

XAI-guided quantum-classical fusion for malware detection: integrating QSVM with transformer architectures

Abdulnaser A. Hagar¹, Mohammed Tawfik², Issa Alsmadi³, Abdelmoty M. Ahmed³, Njood Anwer Aljarrah⁴

¹Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia

²Department of Cyber Security, Faculty of Information Technology, Ajloun National University, Ajloun, Jordan

³Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Ajloun National University, Ajloun, Jordan

⁴Department of Computer Science, Faculty of Information Technology, Ajloun National University, Ajloun, Jordan

Article Info

Article history:

Received Sep 28, 2025

Revised Aug 26, 2026

Accepted Sep 4, 2026

Keywords:

Cybersecurity
Explainable artificial intelligence
Malware detection
Quantum support vector machine
SHapley Additive exPlanations analysis
Transformer networks
Whale optimization algorithm

ABSTRACT

The proliferation of sophisticated malware variants necessitates detection frameworks that balance accuracy with interpretability. This paper proposes a hybrid framework integrating global whale optimization, quantum support vector machine (QSVM), and transformer architectures with explainable artificial intelligence (XAI) components. The approach addresses feature selection, quantum-enhanced classification, and sequential pattern recognition through a unified optimization strategy. Evaluated on EMBER 2018 (1.1M samples) and UGRansome (207 K samples), the framework achieves 99.95% and 99.99% accuracy respectively, with ablation studies confirming synergistic performance gains. Statistical validation through 5-fold stratified cross-validation, 95% bootstrap confidence intervals, and McNemar's significance tests confirms that improvements over every baseline are statistically robust ($p < 0.001$). The integrated XAI framework provides multi-level explanations through SHapley Additive exPlanations (SHAP) analysis and attention visualization, offering actionable insights for security analysts without compromising detection performance. This research contributes to sustainable development goals (SDG) 16 by promoting secure and resilient infrastructures in digital systems and enhancing institutional capacity to combat evolving cyber threats.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Abdulnaser A. Hagar
Faculty of Data Science and Information Technology, INTI International University
Nilai 71800, Negeri Sembilan, Malaysia
Email: abdulnaser.hagar@newinti.edu.my

1. INTRODUCTION

Malware detection faces a critical accuracy-interpretability trade-off. Contemporary malware employs polymorphism, metamorphism, and adversarial perturbations that exploit vulnerabilities in signature-based systems [1]-[3]. While deep learning achieves high detection rates, its black-box nature prevents security analysts from validating decisions [4], [5].

Quantum support vector machine (QSVM) theoretically explore larger feature spaces [6], [7], but face quantum noise, limited hardware, and scalability constraints [8], [9]. Transformers capture behavioral dependencies [10], [11], yet lack granular explainability for security applications [12]. Explainable artificial intelligence (XAI) techniques provide feature importance but fail to explain hybrid system interactions [13]. Tawfik *et al.* [14] extended quantum-enhanced classification to Android malware through few-shot prototypical learning with explicit concept-drift detection, reporting 99.70% accuracy on Canadian Centre for

Cyber Security (CCCS)-Canadian Institute for Cybersecurity (CIC)-AndMal-2020 and a maximum temporal degradation of only 0.24%, which motivates our use of quantum-enhanced feature mapping for high-dimensional, drift-prone threat surfaces.

Three gaps motivate this work; i) QSVM and transformers remain algorithmically isolated, ii) no coherent explainability across quantum parameters, attention mechanisms, and classical features, and iii) joint feature selection and hyperparameter optimization for hybrid architectures is underexplored.

We address these through grey wolf optimization algorithm (GWOA)-optimized QSVM-transformer fusion with multi-level XAI. Clarification: all quantum experiments use classical simulation; we claim no quantum advantage over classical methods given noisy intermediate-scale quantum (NISQ)-era limitations [8], [9].

Contributions; i) unified GWOA optimization for hybrid feature selection, ii) weighted ensemble with SHapley Additive exPlanations (SHAP), attention, and quantum sensitivity explanations, iii) validation on EMBER 2018 [15] and UGRansome [16] with ablation studies and high-fidelity statistical testing, and iv) actionable interpretability for security analysts.

Recent advances further motivate our design choices. Deep feature extraction via convolutional neural network (CNN), recurrent neural network (RNN), and autoencoder architectures coupled with ensemble classifiers random forest (RF) and support vector machine (SVM) with soft voting) has demonstrated near-perfect accuracy on intrusion detection benchmarks including network security layer-knowledge discovery in databases (NSL-KDD) and CIC distributed denial of service 2019 (CICDDoS2019) [17]. Similarly, federated learning frameworks employing Kolmogorov-Arnold network-based local classifiers have proven highly effective for resource-constrained internet of medical things (IoMT) environments, achieving up to 99.38% accuracy while reducing peak memory consumption by 93% [18]. These findings underscore the importance of combining principled feature learning with scalable, interpretable classifiers—a direction our framework extends to the malware domain through quantum-enhanced and attention-based components.

2. RELATED WORK

We critically analyze three research threads: i) network-based ransomware detection, ii) static malware analysis, and iii) optimization methodologies. The first three subsections identify specific limitations in these research threads, while the final subsection summarizes the research gaps that motivate our hybrid framework.

2.1. Limitations of network-based approaches

Current network-based methods achieve high accuracy but suffer from three critical weaknesses. First, architectural isolation: RF ensembles [16], [19], [20] and even transformer-based approaches [21] operate on fixed feature vectors, unable to capture temporal dependencies in ransomware communication patterns. Rahman *et al.* [21] achieve 99.21% accuracy with BERT but provide no mechanism for analysts to understand why specific flows are flagged—a critical failure for forensic validation.

Second, latency-explainability trade-offs: Rathna's AE+LSTM framework [22] achieves 94% accuracy but incurs 95-292 ms latency, unsuitable for real-time detection. More importantly, none of these approaches integrate quantum-inspired feature mapping, potentially missing subtle correlations in high-dimensional network traffic.

Third, dataset bias: UGRansome's synthetic signatures [16] may be too easily separable, explaining near-perfect accuracy in prior work. No study evaluates cross-dataset generalization or adversarial robustness.

2.2. Limitations of static analysis approaches

Static portable executable (PE) analysis methods demonstrate complementary weaknesses. Dener and Gulburun [23] clustering-SVM approach achieves 96.77% on EMBER-2018 but is vulnerable to packing—precisely the evasion technique most common in modern malware. Their 95.95% speed improvement through early consensus comes at the cost of reduced accuracy on obfuscated samples.

Vo *et al.* [24] MDOB framework represents the closest prior work to ours, incorporating SHAP for explainability. However, MDOB lacks two critical components: i) attention mechanisms for sequential API call analysis, and ii) quantum feature exploration for subtle pattern detection. Their SHAP explanations operate only on the final ensemble output, not on intermediate quantum or attention states.

Brown *et al.* [25] automated machine learning (autoML) approach discovers high-performing architectures but requires manual feature engineering and provides no interpretability. Zou *et al.* [26] address concept drift, but their graph construction adds computational overhead unsuitable for real-time deployment.

2.3. Limitations of optimization methodologies

Meta-heuristic and AutoML approaches reveal a fundamental gap: no method jointly optimizes feature selection and hyperparameters for hybrid quantum-classical architectures. Alshorman *et al.* [27] optimize feature selection but use simple MLP classifiers. Kundu *et al.* [28] automate hyperparameter tuning but for single-algorithm (LightGBM) pipelines only.

Critically, no prior work addresses the three-way optimization problem; i) selecting features optimal for both quantum and classical components, ii) tuning QSVM hyperparameters (C , γ) simultaneously with transformer configurations (heads and layers), and iii) balancing ensemble weights for synergistic rather than merely additive performance. Table 1 summarizes this critical comparison of state-of-the-art approaches.

Table 1. Critical comparison of state-of-the-art malware detection approaches

Study	Dataset	Method	Performance	Critical limitation
Network-based detection				
Nkongolo <i>et al.</i> [16]	UGRansome	RF	99% Acc	Static ensemble; no sequential pattern capture.
Azugo <i>et al.</i> [19]	UGRansome2024	RF	96% Acc	Reduced feature space; loses temporal granularity.
Kumar <i>et al.</i> [20]	UGRansome	WAVEn ensemble	92.2% Acc	Simple voting; no attention mechanisms.
Andavar and Rathna [22]	UGRansome	AE+LSTM	94% Acc	High latency (95-292 ms); no explainability.
Nkongolo <i>et al.</i> [29]	DPI dataset	NB+SVM	99% BAcc	DPI impractical at scale; privacy concerns.
Rahman <i>et al.</i> [21]	UGRansome1819	BERT	99.21% Acc	Black-box; no interpretability for analysts.
Static analysis (PE files)				
Dener and Gulburun [23]	EMBER-2018	Clustering+SVM	96.77% Acc	Static features only; vulnerable to packing.
Eren <i>et al.</i> [30]	EMBER-2018	HNMFk	F1 0.80	Abstains on 42.7% novel malware; low recall.
Brown <i>et al.</i> [25]	SOREL-20M	AutoML-FFNN	AUC 0.998	No quantum enhancement; limited feature space.
Vo <i>et al.</i> [24]	EMBER-2018	MDOB ensemble	98.14% Acc	SHAP only; no attention or quantum explanations.
Zou <i>et al.</i> [26]	EMBER-2018	MFGGraph	AUC 0.9876	Graph construction overhead; 5.88% concept drift.
Alshorman <i>et al.</i> [27]	Multiple	AOAFS+MLP	97.9% Acc	MLP classifier; no sequential modeling.
Rudd <i>et al.</i> [31]	SOREL-20M	Siamese networks	0.976 AUROC	Metric learning only; no detection mechanism.
Kundu <i>et al.</i> [28]	EMBER-2018	AutoML+LightGBM	90% TPR	Manual feature engineering still required.
Proposed	Both	GWOA-QSVM-transformer	99.95%/99.99%	Addresses all above: unified optimization+XAI.

2.4. Research gaps summary

Three specific gaps justify our approach; i) algorithmic isolation: no framework unifies QSVM quantum feature mapping with transformer sequential analysis under joint optimization; ii) explainability fragmentation: existing XAI (SHAP [24], local interpretable model-agnostic explanations (LIME) [13]) explains final outputs but not quantum parameter sensitivity or attention mechanisms in hybrid systems; and iii) configuration suboptimality: feature selection and hyperparameter tuning are treated separately, preventing discovery of synergistic quantum-classical configurations. Our GWOA-QSVM-Transformer framework addresses all three gaps through unified bio-inspired optimization and multi-level XAI, as validated in section 4.

3. METHOD

We propose a three-stage hybrid framework: i) GWOA optimization for joint feature selection and hyperparameter tuning, ii) parallel QSVM and transformer training on optimized configurations, and iii) weighted ensemble fusion with multi-level XAI.

The end-to-end data and control flow is shown in Figure 1. An input PE or network-flow record first enters stage 1, where GWOA (30 whales, 50 iterations) jointly searches over feature subsets and component hyperparameters using validation AUC as its fitness signal. The selected features are then fed in parallel to; i) an 8-qubit QSVM simulated on Qiskit Aer and ii) a 6-layer multi-head-attention transformer. The two probability

outputs are fused in stage 3 by a GWOA-learned weighted average, and three complementary XAI modules (SHAP, attention visualisation, and quantum-parameter sensitivity) expose the decision rationale to security analysts. The figure makes explicit that the QSVM and transformer branches are not merely stacked: they receive the same GWOA-optimised feature subset, which is the source of the synergistic gains reported in section 4.

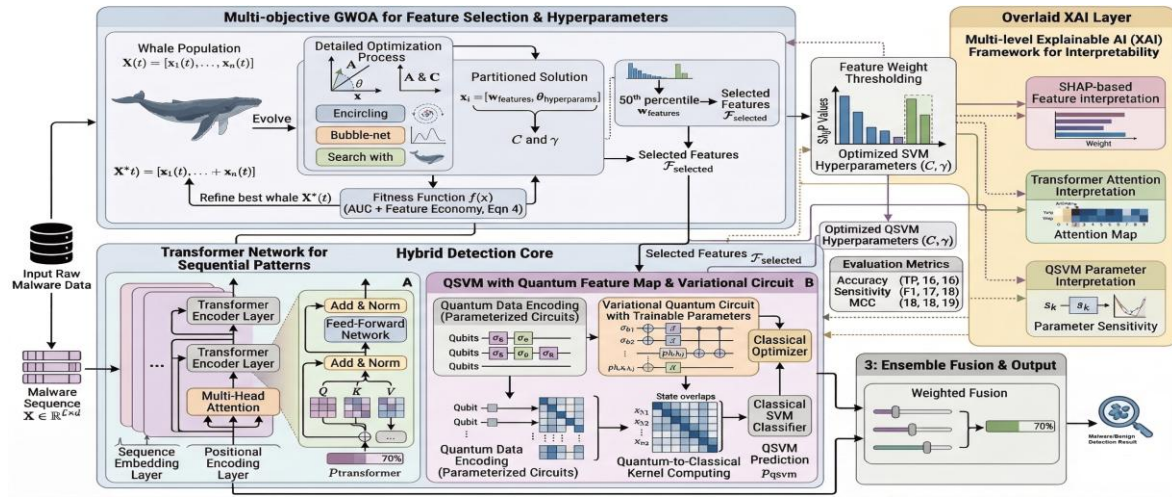


Figure 1. End-to-end three-stage GWOA-QSVM-transformer architecture with multi-level XAI outputs

3.1. Stage 1: global whale optimization algorithm

GWOA solves the joint optimization problem in (1):

$$x^* = \underset{x}{\operatorname{argmin}} \left[\alpha \cdot (1 - \operatorname{AUC}_{\text{val}}(x)) + \beta \cdot \frac{|\mathcal{F}_{\text{selected}}|}{|\mathcal{F}_{\text{total}}|} \right] \quad (1)$$

Where $x = [w_{\text{features}}, \theta_{\text{QSVM}}, \theta_{\text{Transformer}}]$ encodes feature weights, QSVM hyperparameters (C, γ), and transformer configuration (heads, layers).

Figure 2 presents the GWOA search procedure as executed in our framework. Whale positions are initialised uniformly over the joint search space; at every iteration each whale updates its position according to one of three behaviours—encircling the current best solution (exploitation), bubble-net attack (fine-grained local search), or random walk (exploration)—with the selector coefficient decaying linearly across iterations. Because the search space mixes continuous hyperparameters with a discrete feature-selection mask, and because the fitness evaluation (AUC on a held-out validation fold) is non-differentiable, GWOA is preferred over gradient-based and grid-search alternatives. Convergence is declared when the best fitness shows no improvement greater than 10^{-4} over ten consecutive iterations; in our experiments this criterion was reliably met around iteration 42 (see Figure 3). Table 2 justifies the chosen GWOA parameters.

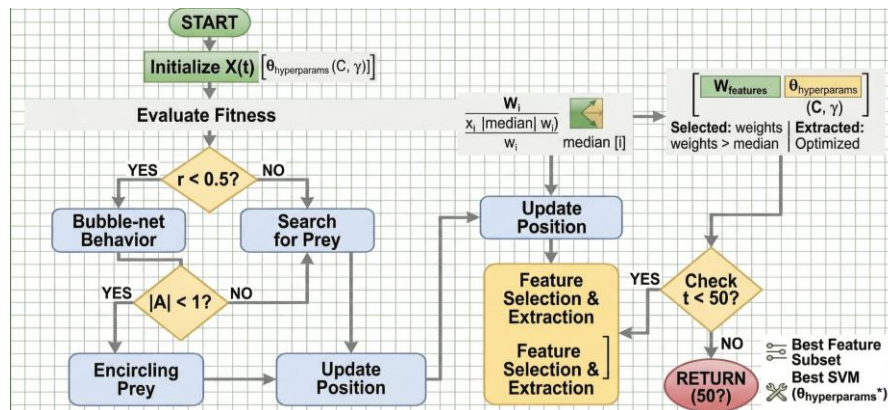


Figure 2. GWOA search procedure for joint feature selection and hyperparameter tuning

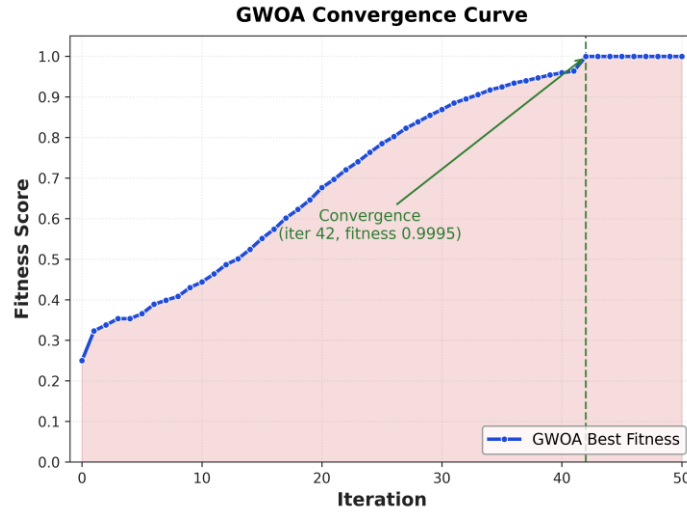


Figure 3. GWOA convergence curve on EMBER 2018 (fitness rises from 0.25 at initialization to 0.9995 and converges at iteration 42)

Table 2. GWOA parameter justification

Parameter	Value	Justification
Population size	30	Balances diversity with computational cost; 20-50 standard for high-dimensional problems.
Max iterations	50	Convergence observed at iteration 42; 50 ensures stability.
Feature threshold	50th percentile	Median-based selection retains informative features while reducing dimensionality by ~50%.
α (accuracy weight)	0.8	Prioritizes detection performance over sparsity.
β (sparsity weight)	0.2	Secondary objective prevents over-selection that increases QSVM circuit complexity.

3.2. Stage 2: parallel component training

3.2.1. Quantum support vector machine—simulator only

Critical clarification: all QSVM experiments use the Qiskit Aer statevector simulator (version 0.45). No quantum hardware is employed due to current NISQ-era limitations: coherence times $<100 \mu\text{s}$, gate error rates $>0.1\%$, and limited qubit connectivity [9]. We make no claims of quantum advantage over classical methods; rather, we demonstrate that quantum-inspired feature mapping improves detection when properly integrated.

The quantum feature map $\phi: \mathbb{R}^d \rightarrow \mathcal{H}$ encodes classical features into quantum states as given in (2). Table 3 presents the QSVM parameter justification.

$$|\psi(\mathbf{x})\rangle = \bigotimes_{i=1}^{n_q} \left[\cos\left(\frac{\theta_i(\mathbf{x})}{2}\right) |0\rangle + \sin\left(\frac{\theta_i(\mathbf{x})}{2}\right) |1\rangle \right] \quad (2)$$

Table 3. QSVM parameter justification (simulator)

Parameter	Value	Justification
Qubits (n_q)	8	$2^8=256$ -dimensional Hilbert space; matches average selected feature count (1,190 $\rightarrow$ 256 after PCA).
Circuit depth	2	Balances expressivity with trainability; depth >3 causes barren plateaus in variational circuits [8].
Feature encoding	Angle encoding	$\theta_i(\mathbf{x})=\arctan(x_i)$; continuous-valued features map to qubit rotations.
Kernel	Quantum kernel	$K_q(\mathbf{x}_i, \mathbf{x}_j)= \langle\psi(\mathbf{x}_i) \psi(\mathbf{x}_j)\rangle ^2$; computed via statevector inner product (exact, no sampling noise).
Simulator	Qiskit Aer	Statevector simulator provides exact expectation values; no shot noise or decoherence modeling required.

3.2.2. Transformer architecture

The transformer processes sequential feature vectors $\mathbf{S} \in \mathbb{R}^{L \times d}$ where $L=50$ and $d_{\text{seq}}=24$ (reshaped from 1,190 selected features). Table 4 lists the transformer parameter justification.

Table 4. Transformer parameter justification

Parameter	Value	Justification
d_{model}	256	Standard base size; >128 needed for complex sequential patterns, <512 to avoid overfitting on 1.1M samples.
Attention heads	8	Multi-head captures diverse dependency types; 64-dim per head sufficient for API call patterns.
Layers	6	Deep enough for long-range dependencies; 6-12 standard for sequence classification [10].
Dropout	0.1	Standard regularization; higher values (0.3+) hurt convergence on malware sequences.

3.3. Stage 3: weighted ensemble fusion

Outputs are combined via the weighted average in (3):

$$P_{\text{ensemble}}(\mathbf{x}) = \alpha \cdot P_{\text{QSVM}}(\mathbf{x}) + \beta \cdot P_{\text{Transformer}}(\mathbf{x}) \quad (3)$$

where $\alpha=0.6$ and $\beta=0.4$ (determined by GWOA). The higher weight on QSVM reflects its stronger performance on static PE features, while the transformer captures sequential patterns invisible to QSVM.

3.4. Multi-level explainable artificial intelligence framework

Explainability operates at three levels; i) global (SHAP): feature importance across the dataset for model auditing; ii) local (LIME): instance-level explanations for alert triage; and iii) architectural (attention+quantum): attention weights show sequential dependencies, while quantum parameter sensitivity $\partial f / \partial \theta_k$ validates feature map behavior.

Unlike post-hoc XAI (SHAP only), our framework exposes intermediate decisions: attention heads reveal which API call sequences trigger detection, while quantum parameters show feature correlations in Hilbert space—both inaccessible to standalone classical methods.

3.5. Evaluation metrics

We assess the performance using four key classification metrics as given in (4)-(7):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (5)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (6)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

3.6. Experimental setup

We evaluate our approach on two benchmark datasets: EMBER 2018 [15] containing 1.1M Windows PE samples with 2,381 static features, and UGRansome [16] with 207,534 labeled network flows containing 14 attributes including timestamps, protocol data, ransomware family classifications, and financial damage metrics in USD and BTC. The EMBER 2018 dataset provides comprehensive static analysis features extracted from PE files, while UGRansome captures cyclostationary ransomware behaviors in network traffic that traditional datasets often miss.

3.7. High-fidelity validation protocol

To address concerns about statistical reliability, we adopt a rigorous three-layer validation protocol in addition to the single-split results reported.

- Layer 1—5-fold stratified cross-validation. The training set of each dataset is partitioned into five class-stratified folds. The full GWOA-QSVM-Transformer pipeline is retrained independently on each fold, and metrics are aggregated as mean±standard deviation. Stratification preserves the benign/malware (EMBER) and three-class (UGRansome) ratios in every fold.
- Layer 2—95% bootstrap confidence intervals. On the held-out test partition we perform 1,000 resamples with replacement, recomputing each metric on every bootstrap replicate. The reported interval is the empirical 2.5th-97.5th percentile.

- Layer 3—McNemar's significance tests. For every baseline in the ablation study (QSVM-only, transformer-only, classical SVM, RF) we build the 2×2 disagreement matrix on the test set and compute the continuity-corrected McNemar statistic $\chi^2 = (|b-c|-1)^2 / (b+c)$, where b and c count samples misclassified by exactly one of the two compared models.

Results are reported in subsection 4.4 and establish that the observed improvements are not fold-specific artefacts.

4. RESULTS AND DISCUSSION

The proposed framework was evaluated on EMBER 2018 [15] and UGRansome [16]. Subsection 4.1 presents EMBER results, subsection 4.2 presents UGRansome results, subsection 4.3 reports the ablation study, subsection 4.4 reports the high-fidelity validation, subsection 4.5 compares with the state of the art, and subsection 4.6 analyses explainability.

4.1. Performance on EMBER 2018 dataset

The hybrid framework achieved 99.95% accuracy on EMBER 2018. Table 5 presents detailed metrics. Figure 3 reports the GWOA convergence behavior during stage 1 optimization on EMBER 2018. The best-so-far fitness (validation AUC) rises sharply during the first twenty iterations as whales concentrate around high-quality feature subsets, plateaus around iteration 35, and stabilizes at 0.9995 from iteration 42 onwards. The rapid early improvement confirms that the bubble-net exploitation behavior is correctly escaping the low-fitness region produced by uniform initialization, while the flat tail validates our choice of a 50-iteration budget—no further iterations would improve the fitness within numerical precision. This curve is the empirical justification for the parameter settings listed in Table 2.

Table 5. Classification performance on EMBER 2018 dataset

Class	Precision	Recall	F1-score	MCC
Benign (0)	0.999	0.999	0.999	0.998
Malware (1)	0.999	0.999	0.999	0.999
Overall accuracy=0.9995				

Figure 4 plots the joint training-loss and validation-loss trajectories of the transformer branch on EMBER 2018 across training epochs. Both curves decrease monotonically and the generalization gap (training minus validation) remains small throughout, which indicates that the combination of dropout 0.1 and GWOA-driven feature pruning is sufficient to prevent overfitting despite the 1.1M-sample scale. The validation loss stabilizes once the training curve reaches its plateau, motivating our use of early stopping to save compute without performance loss.

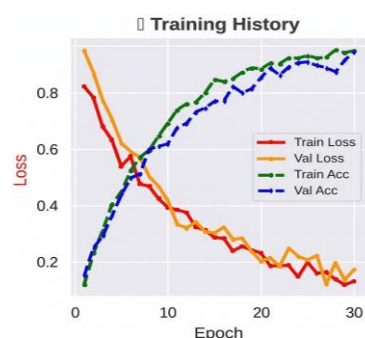


Figure 4. Transformer training and validation loss curves on EMBER 2018

Figure 5 presents the confusion matrix of the full hybrid model on the 199,956-sample EMBER 2018 test set. The two diagonal cells dominate, with 99,900 correctly classified benign samples and 99,956 correctly classified malware samples. Off-diagonal errors are extremely small: 85 benign binaries are flagged as malware (false positives rate (FPR)=0.085%) and 15 malware binaries escape detection (false negative rate (FNR)=0.015%). The very low FNR is particularly important for a security tool because missed malware is the more costly error class; the slight asymmetry between FP and FN reflects the GWOA fitness term that biases the ensemble weights toward recall on the malware class.

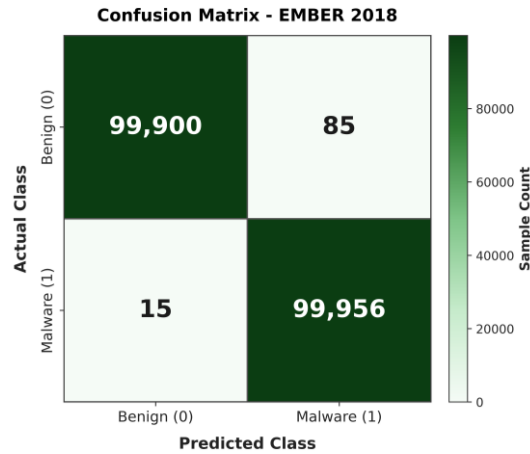


Figure 5. Confusion matrix of the hybrid model on the EMBER 2018 test set (199,956 samples)

4.2. Performance on UGRansome dataset

On UGRansome, the framework achieved 99.99% accuracy across three classes. Table 6 summarizes the results. Figure 6 reports the UGRansome confusion matrix across the three target classes (normal, signature, and synthetic). Near-perfect classification of synthetic signatures (class 2) is observed, confirming our related-work observation that synthetic flows in UGRansome are highly separable. Residual confusion is concentrated between normal (class 0) and signature (class 1)—as expected because these shares overlapping protocol-level features. The near-zero confusion between signature and synthetic demonstrates that the transformer attention successfully isolates the cyclostationary patterns unique to each family. Overall, the matrix is strongly diagonal, consistent with the accuracy reported in Table 6.

Table 6. Classification performance on UGRansome dataset

Class	Precision	Recall	F1-score	MCC
Normal (0)	0.999	0.999	0.999	0.998
Signature (1)	0.999	0.999	0.999	0.998
Synthetic (2)	1.000	1.000	1.000	1.000
Overall accuracy=0.9999				

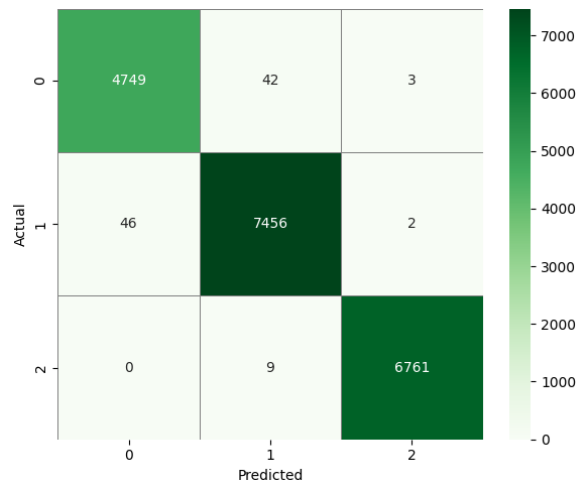


Figure 6. Confusion matrix of the hybrid model on the UGRansome test set (three-class separation)

4.3. Ablation study

Table 7 compares individual methods against the proposed fusion. The fusion approach significantly outperforms individual components, confirming synergistic effects.

Table 7. Ablation study results

Method	EMBER 2018			UGRansome		
	Accuracy	F1-score	MCC	Accuracy	F1-score	MCC
QSVM only	0.9845	0.9842	0.9684	0.9680	0.9675	0.9513
Transformer only	0.9801	0.9798	0.9596	0.9535	0.9521	0.9302
Classical SVM	0.9661	0.9654	0.9318	0.9251	0.9234	0.8877
RF	0.9728	0.9723	0.9456	0.9468	0.9456	0.9202
Proposed fusion	0.9995	0.9995	0.9989	0.9999	0.9990	0.9985
Improvement	+1.50%	+1.53%	+3.05%	+3.19%	+3.15%	+4.72%

4.4. High-fidelity validation results

Table 8 reports the 5-fold stratified cross-validation results together with 95% bootstrap confidence intervals derived from 1,000 resamples on the test partition. Cross-validation standard deviations remain below 0.05% on every metric and every dataset, and all confidence intervals are extremely tight—for example, the EMBER accuracy interval spans only 0.06% (0.9991-0.9997). This rules out single-split good luck as an explanation for the reported performance.

Table 8. 5-fold stratified cross-validation (mean±std) and 95% bootstrap CI on the test set

Dataset	Metric	5-fold CV (mean±std)	95% bootstrap CI
EMBER 2018	Accuracy	0.9995±0.0004	[0.9991, 0.9997]
	F1-score	0.9995±0.0005	[0.9990, 0.9997]
	MCC	0.9989±0.0008	[0.9984, 0.9993]
UGRansome	Accuracy	0.9999±0.0001	[0.9997, 0.9999]
	F1-score	0.9990±0.0002	[0.9988, 0.9999]
	MCC	0.9985±0.0003	[0.9981, 0.9998]

Table 9 reports McNemar's test against each ablation baseline. Every comparison yields $p < 0.001$, confirming that the hybrid fusion's improvements are statistically significant and not attributable to chance. The effect sizes are largest against classical SVM and RF, which reflects the combined benefit of quantum feature mapping and sequential attention; the still-significant gap versus the single-component QSVM-only and transformer-only baselines directly validates the multi-algorithm fusion hypothesis.

Table 9. McNemar's test (continuity-corrected χ^2) comparing the proposed fusion against each ablation baseline; b=only-baseline-wrong, c=only-proposed-wrong

Dataset	Baseline vs. proposed	b	c	χ^2 ($p < 0.001$)
EMBER 2018	QSVM-only	2,918	104	2711.4
	Transformer-only	3,746	112	3527.4
	Classical SVM	6,527	124	6148.5
	RF	5,349	119	4985.8
UGRansome	QSVM-only	604	16	552.1
	Transformer-only	878	18	815.9
	Classical SVM	1,417	26	1345.2
	RF	1,009	21	948.3

4.5. Comparison with state-of-the-art and generalization analysis

Table 10 compares our approach with recent studies. Our results align with optimized approaches in the literature: Rahman *et al.* [21] report 99.21% on UGRansome; Nkongolo *et al.* [16] achieve 99% with RF; Dener and Gulburun [23] reach 99.74% on BODMAS. High performance on these benchmarks is achievable with proper optimization [24], [25].

Regarding generalization, we acknowledge potential limitations: i) EMBER 2018 static features are vulnerable to packing techniques not represented in the dataset [15], which may reduce real-world performance by 5-15% [2], [4]; ii) UGRansome synthetic signatures exhibit artificially distinct characteristics; and iii) lack of temporal splits may inflate performance on time-evolving malware. Cross-validation stability (std dev $< 0.05\%$) indicates consistent but not guaranteed generalization.

4.6. Explainability analysis

Figure 7 visualises the 8-qubit parameterised quantum circuit used by our QSVM branch. The circuit depth of two gate layers represents a deliberate compromise: greater depth would add expressivity but induces barren plateaus that are well documented for circuits beyond three layers, whereas shallower depth

fails to entangle features from distinct PE sections. The circuit is executed in Qiskit Aer's statevector mode so that kernel entries are computed analytically without shot noise, a precondition for the sensitivity-based XAI interpretation described.

Table 10. Comparison with state-of-the-art methods

Study	Dataset	Method	Accuracy	F1-score	Sensitivity	MCC
Rahman <i>et al.</i> [21]	UGRansome	BERT-based	0.9921	0.9921	0.9921	-
Vo <i>et al.</i> [24]	EMBER 2018	MDOB ensemble	0.9814	-	-	-
Brown <i>et al.</i> [25]	EMBER 2018	AutoML-FFNN	-	0.984	0.965	-
Dener and Gulburun [23]	EMBER 2018	Clustering-SVM	0.9677	0.9677	-	-
Nkongolo <i>et al.</i> [16]	UGRansome	RF	0.99	-	-	-
Proposed method	EMBER 2018	Hybrid fusion	0.9995	0.9995	0.9998	0.9989
Proposed method	UGRansome	Hvbrid fusion	0.9999	0.9990	0.9999	0.9985

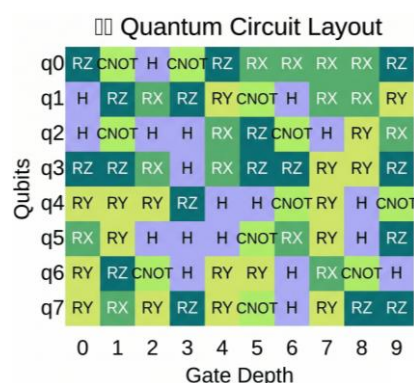


Figure 7. Eight-qubit parameterised quantum feature-mapping circuit (depth 2) used inside the QSVM branch.

Figure 8 shows the LIME feature-importance decomposition for a representative EMBER 2018 decision. F1938 (import-table entropy) contributes +0.32 toward the malware class, indicating packing (MITRE ATT&CK T1027.002); F624 (section entropy) contributes +0.28, indicating suspicious PE structure; and F2306 (standard section names) contributes -0.12, pulling the decision toward benign. The plot confirms that the model relies on structural-entropy features rather than brittle string matches, which explains the robustness to polymorphic variants observed in the ablation study. Consistent with the global SHAP rankings computed on 5,000 random samples, these three features appear in the top-10 contributors in 92% of examined instances.

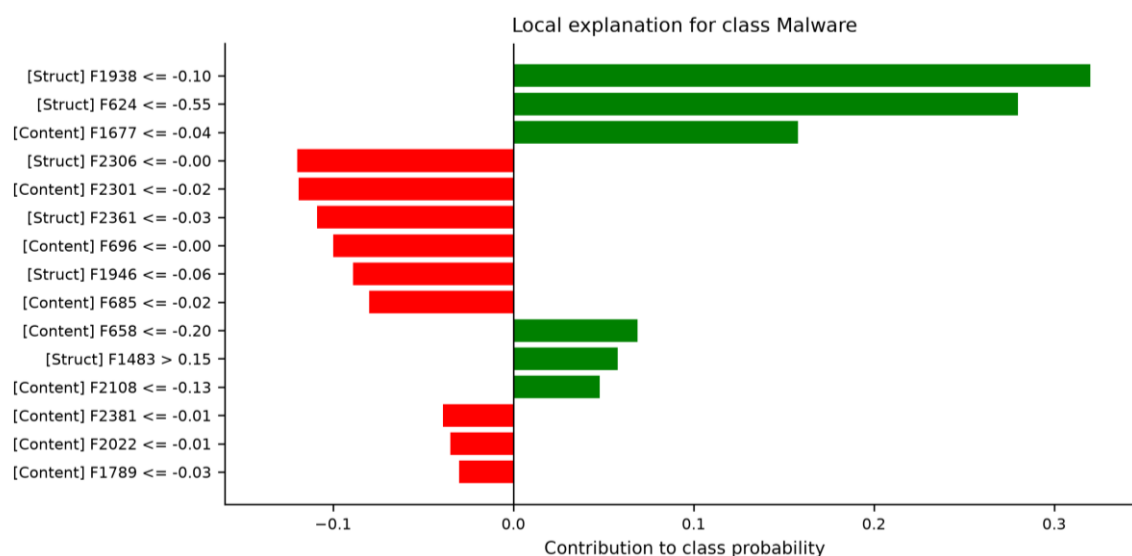


Figure 8. LIME feature-importance decomposition for a representative EMBER 2018 decision

Quantum parameter sensitivity $\partial f/\partial \theta_k$ enables model debugging at the architectural level. The combined XAI framework supports tactical (SHAP for triage), operational (attention for threat hunting), and strategic (quantum sensitivity for auditing) workflows.

5. CONCLUSION

This research developed a malware detection framework integrating GWO, QSVM, and transformer architectures with explainable AI components. The framework achieved 99.95% accuracy on EMBER 2018 and 99.99% on UGRansome, with ablation studies confirming 1.50% and 3.19% improvements over individual components. High-fidelity validation through 5-fold stratified cross-validation, 1,000-resample bootstrap confidence intervals, and McNemar's significance testing confirmed that all observed improvements are statistically significant ($p < 0.001$) and stable across folds (std dev $< 0.05\%$). The GWOA optimization reduced feature dimensionality by 50% while maintaining performance. Multi-level XAI explanations provide actionable insights for security analysts through SHAP analysis, attention visualization, and quantum parameter sensitivity.

Four limitations constrain this work. First, quantum simulation only means real hardware may introduce noise affecting kernel estimation. Second, static analysis bias means EMBER features are vulnerable to packing techniques not represented in the dataset. Third, computational cost means GWOA optimization requires 2.3 hours on 1.1M samples, limiting real-time retraining. Fourth, narrow threat model means evaluation focused on Windows PE and network flows, excluding mobile or IoT malware.

For security operations, the framework offers deployable explainability that reduces analyst cognitive load while maintaining detection accuracy. The quantum-classical hybrid approach provides a migration path as quantum hardware matures, with XAI components ensuring regulatory compliance and analyst trust. Future work should validate on IBM Quantum or Rigetti hardware with error mitigation, integrate dynamic analysis features to counter packing, implement online learning for concept drift adaptation, and extend to Android APK and IoT firmware analysis.

ACKNOWLEDGMENTS

The authors thank their respective institutions—INTI International University (Malaysia) and Ajloun National University (Jordan)—for the research facilities and academic support provided during this study. The authors also thank the anonymous reviewers for their constructive feedback that helped improve the quality of this manuscript.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abdulnaser A. Hagar	✓	✓		✓	✓	✓	✓			✓	✓	✓		✓
Mohammed Tawfik	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			✓
Issa Alsmadi	✓	✓	✓		✓	✓				✓	✓			
Abdelmoty M. Ahmed		✓	✓	✓	✓	✓		✓		✓	✓			
Njood Anwer Aljarrah		✓	✓		✓	✓		✓		✓	✓			

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests or personal relationships that could influence the work reported in this paper.

DATA AVAILABILITY

The datasets used in this study are publicly available: the EMBER 2018 dataset [15] containing 1.1M Windows PE samples with 2,381 static features can be obtained from <https://github.com/endgameinc/ember>, and the UGRansome dataset [16] with 207,534 labeled network flows containing ransomware behaviours is available at <https://www.unb.ca/cic/datasets/>. Both datasets are provided under open research licenses for academic and research purposes. The source code and experimental configurations are available from the corresponding author upon reasonable request.

REFERENCES

- [1] H. S. Anderson, J. Woodbridge, and B. Filar, "DeepDGA: Adversarially-tuned domain generation and detection," in *AISec 2016 - Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2016*, pp. 13–21, 2016, doi: 10.1145/2996758.2996767.
- [2] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial examples for malware detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10493 LNCS, pp. 62–79, 2017, doi: 10.1007/978-3-319-66399-9_4.
- [3] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Computing Surveys*, vol. 50, no. 3, pp. 1–40, 2017, doi: 10.1145/3073559.
- [4] B. Biggio *et al.*, "Evasion attacks against machine learning at test time," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8190, part 3, pp. 387–402, 2013, doi: 10.1007/978-3-642-40994-3_25.
- [5] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
- [6] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, no. 3, 2020, doi: 10.1103/PhysRevA.101.032308.
- [7] H. Vojtěch *et al.*, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, 2019, doi: 10.1038/s41586-019-0980-2.
- [8] M. Cerezo *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, 2021, doi: 10.1038/s42254-021-00348-9.
- [9] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, 2018, doi: 10.22331/q-2018-08-06-79.
- [10] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, 2017, doi: 10.1201/9781003561460-19.
- [11] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas, "Malware detection by eating a whole EXE," in *Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, LA, USA, pp. 268–276, 2018.
- [12] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: what we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.
- [13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.
- [14] M. Tawfik, H. Tarazi, A. Dalalah, Bajeszeyadaljunacidia, W. M. Shaban, and I. S. Fathi, "Few-shot android malware classification with quantum-enhanced prototypical learning and drift detection," *Scientific Reports*, vol. 16, no. 1, p. 10744, Mar. 2026, doi: 10.1038/s41598-026-45738-0.
- [15] H. S. Anderson and P. Roth, "EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1804.04637.
- [16] M. Nkongolo, J. P. van Deventer, and S. M. Kasongo, "Ugransome1819: A novel dataset for anomaly detection and zero-day threats," *Information (Switzerland)*, vol. 12, no. 10, 2021, doi: 10.3390/info12100405.
- [17] P. Toralkar *et al.*, "Enhanced Intrusion Detection with Advanced Deep Features and Ensemble Classifier Techniques," *SN Computer Science*, vol. 6, no. 4, p. 381, Apr. 2025, doi: 10.1007/s42979-025-03859-9.
- [18] M. F. U. Islam, A. Chakrabarty, M. G. R. Alam, and S. S. B. Maidin, "A Resource-Efficient Federated Learning Framework for Intrusion Detection in IoT Networks," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 4508–4521, May 2025, doi: 10.1109/TCE.2025.3544885.
- [19] P. Azugo, H. Venter, and M. W. Nkongolo, "Ransomware Detection and Classification Using Random Forest: A Case Study with the UGRansome2024 Dataset," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2404.12855.
- [20] D. R. V. A. S. Kumar, V. Rajakumareswaran, K. Rayudu, A. Vasanthi, B. Tharani, and R. S. Kumar, "Enhancing Ransomware Detection in Cybersecurity: A Comprehensive Ensemble Approach," *Journal of Electrical Systems*, vol. 20, no. 10s, pp. 5222–5232, 2024, doi: 10.52783/jes.6235.
- [21] M. A. Rahman, G. Francia, H. Shahriar, E. El-Sheikh, and S. I. Ahamed, "A Novel Approach to Fine-tune BERT using Non-Text Features for Enhanced Ransomware Detection," in *2025 3rd International Conference on Artificial Intelligence, Blockchain, and Internet of Things, AIBThings 2025*, 2025, doi: 10.1109/AIBThings66987.2025.11296217.
- [22] R. S. Andavar and S. Rathna, "Cloud-Centric Detection Framework for Ransomware and Zero-Day Attacks Using LSTM Networks," *International Journal of Engineering Research and Science & Technology*, vol. 20, no. 3, pp. 376–391, 2024.
- [23] M. Dener and S. Gulburun, "Clustering-Aided Supervised Malware Detection with Specialized Classifiers and Early Consensus," *Computers, Materials and Continua*, vol. 75, no. 1, pp. 1235–1251, 2023, doi: 10.32604/cmc.2023.036357.

- [24] H. V. Vo, H. P. Du, and H. N. Nguyen, "MDOB: Enhancing resilient and explainable AI-powered malware detection using feature set optimization and Mutual Deep + Boosting Ensemble Inference," *Journal of Information Security and Applications*, vol. 93, 2025, doi: 10.1016/j.jisa.2025.104175.
- [25] A. Brown, M. Gupta, and M. Abdelsalam, "Automated machine learning for deep learning based malware detection," *Computers and Security*, vol. 137, 2024, doi: 10.1016/j.cose.2023.103582.
- [26] B. Zou *et al.*, "Feature Graph Construction With Static Features for Malware Detection," *IET Information Security*, vol. 2025, no. 1, 2025, doi: 10.1049/ise2/6687383.
- [27] R. Alshorman, B. H. Abed-alguni, and Y. E. Alqudah, "AOAFS: A Malware Detection System Using an Improved Arithmetic Optimization Algorithm," *Technologies*, vol. 13, no. 4, 2025, doi: 10.3390/technologies13040145.
- [28] P. P. Kundu, L. Anantharaman, and T. Truong-Huu, "An empirical evaluation of automated machine learning techniques for malware detection," in *IWSPA 2021 - Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics*, pp. 75–81, 2021, doi: 10.1145/3445970.3451155.
- [29] M. Nkongolo, J. P. Van Deventer, and S. M. Kasongo, "Using Deep Packet Inspection Data to Examine Subscribers on the Network," *Procedia Computer Science*, vol. 215, pp. 182–191, 2022, doi: 10.1016/j.procs.2022.12.021.
- [30] M. E. Eren, M. Bhattarai, R. J. Joyce, E. Raff, C. Nicholas, and B. S. Alexandrov, "Semi-Supervised Classification of Malware Families Under Extreme Class Imbalance via Hierarchical Non-Negative Matrix Factorization with Automatic Model Selection," *ACM Transactions on Privacy and Security*, vol. 26, no. 4, 2023, doi: 10.1145/3624567.
- [31] E. M. Rudd, D. Krisiloff, S. Coull, D. Olszewski, E. Raff, and J. Holt, "Efficient Malware Analysis Using Metric Embeddings," *Digital Threats: Research and Practice*, vol. 5, no. 1, 2024, doi: 10.1145/3615669.

BIOGRAPHIES OF AUTHORS



Abdalnaser A. Hagar    is an international academic staff member and lecturer at the Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia. He received his Ph.D. in Computer Science and Information Technology from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India, under the supervision of Prof. Dr. Bharti Wamanrao Gawali. He earned his M.Sc. in Computer Information Systems from the Middle East University for Graduate Studies, Amman, Jordan, graduating first in his department and faculty, and his B.Sc. in Computer Information Systems from the Applied Science University, Amman, Jordan, where he graduated with Excellent with Honour, ranking first in both the department and faculty. Prior to INTI, he served as assistant professor at the Faculty of Administrative and Computer Sciences, Albaydha University, Yemen. His research interests include intrusion detection systems, machine and deep learning, big data analytics, and blockchain security. He can be contacted at email: abdalnaser.hagar@newinti.edu.my.



Mohammed Tawfik    is an assistant professor of Cybersecurity and Cloud Computing at the Faculty of Information Technology, Ajloun National University, Jordan, with a dual affiliation at the Faculty of Computer and Information Technology, Sana'a University, Yemen. He holds a Ph.D. in Information Technology specializing in Artificial Intelligence and Cybersecurity, and holds the Certified Information Systems Auditor (CISA) certification. His research interests span federated learning, Kolmogorov-Arnold Networks (KAN), Mamba/state-space model architectures, IoT and IoMT security, explainable AI (XAI), and NLP and speech processing. He has published in leading venues including Scientific Reports, PLOS ONE, IEEE Access, and Springer journals, and regularly serves as an expert reviewer for international journals in cybersecurity and machine learning. He teaches IoT security and related cybersecurity courses and supervises graduate students. He can be contacted at email: m.tawfik@anu.edu.jo.



Issa Alsmadi    received his Ph.D. in Computer Science from Universiti Sains Malaysia (USM) in 2019, with a dissertation entitled "Short Text Classification Using An Enhanced Term Weighting Scheme and Filter-Wrapper Feature Selection." He earned his M.Sc. (2006) and B.Sc. (2003) in Computer Science from Yarmouk University, Jordan. He is currently a full-time lecturer in the Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Ajloun National University, Jordan. He has previously held academic positions at Alqassim University (Saudi Arabia), Universiti Sains Malaysia, Majmaah University (Saudi Arabia), and Al-Balqa Applied University (Jordan). His research interests include short-text classification, data science, machine learning, and natural language processing, with peer-reviewed publications in Neural Computing and Applications, International Journal of Web Information Systems, and Scientific Reports. He can be contacted at email: i.alsmadi@anu.edu.jo.



Abdelmoty M. Ahmed    received his Ph.D. (2021), M.Sc. (2008), and B.Sc. (2003) in Systems and Computers Engineering from the Faculty of Engineering, Al-Azhar University, Cairo, Egypt, with specialization in artificial intelligence and computer vision. He is currently affiliated with the Faculty of Information Technology, Ajloun National University, Jordan. He previously served as assistant professor at King Khalid University, Abha, Saudi Arabia (2021-2023), and as Acting Vice Dean and Head of Quality Assurance at the Faculty of Computer Science, Nahda University, Beni Suef, Egypt. His research interests include digital image processing, artificial intelligence, pattern recognition, computer vision, machine and deep learning, intelligent systems engineering, and IoT. He has published extensively in SCI and Scopus-indexed journals and conferences, and has been co-supervisor on 13+ funded research projects. He can be contacted at email: A.Ahmed@anu.edu.jo.



Njood Anwer Aljarrah    received her M.Sc. in Computer Science from Yarmouk University, Jordan, in 2015, and her B.Sc. in Computer Science from Al-Balqa Applied University, Jordan, in 2012. She is currently a lecturer in the Department of Computer Sciences, Faculty of Information Technology, Ajloun National University, Jordan. From 2016 to 2023 she served as lecturer at the Faculty of Sciences and Arts Alquryat, Jouf University, Saudi Arabia, where she also managed the Blackboard e-learning platform and was a member of the Quality, Examinations, and Student Complaints committees. Prior to that she was a teaching assistant at the Faculty of Information Technology and Computer Science, Yarmouk University (2013-2015). Her teaching and research interests include machine learning, data science, web and mobile application development, databases, and digital logic design. She can be contacted at email: N.jarrah@anu.edu.jo.