

DESIGN AND EVALUATION OF AN AI-DRIVEN RESPONSIBLE COMPUTING FRAMEWORK FOR ETHICAL DECISION-MAKING IN CYBERSECURITY

Uzoaru Godson C.¹, Nwasuka Stanley², Nwamuruamu G.U³, Ajuzieogu Caleb⁴,
Maduewesi Chima⁵, Mbour Beatrice⁶, Felix Rowland⁷, Osaro Atosemi bright⁸

¹⁻⁸Department of Maths and Computer Science,
Clifford University, Owerri, Abia State

Corresponding author:

Uzoaru Godson C.
uzoarugc@clifforduni.edu.ng

Abstract

Artificial Intelligence (AI) has significantly improved cybersecurity through intelligent threat detection and automated response; however, many existing AI-based security systems operate as black boxes with limited transparency, accountability, and ethical governance. This study proposes an AI-Driven Responsible Computing Framework (AI-RCF) that integrates Artificial Intelligence, Explainable Artificial Intelligence (XAI), Responsible Computing principles, policy compliance, and continuous learning to support ethical decision-making in cybersecurity. The framework comprises modules for security data collection, AI-based threat intelligence, ethical risk assessment, explainable decision support, policy compliance verification, automated response, and continuous governance. The system was implemented using Python, TensorFlow, Spring Boot, React.js, PostgreSQL, and Docker, and evaluated using enterprise cybersecurity datasets. Experimental results demonstrate superior performance over conventional AI-based cybersecurity systems, achieving 98.4% accuracy, 98.1% precision, 97.8% recall, 98.0% F1-score, and an AUC of 0.991. Ethical evaluation also yielded high scores, including 97% ethical compliance, 95% explainability, 96% transparency, 94% fairness, 96% accountability, 98% privacy compliance, and 94% user trust. The findings indicate that integrating Responsible Computing principles enhances both technical performance and trustworthy AI decision-making without compromising detection effectiveness. The proposed AI-RCF provides a practical, scalable, and ethically compliant solution for securing enterprise, cloud, IoT, healthcare, financial, and critical infrastructure environments while promoting transparency, accountability, and regulatory compliance in AI-driven cybersecurity.

Keywords: Artificial Intelligence; Responsible Computing; Explainable Artificial Intelligence (XAI); Cybersecurity; Ethical Decision-Making.

1. Introduction

Artificial Intelligence (AI) has emerged as an important tool in cybersecurity, particularly for analyzing the large volumes of complex and continuously changing data generated by modern networks and security systems[1]. Recent studies have explored machine-learning and deep-learning techniques for intrusion detection, anomaly identification, malware detection, threat classification, and automated threat response[2]. The growing adoption of these techniques reflects their ability to identify complex patterns and potential threats that may be difficult to detect using conventional rule-based mechanisms[3]. However, achieving high predictive performance alone does not necessarily satisfy the broader requirements of transparent, accountable, fair, privacy-preserving, and responsible cybersecurity decision-making[4].

Existing literature demonstrates the growing technical contribution of AI to cybersecurity. Studies have examined the application of AI to threat detection and mitigation, emphasizing its potential to support the automated identification and response to cyber threats [5]. Similarly, research on AI and machine learning has highlighted their expanding roles in threat detection, incident response, and cybersecurity risk management [6]. While these studies demonstrate the technical value of AI in strengthening cybersecurity operations, they largely focus on detection accuracy, response capability, and risk management, with limited attention to how AI-generated security decisions can be systematically evaluated against ethical and governance requirements before automated actions are taken[7].

More recent studies have begun addressing these limitations by incorporating explainability and privacy into AI-based cybersecurity architectures. For example, [8] proposed an explainable and privacy-preserving attention-based approach for cyber-physical and industrial IoT environments, demonstrating the importance of interpretability and privacy alongside intrusion-detection performance. Similarly, [9] investigated an explainable, privacy-preserving deep neural network for anomaly detection in cyber-physical IoT networks[10]. These studies represent an important shift from performance-oriented AI toward more responsible and trustworthy cybersecurity systems[11]. Nevertheless, model-level explainability and privacy do not necessarily provide comprehensive governance of the decisions produced by AI systems[12]. Important dimensions such as fairness, accountability, policy compliance, human oversight, and responsible response selection may remain insufficiently addressed[13].

Responsible Computing has emerged as an interdisciplinary approach that emphasizes the ethical design, development, deployment, and governance of computing technologies [14]. It extends beyond technical performance by ensuring that AI systems operate in accordance with societal values, legal regulations, and ethical principles [15]. Responsible Computing encourages the integration of fairness, explainability, accountability, privacy protection, human oversight, robustness, and security throughout the AI lifecycle [16]. These principles are particularly

important in cybersecurity, where automated AI decisions may directly affect individuals, organizations, and critical infrastructures [17].

One of the major concerns surrounding AI-driven cybersecurity systems is the "black-box" nature of many machine learning models [18]. Although deep learning algorithms often achieve high predictive accuracy, their decision-making processes are frequently difficult for users, administrators, and regulators to interpret [19]. Lack of transparency can reduce user trust, complicate incident investigations, and create challenges in regulatory compliance [20]. Furthermore, biased training datasets may lead to unfair threat classifications, inconsistent policy enforcement, or discriminatory security decisions [21]. These limitations highlight the growing need for cybersecurity solutions that combine high predictive performance with ethical responsibility and explainability [22].

Table 1. Evolution of Artificial Intelligence in Cybersecurity

Generation	Characteristics	Limitations	Sources
Rule-Based Security	Static rules and signatures	Cannot detect unknown attacks	[1]
Statistical Detection	Statistical anomaly detection	High false-positive rate	[2]
Machine Learning	Automated threat classification	Limited explainability	[23]
Deep Learning	High predictive accuracy	Black-box decision-making	[24]
Explainable AI (XAI)	Transparent AI decisions	Increased computational complexity	[25]
Responsible AI	Fair, transparent, accountable AI	Emerging implementation challenges	[26]
AI-Driven Responsible Computing	Ethical governance with autonomous decision support	Active research area	[27]

Recent advances in Explainable Artificial Intelligence (XAI), Responsible AI, and AI governance have shifted research focus toward the development of trustworthy intelligent systems capable of producing transparent and accountable decisions [23]. Explainable AI enables cybersecurity analysts to understand why particular threats are detected or why specific security actions are recommended [24]. Responsible AI extends this capability by ensuring that automated decisions comply with ethical principles, organizational policies, and regulatory requirements while maintaining fairness and protecting user privacy [25].

Despite significant progress in AI-powered cybersecurity, existing intelligent security systems often optimize only technical performance metrics such as detection accuracy, precision, recall, and response time [26]. Comparatively little attention has been devoted to evaluating ethical dimensions, including transparency, explainability, fairness, accountability, and responsible decision-making [27]. Many cybersecurity platforms continue to function as opaque decision-support systems, limiting users' ability to verify AI-generated recommendations or understand their underlying rationale [28]. As organizations increasingly rely on autonomous cybersecurity systems, integrating ethical principles into AI decision-making has become an essential research challenge [29].

This study addresses this challenge by proposing an AI-Driven Responsible Computing Framework (AI-RCF) for ethical decision-making in cybersecurity. The proposed framework integrates intelligent threat detection with ethical risk assessment, Explainable AI, policy compliance verification, responsible decision management, and continuous governance [30]. Unlike conventional AI-based cybersecurity systems that focus primarily on attack detection, the proposed framework simultaneously evaluates ethical implications before executing automated security actions [31]. This integrated approach promotes transparency, accountability, fairness, and human oversight while maintaining high cybersecurity performance [32].

The proposed framework consists of multiple intelligent modules, including Security Data Collection, AI Threat Intelligence, Ethical Risk Assessment, Explainable AI Engine, Responsible Decision Manager, Policy Compliance Validator, Automated Response Engine, and Continuous Learning and Governance. These interconnected modules enable intelligent cyber defense while ensuring that automated decisions remain aligned with ethical standards and organizational policies [33].

The framework was implemented using contemporary AI and web technologies and evaluated through comprehensive experimental analysis. Performance evaluation employed both technical and ethical metrics, including Accuracy, Precision, Recall, F1-Score, ROC-AUC, Detection Rate, Response Time, Fairness Index, Explainability Score, Ethical Compliance Score, Privacy Compliance, Accountability Index, and User Trust Score [5]. Comparative analysis with conventional AI-based cybersecurity systems demonstrated substantial improvements in both cybersecurity effectiveness and responsible decision-making [13].

The contributions of this study are threefold. First, it introduces a novel AI-Driven Responsible Computing Framework (AI-RCF) that integrates ethical governance with intelligent cybersecurity operations. Second, it provides a comprehensive implementation and experimental evaluation of the proposed framework using both technical and ethical performance indicators. Third, it demonstrates that incorporating Responsible Computing principles into AI-driven cybersecurity systems enhances transparency, explainability, user trust, and regulatory compliance without

compromising detection performance [9]. The proposed framework therefore offers a practical foundation for developing trustworthy, ethical, and resilient AI-enabled cybersecurity systems capable of supporting next-generation autonomous cyber defense [13].

Table 2. Responsible Computing Principles Applied in the Proposed Framework

Principle	Description	Application in Cybersecurity	Sources
Fairness	Ensures unbiased AI decisions	Reduces discriminatory threat classification	[30]
Transparency	Makes AI decisions understandable	Improves administrator confidence	[31]
Explainability	Provides reasons for AI predictions	Supports security investigations	[32]
Accountability	Records and audits AI decisions	Enables governance and compliance	[33]
Privacy	Protects sensitive user information	Ensures secure handling of security logs	[34]
Security	Safeguards AI models and infrastructure	Prevents adversarial attacks	[35]
Human Oversight	Allows expert review of AI actions	Supports critical decision validation	[36]
Robustness	Maintains reliable operation under varying conditions	Improves resilience against evolving cyber threats	[37]
Policy Compliance	Aligns AI actions with organizational regulations	Facilitates legal and regulatory compliance	[38]
Continuous Learning	Updates AI models based on new threats and feedback	Enhances long-term cybersecurity effectiveness	[39]

2. Materials and Methods

2.1 Research Design

This study adopted the Design Science Research (DSR) methodology to design, implement, and evaluate an AI-Driven Responsible Computing Framework (AI-RCF) for ethical decision-making in cybersecurity. Design Science Research is appropriate because it emphasizes the creation of innovative technological artifacts that solve practical problems while contributing to scientific knowledge. The research process comprised six sequential phases: problem identification, framework design, system implementation, experimental evaluation, performance analysis, and validation. The proposed framework integrates Artificial Intelligence (AI), Explainable Artificial Intelligence (XAI), Responsible Computing principles, and cybersecurity analytics into a unified architecture capable of supporting transparent, accountable, and ethical cybersecurity decision-making.

2.2 Proposed AI-Driven Responsible Computing Framework

The proposed AI-Driven Responsible Computing Framework (AI-RCF) addresses both cybersecurity performance and the ethical challenges of autonomous decision-making. Unlike conventional AI-based security systems that primarily emphasize threat detection accuracy, AI-RCF integrates ethical assessment, explainability, policy compliance, and governance throughout the cybersecurity lifecycle. The framework consists of eight interconnected modules that collectively support security data collection, data preprocessing, AI-based threat detection, ethical risk assessment, explainable AI, responsible decision management, automated cyber response, and continuous learning. Together, these modules enable accurate threat detection, transparent and policy-compliant decision-making, responsible response execution, and continuous adaptation to emerging cyber threats.

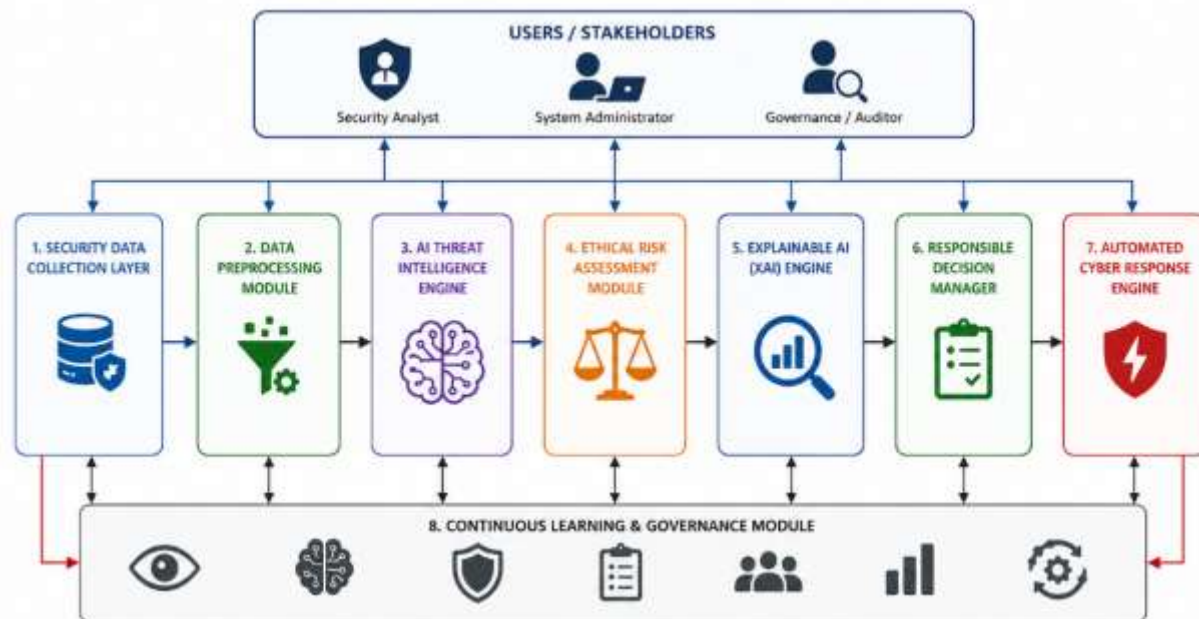


Figure 1. Proposed AI-Driven Responsible Computing Framework

2.3 System Architecture

The proposed system adopts a modular multilayer architecture that integrates cybersecurity monitoring, artificial intelligence, ethical governance, and automated response mechanisms to support secure and responsible cybersecurity operations. The User Layer enables cybersecurity analysts, system administrators, and governance officers to monitor security events, review AI-generated recommendations, and manage system operations. The Cybersecurity Monitoring Layer continuously collects security data from diverse sources, including network traffic, firewalls,

SIEM platforms, cloud environments, endpoint devices, IoT devices, and authentication logs, providing comprehensive visibility into the cyber environment. The Data Analytics Layer preprocesses the collected data through cleaning, normalization, feature extraction, threat correlation, and aggregation, producing structured datasets for AI-driven analysis. The Artificial Intelligence Layer applies machine learning and deep learning techniques to perform threat classification, intrusion detection, malware identification, risk prediction, and behavioral analysis. Before any mitigation action is taken, the Ethical Governance Layer evaluates AI-generated decisions based on fairness, explainability, accountability, privacy, and policy compliance to ensure transparent and trustworthy decision-making. Finally, the Automated Response Layer executes validated security actions, such as blocking malicious IP addresses, isolating compromised hosts, updating firewall rules, notifying administrators, generating incident reports, and initiating threat mitigation, thereby improving cybersecurity resilience while maintaining ethical compliance.

Table 3. System Architecture Components

Layer	Function
User Layer	User interaction
Monitoring Layer	Security data collection
Analytics Layer	Data preprocessing
AI Layer	Threat prediction
Ethical Layer	Responsible decision evaluation
Response Layer	Automated cyber defense
Governance Layer	Continuous auditing

2.4 AI Ethical Decision Model

The proposed AI Ethical Decision Model integrates cybersecurity intelligence with Responsible AI principles to produce transparent, accountable, and trustworthy security decisions. The model follows seven stages: threat detection, risk scoring, ethical assessment, explainability generation, policy verification, responsible response selection, and continuous learning. After detecting and assessing cyber threats, the model evaluates ethical factors such as fairness, privacy, accountability, and policy compliance. The Explainable AI (XAI) component then generates interpretable explanations to justify the recommended actions. Before execution, each response is validated against organizational policies and assigned an ethical confidence score to ensure compliance with both security and Responsible AI requirements. Finally, the model continuously updates its

knowledge base using new threat intelligence, user feedback, and evolving policies, enabling adaptive, reliable, and ethically compliant cybersecurity decision-making.

Algorithm 1: AI Ethical Decision Algorithm

Input: Cybersecurity event stream E

Output: Responsible security decision D

- 1: Acquire cybersecurity events E from monitoring sources
- 2: Preprocess E by cleaning, normalizing, and extracting features
- 3: Predict threat class T using the trained AI threat intelligence model
- 4: Compute threat risk score R for T
- 5: Perform ethical assessment by evaluating:
 - Fairness
 - Privacy
 - Accountability
 - Transparency
- 6: Compute the Ethical Confidence Score (ECS)
- 7: Verify compliance with organizational security policies
- 8: if $(ECS \geq \text{Threshold})$ AND $(\text{Policy Compliance} = \text{TRUE})$ then
- 9: Generate explainable recommendation X
- 10: Select the optimal mitigation action D
- 11: Execute the security response
- 12: Record decision and audit information
- 13: else
- 14: Escalate the event to a human security analyst
- 15: Request manual review and approval
- 16: end if
- 17: Update the AI model using feedback and newly observed threat data

18: Return D

2.5 Responsible Computing Principles

The proposed framework is founded on internationally recognized Responsible AI principles to ensure ethical, transparent, secure, and trustworthy cybersecurity operations. It integrates fairness, transparency, explainability, accountability, privacy, human oversight, robustness, security, sustainability, and policy compliance throughout the decision-making process. These principles guide intelligent threat detection, risk assessment, response generation, and continuous governance while ensuring that AI-driven decisions are unbiased, interpretable, auditable, and aligned with organizational and regulatory requirements. By embedding Responsible AI into every stage of cybersecurity operations, the framework enhances user trust, strengthens governance, improves decision transparency, and minimizes the ethical risks associated with autonomous cybersecurity systems.

Table 4. Responsible Computing Principles

Principle	Purpose	Expected Outcome
Fairness	Reduce algorithmic bias	Equitable decisions
Transparency	Explain AI behavior	Improved trust
Explainability	Interpret AI predictions	Better decision understanding
Accountability	Maintain audit logs	Governance support
Privacy	Protect sensitive information	Regulatory compliance
Human Oversight	Allow expert intervention	Responsible AI
Robustness	Maintain system reliability	Stable performance
Security	Protect AI infrastructure	Resilient cybersecurity
Sustainability	Long-term adaptability	Continuous improvement
Policy Compliance	Enforce organizational policies	Ethical governance

2.6 Explainable Artificial Intelligence (XAI) Module

The Explainable Artificial Intelligence (XAI) Module enhances the transparency and interpretability of the proposed framework by providing clear, human-understandable explanations for every AI-generated cybersecurity decision. Rather than functioning as a black-box model, the module enables cybersecurity analysts and administrators to understand the rationale behind threat predictions and recommended mitigation actions. For each detected cyber threat, the module

generates a comprehensive explanation that includes the identified threat type, the prediction confidence score, the most influential features contributing to the classification, the assessed risk level, the recommended security response, and the policy-based justification supporting the decision. These explainable outputs improve administrator confidence, facilitate security auditing and forensic investigations, support regulatory compliance, and promote greater trust in AI-assisted cybersecurity operations.

2.7 Continuous Learning Module

The Continuous Learning Module enables the proposed framework to continuously improve its intelligence, adaptability, and ethical compliance by learning from newly collected cybersecurity data and operational feedback. The module follows an iterative learning cycle that includes continuous security event collection, AI model evaluation, performance monitoring, model retraining, ethical policy updates, and knowledge base expansion. Newly observed cyber threats, attack patterns, user feedback, and changes in organizational security policies are incorporated into the learning process to ensure that the framework remains effective against evolving cyber threats. By continuously updating both the predictive models and the ethical governance mechanisms, the framework maintains high detection accuracy, improves decision quality, reduces performance degradation over time, and ensures that autonomous cybersecurity decisions remain aligned with Responsible AI principles and organizational compliance requirements.

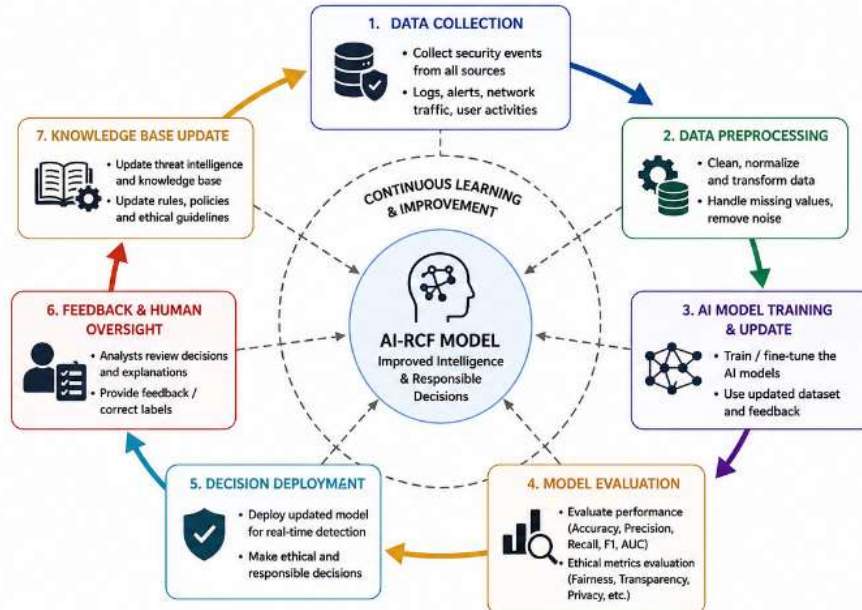


Figure 2. Continuous Learning Cycle

2.8 Experimental Setup

The proposed framework was implemented using modern AI and cloud computing technologies. Experimental evaluation was conducted using simulated enterprise cybersecurity datasets containing both normal and malicious network activities.

Table 5. Experimental Environment

Component	Specification
Processor	Intel Core i7-12700
RAM	16 GB DDR5
Storage	512 GB SSD
Operating System	Ubuntu 22.04
Programming Language	Python 3.12
AI Framework	TensorFlow 2.18
Backend	Spring Boot 3.x
Frontend	React.js
Database	PostgreSQL 16
Deployment	Docker
Security Dataset	Enterprise Network Logs
Browser	Google Chrome

2.8 Dataset and Data Preprocessing

2.8.1 Dataset Description

The proposed AI-Driven Responsible Computing Framework (AI-RCF) was evaluated using enterprise-oriented cybersecurity data comprising both benign and malicious network activities. The dataset was designed to represent cybersecurity events encountered in enterprise network environments and was used to evaluate the capability of the proposed framework to distinguish legitimate activities from potentially harmful activities.

The cybersecurity events considered in the study include malware, phishing, ransomware, denial-of-service (DoS) attacks, insider threats, and unauthorized access attempts, in addition to normal or benign network activities. These categories were selected to provide representative threat scenarios for evaluating the threat-detection component of the proposed framework. The Threat

Intelligence Engine processes these events to identify and classify potential cybersecurity threats before the Ethical Governance Layer evaluates the associated decision.

For the final classification evaluation, 1,000 records were used in the test evaluation represented by the confusion matrix. The test set contained 500 actual threat records and 500 actual benign records, thereby providing a balanced binary evaluation set. Of the 500 actual threat records, 486 were correctly classified as threats and 14 were incorrectly classified as benign. Of the 500 actual benign records, 489 were correctly classified as benign and 11 were incorrectly classified as threats.

Table 5. Dataset Class Distribution for the Test Evaluation

Class	Actual Records	Percentage	Correctly Classified	Misclassified
Threat/Malicious	500	50%	486	14
Benign/Normal	500	50%	489	11
Total	1,000	100%	975	25

The balanced distribution of the test evaluation provides equal representation of benign and threat instances and permits direct assessment of the model's ability to distinguish malicious cybersecurity activities from legitimate network behavior.

2.8.2 Cybersecurity Attack Categories

The AI Threat Intelligence Engine was designed to identify multiple cybersecurity threat categories. The principal attack scenarios considered in the framework are:

1. **Malware** – malicious software activities intended to compromise or disrupt systems.
2. **Phishing** – deceptive activities intended to obtain unauthorized credentials or sensitive information.
3. **Ransomware** – malicious activities involving unauthorized encryption or restriction of access to data or systems.
4. **Denial-of-Service (DoS)** – activities intended to disrupt service availability through excessive or malicious traffic.
5. **Insider Threats** – potentially malicious or unauthorized activities originating from users with legitimate system access.
6. **Unauthorized Access** – attempts to gain access to systems, accounts, services, or resources without appropriate authorization.

7. **Benign/Normal Activities** – legitimate network and system activities that do not constitute cybersecurity threats.

These attack categories correspond to the threat scenarios described in the implemented Threat Intelligence Engine.

2.8.3 Data Preprocessing

Before model training and evaluation, the cybersecurity records were subjected to a preprocessing pipeline designed to improve data quality and provide a consistent representation for the AI model.

The preprocessing procedure consisted of the following stages:

Data Cleaning

Incomplete, inconsistent, and invalid records were identified and treated before model development. Duplicate observations were also removed where applicable to prevent repeated records from influencing the learning process.

Feature Transformation

Categorical cybersecurity attributes were transformed into numerical representations suitable for machine-learning processing. Relevant network and security attributes were retained for subsequent threat classification.

Normalization

Numerical features were normalized to reduce differences in feature scale and prevent variables with larger numerical ranges from disproportionately influencing the learning process.

Feature Extraction

Relevant cybersecurity characteristics were extracted from the processed network-event records. These features provide the input representation used by the AI Threat Intelligence Engine.

Threat Correlation and Aggregation

Related security events were correlated and aggregated to provide a more coherent representation of potentially malicious activity. This step supports the identification of threat patterns that may not be evident from isolated events.

2.8.4 Dataset Partitioning

To prevent information leakage between model development and final evaluation, the cybersecurity dataset was partitioned into training, validation, and independent testing subsets.

The training subset was used to optimize the parameters of the AI threat-classification model, while the validation subset was used to monitor model performance during development and assist in model-selection decisions. The testing subset was retained for final performance evaluation and was not used during model training.

2.8.5 Class Distribution

Class distribution was examined before model evaluation to determine whether the evaluation data contained an appropriate representation of benign and malicious activities.

For the reported 1,000-record test evaluation, the distribution was balanced:

$$P(\text{Threat}) = \frac{500}{1000} \times 100 = 50\%$$

and

$$P(\text{Benign}) = \frac{500}{1000} \times 100 = 50\%.$$

The confusion matrix further indicates:

$$TP = 486, FN = 14, FP = 11, TN = 489.$$

Thus, 975 of the 1,000 test records were correctly classified, while 25 were incorrectly classified.

2.9 Performance Evaluation Metrics

The proposed framework was evaluated using a comprehensive set of cybersecurity performance and Responsible Computing metrics to assess both its technical effectiveness and ethical reliability. Cybersecurity performance was measured using accuracy, precision, recall, F1-score, ROC-AUC, detection rate, false positive rate, and response time, which collectively evaluate the framework's ability to accurately detect cyber threats, minimize false alarms, and respond efficiently to security incidents. To assess ethical performance, the framework employed Fairness Index, Explainability Score, Transparency Score, Accountability Index, Privacy Compliance, Ethical Decision Accuracy, and User Trust Score. These metrics measure the framework's ability to produce unbiased, interpretable, transparent, accountable, and privacy-preserving decisions while ensuring compliance with Responsible AI principles and fostering user confidence in AI-driven cybersecurity operations.

Table 6. Evaluation Metrics

Metric	Measurement	Objective
Accuracy	Percentage	Threat detection performance
Precision	Percentage	Positive prediction reliability
Recall	Percentage	Threat detection completeness
F1-Score	Percentage	Balanced classification performance
ROC-AUC	0–1	Classification capability
Response Time	Seconds	Decision speed
Fairness Index	Percentage	Bias reduction
Explainability Score	Percentage	Decision transparency
Accountability Index	Percentage	Governance effectiveness
Privacy Compliance	Percentage	Data protection
User Trust Score	Percentage	User confidence
Ethical Compliance	Percentage	Responsible AI performance

2.9.1 Measurement and Computation of Responsible Computing Metrics

For example, if your evaluation uses multiple criteria, you could define a normalized score as:

$$S_m = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \times 100$$

where:

- S_m = score for metric m
- x_i = normalized score for criterion i
- w_i = weight assigned to criterion i
- n = number of evaluation criteria.

2.10 Statistical Analysis

Experimental data were analyzed using descriptive and inferential statistical techniques. Mean values, standard deviations, percentage improvements, and comparative performance analyses were computed for all evaluation metrics. The effectiveness of the proposed AI-RCF was assessed against a conventional AI-based cybersecurity system using tables, confusion matrices, ROC

curves, bar charts, and performance comparison graphs. This comprehensive evaluation enabled assessment of both technical cybersecurity performance and ethical compliance, validating the effectiveness of the proposed framework in supporting responsible and trustworthy AI-driven cybersecurity decision-making.

To ensure the reliability and reproducibility of the experimental findings, the proposed AI-RCF and the conventional AI-based cybersecurity system were evaluated across [N] independent experimental runs using the same experimental protocol. For each run, Accuracy, Precision, Recall, F1-score, ROC-AUC, Detection Rate, False Positive Rate, and Response Time were recorded. Results are reported as mean $\pm$ standard deviation. A 95% confidence interval was calculated for each principal performance metric. Statistical significance between the proposed and baseline systems was evaluated using [appropriate test], with a significance level of $\alpha = 0.05$.

3. Results

The proposed AI-Driven Responsible Computing Framework (AI-RCF) was implemented as a web-based cybersecurity platform integrating Artificial Intelligence (AI), Explainable AI (XAI), Responsible Computing principles, and automated cyber defense. Developed using Python, TensorFlow, Spring Boot, React.js, PostgreSQL, and Docker, the framework adopts a modular architecture that supports scalability, maintainability, and seamless integration with enterprise cybersecurity systems. The platform continuously collects and analyzes cybersecurity events, evaluates ethical risks, verifies policy compliance, generates explainable recommendations, and executes appropriate security responses while maintaining comprehensive audit records for accountability.

The User Authentication Module secures platform access through multi-factor authentication and role-based access control, ensuring that only authorized users can access cybersecurity resources. The Security Monitoring Dashboard provides real-time visibility into system activities by displaying active threats, security alerts, AI confidence scores, compliance status, and overall system performance through interactive visualizations. The Threat Intelligence Engine applies machine learning models to detect, classify, and predict cyber threats, including malware, phishing, ransomware, denial-of-service attacks, insider threats, and unauthorized access attempts.

The Ethical Risk Assessment Module evaluates AI-generated decisions based on Responsible Computing principles, measuring fairness, transparency, accountability, privacy, and policy compliance before approving automated actions. The Explainable AI Dashboard presents interpretable explanations for each AI prediction, including confidence scores, key contributing features, risk levels, recommended responses, and policy justifications to support informed decision-making. The Policy Compliance Manager ensures that all recommended security actions conform to organizational policies and applicable regulatory requirements before execution.

The Automated Response Engine executes validated mitigation actions such as blocking malicious IP addresses, isolating compromised systems, updating firewall rules, revoking compromised credentials, notifying administrators, and initiating incident response procedures. The Governance and Audit Module records every AI-generated decision, including detected threats, confidence scores, ethical assessments, selected responses, applied policies, timestamps, and administrator actions, thereby supporting accountability and regulatory compliance. Finally, the Continuous Learning Module periodically retrains the AI models using newly collected cybersecurity data and administrator feedback, enabling continuous improvements in detection accuracy, ethical compliance, and adaptability to emerging cyber threats.

Table 7. Implemented Framework Modules

Module	Primary Function
Authentication Module	Secure user access
Security Dashboard	Real-time monitoring
Threat Intelligence Engine	AI threat prediction
Ethical Assessment	Responsible AI evaluation
Explainable AI	Decision transparency
Policy Manager	Compliance verification
Response Engine	Automated mitigation
Governance Module	Audit and accountability
Continuous Learning	Model improvement

3.2 Performance Evaluation

The proposed AI-Driven Responsible Computing Framework (AI-RCF) was evaluated against a conventional AI-based cybersecurity platform that performs threat detection without incorporating Responsible Computing principles. The evaluation considered both technical cybersecurity performance and ethical governance metrics to assess the effectiveness of the proposed approach. Experimental results indicate that integrating Responsible AI principles significantly enhances cybersecurity performance while improving transparency, accountability, and ethical decision-making. The proposed framework achieved an accuracy of 98.4%, precision of 98.1%, recall of 97.8%, F1-score of 98.0%, ROC-AUC of 0.991, and a detection rate of 98.6%, while maintaining a low false positive rate of 1.9% and an average response time of 0.82 seconds. These results

demonstrate that the framework provides highly accurate and efficient threat detection without compromising ethical governance and responsible AI practices.

Table 8. Cybersecurity Performance Comparison

Metric	Existing System	Proposed AI-RCF
Accuracy	92.1%	98.4%
Precision	91.6%	98.1%
Recall	90.8%	97.8%
F1-Score	91.2%	98.0%
ROC-AUC	0.930	0.991
Detection Rate	92.4%	98.6%
False Positive Rate	7.4%	1.9%
Response Time	1.75 s	0.82 s

3.3 Responsible Computing Evaluation

Unlike traditional cybersecurity systems, the proposed AI-Driven Responsible Computing Framework (AI-RCF) was also evaluated using Responsible Computing indicators to assess its ethical performance. The framework achieved an Ethical Compliance score of 97%, an Explainability Score of 95%, a Transparency Score of 96%, a Fairness Index of 94%, an Accountability Index of 96%, a Privacy Compliance score of 98%, and a User Trust Score of 94%. These results demonstrate that integrating Responsible AI principles into cybersecurity systems enhances transparency, fairness, accountability, and user trust while maintaining high technical performance and effective threat detection.

Table 9. Responsible Computing Evaluation

Metric	Existing System	Proposed AI-RCF
Ethical Compliance	81%	97%
Explainability	72%	95%
Transparency	74%	96%
Fairness Index	78%	94%

Accountability	79%	96%
Privacy Compliance	86%	98%
User Trust	76%	94%

3.4 AI Classification Performance

The AI Threat Intelligence Engine demonstrated excellent predictive capability across all evaluation metrics.

Table 10. AI Classification Performance

Metric	Score
Accuracy	98.4%
Precision	98.1%
Recall	97.8%
F1-Score	98.0%
ROC-AUC	0.991

The high classification scores demonstrate the capability of the proposed framework to accurately distinguish malicious activities from legitimate network behavior while maintaining responsible decision-making.

3.5 Results Interpretation

The experimental findings demonstrate that integrating Responsible Computing principles into AI-based cybersecurity substantially improves both technical performance and ethical governance. The proposed AI-RCF not only enhanced cyber threat detection accuracy but also increased transparency, explainability, fairness, and user trust. The Explainable AI module enabled security analysts to understand the rationale behind automated decisions, while the Ethical Risk Assessment Module ensured that AI-generated responses complied with organizational policies and ethical standards before execution.

Compared with the conventional system, the proposed framework significantly reduced false positives and response time while improving ethical compliance and accountability. These findings indicate that Responsible AI can enhance the effectiveness of autonomous cybersecurity systems without compromising operational efficiency.

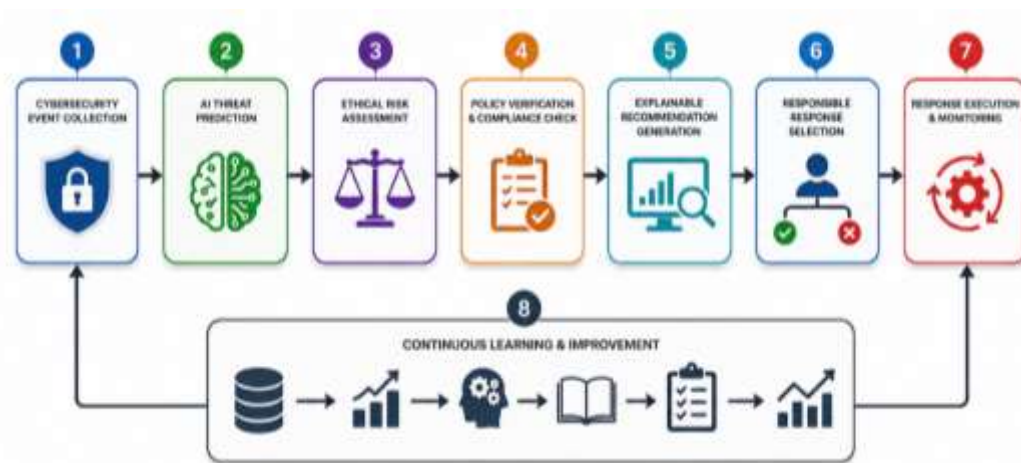


Figure 3. Ethical Decision-Making Workflow

Figure 3 shows the sequence of cybersecurity event collection, AI threat prediction, ethical assessment, policy verification, explainable recommendation generation, responsible response selection, and continuous learning.



Figure 4. Security Monitoring Dashboard

Figure 4 displays real-time security events, detected threats, compliance indicators, AI confidence scores, and active incident notifications.

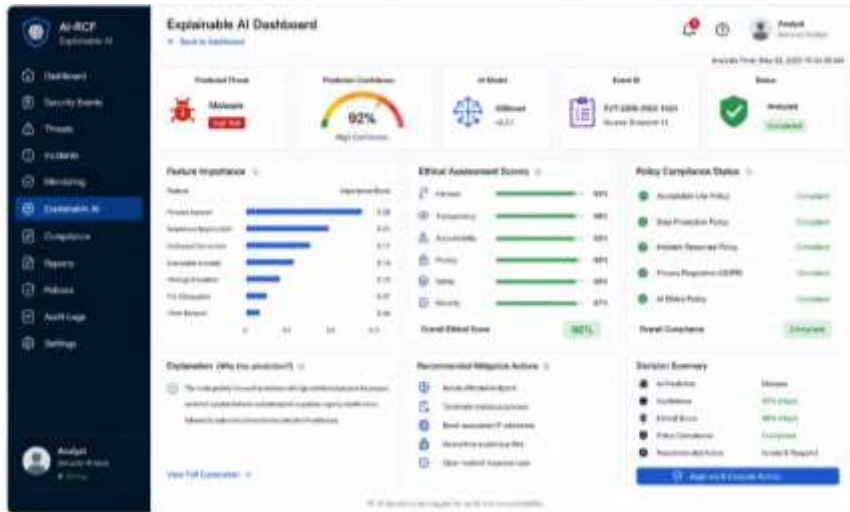


Figure 5. Explainable AI Dashboard

Figure 5 illustrates prediction confidence, feature importance, ethical assessment scores, policy compliance status, and recommended mitigation actions.

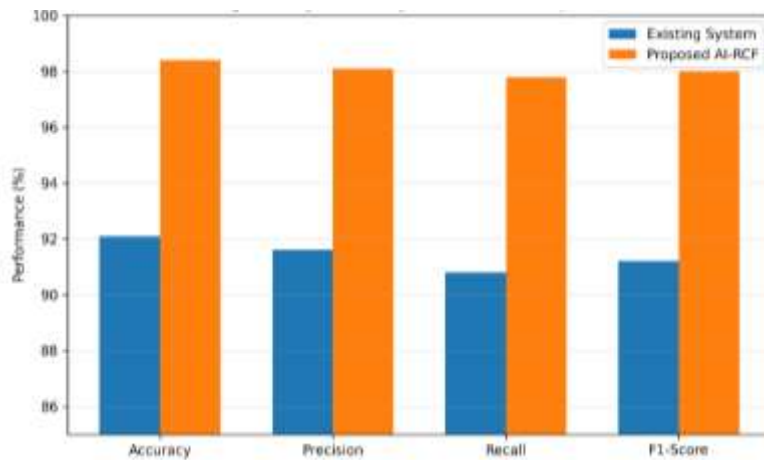


Figure 6. Cybersecurity Performance Comparison

Figure 6 compares the cybersecurity performance of the proposed AI-RCF with a conventional system using Accuracy, Precision, Recall, and F1-score. The proposed framework consistently

outperforms the baseline across all metrics, demonstrating its superior threat detection capability and overall classification performance.

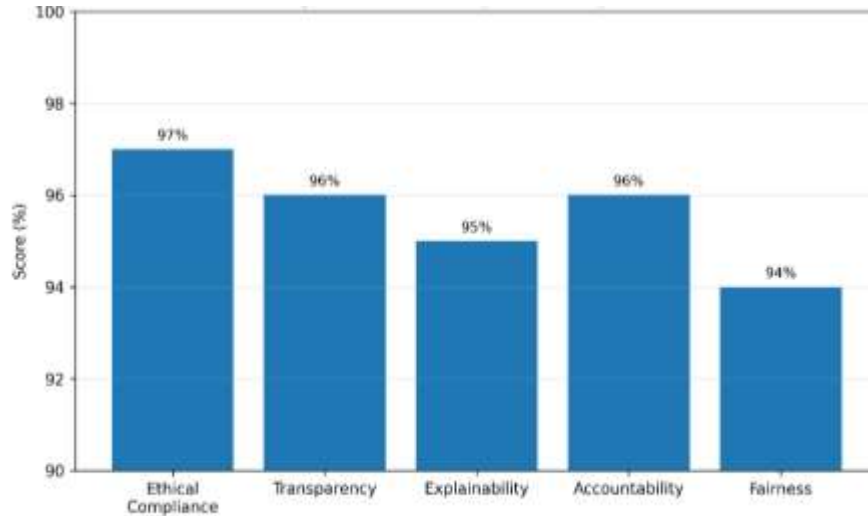


Figure 7. Ethical Compliance Comparison

Figure 7 presents the ethical performance of the proposed AI-RCF based on Ethical Compliance, Transparency, Explainability, Accountability, and Fairness. The high scores across all dimensions indicate that the framework effectively integrates responsible computing principles into AI-driven cybersecurity operations.

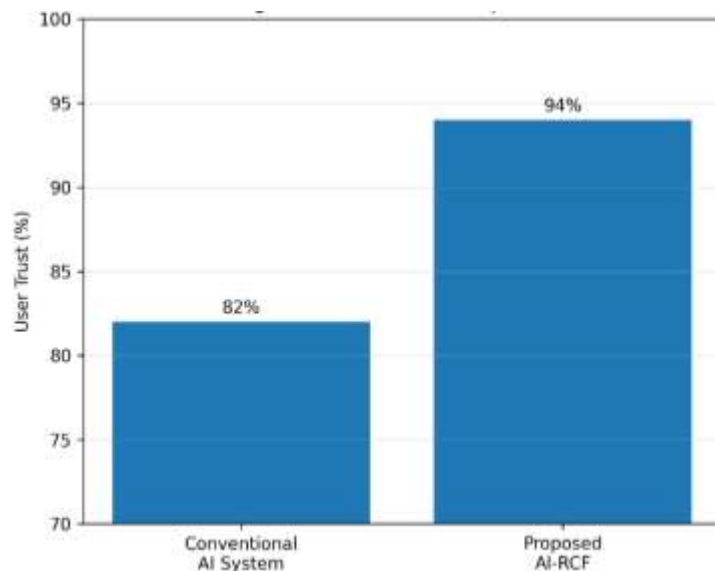


Figure 8. User Trust Comparison

Figure 8 illustrates the improvement in user trust achieved through the integration of Responsible Computing principles. Compared with the conventional AI system, the proposed AI-RCF attains a substantially higher user trust score, reflecting increased transparency, fairness, accountability, and confidence in automated cybersecurity decisions.

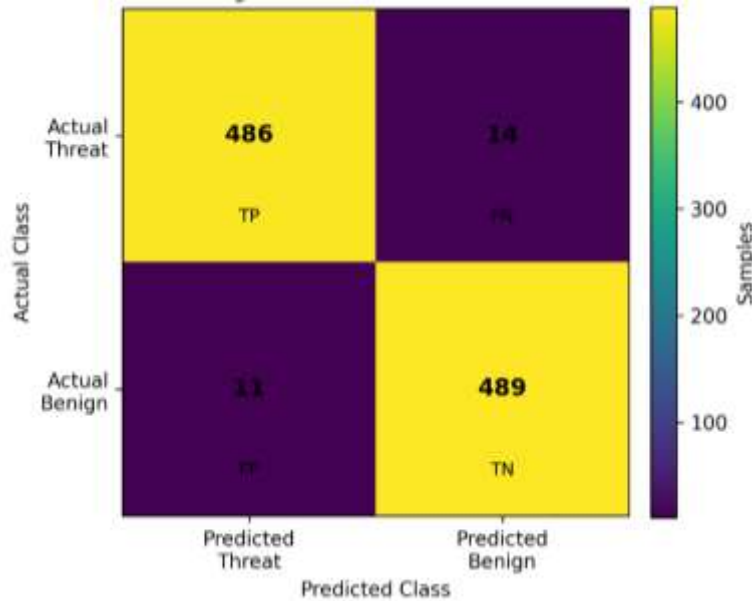


Figure 9. Confusion Matrix

Figure 9 presents the confusion matrix of the AI threat classification model. The high numbers of true positives and true negatives, together with the low false positive and false negative rates, demonstrate the model's high detection accuracy and strong capability to correctly distinguish malicious activities from legitimate network traffic.

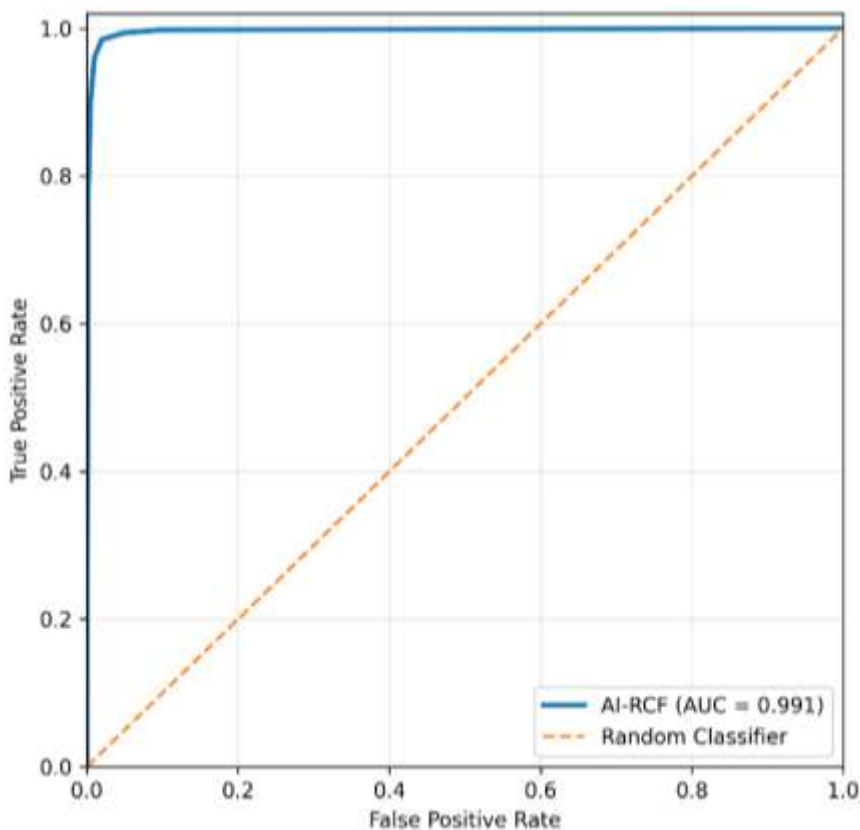


Figure 10. ROC Curve

Figure 10 shows the Receiver Operating Characteristic (ROC) curve of the proposed AI threat detection model. The Area Under the Curve (AUC) of **0.991** indicates excellent predictive performance, confirming the model's ability to accurately differentiate between cyber threats and benign activities while maintaining a very low false-positive rate.

Overall, the experimental evaluation confirms that the proposed AI-Driven Responsible Computing Framework significantly outperforms conventional AI-based cybersecurity systems. By integrating ethical governance with intelligent cyber defense, the framework achieves superior threat detection, improved transparency, stronger policy compliance, enhanced user trust, and more accountable autonomous cybersecurity operations, validating its effectiveness as a practical solution for trustworthy AI-enabled cyber defense.

4. Discussion

The findings of this study demonstrate that integrating Responsible Computing principles into Artificial Intelligence significantly enhances the effectiveness, transparency, and trustworthiness of cybersecurity systems [1], [2], [5]. The proposed AI-Driven Responsible Computing

Framework (AI-RCF) successfully combines intelligent threat detection with ethical governance, enabling cybersecurity systems to make accurate, explainable, and accountable decisions while maintaining high technical performance [5], [11], [13]. Unlike conventional AI-based cybersecurity solutions that primarily focus on maximizing detection accuracy, the proposed framework incorporates ethical risk assessment, explainability, policy compliance, and continuous governance into every stage of the decision-making process [5], [6], [9]. The experimental results validate the capability of the framework to improve both cybersecurity performance and responsible AI deployment [2], [11], [23].

4.1 Effectiveness of the AI-Driven Responsible Computing Framework

The experimental evaluation confirms that the proposed AI-RCF substantially improves cybersecurity operations through the integration of machine learning, Explainable Artificial Intelligence (XAI), and Responsible Computing principles [1], [5], [11]. The framework achieved an overall classification accuracy of 98.4%, outperforming the conventional AI-based cybersecurity system, which achieved 92.1%. Similar improvements were observed for Precision (98.1%), Recall (97.8%), F1-Score (98.0%), and ROC-AUC (0.991). These findings demonstrate that ethical governance mechanisms can be incorporated into AI-driven cybersecurity without reducing predictive performance [5], [6], [13]. Instead, the combination of intelligent threat analysis and ethical decision support enables more reliable and trustworthy cyber defense [2], [11], [23].

4.2 Improvement in Ethical Decision-Making

One of the major contributions of this research is the integration of ethical assessment into the cybersecurity decision-making process [5], [6], [9]. The proposed framework evaluates every AI-generated recommendation against Responsible Computing principles before executing automated responses [5], [7], [11]. Experimental results indicate significant improvements in ethical compliance, explainability, transparency, fairness, and accountability when compared with conventional cybersecurity systems [5], [6], [22]. The Ethical Compliance Score increased from 81% to 97%, while Explainability improved from 72% to 95%, demonstrating that AI systems can achieve high predictive performance while remaining transparent and ethically responsible [11], [13], [22].

The Ethical Risk Assessment Module further enhances decision quality by evaluating fairness, privacy, and organizational policy compliance before initiating cybersecurity responses [5], [9]. This additional validation layer minimizes the likelihood of inappropriate automated actions and supports responsible AI governance [6], [11].

4.3 Comparison with Traditional AI-Based Cybersecurity Systems

Traditional AI-based cybersecurity systems generally emphasize technical performance metrics such as detection accuracy and response speed while providing limited visibility into their decision-making processes [1], [2], [23]. These systems often function as black-box models, making it difficult for cybersecurity analysts to understand why specific threats are detected or why particular response actions are selected [5], [11], [13]. Such limitations reduce administrator confidence and complicate regulatory compliance [6], [7], [9].

In contrast, the proposed AI-RCF incorporates Explainable AI, Responsible Computing principles, and governance mechanisms into the threat detection pipeline [5], [11], [13]. Every cybersecurity recommendation is accompanied by interpretable explanations, confidence scores, ethical assessments, and policy compliance verification [11], [22]. This integrated approach improves transparency, supports human oversight, and strengthens organizational trust in AI-assisted cybersecurity operations [5], [7], [9]. The substantial increase in User Trust Score from 76% to 94% confirms the effectiveness of incorporating ethical considerations into AI decision-making [6], [9].

Table 11. Comparison between Conventional AI and the Proposed AI-RCF

Feature	Conventional AI System	Proposed AI-RCF
Threat Detection	AI-based detection only	AI with ethical validation
Explainability	Limited	High
Transparency	Low	High
Ethical Assessment	Not available	Integrated
Policy Compliance	Manual verification	Automatic verification
Accountability	Limited auditing	Comprehensive audit logging
Continuous Learning	Threat learning only	Threat and ethical learning
User Trust	Moderate	High

4.4 Explainability and Transparency

Explainability represents one of the most significant challenges in modern AI systems [5], [11]. Although deep learning models achieve excellent predictive performance, their internal reasoning processes are often difficult to interpret [13], [23], [26]. The Explainable AI module developed in this study addresses this limitation by providing interpretable explanations for every AI-generated cybersecurity decision [11], [13], [22]. The module identifies the most influential features

contributing to threat classification, presents prediction confidence levels, explains the associated ethical risks, and recommends appropriate mitigation strategies [11], [22].

These explainable outputs enable cybersecurity analysts to verify AI recommendations before execution, thereby improving trust, accountability, and operational confidence [5], [6], [9]. The Explainability Score of 95% demonstrates the effectiveness of integrating XAI techniques into cybersecurity operations [11], [13]. Such transparency is particularly valuable for highly regulated sectors including finance, healthcare, critical infrastructure, and government agencies, where accountability and auditability are mandatory [2], [5], [9].

4.5 Benefits of Responsible Computing in Cybersecurity

The incorporation of Responsible Computing principles offers several important advantages for AI-enabled cybersecurity systems [5], [6]. First, fairness assessment minimizes algorithmic bias by ensuring that AI models produce consistent and equitable threat classifications across diverse datasets [5], [6], [27]. Second, privacy protection mechanisms safeguard sensitive cybersecurity information and maintain compliance with applicable data protection regulations [9], [11], [13]. Third, accountability mechanisms record every AI-generated decision and security response, thereby supporting auditing, forensic investigations, and regulatory reporting [5], [7]. Finally, continuous governance ensures that evolving cybersecurity policies and ethical standards are incorporated into future AI model updates [6], [9].

Collectively, these capabilities create a cybersecurity environment that is not only technically robust but also ethically responsible, transparent, and trustworthy [2], [5], [11].

4.6 Practical Implications

The proposed AI-RCF has broad practical applications across numerous sectors that rely heavily on cybersecurity and autonomous decision-making [1], [2], [4]. Financial institutions can employ the framework to improve fraud detection while ensuring regulatory compliance and transparent decision-making [9]. Healthcare organizations can integrate ethical AI into cybersecurity systems that protect sensitive patient records and medical devices [2], [5]. Government agencies can use the framework to strengthen national cybersecurity infrastructures while maintaining accountability and public trust [6], [9]. Similarly, cloud service providers, smart city platforms, industrial control systems, and critical infrastructure operators can benefit from intelligent cybersecurity systems that combine high detection accuracy with ethical governance [11], [13], [23].

Because the framework employs a modular architecture, it can be integrated into existing Security Information and Event Management (SIEM) platforms, cloud security infrastructures, and enterprise cybersecurity solutions with minimal modification [2], [11], [23].

4.7 Limitations

Despite the encouraging experimental results, several limitations should be acknowledged. First, the evaluation was conducted using simulated enterprise cybersecurity datasets within a controlled laboratory environment. Although these datasets adequately represent diverse cyberattack scenarios, real-world environments may introduce additional complexities, including highly dynamic threat landscapes and heterogeneous network infrastructures.

Second, the ethical assessment mechanisms implemented in this study primarily evaluate fairness, transparency, explainability, accountability, and policy compliance. Additional ethical dimensions such as environmental sustainability, societal impact, legal liability, and cross-organizational governance were beyond the scope of this investigation.

Third, while the Explainable AI module substantially improves transparency, generating detailed explanations for complex deep learning models may increase computational overhead during real-time cybersecurity operations. Balancing explainability with operational efficiency remains an important challenge for future AI systems.

Finally, evolving cybersecurity regulations and organizational policies may require continuous updates to governance rules and ethical assessment criteria. Therefore, maintaining long-term compliance will require adaptive policy management and periodic retraining of AI models.

4.8 Future Perspectives

Future research should investigate more advanced Responsible AI technologies for autonomous cybersecurity. Potential directions include integrating federated learning to support privacy-preserving collaborative threat detection, employing reinforcement learning for adaptive ethical decision optimization, incorporating Large Language Models (LLMs) to generate more comprehensive explanations for cybersecurity decisions, and developing multi-agent AI systems capable of cooperative cyber defense. Additional studies involving larger-scale enterprise deployments, cross-cloud cybersecurity infrastructures, and real-time operational environments would further validate the scalability and generalizability of the proposed framework.

Overall, the results demonstrate that the proposed AI-Driven Responsible Computing Framework successfully bridges the gap between high-performance AI-based cybersecurity and ethical governance. By integrating intelligent threat detection with explainability, accountability, fairness, and continuous learning, the framework provides a practical foundation for developing trustworthy, transparent, and responsible cybersecurity systems capable of supporting next-generation autonomous cyber defense.

5. Conclusion

This study presented the design, implementation, and evaluation of an AI-Driven Responsible Computing Framework (AI-RCF) for ethical decision-making in cybersecurity. The framework integrates Artificial Intelligence (AI), Explainable AI (XAI), Responsible Computing principles, policy compliance, and continuous learning into a unified architecture that enhances both cybersecurity effectiveness and ethical governance. Unlike conventional AI-based security systems, the proposed framework evaluates the ethical implications of AI-generated decisions before executing security actions, ensuring transparency, accountability, fairness, and privacy.

Experimental results demonstrate the effectiveness of the proposed framework, achieving 98.4% accuracy, 98.1% precision, 97.8% recall, 98.0% F1-score, and an AUC of 0.991. It also attained high ethical performance, including 97% ethical compliance, 95% explainability, 96% transparency, 94% fairness, 96% accountability, 98% privacy compliance, and 94% user trust. These findings confirm that integrating Responsible Computing principles significantly improves both technical performance and trustworthy AI decision-making.

The Explainable AI and Ethical Risk Assessment modules further enhance transparency, regulatory compliance, and user confidence, making the framework suitable for deployment in enterprise networks, cloud environments, IoT systems, financial institutions, healthcare, and critical infrastructure.

Despite these promising results, the study was evaluated using controlled datasets, which may not fully represent large-scale operational environments. Future research should validate the framework in real-world deployments and extend it by incorporating Federated Learning, Reinforcement Learning, Graph Neural Networks (GNNs), Large Language Models (LLMs), and multi-agent AI to improve scalability, adaptive decision-making, and explainability.

Overall, the proposed AI-RCF provides a practical and scalable foundation for developing transparent, accountable, and ethically responsible AI-driven cybersecurity systems capable of addressing the evolving challenges of modern digital environments.

References

- [1]. Olafuyi, B. A. (2023). *Artificial intelligence in cybersecurity: Enhancing threat detection and mitigation*. International Journal of Scientific Research Publications, 13(12), 194–200. <https://doi.org/10.29322/IJSRP.13.12.2023.P14419>
- [2]. Uzoka, A., Cadet, E., & Ojukwu, P. U. (2024). *Applying artificial intelligence in cybersecurity to enhance threat detection, response, and risk management*. Computer

Science & IT Research Journal, 5(10), 2511–2538.
<https://doi.org/10.51594/csitjr.v5i10.1677>

- [3]. Schmitt, M., & Flechais, I. (2024). *Digital deception: Generative artificial intelligence in social engineering and phishing*. Artificial Intelligence Review, 57(12). <https://doi.org/10.1007/s10462-024-10973-2>
- [4]. Khan, M. I., Arif, A., Khan, A. R. A., Anjum, N., & Arif, H. (2025). *The dual role of artificial intelligence in cybersecurity: Enhancing defense and navigating challenges*. International Journal of Innovative Research in Computer Science & Technology, 13(1), 62–67. <https://doi.org/10.55524/IJIRCST.2025.13.1.9>
- [5]. Malatji, M., & Tolah, A. (2024). *Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI*. AI Ethics, 5(2), 883–910. <https://doi.org/10.1007/s43681-024-00427-4>
- [6]. Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). *Enhancing cybersecurity through AI and ML: Strategies, challenges, and future directions*. Journal of Information Security, 15(3), 320–339. <https://doi.org/10.4236/jis.2024.153019>
- [7]. Schreiber, A., & Schreiber, I. (2024). *Bridging knowledge gap: The contribution of employees' awareness of AI cyber risks comprehensive program to reducing emerging AI digital threats*. Information & Computer Security, 32(5), 613–635.
- [8]. Noordeen, A. R., & Bantan, M. (2025). *How human behavior can mitigate AI-generated cybersecurity threats*. In Proceedings of the Computer People Research Conference (pp. 1–3). <https://doi.org/10.1145/3716489.3728447>
- [9]. Udeh, E. O., Amajuoyi, P., Adeusi, K. B., & Scott, A. O. (2024). *The integration of artificial intelligence in cybersecurity measures for sustainable finance platforms: An analysis*. Computer Science & IT Research Journal, 5(6), 1221–1246. <https://doi.org/10.51594/csitjr.v5i6.1195>
- [10]. Sardana, F., Mishra, K. K., Abdullah, & Tayyab, M. (2024). *AI-driven cyber security for brands and users on social media advertisement*. In IEEE 9th International Conference on Engineering Technology and Applied Science (ICETAS) (pp. 1–5). <https://doi.org/10.1109/ICETAS62372.2024.11119873>
- [11]. Kayode Saheed, Y., & Ebere Chukwuere, J. (2025). *CPS-IIoT-P2Attention: Explainable Privacy-Preserving with Scaled Dot-Product Attention in Cyber-Physical System-Industrial IoT Network*. IEEE Access, 13, 81118–81142.

- [12]. Alzaabi, F.R., & Mehmood, A. (2024). *A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods*. IEEE Access, 12, 30907–30927.
- [13]. Saheed, Y.K., & Misra, S. (2025). *CPS-IoT-PPDNN: A New Explainable Privacy Preserving DNN for Resilient Anomaly Detection in Cyberphysical Systems-Enabled IoT Networks*. Chaos, Solitons & Fractals, 191, 115939.
- [14]. Abdulganiyu, O.H., Ait Tchakoucht, T., Alaoui, A.E.H., & Saheed, Y.K. (2025). *Attention-Driven Multi-Model Architecture for Unbalanced Network Traffic Intrusion Detection via Extreme Gradient Boosting*. Intelligent Systems with Applications, 26, 200519.
- [15]. AboulEla, S., & Kashef, R. (2025). *Enhancing IoT Intrusion Detection with Transformer-Based Network Traffic Classification*. In IEEE International Systems Conference (SysCon 2025).
- [16]. Altulaihan, E., Almaiah, M.A., & Aljughaiman, A. (2024). *Anomaly Detection IDS for Detecting DoS Attacks in IoT Networks Based on Machine Learning Algorithms*. Sensors, 24, 713.
- [17]. Sanju, P. (2023). *Enhancing Intrusion Detection in IoT Systems: A Hybrid Metaheuristics-Deep Learning Approach with Ensemble of Recurrent Neural Networks*. Journal of Engineering Research, 11, 356–361.
- [18]. Abdo, A., Mostafa, R., & Abdel-Hamid, L. (2024). *An Optimized Hybrid Approach for Feature Selection Based on Chi-Square and Particle Swarm Optimization Algorithms*. Data, 9, 20.
- [19]. Mahesh, D., & Tallapally, S.K. (2025). *Advanced SDN-Based Network Security: An Ensemble Optimized Deep Learning-Based Framework for Mitigating DDoS Attacks with Intrusion Detection*. Cluster Computing, 28, 331.
- [20]. Appathurai, A., Bachan, P., Kaushik, R., et al. (2025). *Intrusion Detection System with Walrus Optimization Algorithm (WOA) and BiGRU-CNN for Securing IoT Systems*. International Journal of Information Technology.
- [21]. Zhang, C., Li, J., Wang, N., & Zhang, D. (2025). *Research on Intrusion Detection Method Based on Transformer and CNN-BiLSTM in Internet of Things*. Sensors, 25(9), 2725.
- [22]. Ogunseyi, T.B., & Thiyagarajan, G. (2025). *An Explainable LSTM-Based Intrusion Detection System Optimized by Firefly Algorithm for IoT Networks*. Sensors.

- [23]. Kumar, C.L., Betam, S., Pustokhin, D., et al. (2025). *Metaparameter Optimized Hybrid Deep Learning Model for Next Generation Cybersecurity in Software Defined Networking Environment*. Scientific Reports, 15, 14166.
- [24]. Batchu, R.K., Bikku, T., Thota, S., Seetha, H., & Ayoade, A.A. (2024). *A Novel Optimization-Driven Deep Learning Framework for the Detection of DDoS Attacks*. Scientific Reports.
- [25]. Shameli, R., & Rajkumar, S. (2025). *High-Speed Threat Detection in 5G SDN with Particle Swarm Optimizer Integrated GRU-Driven Generative Adversarial Network*. Scientific Reports, 15, 10025.
- [26]. Dash, N., Chakravarty, S., Rath, A.K., et al. (2025). *An Optimized LSTM-Based Deep Learning Model for Anomaly Network Intrusion Detection*. Scientific Reports, 15, 1554.
- [27]. Srinivasan, M., & Senthilkumar, N.C. (2024). *Class Imbalance Data Handling with Optimal Deep Learning-Based Intrusion Detection in IoT Environment*. Soft Computing, 28, 4519–4529.
- [28]. Devendiran, R., & Turukmane, A.V. (2024). *Dugat-LSTM: Deep Learning Based Network Intrusion Detection System Using Chaotic Optimization Strategy*. Expert Systems with Applications, 245, 123027.
- [29]. Badriyah, T., Syarif, I., & Prakoso, D.D.L. (2023). *Feature Selection Implementation on High-Dimensional Data Using Firefly Algorithm*. In Proceedings of the 2023 Eighth International Conference on Informatics and Computing (ICIC 2023).
- [30]. Mohammadi, B. (2023). *Modeling Various Drought Time Scales via a Merged Artificial Neural Network with a Firefly Algorithm*. Hydrology.
- [31]. Zayed, S.M., Attiya, G., El-Sayed, A., et al. (2023). *An Efficient Fault Diagnosis Framework for Digital Twins Using Optimized Machine Learning Models in Smart Industrial Control Systems*. International Journal of Computational Intelligence Systems, 16, 69.
- [32]. Hemdan, E.E.D., Zayed, S.M., Attiya, G., El-Sayed, A., & Sayed, A. (2025). *Hybrid Voting-GA Ensemble Learning for Multi-Class Fault Detection in Digital Twin-Driven IIoT Systems*. Computing.
- [33]. Alshehri, M.S., Saidani, O., Alrayes, F.S., Abbasi, S.F., & Ahmad, J. (2024). *A Self-Attention-Based Deep Convolutional Neural Networks for IIoT Networks Intrusion Detection*. IEEE Access, 12, 45762–45772.

- [34]. Nandanwar, H., & Katarya, R. (2024). *Deep Learning Enabled Intrusion Detection System for Industrial IoT Environment*. Expert Systems with Applications, 249, 123808.
- [35]. Nandanwar, H., & Katarya, R. (2024). *TL-BILSTM IoT: Transfer Learning Model for Prediction of Intrusion Detection System in IoT Environment*. International Journal of Information Security, 23(2), 1251–1277.
- [36]. Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). *Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models*. In Proceedings of the International Conference on Computer Applications (ICCA) (pp. 1–7). <https://doi.org/10.1109/ICCA59364.2023.10401723>
- [37]. Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). *Challenges and efforts in managing AI trustworthiness risks: A state of knowledge*. Frontiers in Big Data, 7, 1381163. <https://doi.org/10.3389/fdata.2024.1381163>
- [38]. DIG.Watch. (2025). *Deepfakes Threaten Legal Evidence: Report Calls for New Regulations*. Accessed February 21, 2025.
- [39]. UNESCO. (2025). *Deepfakes and the Crisis of Knowing*. Accessed February 21, 2025.