International Journal of Information Security (2026) 25:51
https://doi.org/10.1007/s10207-026-01218-0

REGULAR CONTRIBUTION

Hybrid AI-Based dynamic risk assessment framework with explainable AI practices for composite product cybersecurity certification

Shareeful Islam[1] · Bilal Sardar[1] · Eleni Maria Kalogeraki[2] · Kostas Lampropoulos[3] · Spyridon Papastergiou[4,5]

Received: 24 November 2025 / Accepted: 15 January 2026

Abstract

Cybersecurity certification generally relies on risk assessment results to identify suitable controls and assess the completeness of these controls for security requirement satisfaction and overall security assurance. Prioritization of relevant vulnerabilities is essential to support the risk assessment and overall conformity assessment. However, the security context has continuously evolved with variations in attack surfaces, vulnerability exploitation, and the regulatory landscape–factors that significantly impact the conformity assessment process. This research proposes a hybrid AI framework integrating ensemble learning with GPT-3.5 for effective risk management within composite product cybersecurity conformity assessment under the European Cybersecurity Certification Scheme. It operationalizes Explainable AI (XAI) practices using SHAP and LIME methods to identify the most influential features affecting vulnerability predictions, and applies marginal analysis to measure the quantifiable gap closure between required and actual security postures to validate security control adequacy and requirement satisfaction based on calculated risk levels. This facilitates the adoption of XAI in the context of cybersecurity certification, extending its utility beyond general AI-enabled application scenarios. An industrial pilot scenario based on the P-NET 5G/6G Testing and Integration Service infrastructure, along with a dataset-based experiment, was conducted to evaluate the proposed framework. The results indicate that the hybrid model achieved 89% accuracy for vulnerability exploitation score prediction, enabling accurate risk calculation for conformity assessment. Furthermore, the XAI analysis revealed that the identified security controls demonstrate adequate performance in satisfying mapped security functional requirements. Ultimately, the framework provides quantifiable validation of security control effectiveness, enabling auditors to trace the logical connections between vulnerability predictions, risk calculations, and security requirement satisfaction for an informed certification decision.

Keywords Hybrid AI model · Explainable AI · Dynamic risk assessment · European cybersecurity certification scheme · Risk-based protection Profile · Composite product

1 Introduction

In the modern digital world, digital systems are adopted in almost every sector to deliver software intensive system and services. Such systems offer significant performance, robust, and scalable benefits but also increase security risks (attack surface) due to the complex and interconnected subsystems, evolving threat landscape and vulnerabilities exploitation [1]. These systems generally include multiple hardware and software products that are often built and provided by different organizations based on diverse security requirements. Many of them are manufactured outside the EU, or incorporate freeware/open-source components, with unknown vulnerabilities and lack of security controls [2]. In this context, EU Cybersecurity Act (EUCSA 881 / 2019) entrenches the

✉ Shareeful Islam
shareeful.islam@aru.ac.uk

1 School of Computing and Information Science, Anglia Ruskin University, Cambridge, U.K.

2 Security Lab Consulting, Cork, Ireland

3 Emerging Networks & Vertical Applications, p-NET, Patras, Greece

4 Research and Innovation, MAGGIOLI S.P.A., Santarcangelo di Romagna, Italy

5 Department of Informatics, University of Piraeus, Piraeus, Greece

Published online: 10 February 2026

cybersecurity compliance framework within the EU which enables to use one mutually recognized certificate throughout the EU and offers a unified certification scheme [3]. Specifically, the purpose of this act is to ensure an adequate level of cybersecurity for Information and Communication Technology (ICT) products and services by evaluating to comply with specified security requirements. The European Cybersecurity Certification Scheme(EUCC) is the Common Criteria(CC) based approach which looks into the compliance assessment using CC evaluation practice, where any ICT product or composite product can be subject to a formal security evaluation also known as Conformity Assessment (CA) through which the compliance against specific selected security objectives and selected applicable requirements, (described either in a Security Profile or in a specific product Target of Evaluation(ToE) and Composite Product Target of Evaluation(CToE) is evaluated following a specific Cybersecurity Certification Scheme and ensures that such system is sufficiently protected from potential threats and risks and trusted to be deployed [4, 5].

There are several recent examples where AI models integral to security have become targets themselves. Recent research highlights the rise of adversarial attacks on XAI, where malicious actors attempt to manipulate explanations to hide threats or mislead auditors. Surveys have categorized the landscape of adversarial attacks and defenses in explainable AI [6], while specific studies demonstrate how XAI in cybersecurity can be 'Explainable but Vulnerable', allowing attackers to perturb inputs to mask malicious behaviour [7]. Additionally, research into cyber threat ontologies and adversarial machine learning has shown how prediction perturbations can compromise system integrity [8]. Understanding these dynamics is crucial for the cybersecurity industry to ensure that the AI systems used for certification are themselves robust against manipulation.

In this context, it is necessary to assess the effectiveness of implemented security controls to protect the Composite Product Target of Evaluation (C-ToEs) by analysing Security Functional Requirements (SFRs) and risks within a defined operating environment [9]. However, the certification processes are complex and characterized by a static approach that doesn't consider the continuous changes, improvements and enhancements of C-ToEs environment nor adoption of AI based support for domain specific compliance checking[5, 10]. The audit artefacts generated by the system context also need proper explanation for the entire conformity assessment[11]. For instance, auditor needs to justify how easy it is for an attacker to exploit a vulnerability given the specific set of applied security controls. Moreover, there are a large number of vulnerabilities published but not all of them are exploited; therefore, we need to prioritise the vulnerabilities that can be exploited and link them with risk and control for an effective conformity assessment. The novel contribution of this research is the hybrid AI framework with Explainable AI practice to support the composite product cybersecurity certification. This approach addresses the challenges of evolving security context by adoption of dynamic risk assessment and Explainable AI to interpret the effectiveness of security control and tracing with the vulnerabilities and controls. The reason for considering XAI for cybersecurity conformity assessment is due to its capabilities to transform opaque AI-driven risk predictions into transparent, audit-ready evidence. By leveraging techniques like SHAP and LIME, XAI provides the necessary interpretability to justify security control selection, ensuring that automated risk assessments meet the rigorous transparency and accountability standards required by certification schemes such as the EUCC. The work presents three novel contributions.

- Firstly, we propose a hybrid AI based dynamic risk assessment framework that leverages Random Forest, Gradient Boosting, and ElasticNet for intelligent feature selection, followed by LLM based GPT-3.5 for vulnerability exploitation prediction to support the risk assessment. The framework includes a systematic process to formulate risk-based protection profile and support the stakeholders with cybersecurity certification specifically for C-ToE based system. The hybrid model addresses the heterogeneous nature of cybersecurity data while maintaining computational efficiency through focused feature selection, directly supporting C-ToE conformity assessment processes.
- Secondly, we implement Explainable AI (XAI) techniques specifically to validate security control adequacy and requirement satisfaction for C-ToE systems. The XAI implementation employs three complementary approaches: marginal feature analysis to test control effectiveness through parameter variation, SHAP analysis to quantify how implemented controls contribute to vulnerability score reduction, and LIME analysis to explain specific vulnerability scenarios and identify control gaps. These techniques enable auditors to trace the logical connection between vulnerability predictions, control implementations, and security requirement satisfaction with interpretable evidence as a part of conformity assessment.
- Finally, we conducted a comprehensive evaluation to demonstrate the effectiveness of hybrid model for C-ToE vulnerability, risk, and conformity assessment, utilizing both a real-world pilot case and data set-based experiment. The pilot case study applies the framework to P-NET's Testing and Integration Service, a 5G/6G telecommunications infrastructure serving as a realistic testbed for composite ICT product certification. The experimental validation employs the CVEJoin dataset [12] to train and evaluate the hybrid ensemble-

LLM model to predict vulnerability exploitation scores. The evaluation includes performance comparison studies between ensemble-only methods, individual model approaches, and the proposed hybrid framework. Moreover, the scope of evaluation also focuses on the utility of XAI in verifying security controls and generating the necessary evidence for audit compliance.

2 Literature review

This literature review explores the recent developments in explainable artificial intelligence (XAI) for cybersecurity, explores AI-driven certification processes, and hybrid AI model for cybersecurity. Moreover, it it provides an overview of the relevant cybersecurity certification schemes.

2.1 Cybersecurity certification scheme

The cybersecurity certification scheme establishes a systematic framework for evaluating software and hardware such as ICT products, services, and processes within the European regulatory landscape. It is a formal procedure that provides independent verification and validation that ICT products meet specified security requirements and can adequately protect against identified threats and vulnerabilities [2]. The certification framework becomes particularly critical for composite ICT product, which consist of multiple interconnected components from different vendors that must work together securely while maintaining individual security properties and collective system integrity. The European Cybersecurity Certification Scheme is the Common Criteria based European candidate cybersecurity certification scheme (EUCC) which follows CC based Security Evaluation, and ISO/IEC 15408 and ISO/IEC 18045 [15]. The EUCC framework provides three assurance levels - Basic, Substantial, and High - with increasing rigor of evaluation methods and depth of security analysis required to demonstrate compliance with security objectives [16]. The EUCC enables any applicable ICT product to undergo a formal Conformity Assessment (CA)–a security evaluation that certifies compliance against specific security objectives and requirements [17]. A Target of Evaluation (TOE) comprises the product's software, firmware, or hardware, whereas for integrated systems, this extends to a Composite ToE (C-ToE) encompasses multiple individual TOEs that function requiring assessment of both individual component security and inter-component security relationships [18]. Security Profiles (SPs) define implementation independent security requirements for categories of ICT products that meet specific needs. The certification process is structured, starting with scope definition using the Target of Evaluation (ToE). There are two types of requirements need to be evaluated through the certification process including Security Functional Requirements (SFRs), which specify the mandatory security capabilities the ToE must implement (e.g., authentication), and Security Assurance Requirements (SARs), which define the necessary evaluation rigor to ensure the SFRs are correctly implemented, with confidence measured by the Evaluation Assurance Level (EAL).

However, the current certification processes are time-consuming and costly, characterized by a static approach that doesn't consider continuous changes within the system context [2, 5]. Specifically, it relies on point-in-time assessments that fail to capture the dynamic security posture of systems that continuously evolve through updates and evolving threats and vulnerability exploitation. This makes maintaining continuous conformity extremely challenging particularity for complex and composite systems. Moreover, the audit evidences generated by the applications, and system also need proper explanation for the entire conformity assessment process, particularly when automated security tools generate evidences . This necessity drives the need for AI based dynamic risk assessment approaches and Explainable AI (XAI) by enabling real-time risk evaluation, automating vulnerability assessments, and providing the transparent, auditable evidence required for ongoing conformity assessment of these complex ICT products. The Figure 1 shows how the certification scheme flows from EU Cybersecurity ACT to certification levels and challenges.

2.2 AI and explainable AI for cybersecurity conformity assessment

The integration of AI and XAI within the conformity assessment processes has recently getting attention particularly for evidence collection, assessment and justification with specific security requirements and objectives. Recent research demonstrates how AI enhances decision-making amidst complex risk scenarios and evolving regulatory landscapes, with AI-driven practices including proactive risk management, precise risk assessment, and real-time monitoring that reshape compliance operations [19]. The integration of explainable AI techniques enables auditors to validate automated compliance decisions and justify the adequacy of evidence against regulatory requirements, particularly in frameworks like GDPR, HIPAA, and PCI-DSS [20]. Papastergiou et al. addressed certification challenges by proposing a Composite Inspection and Certification (CIC) System specifically designed for cybersecurity assessment of interconnected ICT products, services, and processes within Digital Business Ecosystems [1]. The work emphasizes that product-level security assessment needs to consider not only individual component vulnerabilities but also cascading effects of cyber-attacks across interconnected systems. Investigation of regulatory implications incorporating XAI into cybersecurity frameworks highlights challenges of bal-

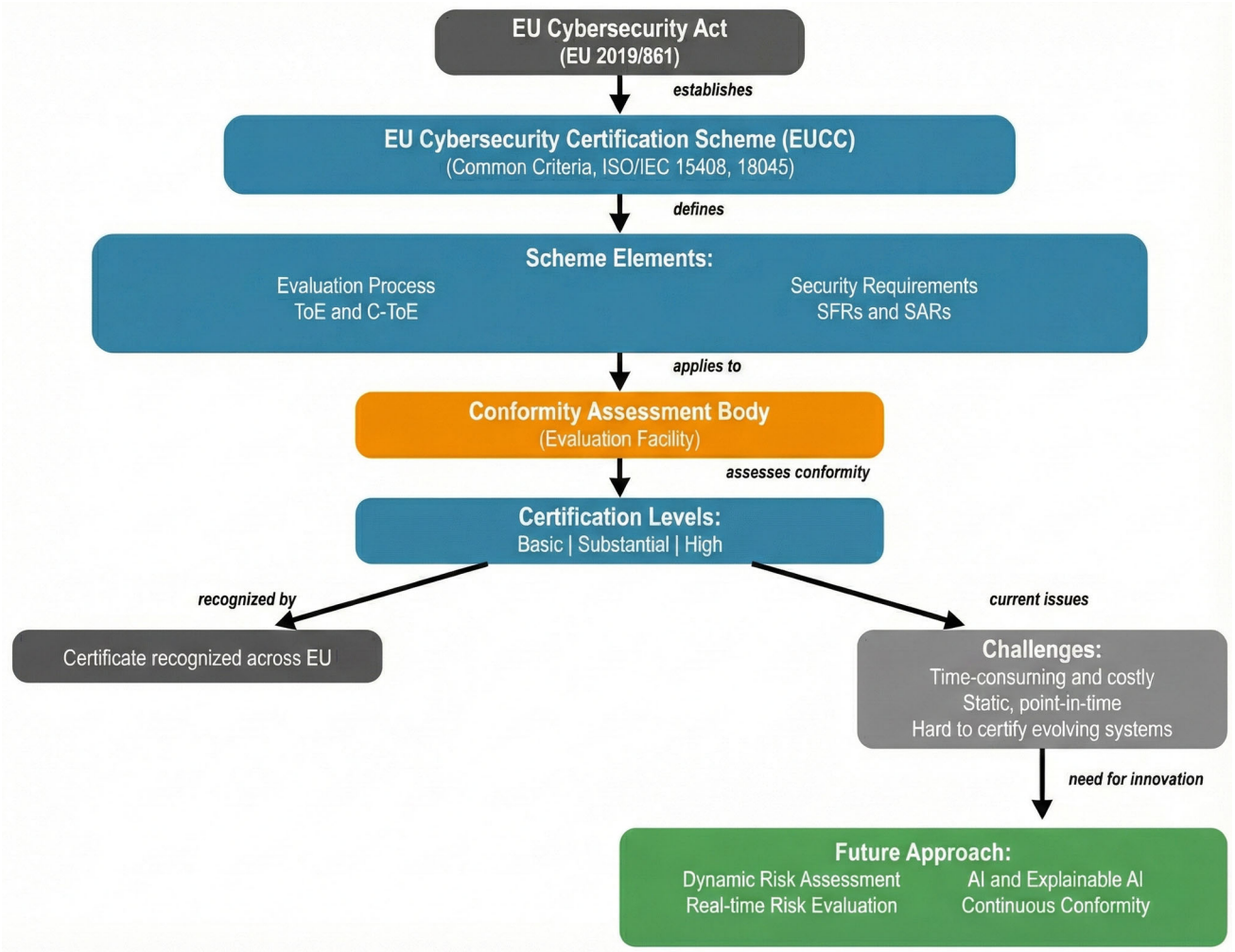

Fig. 1 EU Cybersecurity Certification Framework Flow.

ancing transparency with cybersecurity requirements and evaluates practical methodologies for XAI implementation in audit contexts [21]. Empirical validation comparing LIME and SHAP applicability demonstrates that both techniques effectively explain model decisions, with SHAP providing more consistent global explanations while LIME offers superior local interpretability for individual predictions [22]. This research provides practical evidence that XAI techniques can maintain detection accuracy while providing the transparency required for audit purposes, addressing the critical need for explainable automated decision-making in security compliance frameworks. Another dimension of XAI technique, i.e., Permutation Feature Importance (PFI), is also used to demonstrate feature importance, similar to SHAP. PFI is widely considered for specific AI models including Random Forest (RF) due to its model-agnostic capability to measure global feature significance by evaluating the decrease in model accuracy when a feature's values are randomly shuffled [23]. However, PFI can be biased when features are highly correlated, which necessitates careful application or the use of conditional variants [24, 25].

2.3 Hybrid AI models for cybersecurity risk assessment

The integration of hybrid AI architectures and large language models has emerged as a transformative approach in cybersecurity applications, offering enhanced capabilities for threat detection, vulnerability assessment, and security decision-making. Advanced hybrid frameworks integrate multiple learning paradigms including stacking ensemble methods, Bayesian model averaging, and conditional ensemble approaches to achieve enhanced accuracy in intrusion detection and threat classification [26]. These systems leverage meta-classifier approaches to combine outputs from diverse base classifiers, enabling robust performance across varied attack scenarios while maintaining explainable decision rationale. Large language model applications in cyber-

security demonstrate superior capabilities in understanding both natural language and code semantics, enabling comprehensive analysis of security-related documentation, code repositories, and threat intelligence [27]. Hybrid architectures combining LLMs with traditional machine learning methods have shown enhanced performance in vulnerability assessment applications, leveraging natural language understanding capabilities while maintaining efficiency and interpretability of conventional ML algorithms for structured data analysis [28]. These integrated approaches provide more holistic threat evaluation capabilities than single-paradigm solutions.

In summary, the existing works reveal the significant of AI for cybersecurity with limited focus on the certification. Moreover, the black-box nature of AI models creates fundamental challenges towards transparent and auditable decision-making process. The adoption of XAI techniques mainly focuses on cybersecurity context rather than formal cybersecurity certification schemes. Specifically,there is a lack of systematic comparison regarding the evolution of Large Language Models–from GPT-3's few-shot learning to GPT-4's complex reasoning–to determine the optimal balance of instruction-following and cost-efficiency required for certification tasks. Additionally, while Permutation Feature Importance (PFI) is recognized for demonstrating feature importance in specific models like Random Forest [23], its specific application alongside SHAP and LIME to mitigate correlation biases and validate audit evidence within a certification framework remains strictly limited. This research gap necessitates the development of a comprehensive framework that combines the predictive power of hybrid AI models with explainable techniques specifically tailored for cybersecurity conformity assessment, enabling certification bodies to leverage advanced AI capabilities while maintaining the transparency, traceability, and systematic validation required for formal certification processes in increasingly complex digital ecosystems.

3 Explainable AI for cybersecurity conformity assessment

The AI based cybersecurity solutions for threat detection, risk assessment, and security control selection present "black box" challenges for certification bodies for assessing the completeness of controls. XAI bridges the gap between advanced AI capabilities and certification requirements by making AI-driven security decisions interpretable, traceable, and auditable. The certification process requires clear documentation of security decision-making, influencing factors, and control selection rationale. The integration of XAI supports to this direction and transforms traditional audit methodologies from static checklist-based approaches to dynamic, evidence-based assessments.

3.1 XAI techniques and criteria

The existing XAI techniques can be broadly categorized based on their approach to generating explanations and the type of insights they provide [29]. Three primary explainable AI methods support cybersecurity certification: Marginal Feature Contribution (sensitivity analysis of parameter changes), SHAP (mathematically rigorous global/local explanations), and LIME (instance-specific interpretable models). As shown in Fig 2, these techniques differ in scope and presentation but collectively enable comprehensive AI decision explanations for certification purposes.

To align the evaluation of XAI with the specific assurance requirements of the EUCC and Common Criteria schemes [4, 15], this framework adapts standard interpretability metrics into certification-specific criteria. The evaluation focuses on two dimensions critical for generating valid audit evidence:

- *Fidelity and Technical Accuracy (Evidence Integrity):* In the context of certification, fidelity serves as a proxy for *evidence integrity*. It ensures that the AI-generated explanations accurately reflect the underlying risk logic required for the vulnerability assessment assurance class (AVA_VAN) [30]. Within the Common Criteria framework [4], **AVA_VAN (Vulnerability Analysis)** is the specific requirement family that compels evaluators to ascertain whether identified vulnerabilities are actually exploitable by an attacker with a specific potential (e.g., "Enhanced-Basic"). Technical accuracy requires that these explanations map directly to established security principles (e.g., separating "exploitability" from "impact"). For an auditor, low fidelity translates to "unreliable evidence," which would lead to a failure in verifying the Security Target. For example, accurately attributing a "High Risk" classification to a specific "public exploit availability" creates a verifiable audit trail [15].
- *Scalability (Audit Efficiency):* This criterion addresses the practical constraints of certifying Composite ToEs (C-ToEs), which generate massive volumes of vulnerability data. Scalability is evaluated not just as computational speed, but as *audit efficiency*–the ability to maintain explanation quality while processing thousands of assets [29]. Cognitive scalability is paramount; effective systems must provide hierarchical summaries (e.g., aggregating 300 network vulnerabilities into a single "High Risk" assurance claim) to prevent auditor fatigue and ensure that the certification process remains feasible within reasonable timeframes [5].

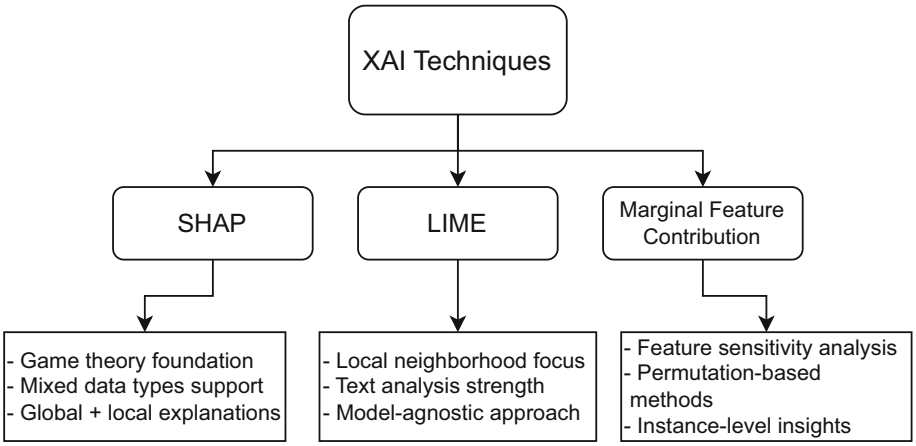

Fig. 2 Taxonomy of Explainable AI (XAI) techniques: SHAP for global model interpretation and LIME for local instance-level verification.

3.2 X-AI for cybersecurity certification

The integration of XAI can effectively support the cybersecurity certification processes and several existing standards and frameworks such as the EU Artificial Intelligence Act, ISO/IEC 42001 and the NIST AI Risk Management Framework – call for transparent, accountable AI-driven security systems that can demonstrate trustworthy operation throughout their lifecycle [31–33]. As XAI provides five critical capabilities for conformity assessment:

- Certification Challenge Resolution: XAI provides transparent insights into AI decision-making logic, enabling certification bodies to verify reasoning behind security recommendations and evaluate both system effectiveness and decision quality.
- Evidence Generation: XAI systematically documents security decisions, influential factors, and recommendation priorities, transforming subjective assessments into evidence-based evaluations with clear, traceable reasoning for audit purposes.
- Compliance Validation: XAI demonstrates alignment between AI-driven decisions and regulatory requirements, showing how recommendations support control objectives and comply with organizational security policies.
- Continuous Monitoring: XAI enables real-time compliance evaluation and multi-stakeholder communication through accessible explanations, supporting dynamic certification processes with early issue detection.
- Risk-Based Certification: XAI provides detailed insights into threat assessment and response prioritization, supporting sophisticated certification methodologies across diverse operational contexts.

3.3 Justification of selected XAI techniques for conformity assessment

The framework adopts a hybrid XAI approach derived from the taxonomy of explainable methods for cybersecurity established by Yan et al. [30] and Elkhawaga et al. [29]. While numerous interpretability methods exist, the selection of SHAP, LIME, and Marginal Analysis is empirically justified by the specific requirements of the Common Criteria (CC) and EUCC schemes. Certification audits require not just model transparency, but distinct types of evidence: global consistency to prove unbiased risk scoring, local fidelity to justify individual asset classifications, and sensitivity analysis to validate the adequacy of security countermeasures. Consequently, the framework integrates these three techniques as follows:

- *SHAP (Shapley Additive Explanations) for Global Consistency:* Selected for its rigorous foundation in cooperative game theory, SHAP provides the mathematical guarantee of consistency that other feature importance methods lack [22]. In a certification context, auditors must verify that the risk model treats similar vulnerability patterns fairly across the entire target system. SHAP values allow auditors to generate a global hierarchy of risk factors, empirically demonstrating which security features (e.g., severity metrics vs. exploit availability) systematically drive the risk assessment [30]. This global view enables the verification of the Security Problem Definition required by the Protection Profile.
- *LIME (Local Interpretable Model-agnostic Explanations) for Instance Verification:* While SHAP provides global consistency, LIME is selected for its superior computational efficiency and fidelity in generating local explanations for specific data points [22]. During conformity assessment, auditors often perform "spot checks" on specific high-risk assets to validate the decision logic. LIME facilitates this by creating a localized linear sur-

rogate model around a specific vulnerability instance, offering a granular explanation that allows for the traceback of evidence required by Assurance Class AVA_VAN without requiring a full re-computation of the global model [29].
– *Marginal Analysis for Control Adequacy Validation:* Standard XAI methods (SHAP/LIME) identify *which* features matter, but do not quantify *how much* improvement is required. To address the certification requirement of validating "Control Adequacy," the framework employs Marginal Analysis as a sensitivity measurement technique [5]. By systematically perturbing influential features identified by SHAP/LIME, this method quantifies the precise reduction in risk score achievable through optimal feature performance. This provides the empirical evidence needed to justify that *any* implemented security countermeasure effectively satisfies its mapped Security Functional Requirement (SFR), quantifying the gap between current and optimal protection [5].

4 Proposed framework

This section presents the proposed framework that integrates dynamic risk assessment with explainable AI techniques to enhance the conformity assessment process for composite products.

4.1 Adoption of hybrid AI model

The hybrid model combines ensemble learning with GPT-3.5 to improve vulnerability exploitation score prediction accuracy while maintaining computational efficiency. This integration provides three key benefits: ensemble methods excel at identifying predictive features from large cybersecurity datasets through statistical analysis [30]; GPT-3.5 provides superior contextual understanding to interpret complex feature relationships, recognizing patterns traditional algorithms miss regarding vulnerability attributes, exploit availability, and temporal factors [31]; and feeding only ensemble-selected features to GPT-3.5 achieves better accuracy while reducing computational overhead [32]. The model employs a two-stage approach:

– *Ensemble Feature Selection Stage:* incorporates Random Forest, Gradient Boosting, and ElasticNet to improve robustness and generalizability of feature selection, with Random Forest handling mixed data types effectively for heterogeneous vulnerability datasets [33], Gradient Boosting detecting complex non-linear interactions [34], and ElasticNet prioritizing sparsity and interpretability while reducing overfitting [35].
– *GPT-3.5 Contextual Analysis Stage:* receives distilled features from ensemble selection [36], concentrating on meaningful characteristics while reducing computational overhead, generating predictive scores based on contextual relationships among selected features and capturing complex dependencies that conventional statistical models may overlook. Figure 3 shows the hybrid AI model architecture where ensemble learning methods first identify important features from vulnerability data, which are then processed by GPT-3.5 for contextual analysis and final vulnerability score prediction.

4.2 Process

The process follows the Common Criteria-based EUCC scheme and consists of four sequential phases where initial phases define the scope of the C-ToEs , calculate the risk level and formulate the risk-based protection profile. Finally, the process operationalises the XAI to support the auditor with the conformity assessment. This process is unique in integrating dynamic risk assessment with explainable AI validation within the established Common Criteria framework, enabling objective quantification of control effectiveness beyond traditional checklist-based evaluations.

4.2.1 Phase 1: Define C-ToE

The C-ToE certification process initiates by defining the C-ToE where relevant ICT products and embedded assets are identified along with security declaration, serving as a necessary prerequisite for conformity assessment through two distinct steps. *Step 1.1: Create C-ToE* creates the C-ToE including ICT product or service under evaluation with key attributes including name, description and EAL (Evaluation Assurance Level) from a 7-level pre-defined scale based on CC assurance scales [15], where C-ToE owners decide the appropriate EAL considering product criticality, components, intended use, and potential impact upon compromising the specific C-ToE (e.g., EAL5 for high-security systems like smartcards, EAL7 for defense-grade environments), with higher EAL indicating higher assurance degree evidence and requiring more extensive testing, documentation, and depth of vulnerability analysis using the AVA assurance class of CC Framework. *Step 1.2: Asset Management and Security Declaration* identifies assets linked with the C-ToE and declares security specifications, where assets include various software, appliances or sub-services required to run the ICT product, recognized through:

– C-ToE Owner specifies cyber-asset technical details with Common Platform Enumeration (CPE) identifier using MITRE-NIST structured naming scheme [41]

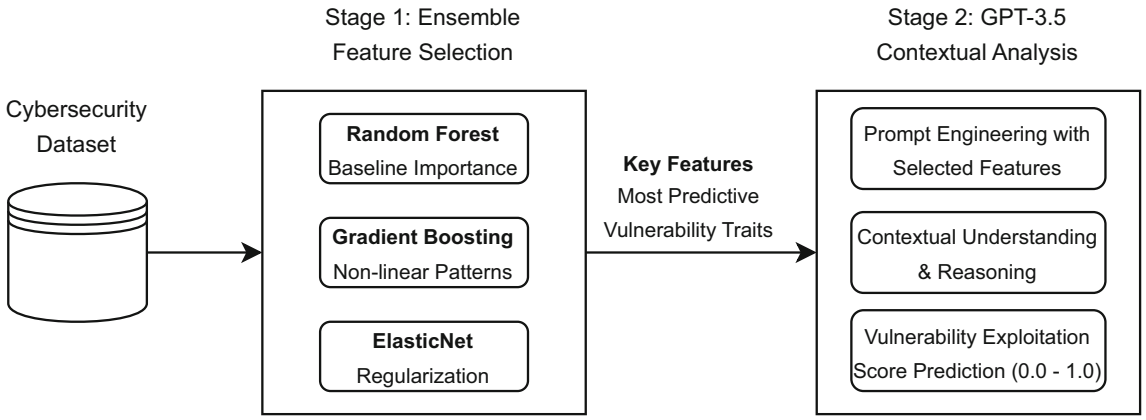

Fig. 3 Hybrid AI Model Architecture.

– C-ToE Owner provides Software Bill Of Materials (SBOM) file describing assets as Package URLs (PURL) for unified software package identification [42]

Security declaration includes security objectives and requirements linking identified assets with Security Functional Requirements (SFRs) based on CC Framework, where each security objective is addressed through SFR satisfaction via proper security controls implementation (authentication, encryption, access control), enabling auditors to evaluate whether specified security objectives and SFRs are met through XAI analysis of security controls implementation, with the phase outcome providing detailed C-ToE specification including ICT product with its EAL, related assets, and security objectives, requirements and controls for each asset.

4.2.2 Phase 2: Hybrid AI based dynamic risk assessment

Once the C-ToE is defined , it is necessary to predict the exploitation of the vulnerabilities related with the identified asset and further use the vulnerability scores to quantify the risk. Therefore, the aim of this phase is to adopt the proposed hybrid AI approach for dynamic risk assessment. The phase includes three tasks that cover vulnerability identification, hybrid model training for vulnerability exploitability prediction, and risk assessment through leveraging both ensemble learning and LLM capabilities.

Task 2.1: Train the hybrid model:

The C-ToE vulnerability dataset requires systematic analysis to identify and prioritize vulnerabilities based on their exploitation potential for risk assessment purposes. The implementation process begins with the systematic gathering of vulnerability and threat data from diverse industry sources. As illustrated in Fig 4, the framework aggregates data from static vulnerability databases (containing severity metrics), dynamic threat intelligence feeds (providing real-time exploitation status), and vendor-specific security advisories. This multi-source aggregation is critical for establishing a ground truth that reflects both technical severity and actual industry risk. Following gathering, the raw data undergoes rigorous preprocessing–including cleaning, filtering, and categorical encoding–to generate the structured training set required for the hybrid model. This task aims to extract key vulnerabilities that demonstrate high exploitation likelihood from the comprehensive vulnerability dataset, enabling effective mapping between vulnerability characteristics and exploitation probability. The task adopts the proposed hybrid model in a systematic manner to predict vulnerability exploitation scores, which are subsequently used to calculate relevant risk levels for conformity assessment. This task consists of four distinct steps that progress from data preprocessing through model evaluation.

*Data Preprocessing:*This step involves cleaning, formatting, and optimizing vulnerability data specifically for hybrid model training, focusing on vulnerabilities affecting C-ToE assets identified in Phase 1.

– **C-ToE Vulnerability Identification and Dataset Filtering:** The preprocessing filters the CVEJoin dataset to extract vulnerability records affecting C-ToE assets. CVEJoin aggregates vulnerability information from multiple sources including NVD, EPSS, and threat intelligence feeds. The filtering uses string-based matching on vendor and product columns with manual validation to ensure accurate asset-vulnerability association.
– **Data Filtering and Cleaning:** The system retains only entries matching C-ToE assets (e.g., Microsoft Windows, Apache HTTP Server). Missing values in critical fields are handled through removing incomplete records or using domain-specific default values.
– **Feature Engineering:** Additional features are created to improve prediction accuracy, including time elapsed

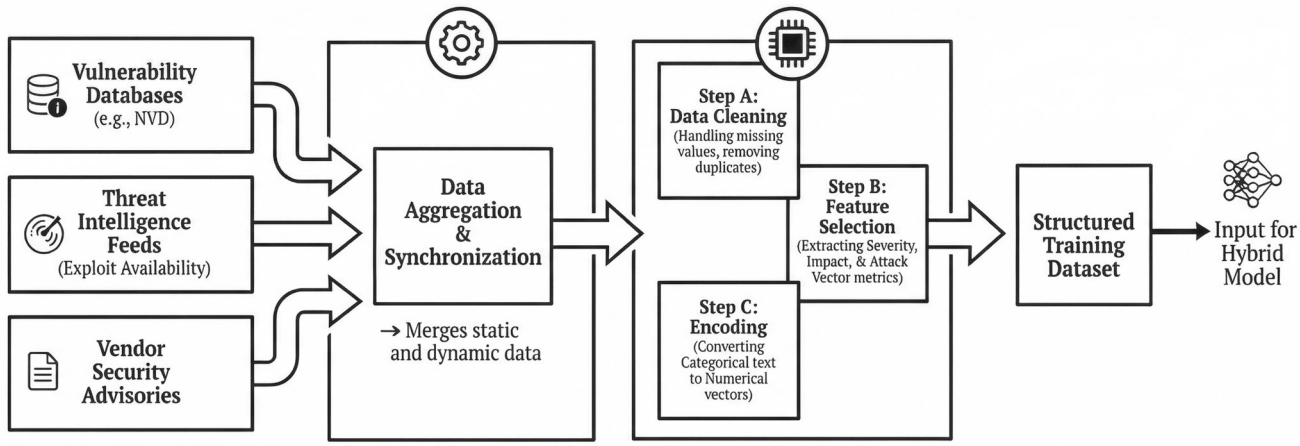

Fig. 4 Implementation Process showing the data gathering, aggregation, and preprocessing stages to generate the structured training dataset.

since vulnerability disclosure, binary indicators for exploit availability, and standardized categorical variables.

– **Dataset Splitting:** Preprocessed data is divided into training and test sets using stratified sampling to ensure representative distributions of vulnerability types, severity levels, and EPSS score ranges.

Key Feature Selection using Ensemble Learning: This step extracts key features using three machine learning algorithms that examine data from different perspectives to identify features predicting vulnerability exploitation probability. Random Forest builds multiple decision trees and calculates feature importance by measuring each feature's contribution to reducing prediction error, Gradient Boosting builds sequential models learning from previous mistakes to identify features most useful for correcting errors and capturing complex patterns, and ElasticNet applies mathematical penalties to irrelevant features while preserving important ones through regularization. Each method has distinct strengths: Random Forest handles mixed data types with stable rankings, Gradient Boosting captures complex non-linear patterns, and ElasticNet provides rigorous linear selection, but combining all three provides robust feature selection that leverages each method's benefits while mitigating individual weaknesses. Individual rankings are combined using weighted averaging with optimal weights determined through cross-validation experiments to optimize prediction performance while maintaining computational efficiency for GPT-3.5 processing.

GPT-3.5 Training for Vulnerability Score Prediction: Following feature selection from previous step, this step trains GPT-3.5 to predict vulnerability exploitation scores using only the ensemble-selected technical features. The rationale for model selection is, while newer models like GPT-4 offer enhanced reasoning capabilities, recent benchmarks indicate that for structured classification tasks, the performance gap is often negligible compared to the increased computational cost and latency. GPT-3.5 Turbo remains a highly efficient choice for specific, fine-tuned tasks where the goal is pattern recognition within defined technical constraints rather than open-ended creative generation. In this context, the selected features are converted into structured prompts that focus solely on technical vulnerability attributes, with prompt engineering creating templates that effectively communicate vulnerability characteristics to GPT-3.5 without organizational-specific context. This approach enables GPT-3.5 to learn general patterns in vulnerability exploitation without being tied to specific organizational contexts, making the model broadly applicable across different C-ToE configurations while developing understanding of how selected technical vulnerability characteristics influence exploitation probability in general cybersecurity contexts.

Model Performance: The hybrid model undergoes validation testing to assess its generalization capability. This helps determine whether it can correctly predict vulnerability scores in realistic scenarios – not just on the training data. To measure how well the model performs, three well-known regression metrics are used:

Coefficient of Determination($\mathbf{R}^2$): R^2 shows how much of the actual variation in vulnerability scores is explained by the model as mathematically presented in equation 1. As presented in equation 1, R^2 value close to 1 means the model's predictions align closely with real-world patterns, while a value close to 0 indicates poor predictive capability [43].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

Where:

– y_i the true vulnerability score for sample i

– $\hat{y}_i$ the predicted score from the model for sample i
– $\bar{y}$ the mean of all true scores in the test set
– n = total number of samples

Root Mean Square Error(RMSE): RMSE represents the average size of the error between predicted and true scores, with a higher penalty for larger mistakes as mathematically presented in equation 2. Lower RMSE values indicate that the model's predictions are generally close to the true values. RMSE is expressed in the same units as the target variable, making it directly interpretable [44].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

Where:

– y_i the true vulnerability score for sample i
– $\hat{y}_i$ the predicted score from the model for sample i
– N = total number of samples

Mean Absolute Error (MAE): MAE calculates the average absolute difference between predicted and actual scores, giving a simple, easy-to-interpret measure of how far off the predictions are on average as mathematically presented in equation 3. Unlike RMSE, MAE treats all errors equally without penalizing larger errors more heavily [45].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

Where:

– y_i the true vulnerability score for sample i
– $\hat{y}_i$ the predicted score from the model for sample i
– N = total number of samples

The above-mentioned metrics provide the capabilities to confirm how accurately the model is trained. That are not linked with each other, however collectively offer a multidimensional evaluation of model performance. R^2 indicates how well the hybrid approach captures vulnerability exploitation patterns by measuring the proportion of variance explained. RMSE and MAE provide complementary perspectives on prediction accuracy, with RMSE being more sensitive to outliers and MAE offering a more robust measure of typical error magnitude. Cross-validation ensures consistent performance across different data splits, validating the model's robustness for vulnerability score prediction applications. The integration of these metrics provides a comprehensive picture of hybrid model performance and serves as evidence for the added value of combining ensemble learning with GPT-3.5 contextual analysis, ultimately enhancing predictive accuracy and reliability in real-world cybersecurity scenarios.

Task 2.2: C-ToE vulnerability exploitation score

This task ranks the vulnerabilities based on the predicated exploitability score by using the proposed hybrid model. The rank is scaled into five priority levels based on their exploitation probability ranges:

– Very High (> 0.9): Vulnerabilities in this range are extremely likely to be exploited, and the associated risk can be materialized.
– High (0.7 - 0.9): Vulnerabilities in this range are highly likely to be exploited, and the associated risk can be materialized especially in unattended systems.
– Medium (0.4 - 0.69): Vulnerabilities in this range have a moderate chance of being exploited. The possibility of the associated risk being materialized depends on specific conditions.
– Low (0.2 - 0.39): Vulnerabilities in this range are unlikely to be exploited. Risk materialization is less likely to happen.
– Very Low (≤ 0.2): Vulnerabilities in this range are extremely unlikely to be exploited. They often possess minimal risk.

Task 2.3: Calculate risk level and controls declaration

This final task calculates the risk level for the identified asset so that suitable control can be selected according to the defined security declaration and risk level. The individual risk level calculation depends on the related vulnerability score for each ToE. The assessment generates a comprehensive risk profile that lists all vulnerability-asset combinations with their individual risk levels. This detailed approach enables security teams to prioritize specific vulnerabilities affecting specific assets, supporting more targeted and effective risk mitigation strategies Hence, each vulnerability-asset combination receives a risk classification based on the vulnerability's exploitation probability and potential impact:

– Very High Risk: Vulnerabilities that pose immediate and severe threats to the specific asset
– High Risk: Vulnerabilities that pose significant threats to the asset requiring prompt attention
– Medium Risk: Vulnerabilities that pose moderate threats to the asset under certain conditions
– Low Risk: Vulnerabilities that pose limited threats to the specific asset

– Very Low Risk: Vulnerabilities that pose minimal threats to the asset

Once the risks are prioritized, it is necessary to declare the controls for mitigating the risks. This control selection is a decision-making process to choose the right control strategy and suitable recommended controls. Hence, the C-ToEs Owner, based on the defined security objectives, the respective SFRs and the risk assessment results, shall undertake optimal decisions to apply security controls that mitigate vulnerabilities and/or threats and satisfy the defined SFRs on the C-ToEs. The control selection reviews the identified vulnerabilities as risk factor for the risk event. Therefore these risk factors allow to considers the most common vulnerabilities that have impacted on the most assets which require immediate attention. The control selection uses mapping procedure to recommends security controls provided by the NIST SP 800-53 catalogue per specific SFR [46]. The catalogue provides list of security and privacy controls to safeguard the assets of the organization

4.2.3 Phase 3: Generation of risk-based composite protection profile specification

This phase focuses on generation of risk-based Composite Protection Profile (C-PP) to the corresponding C-ToEs based on the identified assets, security objectives and requirements, risk assessment results from the previous phases. It is risk-based, as it addresses the security problem definition upon the produced risk assessment results from the previous phase. This provides the auditor to assess the possible claims for a product through this profile specifications. The generation of PP is supported by a unified Protection Profile structure and template consists of security objectives, SFRs and security controls by following the Common Criteria(CC) specification. The phase allows the assessor to evaluate the claims of the C-PP and investigate whether the corresponding C-ToEs meets specific requirements that rely on an adopted cybersecurity certification scheme. Hence, the assessor reviews the C-PP to validate whether the defined SFRs are satisfied based on the adopted EAL, the C-ToEs Owner has the capability to add evidence and additional information to justify for each SFR why and how the use of the corresponding security control(s) satisfy(-ies) it. The proposed risk-based C-PP is built in the three structure format and each level provides a distinct description of the features

– First level provides informative features related to the C-PP identification
– Second and third level sub-features further analyses the 1st and 2nd level consequently

– **Composite PP Introduction feature and sub-features**
  - *Composite PP Reference:* To provide informative content that identify the Composite PP using 3rd level sub-features C-PP Name, ID, Version, Owner, Publication date
  - *C-ToEs Overview:* To specify the content of the C-ToEs with 3rd level sub-features including C-ToEs Reference with ID, product, description and individual ToE detailed with name, product, version, CPE, etc
  - *Certification-related information:* To declare the C-PP certification status and past certification-related information including Declaration of Conformity and previous certification detail.

– **Composite PP Conformance Claims features and sub-features**
  – *EUCC Conformance claim:* To declare the conformance claim from EUCC
  – *CC Conformance claim:* To declare the conformance claim from CC including EAL claim

– **Security Declaration**
  – *Security Objectives and Security Functional Requirements per ToE:* To declare list of security objective ID and name, security functional requirements ID and name, and description by following the CC.
  – *Security Risks:* To specify list of security risks
  – *Security Controls:* To declare security control ID, name and description
  – *Security Assurance Requirements:* To declare composite CC EAL, Composite EU Assurance Level (EUCC), Attack Potential Level, and Security Assurance Rationale

4.2.4 Phase 4: Operationalise XAI for conformity assessment

This final phase adopts XAI techniques to support the conformity assessment process by validating whether security controls adequately satisfy the security requirements. The operationalization of XAI in this phase explains which features are most influential in the model's predictions and translates these insights into formal audit evidence. As illustrated in Figure 5, this phase follows three sequential steps:

4.2.4.1 Influential features extraction using SHAP and LIME

This step extracts the most influential features from the key features selected by the ensemble learning component in Phase 2, identifying which specific factors have the strongest

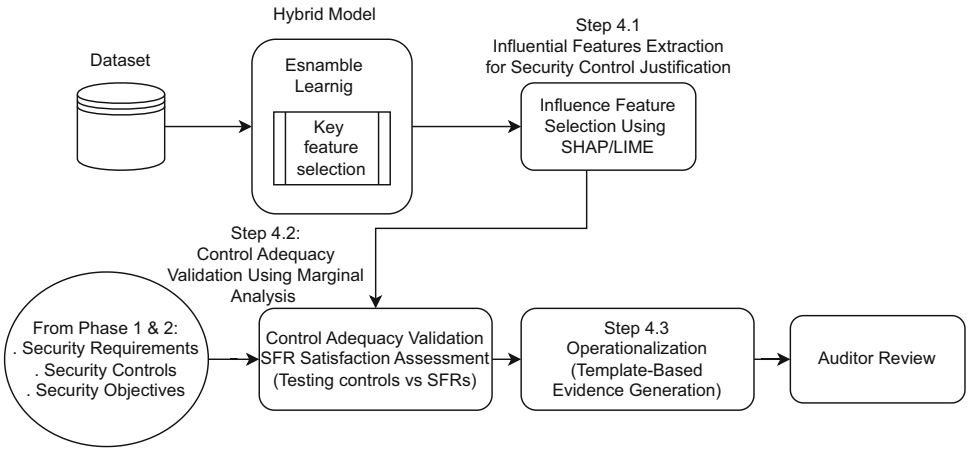

Fig. 5 Phase 4 XAI-based Conformity Assessment Process Flow.

impact on vulnerability score predictions and should be the focus for security control validation. SHAP analysis calculates Shapley values for each feature across all predictions to determine average contributions to final vulnerability scores, providing consistent global importance rankings by examining feature contributions across the entire dataset and identifying features with the most systematic impact on model predictions. LIME analysis generates local explanations for specific vulnerability predictions by creating simplified interpretable models around individual instances, perturbing input features of high-scoring vulnerabilities and observing how changes affect predictions to identify features most responsible for specific high-risk predictions. The combination of SHAP and LIME produces a refined set of influential features that have the strongest impact on vulnerability score predictions, providing a focused list of critical vulnerability characteristics that should be addressed by security controls.

4.2.4.2 Control adequacy validation via counterfactual sensitivity

This step validates security control adequacy by testing performance gaps between current and optimal control implementations. To ensure the XAI explanations are faithful and actionable, the framework operationalizes Marginal Analysis as a Counterfactual Sensitivity Test. The influential features from previous step connect to security controls through a clear relationship chain: features represent vulnerability characteristics, which affect assets protected by specific controls. When features show high influence, the related controls require validation. To achieve this, the framework simulates counterfactual scenarios asking the question: *"If the security control were fully optimized, how would the risk prediction change?"*

The process systematically varies each influential feature across its value range while holding other features constant. This allows the system to observe the impact on vulnerability score predictions and identify the values that produce the lowest risk scores (the "Optimal Counterfactual State"). By comparing the current feature distribution against this optimal state, the framework measures the potential improvement available through better control implementation. To ensure the assessment is rigorous, this step employs a defined decision framework based on statistical distribution and residual risk management:

1. **Risk Significance Classification (Finding the Drivers):** *Need:* Traditional risk assessments often rely on static thresholds (e.g., "Score > 0.5"), failing to account for varying risk baselines. *Solution:* The framework calculates dynamic materiality thresholds (T) based on the statistical distribution of SHAP values across the training set (Z-score principle). *Formula:* $T_{crit} = \mu + \sigma$. Features exceeding this threshold are classified as **"Critical Risk Drivers"** and selected for validation.
2. **The "Required" Gap Calculation (Setting the Standard):** *Logic:* Before simulating the counterfactual, we must determine how much improvement is necessary to make the asset safe.
   Formula: The **Adequacy Threshold** ($\tau_{required}$) is calculated as the percentage reduction needed to lower the asset's risk from its *Inherent Score* ($R_{inherent}$) to the *Acceptable Safety Limit* ($R_{acceptable}$) defined in the Security Target.

$$\tau_{required} = \frac{R_{inherent} - R_{acceptable}}{R_{inherent}} \times 100 \tag{4}$$

Theoretical Justification (Why it Works): This formula operationalizes the principle of **Residual Risk Management** required by ISO/IEC 27005 [33]. The numerator represents the "Excess Risk" that violates the safety standard. Dividing this by the total risk calculates the **Proportion of Risk** that must be eliminated. Therefore,

setting the threshold equal to this percentage mathematically guarantees that any "Adequate" control will reduce the residual risk to a level at or below the acceptable limit. This replaces subjective auditor opinions with an objective, calculated sufficiency criterion.

3. **The "Actual" Gap Measurement (Counterfactual Simulation):** *Logic:* The framework then uses Marginal Analysis to simulate the counterfactual "Optimal State" of the implemented control. It measures the **Actual Gap Closure** (ΔG_{actual})–the precise percentage difference between the *observed risk* and the *counterfactual risk* if the Critical Driver were fully mitigated.
4. **The Adequacy Check (The Comparison):** Finally, the framework compares the counterfactual performance against the *Required Standard*:
   - **Adequate** ($\Delta G_{actual} \geq \tau_{required}$): The control bridges the safety gap.
   - **Moderate** ($10\% \leq \Delta G_{actual} < \tau_{required}$): The control helps, but the asset remains above the safety limit.
   - **Inadequate** ($\Delta G_{actual} < 10\%$): The control is ineffective.

Operationalization via structured template-based generation

To operationalize the evidence translation, the final step employs a deterministic, template-based generation mechanism. This step functions strictly as a reporting engine; it consumes the *Classifications* produced in Step 4.2 (e.g."Adequate") and maps them to rigid string-based templates to ensure that every generated audit claim is technically accurate and legally reproducible.

The translation engine functions through a simple mapping pipeline:

- **Input Ingestion:** The system accepts the tuple (Risk Class, Control Status) derived in Step 4.2.
- **Feature Mapping:** Raw feature names are mapped to standardized regulatory terminology (e.g., `has_public_ exploit` → "Public Exploit Availability").
- **Template Instantiation:** The engine selects the pre- validated regulatory text block matching the input pair.

For example, a *"Adequate"* pair triggers the following template structure:

Template ID: TPL-CRIT-ADQ
Structure: *"The risk classification for [Asset Name] is driven by [Mapped Feature Name] (Importance: [SHAP Value]), which exceeds the critical threshold of* $[\mu + \sigma]$. *While the attack potential is elevated, the implemented control [Control ID] demonstrates a gap closure of* $[\Delta G\%]$, *satisfying the adequacy threshold for AVA_VAN compliance."*

By strictly adhering to these string-based definitions, the framework ensures that the generated evidence satisfies the *Repeatability* and *Reproducibility* requirements of the ISO/IEC 15408 evaluation methodology [4].

5 Evaluation

The evaluation considers a combination of pilot use case and experiment linked with security datasets to demonstrate the applicability of the proposed framework. This section presents the systematic implementation and evaluation of our proposed approach, demonstrating its effectiveness in supporting dynamic risk assessment for composite ICT product. The evaluation aims to validate using XAI techniques whether applied security controls adequately fulfil established security requirements, demonstrate the framework's capability to support cybersecurity certification activities.

5.1 Pilot use case scenario

P-NET is a pioneering Competence Center dedicated to advancing 5G/6G and emerging digital technologies, operating as an advanced 5G infrastructure facility providing carrier-grade and commercial-grade private 5G Standalone (SA) networks [48]. The organization serves as a dynamic testing and experimentation environment for pre-commercial verification, technological innovation and tailored vertical deployments under the Network as a Service (NaaS) model with customizable and programmable network capabilities. P-NET bridges the gap between academic research and industrial implementation of next-generation telecommunications technologies, providing comprehensive research and development services for advanced wireless communications, edge computing applications, IoT integration, and XR applications while collaborating with major telecommunications operators, equipment manufacturers, and academic institutions to accelerate 5G/6G solution development and deployment.

The Testing and Integration Service provides advanced software-intensive facilities for users to integrate and test new 5G/6G technologies in controlled, pre-commercial validation environments replicating real-world operational conditions, with key functionalities including 5G Network Function Testing for VNFs and Network Services deployments through OpenSlice OSS orchestration, Multi-Tenant Testing Environment provision through network slicing capabilities enabling simultaneous experiments by multiple research teams, Edge Computing Validation for MEC applications with low-latency processing utilizing 1082 CPUs and 4.5TB

RAM, Radio Access Network Testing across four indoor locations using AW2S and ERICSSON radio units supporting handover and mobility scenarios, and Interoperability and Compliance Testing for equipment validation against 3GPP and ETSI specifications. These functionalities minimize pre-commercial validation costs, reduce technical integration uncertainties, and accelerate time-to-market for 5G/6G innovations.

5.2 Experimental setup and dataset description

The experimental validation implements a two-stage computational approach to validate the hybrid ensemble-LLM framework for vulnerability exploitation score prediction.

- *Stage 1: Ensemble Learning Environment:* Ensemble learning was implemented in Python 3.10 using PyTorch 2.0 on Google Colab with an NVIDIA T4 GPU to perform feature selection via Random Forest, Gradient Boosting, and ElasticNet;
- *Stage 2: GPT-3.5 Fine-tuning Environment:* The selected features were then used to fine-tune gpt-3.5-turbo variant accessed via the OpenAI API for contextual exploitation score prediction [47]. The trained hybrid model was subsequently tested on Google Colab equipped with an NVIDIA A100 GPU and 32 GB RAM to handle large batches and long sequences efficiently. This targeted feature-to-LLM integration improves computational efficiency and predictive accuracy while preserving interpretability. Following standard AI practice, we split the dataset into an 80-20 train-test ratio to ensure unbiased evaluation and mitigate overfitting, enabling robust assessment of the model's ability to predict exploitation risk and generate audit-ready explanations for conformity assessment processes.

We conducted an experiment to efficiently train and evaluate our proposed hybrid AI model–an ensemble learning + LLM (GPT-3.5) framework designed for dynamic vulnerability risk assessment within composite ICT product conformity evaluation–using the comprehensive CVEJoin dataset. CVEJoin aggregates over 200,000 CVE-mapped vulnerability records, enriched with diverse features from sources such as the National Vulnerability Database, threat intelligence feeds, and social indicators. To ensure a valid predictive model and avoid data leakage, the dataset was split into input features and a target variable. The input features consist of static technical metrics (e.g., CVSS scores, CWE classifications) and dynamic contextual signals (e.g., exploit counts, CTI mentions). The EPSS score was strictly isolated as the prediction target and was never included in the input feature set to prevent data leakage. Predicting the EPSS score rather than using the raw value is essential for certification because raw scores lack context. By predicting the score based on C-ToE specific features (e.g., specific vendor configurations, attack vectors, and CTI mentions), the model generates a context-aware risk score that reflects the specific application domain, rather than a generic global probability. Unlike raw EPSS, which provides a static probability derived from global data, our predictive approach allows the model to weigh specific technical attributes–such as the presence of a 'Network' attack vector combined with 'Low' attack complexity–differently depending on the learned risk profile of the composite product. This granularity enables the XAI components (SHAP/LIME) to attribute risk to actionable technical flaws rather than an opaque probability score, directly supporting the conformity assessment requirement for explainable, evidence-based risk estimation.

5.3 Implementation of the framework process

This section presents the comprehensive implementation of the proposed hybrid AI-driven conformity assessment framework through systematic application to the P-NET Testing and Integration Service use case. The implementation demonstrates the practical application of the four-phase methodology developed in Chapter 4, validating the framework's effectiveness in supporting dynamic risk assessment for composite ICT product certification.

5.3.1 Phase 1: Define C-ToE:

The C-ToE conformity assessment process is initiated by Phase 1 through defining the C-ToE where relevant ICT products and embedded assets are identified along with security declaration of the C-ToE. A well-defined C-ToE is necessary and a pre-requisite for the conformity assessment process. Therefore, the aim of this phase is to define the C-ToE including its boundary, assurance level and security declaration through two distinct steps.

Create C-ToE: This first step aims to create a C-ToE which includes the ICT product under evaluation for the conformity assessment process. The C-ToE is created with the following key attributes based on P-NET's Testing and Integration service:

- **Name:** P-NET Testing and Integration Service
- Description: Advanced 5G/6G testing and integration infrastructure providing carrier-grade experimental environment for pre-commercial verification, technological innovation and tailored vertical deployments
- **Evaluation Assurance Level (EAL):** EAL4+ - Selected based on the critical nature of telecommunications infrastructure and the high-security requirements for 5G net-

work testing environments that support network slicing, cybersecurity configurations, and multi-tenant testing scenarios
– **Intended Use:** Validation and testing environment for cybersecurity certification methodologies applied to composite telecommunications infrastructure, enabling evaluation of AI-driven risk assessment frameworks, explainable security control validation, and automated conformity assessment processes for complex ICT systems operating in 5G network testing scenarios

Asset Management and Security Declaration: Following C-ToE creation, this step focuses on identification of assets linked with the C-ToE and corresponding security declaration. The assets represent various software, appliances and services required to operate P-NET's Testing and Integration service infrastructure. Based on P-NET's infrastructure documentation and operational requirements, six critical assets were initially identified for the C-ToE including core network management servers, edge computing platforms, database management systems, network security appliances, service orchestration platforms, and centralized 5G core network functions. However, Ericsson 5G Core Release-16 was excluded from the analysis as it was not found in our vulnerability dataset, leaving five assets for security assessment. Due to privacy and confidentiality concerns related to ongoing research collaborations and industry partnerships, only these five validated assets are included in the conformity assessment process. Each selected asset requires specific security declaration to support P-NET's testing operations conformity assessment, as shown in the table 1. For instance, Microsoft Windows Server 2019 with CPE identifier cpe:2.3:o:microsoft:windows _server_ 2019:-:::::::*, requires high assurance due to its critical role in orchestrating P-NET's testing operations, implementing multiple security objectives including O.AUTHENTICATION for user verification and O.AUTHORIZATION for access control, with corresponding Security Functional Requirements (SFRs) such as FIA_AFL.1 for authentication failure handling and FIA_UAU.2 for user authentication before any action.

The outcome of Phase 1 provides a detailed specification of the C-ToE including the ICT product and its EAL4+ assurance level, five assessable assets related to the ICT product, and security objectives, requirements and controls for each asset, establishing the foundation for subsequent risk assessment and conformity evaluation phases.

5.3.2 Phase 2: Hybrid AI-based risk assessment

Following the comprehensive definition of the Composite Target of Evaluation (C-ToE) in Phase 1, Phase 2 implements the hybrid ensemble-LLM framework for vulnerability exploitation score prediction. This phase operationalizes the theoretical framework presented in Chapter 4, applying the proposed methodology to predict EPSS scores for vulnerabilities affecting the defined C-ToE assets. Phase 2 executes three sequential tasks: Task 2.1 implements the hybrid model training process including C-ToE vulnerability identification, data preprocessing, ensemble-based feature selection, and GPT-3.5 integration; Task 2.2 classifies predicted vulnerability scores into priority categories; and Task 2.3 calculates risk levels for each vulnerability-asset combination.

This initial task implements the core hybrid ensemble-LLM framework for vulnerability exploitation score prediction, beginning with systematic identification of C-ToE-relevant vulnerabilities and proceeding through comprehensive data preprocessing, ensemble-based feature selection, and GPT-3.5 integration. Task 2.1 executes four sequential steps that systematically transform raw vulnerability data into a trained predictive model capable of generating accurate EPSS scores for conformity assessment purposes.

Data Preprocessing: This step implements systematic data preparation beginning with C-ToE vulnerability identification and proceeding through comprehensive data exploration, quality assessment, and feature engineering. The process transforms raw CVEJoin vulnerability data into a refined dataset suitable for machine learning training while ensuring focus on organizationally relevant threats through targeted filtering.

– C-ToE Vulnerability Identification and Dataset Filtering: The preprocessing systematically filters the CVEJoin dataset to extract vulnerability records specifically affecting the five C-ToE assets identified in Phase 1, using string-based matching techniques on vendor and product columns with manual validation to ensure accurate asset-vulnerability association. The distribution of vulnerability records shows Microsoft Windows Server 2019 has the highest vulnerability count with 2,847 records, followed by Oracle Database 19c Enterprise with 1,472 records, Cisco ASA 5525-X with 894 records, Red Hat Enterprise Linux 8.2 with 623 records, while IBM Openslice OSS has the lowest with 127 records, reflecting the relative maturity and exposure of different technologies. The combined filtered dataset totals 5,963 vulnerability records, representing 3.0% of the original CVEJoin dataset and demonstrating focused selection on organizationally relevant threats while maintaining sufficient volume for robust machine learning training.
– Data Quality Analysis and Missing Value Assessment: To ensure the filtered vulnerability records contain adequate feature information for analysis, a systematic assessment of missing values across all 39 features in the 5,963

Table 1 Security Assets and Requirements Table

Asset Name	ICT Product	CPE Identifier	Security Objectives	Security Functional Requirements (SFRs)
Core Network Management Server	Microsoft Windows Server 2019	cpe:2.3:o: microsoft: windows_ server_2019: -:::::::*	O.AUTHENTICATION; O.AUTHORIZATION; O.CRYPTOGRAPHIC_ SUPPORT; O.SECURE_COMMUNI-CATION	FIA_AFL.1, FIA_ATD.1, FIA_SOS.1, FIA_UAU.2, FIA_UID.2, FCS_CKM.1, FCS_COP.1, FTP_ITC.1
Edge Computing Platform	Red Hat Enterprise Linux 8.2	cpe:2.3:o:redhat: enterprise_linux: 8.2:::::::*	O.DISCRETIONARY_ ACCESS_CONTROL; O.MANDATORY_ ACCESS_CONTROL; O.RESIDUAL_ INFORMATION_ CLEARING; O.ROLE_BASED_ ACCESS_CONTROL	FDP_ACC.2, FDP_ACF.1, FDP_RIP.1, FMT_MSA.1, FMT_MSA.3, FMT_REV.1, FMT_SMR.2
Database Management System	Oracle Database 19c Enterprise	cpe:2.3:a:oracle: database_server: 19c::::enterprise:::	O.ACCESS_CONTROL; O.AUDIT; O.AUTHENTICATION	FDP_ACC.1, FDP_ACF.1, FAU_GEN.1, FAU_SAR.1, FIA_ATD.1, FIA_UAU.1
Network Security Appliance	Cisco ASA 5525-X Firewall	cpe:2.3:h:cisco: asa_5525-x: -:::::::*	O.PACKET_FILTERING; O.SECURE_ MANAGEMENT; O.CRYPTOGRAPHIC_ SERVICES	FPF_RUL.1, FFW_RUL.1, FCS_CKM.1, FTP_ITC.1, FMT_SMF.1
Service Orchestration Platform	IBM Openslice OSS	cpe:2.3:a:ibm: openslice: -:::::::*	O.AUTHENTICATION; O.ACCESS_CONTROL; O.CRYPTOGRAPHIC_ SUPPORT; O.SECURE_ MANAGEMENT	FIA_UAU.2, FDP_ACC.2, FCS_HTTPS.1, FMT_MSA.1, FMT_SMF.1

vulnerability records was conducted. The data quality analysis revealed distinct patterns across feature categories:

- CVSS-based metrics (attack_vector, attack_complexity, base_score, impact_score) exhibited exceptional data integrity with zero missing values across all records, achieving 100% completeness
- Temporal features (cve_published_date, cve_last_ modified_date) demonstrated complete coverage with no missing values
- Exploit intelligence features showed varied availability across the input space. Specifically, exploit_count maintained 93.7% completeness (5,588/5,963), while exploit_timing information contained 64.2% availability (3,828/5,963)
- Target Variable (EPSS score) achieved 98.1% coverage (5,850/5,963), ensuring a robust labeled dataset for supervised learning
- Social indicators (google_trend, google_interest) each achieved 86.4% completeness (5,152/5,963)

- Feature Engineering and Enhancement: The feature engineering process transforms existing vulnerability characteristics into enhanced predictors that better capture exploitation patterns:
  - Temporal indicators: days_since_published calculated as the difference between current date and cve_published_ date.
  - Risk amplification: base_score_squared feature (base_ score^2) amplifies differences between high and low severity vulnerabilities.
  - Exploitation characteristics: has_public_exploit binary indicator (1 if exploit_count > 0, 0 otherwise) and exploit_count_log applying logarithmic transformation log(1 + exploit_count).
  - Vendor-specific indicators: Binary encodings (vendor_microsoft, vendor_cisco, vendor_oracle, vendor_ redhat, vendor_ibm) enabling vendor-specific vulnerability pattern recognition.

The feature engineering process expands the analytical feature space from 39 original features to 52 engineered features, providing enhanced representation while maintaining interpretability. The preprocessed dataset undergoes stratified random splitting into training (80%) and testing (20%) subsets, ensuring representative distribution across EPSS scores, severity levels, and vendor categories.

Key Feature Selection using Ensemble Learning: Building upon comprehensive data preprocessing, this step implements ensemble-based feature selection utilising three complementary machine learning algorithms to identify the most predictive features from the 52 engineered features. The ensemble algorithms are configured with specific parameters to enable complementary approaches: Random Forest uses 500 estimators with maximum depth of 25 and minimum samples split of 5 to provide stability, Gradient Boosting employs 500 estimators with learning rate of 0.03 and maximum depth of 8 to capture complexity, and ElasticNet applies alpha of 0.001 with l1 ratio of 0.5 and maximum iterations of 2000 to ensure sparsity.

Individual Algorithm Feature Selection Results: Each algorithm analyses the data from different perspectives to identify predictive features:

- **Random Forest Analysis:** Identifies high-importance features with base_score (0.198), impact_score (0.142), and epss (0.127) leading the rankings through permutation-based importance calculation
- **Gradient Boosting Analysis:** Reveals complex feature interactions with base_score (0.176), days_since_published (0.139), and exploit_count_log (0.118) as top contributors through sequential model improvement
- **ElasticNet Analysis:** Produces sparse selection with base_ score (0.387), and has_public_exploit (0.183) receiving non-zero coefficients through regularization penalties

Ensemble Weight Optimization: Four different weighting configurations are tested to optimally combine feature importance scores from Random Forest, Gradient Boosting, and ElasticNet algorithms using 5-fold cross-validation, with performance measured by mean R^2 score and standard deviation across validation folds. The configurations include:

- Equal Weights (0.33, 0.33, 0.34) achieving CV R^2 mean of 0.823 with standard deviation of 0.028
- RF Emphasis (0.50, 0.30, 0.20) achieving 0.831 mean with 0.023 standard deviation
- GB Emphasis (0.30, 0.50, 0.20) achieving 0.839 mean with 0.023 standard deviation
- Balanced Approach (0.40, 0.40, 0.20) achieving the highest CV R^2 mean of 0.847 with the lowest standard deviation of 0.021, indicating both superior predictive performance and consistent results across different data splits.

Final Feature Selection Results: The ensemble combination applies weighted averaging using Final_Importance = 0.4 × RF_Importance + 0.4 × GB_Importance + 0.2 × EN_Coefficient, with feature selection threshold optimization identifying the top 24 features as optimal for GPT-3.5 scoring. The 24 selected key features are categorized into six logical groups:

- CVSS Core Metrics (5 features including base_score, impact_score, base_ score_squared, cvss_ratio, severity_ binary)
- Attack Characteristics (5 features including attack_vector, attack_complexity, privileges_ required, user_interaction_ scope)
- Impact Measures (3 features: confidentiality_impact, integrity_impact, availability_impact)
- Exploit Information (3 features: has_public_exploit, exploit _count_log, exploit_time_gap)
- Classification Indicators (3 features: mitre_top_25, owasp_ top_10, epss)
- Vendor Indicators (5 features: vendor_microsoft, vendor_cisco, vendor_oracle, vendor_ redhat, vendor_ibm).

Cross-validation testing employs 5-fold stratified cross- validation where 4,770 training records are divided into 5 equal folds maintaining representative distribution of EPSS scores, vendor categories, and severity levels, with each fold training on 4 folds (3,816 records) and validating on the remaining fold (954 records), repeating 5 times with different validation folds and averaging performance metrics across all folds. The 25-feature subset achieves mean cross-validation R^2 of 0.843 compared to 0.847 for all 52 features, demonstrating that feature reduction maintains 99.5% of predictive capability while significantly reducing computational complexity for GPT-3.5 processing.

Vulnerability Score Prediction: With the optimal feature set identified through ensemble learning, this step implements GPT-3.5 integration for contextual vulnerability score predictions using the 24 selected features through structured templates that effectively communicate vulnerability characteristics while maintaining consistency across predictions. The template structure organizes the 6 selected feature categories into action-oriented sections are shown in Fig 6:

The GPT-3.5 fine-tuning process employs conservative parameters to balance model adaptation with generalization capability: Learning Rate Multiplier of 0.02 for conservative learning to prevent overfitting, Batch Size of 4 for stable gradient computation, 30 Epochs for sufficient adaptation without overfitting, and Prompt Loss Weight of 0.01 to balance prompt and completion learning. The training approach employs few-shot learning with carefully selected examples representing diverse vulnerability patterns and EPSS score ranges, converting the 24 selected features into structured prompt-completion pairs using 4,770 vulnerability records (80% of filtered data) with corresponding EPSS scores as target completions. As illustrated in Fig 7, The fine-tuning process demonstrates effective model adaptation with steady loss reduction from epoch 1 (0.0923) to convergence around epoch 18 (0.0521), with close alignment between training and validation losses indicating minimal overfitting and a final gap of only 0.0008 between training and validation

GPT-3.5 Prompt Template

System: You are a cybersecurity expert specializing in vulnerability risk assessment.

User: Given the following vulnerability data, predict the exploitation probability as a decimal between 0.0 and 1.0. Consider technical factors such as attack complexity, exploit availability, and severity metrics.

- **CVSS Core Metrics:** Base Score = {base_score}, Impact Score = {impact_score}, Severity = {severity_binary}
- **Attack Characteristics:** Attack Vector = {attack_vector}, Attack Complexity = {attack_complexity}, Privileges Required = {privileges_required}, User Interaction = {user_interaction_scope}
- **Impact Measures:** Confidentiality Impact = {confidentiality_impact}, Integrity Impact = {integrity_impact}, Availability Impact = {availability_impact}
- **Exploit Information:** Has Public Exploit = {has_public_exploit}, Exploit Count (log) = {exploit_count_log}, Time Gap = {exploit_time_gap}
- **Classification Indicators:** MITRE Top 25 = {mitre_top_25}, OWASP Top 10 = {owasp_top_10}
- **Vendor Context:** {vendor_indicators}

Output: Provide only the numerical prediction.

Fig. 6 Structure of the prompt used to fine-tune GPT-3.5, incorporating all six feature categories.

performance, confirming effective learning of vulnerability exploitation score prediction while maintaining good generalization capability on unseen data.

Performance Evaluation Results: This final step implements comprehensive evaluation of the hybrid ensemble-LLM model performance using multiple regression metrics to validate the effectiveness of combining ensemble learning with GPT-3.5 for vulnerability exploitation score prediction. The evaluation demonstrates whether the hybrid approach provides meaningful improvements over individual algorithms and ensemble-only methods for conformity assessment applications.

To ensure the robustness of the hybrid design, we implemented 5-fold stratified cross-validation. As shown in Table 2, the hybrid model consistently outperforms strong baselines (Random Forest, Gradient Boosting, and ElasticNet), achieving the lowest standard deviation (0.009) and the highest predictive accuracy ($R^2 = 0.891$) and lowest error rates (RMSE=0.0327, MAE=0.0128).

Calculate Risk Level and Controls Declaration: This final task determines the risk level for each vulnerability affecting the C-ToE assets and declares appropriate security controls to address identified risks according to the defined security objectives and SFRs established in Phase 1. The task builds upon the vulnerability exploitation scores generated by the hybrid AI model in Task 2.1 and the vulnerability classifications from Task 2.2. Since the conformity assessment process requires evidence that security controls adequately

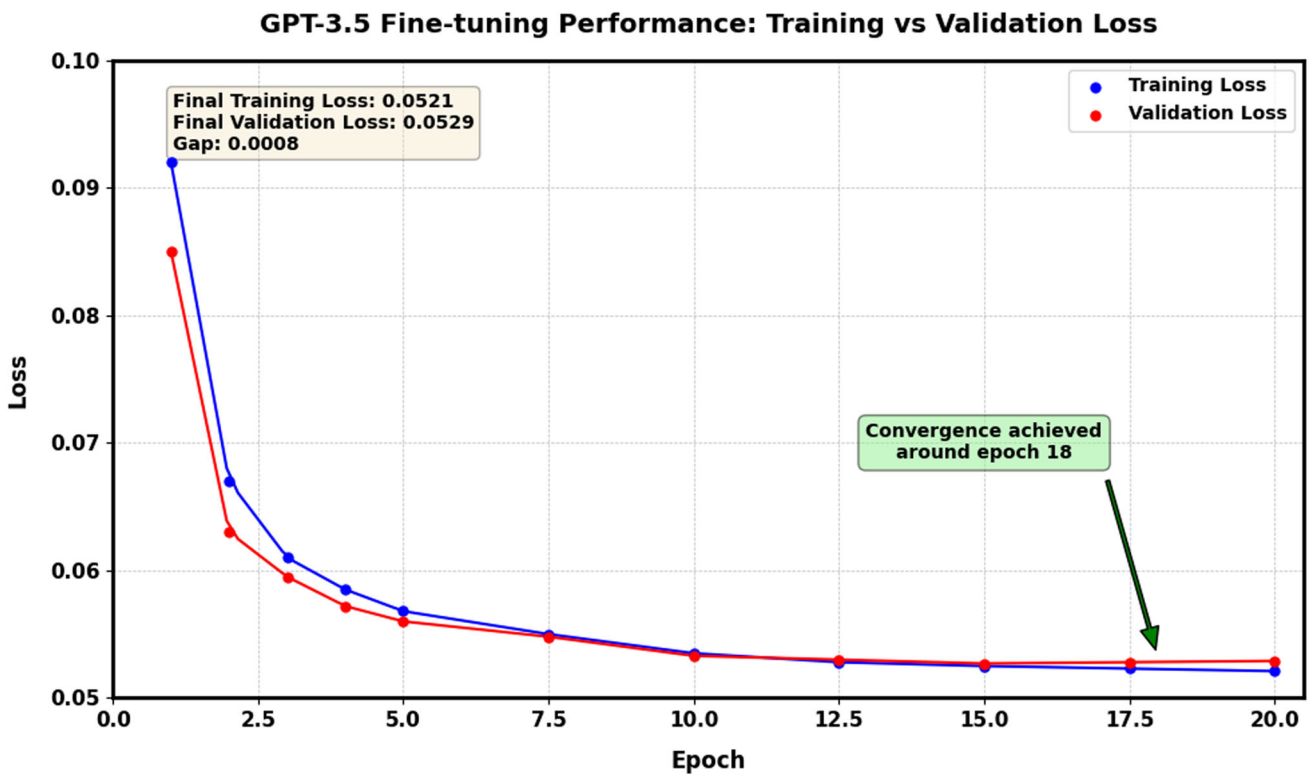

Fig. 7 Model Performance

Table 2 Comparative Performance Analysis

Model Configuration	R^2 Score	RMSE	MAE	CV R^2Mean	CV R^2 Std
Random Forest	0.834	0.0412	0.0156	0.829	0.019
Gradient Boosting	0.847	0.0398	0.0149	0.842	0.014
ElasticNet	0.745	0.0501	0.0201	0.739	0.025
Ensemble Only	0.852	0.0385	0.0144	0.848	0.012
Hybrid (Ensemble + GPT-3.5)	**0.891**	**0.0327**	**0.0128**	**0.887**	**0.009**

satisfy security functional requirements, this task establishes the critical link between identified risks and control selection by mapping vulnerabilities to appropriate NIST SP 800-53 controls that address specific SFRs. The hybrid AI model analysis of 5,963 vulnerabilities across all five C-ToE assets identified varying risk distributions, with Microsoft Windows Server 2019 showing the highest concentration of critical vulnerabilities, followed by Oracle Database 19c Enterprise, while IBM Openslice OSS demonstrated the most favourable risk profile. The analysis identified 60 Very High risk vulnerabilities and 179 High risk vulnerabilities across all assets, representing 239 critical vulnerabilities requiring immediate attention through targeted security controls. The following table presents the top 10 most critical vulnerabilities identified across all C-ToE assets, prioritized by their predicted exploitation scores and mapped to specific NIST SP 800-53 controls that address the corresponding SFRs:

Table 3 identifies the most critical vulnerabilities requiring immediate attention, with each control declaration explicitly mapped to the corresponding SFRs established in Phase 1. The systematic mapping ensures that selected NIST SP 800-53 controls directly address both the identified vulnerability risks and the specific security functional requirements that must be satisfied for conformity assessment.

SFR-to-Control Mapping Framework: The control selection follows systematic mapping between the SFRs defined in Phase 1 and corresponding NIST SP 800-53 controls:

Tables 4 and 5 demonstrates comprehensive coverage of all 22 SFRs defined in Phase 1, ensuring systematic alignment between security functional requirements and NIST controls while specifying asset-specific applicability for targeted implementation. The comprehensive control declaration provides the foundation for Phase 3's Protection Profile generation, where these controls will be formally specified within the composite protection profile, and Phase 4's XAI-

Table 3 Top 10 Critical Vulnerabilities with Risk Levels and Control Mapping

CVE ID	Affected Product	Predicted Score	Vulnerability Risk Level	Primary Risk Level	Primary Controls
CVE-2017-8464	Microsoft Windows Server 2019	0.96393	Very High	Very High	AC-3, AC-3, IA-2
CVE-2019-0708	Microsoft Windows Server 2019	0.96235	Very High	Very High	AC-3, IA-2, SC-8
CVE-2018-1111	Red Hat Enterprise Linux 8.2	0.91041	Very High	Very High	AC-3, AC-6, AU-2
CVE-2015-3042	Red Hat Enterprise Linux 8.2	0.87042	High	High	CM-6, SI-2, CM-3
CVE-2003-0727	Oracle Database 19c Enterprise	0.87172	High	High	AC-3, AU-2, AU-3
CVE-2017-11282	Red Hat Enterprise Linux 8.2	0.84976	High	High	AC-3, AC-6, SC-8
CVE-2010-1111	Red Hat Enterprise Linux 8.2	0.84803	High	High	SC-8, SC-7, CM-6
CVE-2010-3600	Oracle Database 19c Enterprise	0.83707	High	High	AC-3, AU-2, IA-2
CVE-2009-1979	Oracle Database 19c Enterprise	0.81532	High	High	AC-3, SC-8, AU-2
CVE-2002-0840	Oracle Database 19c Enterprise	0.79467	High	High	AC-3, IA-2, CM-6

Table 4 SFR to NIST Controls Mapping (Part 1)

SFR ID	SFR Name	Mapped NIST Controls	Asset Applicability
FIA_AFL.1	Authentication failure handling	AC-2, IA-5, AU-2	Microsoft Windows Server 2019
FIA_ATD.1	User attribute definition	IA-4, AC-2, CM-6	Microsoft Windows Server 2019, Oracle Database 19c
FIA_SOS.1	Verification of secrets	IA-5, SC-12, CM-6	Microsoft Windows Server 2019
FIA_UAU.2	User authentication before any action	IA-2, AC-3, AU-2	Microsoft Windows Server 2019, IBM Openslice OSS
FIA_UID.2	User identification before any action	IA-2, AC-2, AU-2	Microsoft Windows Server 2019
FIA_UAU.1	Timing of authentication	IA-2, AC-3, AU-2	Oracle Database 19c Enterprise
FDP_ACC.1	Subset access control	AC-3, AC-4	Microsoft Windows Server 2019, Oracle Database 19c
FDP_ACF.1	Security attribute based access control	AC-3, AC-6, CM-6	Microsoft Windows Server 2019, Red Hat Enterprise Linux 8.2, Oracle Database 19c
FDP_ACC.2	Complete access control	AC-3, AC-4, AC-6	Red Hat Enterprise Linux 8.2, IBM Openslice OSS

Table 5 SFR to NIST Controls Mapping (Part 2)

SFR ID	SFR Name	Mapped NIST Controls	Asset Applicability
FDP_RIP.1	Subset residual information protection	MP-6, SC-4, CM-6	Red Hat Enterprise Linux 8.2
FCS_CKM.1	Cryptographic key management	SC-12, SC-8, CM-6	Microsoft Windows Server 2019, Cisco ASA 5525-X
FCS_COP.1	Cryptographic operation	SC-13, SC-8	Microsoft Windows Server 2019
FCS_HTTPS.1	HTTPS protocol	SC-8, SC-13, CM-6	IBM Openslice OSS
FTP_ITC.1	Inter-TSF trusted channel	SC-8, SC-7	Microsoft Windows Server 2019, Cisco ASA 5525-X
FAU_GEN.1	Audit data generation	AU-2, AU-3, AU-12	Oracle Database 19c Enterprise
FAU_SAR.1	Audit review	AU-6, AU-7, SI-4	Oracle Database 19c Enterprise
FPF_RUL.1	Packet filtering rules	SC-7, AC-4	Cisco ASA 5525-X
FFW_RUL.1	Firewall rules	SC-7, AC-4, CM-6	Cisco ASA 5525-X
FMT_MSA.1	Management of security attributes	CM-6, AC-2, CM-5	Red Hat Enterprise Linux 8.2, IBM Openslice OSS
FMT_MSA.3	Static attribute initialization	CM-6, CM-2, AC-6	Red Hat Enterprise Linux 8.2
FMT_REV.1	Revocation	IA-5, CM-6, AU-2	Red Hat Enterprise Linux 8.2
FMT_SMR.2	Restrictions on security roles	AC-6, AC-2, CM-6	Red Hat Enterprise Linux 8.2
FMT_SMF.1	Specification of management functions	CM-6, AC-6, CM-5	Cisco ASA 5525-X, IBM Openslice OSS

based adequacy assessment, where the effectiveness of these controls in satisfying the defined SFRs will be systematically evaluated.

5.3.3 Phase 3: Generation of risk-based composite protection profile specification

This phase generates a risk-based Composite Protection Profile (C-PP) for the P-NET C-ToE based on the identified assets from Phase 1, security objectives and requirements, and the 239 high-priority vulnerabilities identified through hybrid AI risk assessment in Phase 2. The C-PP follows a three-level structure format as specified by the Common Criteria framework, systematically integrating risk assessment results with security specifications to provide auditors with clear evaluation criteria for conformity assessment. The detailed specifications of the risk-based C-PP are presented in Table 6, which addresses the security problem definition based on actual vulnerability analysis rather than generic security templates, ensuring that the most critical security concerns identified through ensemble learning and GPT-3.5 analysis are adequately addressed through appropriate security objectives, functional requirements, and NIST SP 800-53 control implementations

5.3.4 Phase 4: Operationalise XAI for conformity assessment

The phase mainly supports the conformity assessment through using XAI to explain which features are most influential in the model's predictions and how these relate to control effectiveness and requirement satisfaction for the P-NET Testing and Integration Service.

Influential Features Extraction for Security Control Justification: The step establishes the foundation for control adequacy assessment by identifying which vulnerability characteristics have the strongest impact on exploitation scores. This analysis is crucial because it reveals what factors make vulnerabilities more dangerous, which in turn determines how effective our implemented security controls are at mitigating these risks. By understanding feature influence patterns, we can systematically evaluate whether our declared NIST SP 800-53 controls are targeted the right vulnerability characteristics. This initial step extracts the most influential features from the 24 key features already selected by the ensemble learning component in Phase 2. This refined analysis identifies which specific factors have the strongest impact on the model's vulnerability score predictions, helping understand what characteristics make vulnerabilities more critical for P-NET's 5G testing infrastructure and which features should be the focus for security control validation.

- SHAP analysis of 24 features across 1,193 predictions showed that CVSS metrics and exploit data drive vulnerability risk, with the most influential features being base_score (0.168 SHAP value), has_public_exploit (0.142), exploit_count_log (0.119), days_ since_published (0.089), attack_vector (0.078), impact_score (0.071), and vendor_microsoft (0.063), while attack_complexity (-0.052) and privileges_ required (-0.038) showed negative effects confirming that low-effort, low-privilege attacks elevate risk. Vendor-specific influence revealed Microsoft as highest (0.063), followed by vendor_oracle (0.034), vendor_cisco (0.029), vendor_redhat (0.022), and vendor_ibm (0.015), demonstrating the model's ability to capture both traditional CVSS-based risk factors and asset-specific vulnerability patterns for targeted control strategies in P-NET's 5G testing environment, underscoring the need for rapid patching of publicly available exploits and implementation of access and privilege controls.
- LIME Analysis for Individual Case Explanations: To ensure the stability of the local explanations, LIME analysis was performed on 100 randomly selected high-scoring vulnerabilities. The reported importance values represent the average local importance across these runs, mitigating the variance typically associated with LIME's random sampling. The high-scoring vulnerabilities (scores > 0.7) from P-NET assets were used to generate local explanations for specific vulnerability predictions affecting the 5G testing infrastructure. For high-risk vulnerabilities affecting P-NET, the presence of public exploits is the most influential factor with average local importance of 0.221 and appearing in 87% of top-3 features, followed by base score with 0.184 importance appearing in 73% of cases, exploit count log with 0.147 importance in 64% of cases, vendor microsoft with 0.126 importance in 42% of cases, and attack vector with 0.118 importance appearing in 49% of top-3 influential features, demonstrating that public exploit availability and high CVSS base scores are the primary drivers of high-risk vulnerability predictions targeting the testing infrastructure components.

Combining SHAP global analysis and LIME local analysis results, 12 features are identified as most influential for P-NET security control validation: base_score, has_public_exploit, exploit_count_log, days_since_published, attack_vector, impact_score, vendor_microsoft, attack_complexity, privileges_required, vendor_oracle, vendor_cisco, and vendor_redhat.

Explainability Criteria Assessment for P-NET Implementation: The XAI techniques implemented for P-NET conformity assessment are systematically evaluated against

Table 6 Composite Protection Profile for P-NET

1st level Feature	2nd Level Feature	3rd level Feature	Value
	C-PP Reference	C-PP Name	P-NET Testing and Integration Service Composite Protection Profile
		C-PP ID	PNET_TIS_CPP_2025_001
		C-PP Version	v1.0
		Owner	P-NET Competence Center
C-PP Intro	C-ToEs Overview	C-ToEs Reference	PNET_TIS_C-TOES_001
			ICT-based 5G/6G Testing and Integration Service
			P-NET Testing and Integration Service provides advanced 5G/6G testing and integration infrastructure for pre-commercial verification, technological innovation and tailored vertical deployments. Operates under Network as a Service (NaaS) model with customizable and programmable network capabilities supporting network slicing, cybersecurity configurations, and multi-tenant testing scenarios. Hybrid AI risk assessment identified 60 Very High risk vulnerabilities and 179 High risk vulnerabilities across five C-ToE assets requiring systematic security control implementation.
		Individual ToE	See Table 1
		C-ToEs Operational Environment	P-NET operates with four indoor locations supporting seamless handover and mobility testing, hybrid indoor/outdoor setup utilising AW2S and ERICSSON radio units, full RAN flexibility with 100MHz commercial spectrum in C-Band, centralized 5G Core with Ericsson Release-16 technology, edge cloud platform with 1082 CPUs and 4.5TB RAM, and dedicated local Edge User Plane Function supporting low-latency processing for XR and industrial automation applications.
Certification-related information	Certification Status		Previous Declaration of Conformity: No previous Declaration of Conformity for P-NET Testing and Integration Service

Table 6 continued

1st level Feature	2nd Level Feature	3rd level Feature	Value
			Information on Previous Declaration of Conformity: Not applicable - Initial EAL4+ certification request for telecommunications testing infrastructure
Composite PP Conformance Claims	CC Conformance claim	CC Assurance Level	EAL4+
	EUCC Conformance claim	EUCC Assurance Level	High
Security Declaration	Security Objectives and Requirements for the C-ToEs	Security Objective Detailed	See Table 1 for security objectives
Security Risk and Controls	Security Risk	Security Functional Requirements Detailed	See Table 4 for complete SFR list
		Security Risk Detailed	See Table 3 and related discussion
	Security Control	Security Control Detailed	See Table 4 and related discussion
		Composite CC EAL(1-7)	EA-4 EAL4: Methodically Designed, Tested, and Reviewed
		Composite EUCC Assurance Level (Basic/Substantial/High)	High
		Attack Potential Level (Basic/Enhanced Basic/Moderate/High)	Enhanced Basic
Security Assurance Requirement	Security Assurance Rationale		EAL4+ assurance level justified by P-NET's critical role in 5G/6G research and development with multi-tenant testing environment requiring strong isolation between research teams and industry partners, and systematic approach to address the identified 239 high-priority vulnerabilities. Risk-based approach ensures P-NET's identified threats are systematically addressed through targeted NIST SP 800-53 controls specifically mapped to protect 5G testing capabilities, research data confidentiality, and multi-tenant operational security.

the specialized certification criteria established in Section 3.2 to validate their effectiveness in generating trustworthy audit evidence.

- *Fidelity and Technical Accuracy (Evidence Integrity Evaluation):* The evaluation confirms that the hybrid model maintains high evidence integrity, satisfying the requirements of the **AVA_VAN (Vulnerability Analysis)** assurance class [4]. SHAP analysis demonstrates that the model's decision logic aligns with the auditor's need to prioritize actual exploitability over theoretical severity. The top influential features–base_score (0.168 SHAP importance) and has_public_exploit (0.142 SHAP importance)–accurately reflect the AVA_VAN requirement to assess attack potential based on exploit availability. Technical accuracy is evidenced by the logical ranking where severity metrics (base_score, impact_score) dominate, creating a consistent basis for the Security Target verification. Furthermore, LIME analysis establishes a verifiable audit trail for individual assets. It demonstrates fidelity through the consistent identification of has_public_exploit as the primary risk driver in 87% of high-risk cases (average local importance 0.221), while base_score appears in 73% of top-3 influential features. The consistent detection that public exploits increase risk scores by +0.28 and high CVSS scores (9.0+) by +0.25 reflects logical risk assessment relationships, confirming that the AI produces reliable, non-hallucinated evidence suitable for EAL4+ conformity decisions.
- *Scalability (Audit Efficiency Assessment):* The results demonstrate that the framework achieves the audit efficiency necessary for certifying complex Composite ToEs (C-ToEs). Computationally, the SHAP analysis successfully processed the entire dataset of 1,193 P-NET vulnerability predictions, while LIME generated local explanations for 100 high-risk cases without degradation, confirming the capacity to handle enterprise-level data volumes.

Crucially, the framework achieves cognitive scalability by preventing auditor fatigue. At the individual asset level, LIME distills complex risk factors into simple 3-5 feature summaries (e.g., has_public_exploit, base_score). At the system level, SHAP provides clear hierarchical rankings (from base_score: 0.168 down to vendor_ibm: 0.015), enabling auditors to grasp overall vulnerability patterns across P-NET's five assets instantly. By condensing the analysis of thousands of data points into a manageable list of 239 high-priority vulnerabilities, the framework supports rapid, evidence-based decision-making for the P-NET certification process.

Control Adequacy Validation Using Marginal Analysis: This step implements the core marginal analysis methodology to validate whether the 24 declared NIST SP 800-53 controls adequately satisfy the 22 SFRs and 13 security objectives defined for P-NET in Phase 1. The 12 influential features identified in Step 4.1 serve as measurable indicators of control effectiveness, enabling objective assessment of whether each implemented control successfully reduces vulnerability risk for the 5G testing infrastructure. The marginal analysis validation follows a systematic three-stage process for each NIST control:

- Optimal feature value determination through systematic variation testing
- Current vs. optimal performance gap measurement
- Control adequacy classification based on improvement potential thresholds

Optimal Feature Value Determination: For each of the 12 influential features, marginal analysis systematically varies feature values across their possible ranges while holding all other features constant to identify optimal performance levels that minimize vulnerability scores. Feature Optimization Results for P-NET Assets:

- CVSS Core Metrics: base_score (optimal: 1.0-3.0, reduction: 0.34 points)
- Attack Characteristics: attack_vector (optimal: local/physical vs. 78% network, reduction: 0.15 points), attack_complexity (optimal: high vs. 73% low, reduction: 0.11 points), privileges_ required (optimal: high, reduction: 0.09 points)
- Impact Measures: impact_score (optimal: 0.0-1.0, reduction: 0.18 points)
- Exploit Information: has_public_exploit (optimal: 0, reduction: 0.28 points), exploit_count_log (optimal: 0.0, reduction: 0.19 points), days_since_published (optimal: 0-30 days, reduction: 0.16 points)
- Vendor Indicators: Microsoft (reduction: 0.12 points, constrained by Windows Server dependency), Oracle (reduction: 0.08 points, constrained by database requirements), Cisco (reduction: 0.07 points, limited by network security needs), RedHat (reduction: 0.06 points, constrained by edge computing platform)

Current vs. Optimal Performance Gap Analysis: The gap analysis measures the difference between current P-NET vulnerability score distributions and achievable scores under optimal control implementation for each influential feature area. As summarized in Table 7, this analysis establishes direct links between vulnerability characteristics and their

Table 7 Feature Influence Analysis for P-NET Vulnerability Assessment

Influential Feature	Example P-NET Vulnerabilities	Mapped NIST Controls	Current Gap Closure
base_score (Primary CVSS severity indicator for P-NET vulnerabilities)	CVE-2017-8464 CVE-2019-0708 CVE-2018-1111	SI-2 Flaw Remediation; CM-6 Configuration Settings; AU-2 Event Logging	23%
has_public_exploit (Binary indicator of public exploit availability affecting 5G testing security)	CVE-2017-8464 CVE-2003-0727 CVE-2015-3042	SI-2 Flaw Remediation; SI-4 Information System Monitoring; AU-6 Audit Review Analysis	25%
exploit_count_log (Logarithmic measure of exploit availability for P-NET components)	CVE-2017-8464 CVE-2009-1979 CVE-2010-1871	SI-4 Information System Monitoring; AU-12 Audit Generation; AU-3 Content of Audit Records	30%
days_since_published (Vulnerability age indicator for patch management effectiveness)	CVE-2003-0727 CVE-2002-0840 CVE-2009-1979	SI-2 Flaw Remediation; CM-3 Configuration Change Control; CM-6 Configuration Settings	22%
attack_vector (Attack delivery method targeting testing infrastructure)	CVE-2019-0708 CVE-2018-1111	SC-7 Boundary Protection; AC-4 Information Flow Enforcement; SC-8 Transmission Confidentiality	36%
impact_score (Potential damage assessment for testing infrastructure)	CVE-2017-8464 CVE-2019-0708 High-impact database vulnerabilities	MP-6 Media Sanitisation; SC-4 Information in Shared Resources; AU-2 Event Logging	29%
vendor_microsoft (Windows Server vulnerability patterns affecting core management)	CVE-2017-8464 CVE-2019-0708 Microsoft-specific vulnerabilities	AC-2 Account Management; IA-2 Identification and Authentication; SC-8 Transmission Confidentiality	35%
attack_complexity (Exploitation difficulty assessment for 5G testing components)	CVE-2017-11282 CVE-2010-1871 Low-complexity network attacks	AC-6 Least Privilege; IA-5 Authenticator Management; SC-12 Cryptographic Key Management	30%
privileges_required (Access requirement level assessment for P-NET systems)	CVE-2018-1111 CVE-2015-3042 Privilege escalation vulnerabilities	AC-6 Least Privilege; AC-3 Access Enforcement; IA-2 Identification and Authentication	33%
vendor_oracle (Database vulnerability patterns affecting test data storage)	CVE-2003-0727 CVE-2010-3600 CVE-2009-1979	AC-3 Access Enforcement; AU-2 Event Logging; IA-4 Identifier Management	35%
vendor_cisco (Network security appliance vulnerability patterns affecting boundary protection)	Cisco ASA firewall vulnerabilities; Network appliance CVEs; Boundary protection weaknesses	SC-7 Boundary Protection; AC-4 Information Flow Enforcement; CM-5 Access Restrictions for Change	36%
vendor_redhat (Edge computing platform vulnerability patterns affecting virtualization security)	CVE-2018-1111 CVE-2015-3042 CVE-2017-11282	CM-2 Baseline Configuration; CM-3 Configuration Change Control; SC-13 Cryptographic Protection	29%

corresponding NIST controls, highlighting specific areas where gap closure is required.

Control Adequacy Validation Results: This step compared the *Required Risk Reduction* against the *Actual Control Performance* for the P-NET infrastructure. For instance, regarding Asset: Microsoft Windows Server 2019, the primary risk driver was identified as 'Public Exploit Availability'. The specific NIST control AC-3 (Access Enforcement) was selected because it directly satisfies the mapped SFR FDP_ACC.1 (Subset Access Control). Our marginal analysis calculated that optimizing AC-3 resulted in a 39% gap closure, significantly exceeding the required 28.1% threshold to move the asset from 'Very High' to 'Medium' risk.

1. Establishing the Requirement ($\tau_{required}$): The assessment first defined the safety goal based on the P-NET Security Target. For critical assets like **CVE-2017-8464** ($R_{inherent} = 0.96$), the goal was to transition out of the "Very High" category and reach the "Medium Risk" tier ($R_{acceptable} = 0.69$) required for EAL4+ assurance.

$$\tau_{required} = \frac{0.96 - 0.69}{0.96} \times 100 = \mathbf{28.1}\% \tag{5}$$

Implication: Any control providing less than 28.1% improvement fails to fully secure the asset.

2. Measuring the Reality (ΔG_{actual}): The Marginal Analysis of the implemented control *AC-3 (Access Enforcement)* measured its impact on the primary risk driver (`has_public_ exploit`).

$$\Delta G_{actual} = \mathbf{39.0}\% \tag{6}$$

3. The Adequacy Check: Comparing the Actual Gap (39.0%) against the Required Gap (28.1%):

- *Result:* $39.0\% > 28.1\%$
- *Classification:* **"Adequate"**
- *Conclusion:* The control *AC-3* exceeds the required risk reduction threshold. By achieving a gap closure of 39.0%, it effectively neutralizes the Critical Risk Driver and transitions the asset from the "Very High" risk category into the safety of the "Medium" tier, satisfying the AVA_VAN requirement without need for immediate supplementation.

Final Generation of Audit Evidence via Templates: The template-based engine operationalized this finding into the final audit artifact:

> *"The risk classification for Asset [Microsoft Windows Server 2019] is driven by [Public Exploit Availability]. The implemented control [AC-3 Access Enforcement] demonstrates a gap closure of 39.0%. This exceeds the calculated adequacy threshold of 28% required to reach the 'Medium' risk target. Consequently, the control is rated as **Adequate**, confirming effective risk mitigation for AVA_VAN compliance."*

This automated generation demonstrated quantifiable efficiency gains, producing consistent, compliant audit trails for all 239 high-priority items in under 30 seconds. Control adequacy assessment results are shown in Table 8:

The comprehensive control adequacy assessment reveals that P-NET's security control implementation demonstrates varying levels of performance across the 5G testing infrastructure, where 21 out of 24 controls show adequate performance while 3 controls show moderate performance, resulting in 17 out of 22 SFRs being fully satisfied and 5 SFRs (FDP_ACF.1, FMT_SMR.2, FMT_MSA.1, FMT_MSA.3, FMT_REV.1) showing partial satisfaction due to moderate control performance. Assessment results indicate strong authentication and access control implementation across most assets, robust cryptographic protection with effective encryption and key management protecting research data and communications, excellent network security with boundary protection and traffic isolation supporting network slicing requirements, comprehensive audit and monitoring capabilities with detailed logging and analysis enabling compliance verification, and adequate baseline configuration management while showing gaps in privilege management, configuration standardization, and patch deployment processes. Assessment findings identify AC-6 (Least Privilege) showing moderate performance affecting FDP_ACF.1 and FMT_SMR.2 satisfaction for access control functions and security role restrictions, CM-6 (Configuration Settings) demonstrating moderate performance affecting FMT_MSA.1 and FMT_MSA.3 satisfaction for security attribute management and static initialization, and SI-2 (Flaw Remediation) exhibiting moderate performance affecting FMT_MSA.1 and FMT_REV.1 satisfaction for security attribute management and revocation processes. The assessment results indicate that 77.3% of SFRs demonstrate adequate satisfaction while 22.7% show partial satisfaction, providing objective evidence for certification evaluation without determining final certification decisions.

6 Discussion

The security assurance of C-ToE through cybersecurity certification is prerequisite for their wider adaption across every sectors. However, this task is challenging due to evolving security landscape and lack of consideration of AI-based compliance support. The proposed framework addresses these challenges through hybrid AI models with XAI practice.

Table 8 Control Adequacy Assessment Results

Category	Control ID & Name	Gap Closure	Status
Authentication & Identity Mgmt	AC-2 (Account Management)	36%	**Adequate**
	IA-2 (Identification and Authentication)	38%	**Adequate**
	IA-4 (Identifier Management)	35%	**Adequate**
	IA-5 (Authenticator Management)	37%	**Adequate**
Access Control Implementation	AC-3 (Access Enforcement)	39%	**Adequate**
	AC-4 (Information Flow Enforcement)	36%	**Adequate**
	AC-6 (Least Privilege)	22%	**Moderate**
Cryptographic Protection	SC-8 (Transmission Confidentiality)	33%	**Adequate**
	SC-12 (Cryptographic Key Management)	31%	**Adequate**
	SC-13 (Cryptographic Protection)	32%	**Adequate**
Network & Boundary	SC-7 (Boundary Protection)	36%	**Adequate**
	SC-4 (Information in Shared Resources)	28%	**Adequate**
Audit & Monitoring	AU-2 (Event Logging)	31%	**Adequate**
	AU-3 (Content of Audit Records)	29%	**Adequate**
	AU-6 (Audit Review Analysis)	30%	**Adequate**
	AU-7 (Audit Reduction and Report Gen.)	30%	**Adequate**
	AU-12 (Audit Generation)	32%	**Adequate**
	SI-4 (Information System Monitoring)	29%	**Adequate**
Configuration Management	CM-2 (Baseline Configuration)	28%	**Adequate**
	CM-3 (Configuration Change Control)	29%	**Adequate**
	CM-5 (Access Restrictions for Change)	32%	**Adequate**
	CM-6 (Configuration Settings)	21%	**Moderate**
	SI-2 (Flaw Remediation)	18%	**Moderate**

6.1 Adoption of hybrid model for dynamic risk assessment

The proposed hybrid model demonstrates significant advancement in vulnerability exploitability prediction through strategic integration of traditional machine learning algorithms with large language model capabilities. The framework achieves superior predictive performance with R^2 score of 0.891 compared to ensemble-only approaches at 0.852 and individual algorithms ranging from 0.745 to 0.847, representing 4.6% improvement over ensemble methods and 19.6% enhancement over the weakest individual algorithm. This performance improvement translates directly to more accurate risk assessment capabilities for composite ICT product certification, enabling certification bodies to make evidence-based decisions regarding vulnerability impact and control adequacy. The ensemble learning component provides robust feature selection through integration of Random Forest, Gradient Boosting, and ElasticNet algorithms. The systematic reduction from 52 engineered features to 24 key features through ensemble consensus demonstrates the framework's ability to distil complex vulnerability datasets into manageable feature sets suitable for GPT-3.5 processing without sacrificing predictive accuracy. GPT-3.5 integration provides contextual analysis capabilities that traditional machine learning approaches cannot achieve through statistical pattern recognition alone, with training convergence from initial loss of 0.0923 to final validation loss of 0.0521 over twenty epochs demonstrating effective model adaptation with minimal overfitting. The economic feasibility represents a significant practical advantage, with structured prompt templates requiring approximately 300 tokens per prompt. For the training dataset of 4,770 records, total training cost is estimated at $2.91 for complete GPT-3.5 fine-tuning. Furthermore, the fine-tuning process was computationally efficient, completing in approximately 45 minutes on the specified NVIDIA A100 hardware. Cross-validation performance demonstrates consistent model behaviour with mean R^2 of 0.887 and standard deviation of 0.009, indicating superior stability compared to individual algorithms and ensemble-only methods.

6.2 XAI for conformity assessment

The integration of explainable AI techniques effectively supports cybersecurity conformity assessment by providing clear, understandable explanations of how security decisions are made. Rather than relying on traditional checklist

approaches, the framework enables auditors to see exactly which factors influence risks and justify the chosen security controls. In particular, SHAP analysis identified the key features, i.e., Base_score (0.168), has_public_exploit (0.142), whereas LIME analysis shows that public exploit availability is the key feature for almost 87% critical vulnerabilities, with an average influence of 0.221. This consistency across different analysis methods builds auditor confidence in the framework's reliability. The adoption of XAI met the established explainability criteria. Specifically, fidelity was demonstrated through consistent feature rankings that aligned with cybersecurity principles - vulnerabilities with higher CVSS scores and available exploits logically received higher risk ratings. Scalability was achieved both computationally, processing 1,193 vulnerability predictions efficiently, and cognitively, maintaining clear explanations whether analysing individual vulnerabilities or summarizing risks across all five P-NET assets. The marginal analysis methodology enabled systematic validation of security control effectiveness beyond simple pass-fail decisions. For P-NET, the evaluation revealed that 21 out of 24 implemented NIST controls adequately satisfy their security requirements, while identifying specific areas needing improvement in privilege management, configuration standards, and patch deployment. This detailed assessment helps organizations understand not just whether their controls work, but how well they work and where improvements are needed.Moreover, beyond the existing practice of XAI presented in this paper, future work can consider the framework as a defense mechanism against adversarial attacks and human errors. By exposing the precise feature contributions driving a risk score, techniques like SHAP could allow auditors to detect adversarial perturbations where an attacker might subtly alter input features to artificially lower a risk classification; the resulting explanation would reveal the inconsistency between the hidden high-risk attributes and the low score. Furthermore, future research may investigate how the framework mitigates human errors leading to attacks by transforming complex, opaque risk data into interpretable evidence, thereby reducing the cognitive load on auditors and minimizing the likelihood of overlooking vital misconfiguration during the certification process.

6.3 Works in context and practical implications

In comparison with existing works, research in regulatory requirements engineering demonstrates the growing need for automated compliance approaches, as evidenced by Kosenkov et al. [49] who identified a lack of structured cybersecurity certification approaches in the current state of the art. While recent studies explore automated analysis of GDPR requirements [50], compliance verification in Android apps using LLMs [51], and generative AI for traceability [52], our research uniquely addresses cybersecurity conformity assessment by applying AI-driven vulnerability analysis to validate security control adequacy. Unlike prior XAI-based dynamic risk assessments using Deep Q-Network (DQN) and LIME [5], which lacked comprehensive auditor support, this framework offers significant practical implications by addressing the fundamental challenge of scalability; the systematic processing of 5,963 vulnerability records demonstrates the capability to handle enterprise-scale certification scenarios while reducing labor requirements. Crucially, to ensure the XAI explanations are not merely plausible but mathematically faithful to the model's behavior, the framework integrates *Counterfactual Sensitivity Tests* via Marginal Analysis. Unlike static heatmaps, this approach verifies the causal link between features and risk scores by simulating the counterfactual state (i.e., "What if the control was optimized?"), thereby providing auditors with actionable "recourse" to alter decision boundaries from "High Risk" to "Acceptable." Consequently, stakeholders including auditors, AI researchers, and C-ToE owners benefit from a robust, reproducible method for verifying compliance against evolving standards. The practical implementation of the work offers significant implications for cybersecurity certification. The framework addresses the fundamental challenge of scalability in traditional certification processes, where manual assessment of complex composite ICT products requires substantial time and specialized expertise. The systematic processing of 5,963 vulnerability records across five critical assets demonstrates the framework's capability to handle enterprise-scale certification scenarios while maintaining detailed traceability at the individual asset level. The economic advantages extend beyond the low-cost GPT-3.5 fine-tuning to encompass reduced labor requirements and accelerated certification timelines. The framework's ability to provide automated control adequacy validation and systematic XAI explanations reduces the time required for evidence collection and analysis, enabling certification bodies to process more certification requests with existing personnel. Stakeholders including auditors, AI researchers, certification bodies, policymakers, manufacturers, CToE owners, and users from any cybersecurity application domain can benefit from using this framework.

6.4 Threats to validity

We exhibit potential threats and limitations that may affect the validity of the proposed approach and findings. The operationalisation of XAI practice using counterfactual based marginal feature analysis, SHAP , and LIME is one of the key contributions of this work for justification of security control and explaining audit evidence. However, SHAP often considers that the features are independent, while outcome from LIME may varies due to the random sampling. Moreover,

counterfactual analysis often lacks diversity due to consideration of single possible option for the correlation among features. This certainly impact to create consistent and deterministic audit evidence. In future, we aim to correlate the outcome from the multiple different explainers to minimise the limitation of each technique. We have focused only vulnerability exploitation as dynamic security parameters for the risk assessment. However, there are other parameters such as assets dependencies to determine the criticality of the compromise dependent assets, and threat intelligence data, which also evolved. To address this, we plan to consider additional parameters for the dynamic risk assessment. The evaluation of our approach considers a single use case scenario with vulnerability related CVEjoin dataset and one of the co-authors from the pilot scenario involved for the evaluation. This may not generalise our findings due to the consideration of single use case and dataset context and lack of availability of auditor feedback. To address this, we plan to consider pilot cases from different application context, inclusion of different security dataset such as threat intelligence data and involvement of other relevant stakeholder for improving the applicability of our approach and generalisation of the findings.

The framework's reliance on a third-party LLM API (OpenAI) introduces potential challenges regarding long-term reproducibility and future model availability. Furthermore, vulnerability ecosystems are inherently dynamic, with EPSS scores and exploitability metrics evolving over time; consequently, the current model represents a specific temporal snapshot, and operational deployment would necessitate periodic retraining to mitigate this drift. Finally, although the framework successfully generates audit-ready evidence, we have not yet conducted empirical user studies with professional auditors to quantitatively measure the resulting reduction in cognitive load.

To address potential biases in XAI explanations, we employed a multi-faceted validation approach. However, SHAP often considers that the features are independent, while the outcome from LIME may vary due to the random sampling. To mitigate these biases, we integrated Permutation Feature Importance (PFI) during the model training phase as a global baseline. This triangulation–comparing PFI's global rankings with SHAP's attribution and LIME's local insights–helps identify and reduce algorithmic bias, ensuring that the generated 'Audit Evidence' is robust and not merely an artifact of a single explanation method's assumptions.

7 Conclusion

The cybersecurity certification is paramount important for composite products to ensure their secure and safe operations. However, security posture of these products and their operating context constantly evolved, brings the necessity to dynamically assess the security risks and systematically assess the adequacy the chosen security control for overall security assurance. This paper presents a novel hybrid AI framework that integrates ensemble learning with GPT-3.5 for dynamic risk assessment and utilizes XAI to operationalise the Explainable AI practice to validate the adequacy of security control and requirement satisfaction. The integration of SHAP, LIME, and marginal analysis enables systematic validation of security control adequacy through quantifiable evidence rather than traditional checklist-based evaluations. The comprehensive evaluation through the pilot use case scenario and experiment with a vulnerability related dataset demonstrates the framework's practical effectiveness in processing large-scale vulnerability datasets and providing evidence of security control performance against established security functional requirements. The methodology successfully bridges the gap between advanced AI capabilities and certification body requirements for transparent, traceable decision-making processes. As a future direction of this work, we intend to consider combination of XAI techniques for explaining audit evidence. We would also like to investigate real-time monitoring capabilities to enable continuous conformity assessment that automatically adapts with evolving security posture throughout composite product lifecycles. To generalize the findings, we are also aiming to validate the framework through different application domains and dataset as a part of the future works.

Acknowledgements This work was supported by the European Union's Horizon Europe Programme, CUSTODES - A Certification approach for dynamic, agile and reUSable assessmenT fOr composite systems of ICT proDucts, servicEs, and processeS [grant number 101120684]; the European Union's Digital Europe Programme, EuDoros - prEparedness and mUtual assistance support by Deployment Of Ready-to-be-Offered cybersecurity Services [grant number 101158605]; the European Union's Digital Europe Programme, CURIUM-Cra sUppoRt continuum [Grant Agreement No. 101190372]; European Union's Horizon Europe Programme, CyberSecDome – An innovative Virtual Reality-based intrusion detection, incident investigation, and response approach for enhancing the resilience, security, privacy, and accountability of complex and heterogeneous digital systems and infrastructures, [grant agreement No. 101120779].

Author Contributions Shareeful Islam: Conceptualization, Writing – original draft, Writing – review and editing, Methodology, Investigation, Validation, Project administration, Supervision, Funding acquisition. Bilal Sardar: Conceptualization, Writing – original draft, Writing – review and editing, Methodology, Software, Investigation, Validation , Visualization. Spyridon Papastergiou: Conceptualization, Writing – original draft, Writing – review and editing, Methodology, Validation,Funding acquisition. Eleni Maria Kalogeraki: Writing – review and editing, Funding acquisition, Resources, Project administration. Kostas Lampropoulos: Writing – review and editing, Visualization, Software Investigation, Resources, Funding acquisition.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors would like to announce no conflict of interest.

Ethical approval This article does not contain any examinations with human members or animals performed by any of the authors.

Competing interests The authors declare no competing interests.

Source code availability The complete source code generated during this work, implementing the hybrid AI-based dynamic risk assessment framework and the associated Explainable AI (XAI) operationalization techniques, is publicly available to support reproducibility. The repository, which includes the ensemble learning modules, GPT-3.5 integration logic, and automated audit evidence generation scripts, can be accessed at: GitHub Repository.

References

1. Papastergiou, S., Islam, S., Kalogeraki, E.M., Chatzopoulou, A., Bountakas, P., Pournaras, K., Beena, S., Polemi, N.: Composite Inspection and Certification (CIC) System for Cybersecuity Assessment of ICT Products, Services, and Processes. In: 2024 IEEE International Conference on Cyber Security and Resilience (CSR). IEEE (2024)
2. ENISA: Industry 4.0 Cybersecurity Challenges and Recommendations. European Union Agency for Cybersecurity (2020)
3. Union, E.: Regulation (EU) 2019/881 of the European Parliament and of the Council on ENISA. Off. J. Eur, Union (2019)
4. Common Criteria Portal: Common Criteria for Information Technology Security Evaluation, Part 1, ISO/IEC 15408 (2024)
5. Basheer, N., Islam, S., Papastergiou, S., Kalogeraki, E.M.: Composite Product Cybersecurity Certification Using Explainable AI Based Dynamic Risk Assessment. In: IEEE International Conference on Cyber Security and Resilience (CSR) (2025)
6. Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion **107**, 102303 (2024)
7. Mia, M., Pritom, M.M.A.: Explainable but Vulnerable: Adversarial Attacks on XAI Explanation in Cybersecurity Applications. arXiv preprint arXiv:2510.03623 (2025)
8. Yeboah-Ofori, A., Ismail, U.M., Swidurski, T., Opoku-Boateng, F.: Cyber threat ontology and adversarial machine learning attacks: Analysis and prediction perturbance. In: IEEE International Conference on Cyber Security and Resilience (CSR), pp. 71–77 (2021)
9. Srinivas, J., Das, A.K., Kumar, N.: Government regulations in cyber security: Framework, standards and recommendations. Future Gener. Comput. Syst. 92, 178–188 (2019). https://doi.org/10.1016/j.future.2018.09.063
10. Marotta, A., Madnick, S.: Analyzing and Categorizing Emerging Cybersecurity Regulations. ACM Comput. Surv. 58(2), Article 51 (2025). https://doi.org/10.1145/3757318
11. Basheer, N., Islam, S., Alwaheidi, M.K.S., Mouratidis, H., Papastergiou, S.: Large language model based hybrid framework for automatic vulnerability detection with explainable AI for cybersecurity enhancement. Integr. Comput.-Aided Eng. (2025). https://doi.org/10.1177/10692509251368663
12. da Ponte, F.R.P., Rodrigues, E.B., Mattos, C.L.C.: CVEjoin: An Information Security Vulnerability and Threat Intelligence Dataset. https://figshare.com/articles/dataset/CVEjoin_A_Security_Dataset_of_Vulnerability_and_Threat_Intelligence_Information/21586923 (2025). Accessed 2025
13. Mohale, V.Z., Obagbuwa, I.C.: A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Front. Artif. Intell. 8 (2025)
14. Union, E.: Regulation (EU) 2019/881 of the European Parliament and of the Council on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification. Off. J. Eur, Union (2019)
15. European Commission: European Common Criteria-based cybersecurity certification scheme (EUCC), Commission Implementing Regulation (EU) 2024/482 (2024)
16. ENISA: EUCC Cybersecurity Certification Scheme - Assurance Levels and Evaluation Methodology. European Union Agency for Cybersecurity (2024)
17. Common Criteria Portal: Common Criteria for Information Technology Security Evaluation (2024)
18. ENISA: Security Profiles for ICT Products. European Union Agency for Cybersecurity (2023)
19. Tillu, R., Muthusubramanian, M., Periyasamy, V.: From data to compliance: The role of AI/ML in optimizing regulatory reporting processes. J. Knowl. Learn. Sci. Technol. 2 (2023)
20. Folorunso, A., Adewumi, T., Adewa, A., Okonkwo, R., Olawumi, T.N.: Impact of AI on cybersecurity and security compliance. Glob. J. Eng. Technol. Adv. 21 (2024)
21. James, C.: Regulatory implications of explainable AI in cybersecurity frameworks. Research Publication, Stanford University (2022)
22. Gaspar, D., Silva, P., Silva, C.: Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. IEEE Access 12 (2024)
23. Breiman, L.: Random Forests. Mach. Learn. **45**(1), 5–32 (2001)
24. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)
25. Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinformatics **9**, 307 (2008)
26. Ahmed, S., Al-Shareeda, M., Alturjman, F.: Explainable AI-based innovative hybrid ensemble model for intrusion detection. J. Cloud Comput. 13 (2024)
27. Zhang, J., Bu, H., Wen, H., Chen, Y., Li, L., Zhu, H.: When LLMs meet cybersecurity: A systematic literature review. Cybersecurity 8 (2025)
28. Li, M., Wang, S., Zhang, Q.: Large language model for vulnerability detection and repair: Literature review and the road ahead. arXiv preprint arXiv:2404.02525 (2024)
29. Elkhawaga, G., Elzeki, O., Abuelkheir, M., Reichert, M.: Evaluating Explainable Artificial Intelligence Methods. Electronics 12 (2023)

30. Yan, F., Wen, S., Nepal, S., Paris, C., Xiang, Y.: Explainable machine learning in cybersecurity: A survey. Int. J. Intell. Syst. 37 (2022)
31. European Commission: Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Brussels (2021)
32. ISO: ISO/IEC 42001: Artificial Intelligence – Management System. Geneva (2023)
33. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD (2023)
34. Razzak, I., Imran, M., Xu, G.: An optimized ensemble model with advanced feature selection for network intrusion detection. PeerJ Comput. Sci. 10 (2024)
35. Wang, W., Zhao, Y., Liu, Q.: How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks. arXiv preprint arXiv:2303.00293 (2023)
36. Anwer, M.S., Imran, M.: A feature selection-driven machine learning framework for anomaly-based intrusion detection systems. Peer Netw, Appl (2025)
37. Simon, S.M., Glaum, P., Valdovinos, F.S.: Interpreting random forest analysis of ecological models to move from prediction to explanation. Sci. Rep. 13 (2023)
38. Liu, M., Cen, L., Ruta, D.: Gradient boosting models for cybersecurity threat detection with aggregated time series features. In: 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS). IEEE (2023)
39. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. 67 (2005)
40. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33 (2020)
41. NIST MITRE CPE: Common Platform Enumeration (2025)
42. Software Bill of Materials (SBOM): NTIA Software Bill of Materials (2025)
43. Kvålseth, T.O.: Cautionary note about R^2. Am. Stat. **39**(4), 279–285 (1985)
44. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7 (2014)
45. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**(1), 79–82 (2005)
46. National Institute of Standards and Technology: Security and privacy controls for information systems and organizations, NIST SP 800-53r5 (2020)
47. OpenAI: GPT-3.5 Turbo fine-tuning and API updates (2025)
48. P-Net: P-Net Testing Infrastructure (2025)
49. Kosenkov, O., Elahidoost, P., Gorschek, T., Fischbach, J., Mendez, D., Unterkalmsteiner, M., Fucci, D., Mohanani, R.: Systematic mapping study on requirements engineering for regulatory compliance of software systems. Inf. Softw. Technol. 178 (2025)
50. Abualhaija, S., Ceci, M., Briand, L.: Legal Requirements Analysis: A Regulatory Compliance Perspective. In: Handbook on Natural Language Processing for Requirements Engineering (2025)
51. Alecci, M., Sannier, N., Ceci, M., Abualhaija, S., Samhi, J., Bianculli, D.: Toward LLM-Driven GDPR Compliance Checking for Android Apps. In: Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering (2025)
52. Etezadi, R., Abualhaija, S., Arora, C., Briand, L.: Classification or Prompting: A Case Study on Legal Requirements Traceability. arXiv preprint arXiv:2502.04916 (2025)
53. Abdeen, W., Wnuk, K., Unterkalmsteiner, M., Chirtoglou, A.: Challenges of Requirements Communication and Digital Assets Verification in Infrastructure Projects. arXiv preprint arXiv:2504.20511 (2025)
54. Abdeen, W., Unterkalmsteiner, M., Wnuk, K.: Auxiliary Artifacts in Requirements Traceability: A Systematic Mapping Study. arXiv preprint arXiv:2504.19658 (2025)
55. Frattini, J., Unterkalmsteiner, M., Fucci, D., Mendez, D.: NLP4RE Tools: Classification. Handbook on Natural Language Processing for Requirements Engineering, Overview and Management. In (2025)