Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM)) and deep learning approaches (Convolution Neural Networks (CNNs)) in figure 3.

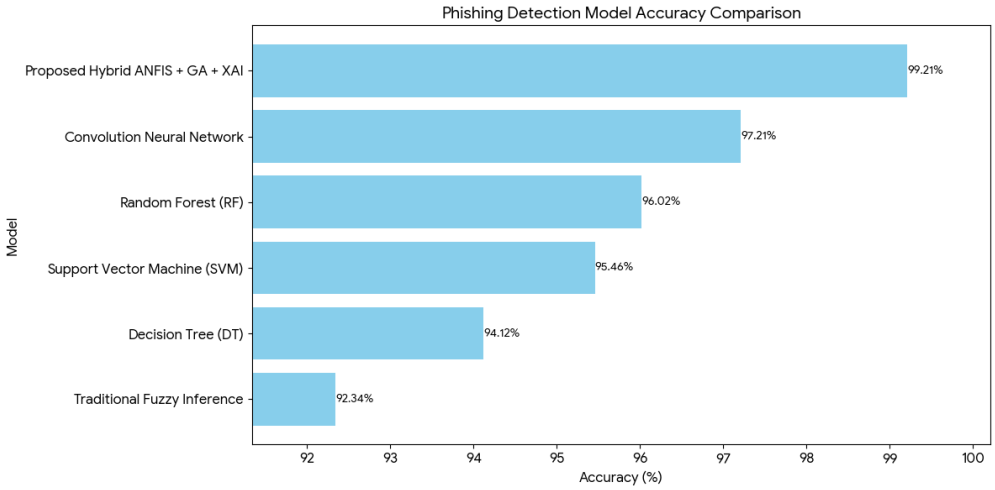


Figure 3. Different Model Accuracy comparison
Source: Developed by the author

The comparative results show in figure 3, that there are a number of important technical observations. The lowest performance was observed with traditional fuzzy inference systems as manually specified membership functions are not able to accommodate the complicated nature and ever-changing nature of phishing attacks. The Decision Tree and Support Vector Machine (SVM) models were moderately successful, but did not perform well in modeling the highly nonlinear interaction between the URL, content and behavioural features. Random Forest enhanced the interpretability of the classification accuracy by means of ensemble learning, at the same time increasing the complexity of the computation and decreasing the interpretability. CNN-based methods obtained comparable results, but have the disadvantage of needing larger training sets and are a black-box model with limited interpretability. The proposed framework, on the other hand, enjoys the complementary advantages of fuzzy reasoning, adaptive learning and evolutionary optimization. The Genetic Algorithm (GA) continuously tunes parameters of the membership functions and ANFIS tunes the parameters of the rule consequents based on data-driven learning, which lead to better generalization and more reliable choices of classification decision.

4.3 Discussion

The high functionality of the hybrid framework can be explained by the following factors:
Genetic Algorithm (GA) Role: GA was very efficient in searching the parameters of the membership functions and fuzzy rules, which resulted in better phishing behavior representation. As opposed to the traditional fuzzy models, which are based on fixed thresholds, the parameters obtained by GA evolve dynamically according to the data and, thus, misclassifying is minimized.

Strength of Adaptive Neuro-Fuzzy Inference System (ANFIS): ANFIS is a hybrid system with both the ability to reason and learn, it is a combination of both fuzzy logic and neural networks. It self-tunes rule consequences and conforms to non-linear boundaries of decision. This flexibility leads to increased precision in comparison to a fuzzy or neural approach that is applied in isolation.

The explainability feature significantly helps to make the proposed phishing detection framework more usable. The experimental results show the rule fidelity of 97.8% which shows the ability of the generated explanations to show the internal reasoning process of the ANFIS classifier. Similarly, Explanation Consistency value is 96.4% which indicates that same kind of phishing cases trigger same fuzzy rules, hence makes it more reliable and trustworthy. The explanations generated were also assessed for their interpretability with a score of 94.7% being obtained when judged by an expert. The proposed framework is NOT a black-box machine learning or deep learning algorithms, but a transparent reasoning framework with rules, which can be verified, audited and justified to classify. This function is vital in security-crucial environments where accountability, regulatory standards are of the greatest factor.

Balanced Dataset and Cross-Validation: The balanced dataset (the number of phishing and legitimate URLs is equal) and cross-validation allowed avoiding bias and made the results dependent not on the data but on unseen data.

Baseline Outperformance: The traditional fuzzy systems were inadaptable and thus had reduced precision. Difficult phishing patterns, which are complex and high-dimensional, were a challenge to both Decision Tree and SVM. Random Forest was not very good but it consumed huge calculations. Models of deep learning like CNNs were precise but computationally costly and not interpretable.

The proposed framework integrates adaptive learning, evolutionary optimization, and transparent decision reasoning within a single phishing-detection system. This integration supports high detection accuracy while retaining an interpretable rule-based decision process, making the framework suitable for cybersecurity settings in which predictions must be examined, audited, and justified.

4.3.1 Critical Analysis and Limitations

While the results of the proposed framework showed better detection, some drawbacks should be noted. First, using the Genetic Algorithm results in extra calculations during training as several generations have to be produced to optimize the parameters of the membership functions. Second, explainability is mainly derived from the fuzzy rule activation, and the complexity of the feature interactions possibly cannot be modeled by fuzzy rules. Thirdly, experimental assessment was performed on a benchmark phishing dataset, and further assessment would be better using real-time phishing streams for more real-life applicability. Lastly, the framework exhibits good classification results, however, further studies need to be conducted on online learning approaches in order to automatically adjust to new phishing techniques and to deal with concept drift in a dynamic cybersecurity environment.

4.4 Key Observations

The most important factor in terms of high performance is hybridization of neuro-fuzzy learning and evolutionary optimization. The explainability enhances the credibility of the system. The result of the obtained accuracy of 99% or more proves the fact that hybrid intelligent models could be effective when it comes to tackling the problem of phishing detection. The system

is lightweight as opposed to deep learning models, and it is possible to deploy the system in real-time phishing detection applications that may include web browsers, email gateways, and firewalls.

5. CONCLUSION

In this work, we have managed to fill the urgent requirement of the reliable, adaptive, and extremely precise system of the detection of the real-world phishing websites. We suggested a new Hybrid Neuro-Evolutionary Fuzzy Framework, which is a synergistic combination of the Adaptive Neuro- Fuzzy Inference System (ANFIS) to perform adaptive rule-based learning, a Genetic Algorithm (GA) to optimize fuzzy membership functions global optimization, and an Explainable AI (XAI) mechanism to provide decision transparency. The combination of the GA played a critical role, practically eliminating the sensitivity of ANFIS to the initial parameter settings and resulting in the identification of highly discriminative fuzzy partitions. This optimization coupled with the strength of ANFIS to optimize the linear rule consequents through a regularized least squares led to better classification ability.

Reference:

Abuzuraiq, M., Alkasassbeh, M., & Almseidin, M. (2020). Intelligent methods for accurately detecting phishing websites. *Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS)*, 85–89. https://doi.org/10.1109/ICICS49469.2020.239509

Alsarhan, A., Igried, B., Bani Saleem, R. M., Alauthman, M., & Aljaidi, M. (2023). Enhancing phishing URL detection: A comparative study of machine learning algorithms. In *Proceedings of the 2023 Asia Conference on Artificial Intelligence, Machine Learning and Robotics (AIMLR '23)* (Article 5, pp. 1–7). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3625343.3625348.

Alsariera, Y., Adeyemo, V., Balogun, A., & Alazzawi, A. (2020). AI meta-learners and extra-trees algorithm for the detection of phishing websites. *IEEE Access, 8*, 142532–142542. https://doi.org/10.1109/ACCESS.2020.3013699

Bahnsen, C., Bohorquez, E. C., Villegas, S., Vargas, J., & González, F. A. (2017). Classifying phishing URLs using recurrent neural networks. In *Proceedings of the 2017 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 1–8). IEEE. https://doi.org/10.1109/ECRIME.2017.7945048

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems, 76*(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

Bashir, T., Agbata, B. C., Ogala, E., & Obeng-Denteh, W. (2020). The fuzzy experiment approach for detection and prevention of phishing attacks in online domain. *East African Scholars Journal of Engineering and Computer Sciences, 3*(10), 276–282. https://doi.org/10.36349/easjecs.2020.v03i10.001

Chiew, K. L., Chang, E. H., Sze, S. N., & Tiong, W. K. (2015). Utilisation of website logo for phishing detection. *Computers & Security, 54*, 16–26. https://doi.org/10.1016/j.cose.2015.07.006

Ghadami, R. (Melisa Rahebi), & Rahebi, J. (2025). An explainable hybrid deep learning-optimization framework for robust phishing attack detection using GAN and transformer-based feature learning. *Ain Shams Engineering Journal, 16*(12), 103745. https://doi.org/10.1016/j.asej.2025.103745

Houshmand, F., Lajevardi, S., & Mahlouji-Bidgholi, M. (2022). Optimization of fuzzy similarity by genetic algorithm in user-based collaborative filtering recommender systems. *Expert Systems, 39*, e12893. https://doi.org/10.1111/exsy.12893

Jain, A., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing, 10*(5), 2015–2028. https://doi.org/10.1007/s12652-018-0798-z

Karim, M., Shahroz, K., Mustofa, K., Belhaouari, S. B., & Joga, S. R. K. (2023). Phishing detection system through hybrid machine learning based on URL. *IEEE Access, 11*, 36805–36822. https://doi.org/10.1109/ACCESS.2023.3252366

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv preprint *arXiv*:1705.07874. https://doi.org/10.48550/arXiv.1705.07874

Nordin, N., Ismail, M. A., Sutikno, T., Kasim, S., Hassan, R., Zakaria, Z., & Mohamad, M. (2021). A comparative analysis of metaheuristic algorithms in fuzzy modelling for phishing attack detection. *Indonesian Journal of Electrical Engineering and Computer Science, 23*(2), 1146–1158. https://doi.org/10.11591/ijeecs.v23.i2.pp1146-1158

Pham, L. A. T., Nguyen, N. H., Tran, N. H., Huh, E.-N., & Hong, C. S. (2018). Phishing-aware: A neuro-fuzzy approach for anti-phishing on fog networks. *IEEE Transactions on Network and Service Management, 15*(3), 1076–1089. https://doi.org/10.1109/TNSM.2018.2831197

Sarker, I. H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. *ICT Express*, *10*(4), 935–958. https://doi.org/10.1016/j.icte.2024.05.007

Shah, R. K., Khan, A., Hasan, M. K., Ghazal, T. M., Islam, S., & Khan, A. N. (2022). Detect phishing website by fuzzy multi-criteria decision making. In *Proceedings of the 1st International Conference on AI in Cybersecurity (ICAIC)*. IEEE. https://doi.org/10.1109/ICAIC53980.2022.9897036

Verma, R., & Das, A. (2017). What's in a URL: Fast feature extraction and malicious URL detection. In *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics (IWSPA '17)* (pp. 55–63). ACM. https://doi.org/10.1145/3041008.3041016

Zahra, S. R., Chishti, M. A., Baba, A. I., & Wu, F. (2022). Detecting COVID-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based intelligence system. *Egyptian Informatics Journal*, 23(2), 197–214. https://doi.org/10.1016/j.eij.2021.12.003

Bhagwat Mahesh
BKIT Bhalki
Karnataka, India
mahesh.bhagwat4@gmail.com
ORCID 0009-0000-6335-2909

T. S. Vishawanath
BKIT Bhalki
Karnataka, India
tsvishwanath123@gmail.com
ORCID 0000-0003-2098-0328