International Journal of Information Security (2026) 25:36
https://doi.org/10.1007/s10207-025-01189-8

REGULAR CONTRIBUTION

Explainable AI based dynamic cybersecurity risk management for cyber insurability

Spyridon Papastergiou[1,2] · Nihala Basheer[3] · Kostas Lampropoulos[4] · Panayiotis Verrios[5] · Shareeful Islam[3]

Received: 2 October 2025 / Accepted: 8 December 2025
© The Author(s) 2026

Abstract
Cybersecurity risk is one of the primary and growing concerns for ensuing security and resilience of organizations, regardless of their size and type. While proactive risk management is effective, it is challenging due to the evolving and sophisticated threat landscape, exploitation of known and unknown vulnerabilities, and a dynamic security context. The dynamic security context further complicates to calculate the accurate risk level, leading to risk perception that can vary between different stakeholders. However, the demand for adopting cyber insurance is increasing as an effective risk mitigation strategy to avoid any potential loss. In this context, this paper proposes an Explainable AI (XAI) based dynamic cybersecurity risk management approach for informed cyber insurability decision making. The approach utilizes an Large Language Model (LLM) based framework for real-time, contextualized risk level assessment and adopts XAI techniques such as feature contribution and correlation, to justify the decision making. A comprehensive evaluation using an industrial use case and experiment demonstrates the applicability of the proposed approach. The experiment part uses a widely used vulnerability dataset to predicate high exploitable vulnerabilities and links them with the identified assets of the use case scenario. The result shows 96.9% accuracy for the exploitable vulnerability identification and XAI operationalisation justifies the selection of right security control and the cyber insurability decision based on the residual risk.

Keywords Explainable AI · LLM model · Cyber Insurance · Case study · Dynamic Cybersecurity Risk Management · Vulnerability

1 Introduction

In today's data-driven digital world, organizations heavily rely on digital infrastructure and data to support their operations. Security protection in this digital environment is more crucial than ever, threats originating in any part of the system can rapidly cascade to other part, resulting in devastating and wide spread business impact. In a recent cyber-attack on one of the largest retailers in UK led to millions of pounds in lost sales [1]. Therefore, proactive cyber security risk management practice is essential in safeguarding against cyber-attacks that pose for data breach, disruption of services, violation of regulatory compliance, and many more. Due to the rapid growth of cyber security risks and evolving nature of cyber-attacks, businesses are constantly looking for mechanisms to effectively manage these risks.

Cyber insurance can be an effective risk mitigation strategy to provide coverage against losses from these risks. In the insurance domain, cyber insurance represents the most rapidly expanding area [2]. The adoption of cyber insurance

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

Spyridon Papastergiou
spyros.papastergiou@maggioli.gr

Nihala Basheer
nihala.basheer@aru.ac.uk

Kostas Lampropoulos
klampropoulos@p-net.gr

Panayiotis Verrios
panayiotis.verrios@acta.com.gr

1 Department of Informatics, University of Piraeus, Piraeus, Greece

2 Research and Innovation, MAGGIOLI S.P.A., Romagna, Italy

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

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

5 ACTA LTD, Patras, Greece

Published online: 22 January 2026

requires a comprehensive understanding of the overall organizational cyber security posture to determine appropriate level of coverage and scope. However, the organizational security context is constantly evolving due to the uncertainty related to the exploitation of vulnerabilities, the reconfiguration of assets and their dependencies, dependencies with third-party service providers and many more [3]. This certainly impacts the determination of the insurance level [4]. For instance, within the given insurance period, there could be no severe risks faced by the business due to the new control being implemented or reduction of attack surface. However,the business could later face multiple cyber incidents which makes it difficult to determine the insurance value [5–7]. To address this complexity, there is a growing need for dynamic risk management approaches that continuously assess security risks in real-time and align risk values with appropriate insurance premiums. Such approaches enable organizations to adapt to evolving threats and make more informed decisions in regards to security controls and insurance coverage.

Within this context, this research work aims to support informed decision making and prevent adverse selection for the acquisition of cyber insurance as a risk mitigation strategy. Our work combines both forward-looking and backward-looking strategies so that risk can be identified and predicated by considering the existing organization data and dynamic security context. The work presents four novel contributions.

- Firstly, we propose an AI-based dynamic risk management approach that incorporates temporal security parameters, such as vulnerability exploitation likelihood, and maps them with relevant organizational assets. This enables the identification and quantification of context-specific risks, supporting informed cyber insurance acquisition decisions. The approach includes a conceptual framework to understand cybersecurity and cyber insurance domain knowledge, along with a systematic process to guide risk assessment based on existing assets, control selection, and insurance justification.
- Secondly, we leverage the capabilities of large language models, specifically CodeBERT (Bidirectional Encoder Representations from Transformers), to enhance dynamic risk management by identifying vulnerabilities with high exploitation potential. CodeBERT, a bidirectional transformer pre-trained on both source code and natural language, captures complex semantic patterns in vulnerability descriptions. By incorporating Exploit Prediction Scoring System (EPSS) scores, the approach prioritizes vulnerabilities based on their exploitation probability, which directly contributes to dynamic and data-driven risk quantification.
- Thirdly, we operationalise Explainable AI (XAI) practice to derive suitable security controls by identifying key features using both feature correlation and contribution techniques. While enhancing transparency in model decisions, the main objective was to extract actionable features for control declaration. These common key features were mapped to corresponding controls following the National Institute of Standards and Technology (NIST) 800-53 framework. The controls, together with a quantified residual risk, informed cyber insurance acquisition decisions, allowing organizations to justify their security posture and coverage requirements with interpretable evidence.
- Finally, a comprehensive evaluation is performed to validate the applicability of the approach, including the use of XAI in real-world security decision-making. The evaluation integrated an industrial use case with experimentation on p-NET 5G infrastructure – an advanced competence centre for digitalization services. Experiment results confirmed the effectiveness of CodeBERT in vulnerability prioritization, while SHapley Additive exPlanations (SHAP) and correlation heatmaps revealed key risk factors such as exploitability, Exploit Prediction Scoring System (EPSS), and confidentiality impact. These insights directly supported control selection and residual risk analysis, informing actionable cyber insurance decisions.

2 Related works

The existing works focus on risk management and AI adoption for cyber insurance. This section provides an overview of the existing works that are relevant to our approach.

2.1 Dynamic Risk Management

The Dynamic Risk Management Response System (DRMRS) is proposed as a quantitative and risk-aware approach aiming to protect critical infrastructure [8]. DRMRS correlates the likelihood of threat success with impact, cost, and response side effects, producing a proactive risk profile detailing attack scenarios, threat risk quantification to measure risk and the response operational impact assessor to measure collateral damage. Another method addresses the safety assurance challenges the cyber physical system considering autonomy, interconnectedness and Frame Conditions perspectives [9]. The aim of this approach is to enhance safety awareness by introducing safety intelligence layer to adjust safety related parameters. It adopts Dynamic Capability Assessment (DCA) to assess system safety related capabilities and Dynamic Risk Assessment (DRA) to measure the potential consequence of safety concerns and trade-off between

performance and safety risk. A systematic literature review of 50 dynamic risk management methods is performed by [10] considering three areas: AI/ML, mathematical-model-based, and unclassified. The review outcome confirms that most models adopt quantitative risk analysis and assess risk level by following the CTI data such as recently discovered vulnerabilities and emerging threats. A dynamic risk assessment framework examines the evolution of risk vectors and their exposure over time [11]. The work demonstrates high level causal relationship among five internal and seven external drivers including cyber defence maturity, regulatory requirements, and technology, yet it lacks specific method for quantifying the risk from these drivers.

2.2 Risk Management and Cyber insurance with AI Adoption

Effective risk management is critical for cyber insurability and AI is now widely adopted in risk management to calculate risk or identify threat patterns due to its capabilities to correlate data . As businesses increasingly adopt digital services and the rise of sophisticated cyber attacks demand the necessity of flexible, data-informed, and economically viable cyber insurance solutions. Recent research highlights the importance of leveraging behavioral patterns, economic models, and artificial intelligence to improve the accuracy of risk assessment and premium calculations. A number of studies and research have contributed towards the development and implementation of various insurance frameworks. For instance, the study by Panou et al. [12] introduced the RiSKi framework which incorporates econometric modelling and serious gaming for estimating asymmetric premiums and addresses the challenges relating to real-time risk dynamics and data scarcity. However, the practical applicability is limited by the absence of a real-world case study. Similarly, A study by [13] presents an integrated framework that optimizes information security investment along with itemized cyber insurance that were designed for SMEs considering an Innovative Cyber Insurance model where insurers set aside part of the premiums to enhance the cybersecurity defences of the insured companies. The framework is notable in illustrating how insurance can act not only as a means of transferring risk but also as an incentive for proactive risk mitigation. Zhang and Zhu [14] are more concerned with contract design as they focused on temporal risk and user behaviour to determine optimal insurance contracts. They also incorporated aspects like moral hazard and incentive compatibility using dynamic Markov Decision Processes which is important for dynamic model formulation. Thlon and Strupcvewski [15] employed statistical and machine learning techniques to show how organizational size alongside history of cyber incidents and perceived risk can help determine the required insurance. Their study also emphasized the importance of availability heuristics as a key decision determinant. With practical application of AI, Jawhar et al. [16] provide automated cyber risk profiling and insurance recommendation generation. It used the capabilities of GPT 4 to analyse the responses of customers structured cybersecurity questionnaires. The model was tested on 100 samples and the results highlight the ability of the model to maintain reliability and precision while creating dynamic risk profiles and generating policy recommendations. Building upon this work, Shaikh et al. [17] examine risks associated with ransomware, proposing a model that combines technical mitigation strategies with insurance design along with organization-wide barriers concerning implementation cost and complexity. The authors explain the gaps in adaptive insurance policies, illustrating the lack of clear cost allocation and coverage highlights. Their findings point to an increasing demand for specialized coverage as ransomware persists in its evolution as a top-tier cyber threat.

Several recent studies have examined the interplay between risk assessment, data availability, and contract design in the cyber insurance domain. Zeller and Scherer [7] investigated how insurers can optimally integrate risk mitigation services into cyber insurance contracts, demonstrating that the interaction between risk transfer and proactive security measures significantly influences pricing and portfolio management. Their findings highlight that insurers may be incentivized to subsidize risk reduction activities depending on dependency structures among insured entities. Complementing this, Cremer et al. [8] conducted a systematic review of data availability for cyber risk and cybersecurity, underscoring the persistent challenge of scarce and fragmented datasets that hinder accurate premium pricing and sustainable insurance product development. The authors argue that improved access to open datasets is critical to overcoming information asymmetry and fostering more precise cyber risk modelling. Chong et al. [9] proposed an incident-specific cyber insurance model, addressing the complexity of designing indemnities for distinct types of cyber incidents such as ransomware or data breaches. Their work establishes an economic foundation for policies that differentiate between incident types, thereby aligning indemnities with the statistical properties of diverse cyber perils. Collectively, these studies advance the understanding of how cyber insurance can evolve into more adaptive, data-driven, and context-sensitive instruments, capable of responding to the uncertainties of a rapidly changing threat landscape.

Moreover, recent research has further explored attack-specific risk assessment and adoption of explainable AI techniques for insurance decision-making. Biswas et al. [18] propose a hybrid explainable AI framework targeting phishing attacks and consider dark forum analysis to estimate attacker expertise, machine learning-based URL classification, and Archimedean Copula to quantify correlated attack losses.

This allows to undertake optimised investment decision considering IT security and cyber insurance. Mukhopadhyay and Jain [19] proposed the Ransomware Risk Management Model (R2M2) based on Protection Motivation theory, threat appraisal components from NIST guidelines for calculating attack severity. The approach is validated through interviews with targeted audience such as firms, security service providers, and cyber insurers, emphasizing the importance of combining perimeter security investments with cyber insurance to address residual ransomware risk. Pavlík et al. [20] extended the temporal dimension of cyber threat assessment by designing algorithms that emphasize time-based financial impact modelling for insurance contexts, recognizing that the detection and expression of cyber threat scenarios over time significantly influences loss determination. A case study is used to demonstrate the applicability for insurance, where the risk level is calculated beyond considering the impact.

Parallel developments in optimization strategies and theoretical foundations have advanced the understanding of security investment and insurance coverage decisions in interconnected risk environments. Uuganbayar et al. [21] proposed cost-effective security controls, selection based algorithm alongside with insurance coverage optimization. The work emphasizes how risk-averse organizations should distribute cyber security budgets between self-protection measures and insurance premiums while considering multiple threats simultaneously. Boonen et al. [22] adopt a non-cooperative game theory to model cybersecurity investments and insurance purchases among interconnected firms, deriving pure-strategy Nash equilibrium conditions and demonstrating that security investments and insurance function as strategic complements rather than substitutes. Awiszus et al. [23] provided a comprehensive actuarial framework using risk types including idiosyncratic, systematic, and systemic risks, emphasizing that while classical actuarial mathematics prove adequate for the first two categories, systemic cyber risks necessitate sophisticated approaches capturing network effects and strategic interactions through risk-neutral valuation. French [24] focus on the broader cyber insurance market structure, existing challenges - including market fragmentation - policy exclusions, and insufficient coverage limits relative to potential loss magnitudes.

In summary, existing work on risk management and cyber insurance focus on AI-driven risk profiling, attack-specific frameworks, and theoretical investment optimization with insurance decision. However, these approaches exhibit several limitations such as emphasis on isolated attack types rather than comprehensive vulnerability assessment within dynamic security context, lack of justification for chosen controls and residual risk in relation to the insurability decision. There is need to consider the dynamic security context for risk assessment and provide appropriate justification for chosen security control and insurability decision. The proposed approach contributes towards this direction. A Comparison of the existing work and proposed work is presented in Table 1.

3 Cyber Insurance

Cyber insurance is a specialized form of coverage designed to protect organizations from financial, legal, and operational consequences of cyber events including data breaches, ransomware attacks, and system compromises [25]. As organizations increasingly depend on digital infrastructure, cyber insurance has emerged as a critical risk management instrument to address residual risks that cannot be fully mitigated through technical controls alone [26]. This section provides background on cyber insurance mechanisms, policy structures, and the technical challenges that motivate our dynamic risk management approach for informed decision making.

3.1 Cyber Insurance Fundamentals

Cyber insurance operates as a risk transfer mechanism where organizations pay premiums to transfer potential cyber-related losses to insurers. Understanding how insurers evaluate security posture and price coverage is essential for connecting vulnerability management with insurance decisions. This evaluation occurs through underwriting processes that face unique actuarial challenges distinguishing cyber insurance from traditional lines.

- **Risk Assessment and Underwriting:** Underwriting is the process by which insurers evaluate organizational risk and determine coverage terms, pricing, and eligibility [27]. Insurers employ three primary assessment mechanisms to evaluate security posture. Firstly, security questionnaires aligned with frameworks like NIST CSF and ISO 27001 assess control implementation across 50-200 control points including patch management, access controls, and incident response. Secondly, external vulnerability scanning identifies exposed weaknesses in internet-accessible assets. Finally, analysis of prior incident history examines past security failures and loss experience. However, these mechanisms capture only point-in-time security posture without continuous monitoring, creating temporal gaps where organizational risk profiles may change significantly throughout the policy period while premiums remain static. This temporal limitation directly connects to broader actuarial challenges insurers face when quantifying cyber risk beyond initial assessments.
- **Actuarial Challenges**: Building on the temporal limitation identified in underwriting, actuarial challenges refer to difficulties applying traditional insurance mathematics

Table 1 Comparison of the Proposed Work with Existing Works

Criteria	Existing Works	Proposed Work
Cyber Security Risk Management Process for Cyber Insurability	Biswas et al. (2024) [18] propose a three-phase framework to manage phishing-specific cyber risks based on life cycle of phishing attack and adopt utility theory and copula to balances investment in technology and cyber insurance as a part of risk mitigation. ML model is adopted for URL classification. It is mainly a theoretical based work and lacks demonstration of the applicability of the proposed approach. Mukhopadhyay & Jain (2024) [19] presents R2M2 model for managing ransomware-specific cyber risks through three modules (assessment, quantification, and mitigation), where PMT theory combined with NIST guidelines is used to identify vulnerabilities and quantify risks in the risk management process. The work advocates mitigation of residual risk though cyber insurance but lacks justification with emphasis only on interview responses. Pavlík et al. (2022) [20] propose an algorithm to determine the financial impacts of selected threats over time. The risk assessment considers time frame of threat, related vulnerability, probability of risk and risk degree. A case study is used to demonstrate its applicability and benefits for insurance and cybersecurity.	The proposed work employs dynamic risk management through vulnerability exploitability prediction using a CodeBERT-based Large Language Model (LLM) integrated with EPSS scores to calculate real-time risk levels. This approach considers the selection of security controls to determine cyber insurability decisions based on residual risk. A comprehensive evaluation of the approach is performed using an industrial use case and experiment.
Incident Specific Cyber Insurance	Mukhopadhyay & Jain (2024) [19] employs collective risk modelling to compute incident severity specifically for ransomware incidents, proposing residual risk-based insurance as a mitigation strategy for incident-specific cyber insurance coverage decisions. The framework is validated through interviews but remains focused solely on ransomware. Chong et al. (2025) [9] apply Pareto optimality framework to determine incident-specific cyber insurance coverage for multiple incident types (Privacy Violation, Data Breach, Fraud/Extortion, etc), where separate deductibles and limits are optimized for each incident type to make incident-specific insurance decisions that benefit both buyer and seller. Shaikh et al. (2024) [17] consider ransomware insurance strategies including coverage aspects, premium costs, and exclusions to provide financial protection and incident response support for incident-specific cyber insurance, where insurance complements technical security measures for comprehensive ransomware risk mitigation.	The proposed work considers the exploitation of specific vulnerabilities what are relevant to the organizational asset such as Linux, MySQL, Tomcat, and PostgreSQL from the pilot context. Key exploitable features are taken into consideration EPSS, attack_vector, user_interaction, and privileges_required to determine the suitable control and insurability decision. This is a technical vulnerability-based approach which addresses root causes rather than incident symptoms to determine appropriate cyber insurance coverage requirements.
Cyber Insurance Coverage	Uuganbayar et al. (2021) [21] develops an exact algorithm with utility theory to optimize the distribution of cybersecurity budget between insurance coverage and self-protection measures, selecting cost-effective controls to maximize coverage efficiency for cyber insurability. The approach is black box without explainability and represents a static, one-time optimization. Boonen et al. (2023) [22] employ non-cooperative game theory to optimize both cybersecurity investment levels and insurance coverage ratios simultaneously, finding that these two strategies are strategic complements through Nash equilibrium analysis for coverage optimization decisions. The limitation is that this approach is purely theoretical without practical implementation guidance. Chong et al. (2025) [9] use the cross-entropy method to solve the incident-specific coverage parameters (deductibles and limits), balancing buyer and seller risk preferences measured by VaR/TVaR to optimize cyber insurance coverage for mutual benefit.	The proposed work employs XAI-driven SHAP analysis to justify the insurance coverage based on trade-off analysis between implementing security control and residual risk based on the key features of the exploitation.

Table 1 continued

Criteria	Existing Works	Proposed Work
Security Control and Insurability Decision	Uuganbayar et al. (2021) [21] proposes a cost-effective control selection algorithm for decision making between self-protection through controls versus purchasing insurance coverage. Lacks justification for control selection along with linking between control effectiveness and residual risk for insurability decisions. Awiszus et al. (2023) [23] present mathematical models that distinguish between idiosyncratic, systematic, and systemic cyber risks for pricing cyber insurance using actuarial and financial mathematics. However, it does not provide a practical framework for how organizations should select security controls or assess insurability for making security control and insurance decisions.	The proposed work justifies the chosen security control based on the key exploitable features and how specific controls mitigate a feature to be exploited for cyber-attacks. The identified controls are mapped to the NIST 800-53. The result from the pilot case study shows that controls such as FIA_UID.2 (User identification before action) is selected for the CVE-2014-0160 vulnerability related with Linux asset, however with the high EPSS score cyber insurance is recommended.
Adoption of Explainable AI for Cyber Insurance	Biswas et al. (2024) [18] apply XAI techniques to explain phishing URL classification decisions and to provide transparent recommendations for optimal investment allocation between IT security and cyber insurance. However, the XAI application is limited on explaining classification results and is not used for security control selection or insurability decisions based on residual risk analysis. Jawhar et al. (2024) [16] discusses general AI mechanisms to enhance the entire cyber insurance lifecycle including risk assessment, policy pricing, and claims processing. However, it does not employ specific XAI techniques to provide decision transparency or explainability for stakeholders in the insurance process.	The proposed work comprehensively operationalizes XAI for dual purposes in cyber insurance: Firstly, SHAP (SHapley Additive exPlanations) analysis is used for identification of key features' contributions; Secondly, correlation heatmaps are employed to reveal feature interdependencies for comprehensive residual risk assessment that justifies insurability decisions. The XAI insights are mapped to NIST 800-53 standardized controls providing transparent and explainable security measures.

to unique characteristics cyber risk's [26]. These challenges manifest across different dimensions that complicate risk pricing. Loss frequency modelling struggles because current methods rely on vulnerability severity rather than exploitation likelihood–only 2-7% of published vulnerabilities are actively exploited [28], yet assessments treat all high-severity vulnerabilities as equally probable threats, producing inaccurate predictions. Loss severity estimation, faces wide incident cost variance ranging from thousands to millions depending on data sensitivity and regulatory jurisdiction. Most critically, portfolio risk management encounters fundamental problems because cyber risks exhibit high correlation –a single vulnerability can simultaneously affect thousands of policyholders [26], violating the independence assumption underlying traditional insurance pooling and creating catastrophic aggregate loss potential. These actuarial challenges directly shape how policies structure coverage, exclusions, and pricing to manage insurer exposure.

3.2 Policy Structure and Coverage Mechanisms

It is necessary for organization to understand the cyber insurance policy structure and coverage mechanism for informed decision making regrading security control and coverage acquisition. Hence, a cyber insurance policy consists of coverage scope, limits, deductibles/retention levels, technical prerequisites, exclusions, and coinsurance provisions, which collectively determine the risk transfer arrangement. This section provides an overview of the relevant areas.

– **Coverage Types and Policy Limits:** Generally, cyber insurance policy is divided on two main categories depending on the loss. First-party coverage addresses direct organizational losses and covering cost for the loss such as data breach response (forensic investigation, notification), business interruption (revenue loss during system downtime), ransomware (ransom payments and restoration cost which is often restricted by insurers), and data restoration (recovery of corrupted data). First-Party coverage often requires the organization to implement specific controls for coverage eligibility, such as multi-factor authentication (MFA) and network segmentation for ransomware coverage, or encryption for data breach response[27]. Third-party coverage protects organization against liability claims brought by external parties including privacy liability (legal defense and settlements related data protection violations) and regulatory defense (costs for regulatory investigations and potential fines where insurable). Both First-Party and Third-Party coverages are subject to aggregate policy limits, which define the

maximum cap the insurers can payout over the policy period, and specific incident types.

- **Deductibles and Retention:** Deductibles define organization's out-of-pocket costs that must met before insurer coverage activates. It is structured as per-occurrence such as for first-party losses,deductibles are applied per-occurrence, while it is for per-claim in case of third-party losses. organizations certainly face a trade-off where higher deductibles reduce premiums by lowering the insurer's liability but increase organizational financial exposure during the incidents. This structure serves dual purposes of reducing insurer's loss exposure while altering moral hazard, bby ensuring the organization bears a meaningful financial consequence. Some policies use aggregate deductibles across all claims instead of per-incident, fundamentally changing economics for organizations experiencing multiple incidents.
- **Exclusions:** While coverage defines insurer obligations, exclusions define boundaries where the coverage ends, directly enforcing the security practices assessed during underwriting. Infrastructure deficiency exclusions deny claims for the known unpatched vulnerabilities beyond specified time frames (30-90 days post-disclosure), enforcing the patch management practices evaluated in underwriting questionnaires. Nation-state attack exclusions deny claims attributed to state actors based on technical indicators, which gained prominence after NotPetya (2017) triggered coverage disputes [27]. Prior knowledge exclusions deny claims for known vulnerabilities but left unremediated, placing continuous burden on organizations to maintain the vulnerability management practices insurers assess. Coinsurance provisions may require organizations to bear 10-20% of covered losses above deductibles, further aligning incentives.
- **Premium Estimation and Rating:** Generally, insurers translate policy structures, deductibles, and exclusions into premium costs using underwriting assessments based on three primary rating approaches: experience rating (loss history), exposure rating (organizational characteristics like revenue and industry), and schedule rating (adjustments for specific risk factors)[27]. The schedule rating adjustments is also significantly influenced by three factors connecting to underwriting mechanisms: vulnerability exposure using CVSS scores determines baseline hazard; exploitation indicators (like CISA KEV listing) signal heightened probability requiring premium adjustments; and implemented security controls reported via questionnaires can earn premium credits. However, specific formulas and credit schedules remain proprietary, creating opacity preventing organizations from understanding how security investments affect costs. This opacity creates critical challenges for organizations attempting to optimize security spending for insurance value.
- **Control Selection:** The proprietary nature of insurer rating methodologies indicates that organizations lack crucial visibility into which security controls (like MFA or network segmentation) influence premium rates or coverage eligibility, making it difficult to justify implementation costs against potential premium reductions. Research shows that substantial premium variance for identical controls across insurers [27] , indicating assessment depends on underwriter judgment rather than standardized methodologies. This forces generic control selection driven by compliance rather than insurance optimization. Moreover, discovering essential control requirements during the quoting or claims process instead of planning phase, preventing cost-benefit analysis and rational security budget allocation between self-protection and risk transfer.This opacity directly connects to the broader challenge of quantifying what risk remains after controls are implemented.
- **Residual Risk:** organizations need to identify the existing residual risk after implementing the controls in order to determine the coverage limit [6]. Without a detailed understanding of the residual risk exposure, organizations need to rely on arbitrary benchmarks like annual revenue, or peer limits to select the coverage. This can lead to over-insurance, i.e., wasting capital on excessive limits or under-insurance, i.e., leaving the organization vulnerable to catastrophic loss. If the organizations lack evidence-based justification in premium negotiations then they are unable to demonstrate risks that are transferred or eliminated using controls. This leads to actuarial challenge as insurers price policies based on perceived risk rather than accurately measure the actual risk reduction achieved by specific security controls.

4 Explainable AI Practice for Cybersecurity

Explainable AI(XAI) plays an important role in ensuring that cybersecurity decisions made utilizing AI are reliable, and understandable to all the user groups. It provides insights into how and why a model made a particular prediction or classification [38]. XAI enhances the interpretability of a model, enables compliance and auditability requirements, and allows human analysts to check or challenge AI decisions with confidence. In the context of cybersecurity risk management, XAI is particularly essential in enabling transparent risk-informed decision-making. It allows organizations to understand the factors enabling risk assessments, substantively prioritize mitigation measures, and link security controls with actual risk drivers [29]. We provide an

overview of the two key components of XAI: feature contribution and feature correlation.

– Feature correlation refers to the statistical relationship between input features in a model, showing how changes in one feature are associated with changes in others [30]. It is vital to ensure that the model is understanding the meaningfully patterns from the data and not learning relying on redundant or misleading features. It is also necessary to check multicollinear ,which can skew model predictions and lower interpretability when high feature correlation is identified. To effectively view and understand the correlations of features, it is essential to visualize them using various tools, as detailed below.

1. **Correlation Heatmap** is a visual tool for showing the correlation of numerical variables in a dataset. It is built from a correlation matrix, with each cell holding the strength and direction of relationship between two variables [31]. It is especially useful in pattern detection, multicollinearity, or feature dependency, which affects the performance of the model and the selection of features in machine learning operations. The hues of the heatmap typically range from dark blue (high negative correlation) to dark red (high positive correlation), while pale hues indicate weak correlations or no correlation. The most commonly used metric to compute correlation in heatmaps is the Pearson correlation coefficient [32]. It measures the linear relationship between two continuous variables and can be computed by the formula 1:

$$r_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \times \sqrt{\sum(y_i - \bar{y})^2}} \tag{1}$$

Here, x and y are single sample points, $\bar{x}$ and $\bar{y}$ are the means of the respective variables, and the summation runs over all observations. The value, r, ranges from between -1 and 1, where values close to 1 indicate strong positive linear correlation, values close to -1 indicate strong negative linear correlation, and values close to 0 indicate no linear relationship.

2. **Feature Contribution** is the mechanism of analysing how every input feature (token or variable) contributes to the prediction of an AI model. This is particularly crucial in complicated or black-box models such as transformes, where it is not clear how the inputs influence the output. By assigning contribution scores to individual features, we gain insights into model behavior, detect potential biases, and build trust with stakeholders and users [33]. In this work, we have considered a widely used techniques to evaluate feature contributions: SHAP (SHapley Additive exPlanations).
3. **SHAP** relies on cooperative game theory and employs the Shapley values concept to distribute the difference between the model's prediction and the average (baseline) prediction proportionally and fairly across all input features. The central idea is to quantify the contribution of a feature by calculating its marginal effect over all possible feature subsets it might be part of. For a model f, an input instance $x \in \mathbb{R}^n$, and a baseline value x' (often the dataset means or a neutral input), SHAP computes the feature attribution ϕ_i for each feature i as in Equation 2:

$$\phi_n = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \times \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{2}$$

where N is the set of all features, S is a subset of features not containing i, and f_S is the model trained or evaluated with only features in subset S. The Shapley value ensures additivity, meaning the sum of all feature attributions equals the difference between the model's output and its expected output as mathematically presented in Equation 3.

$$f(x) = E[f(x)] + \sum_{i=1}^{n} \phi_i \tag{3}$$

SHAP values provide both local explanations of individual predictions and global insights into feature importance across the data set.

5 Dynamic Cybersecurity Risk Management for Insurability

In the context of cybersecurity, dynamicity implies the ever-changing digital world and how it impacts the cybersecurity domain, encompassing the new attack surface volume, new possible vulnerabilities, evolving opponent methodologies, etc. This unpredictability would necessitate an equally adaptive security approach, which gives rise to dynamic risk management. Dynamic risk management includes continuous and real-time supervision, evaluation and risk mitigation in terms of temporal security parameters [34]. It employs methods such as threat intelligence, security telemetry, and AI-based risk scoring for rebalancing exposure levels and dynamic control adjustment based on evolving threats. This approach is preferred over periodic assessments as it enables constant evaluations and organizations can act quickly and without restriction to emerging threats. From an insurance

point of view this form of managed risks improves organizational profile greatly. Cyber Insurance acts as a third party mitigative risk strategy providing financial sustenance and resilience in situations where technical controls cannot keep up with new threats. A dynamic approach articulates positive security governance, reducing the likelihood and severity of claims. It gives insurers timely visibility into the risk exposure of the organization, enables appropriate premium pricing, and enhances the chances of claim validation. Finally, adopting dynamicity through explainable, intelligent, and adaptive risk management enhances both cybersecurity outcomes and trust in insurability.

The proposed approach adopts a dynamic landscape using an LLM-based framework for real-time contextualized risk level assessment. Moreover, trusting model-driven decisions requires them to be transparent as well as auditable, thus XAI techniques like feature contribution and feature correlation are integrated with risk management to maintain decision-driven transparency. These explainability tools help verify the key features and their interactions that drive predictive algorithms thereby allowing more precise control implementation [35]. This combined approach ensures resilient, yet insurable cybersecurity strategies designed to adapt to risks while maintaining transparent accountability in decision processes. In this section, these are further described along with the conceptual view and the unique process intertwining LLM Model and XAI practices within risk management frameworks.

5.1 Conceptual view

A conceptual model is an abstract description of an idea or process that does not go into technical details such as computation and algorithms [36]. It serves to illustrate the interaction of the different components within a system. This also allows better understanding among stakeholders, team members, or researchers due to the need to focus on key ideas rather than implementation details. A conceptual model aims to shed light on intricate systems using logical relationships and hierarchies among significant ideas within relevant frameworks. The essential concepts required for the proposed approach are extracted from the domains of cybersecurity, risk management in information technologies systems, XAI, and cyber insurance domains which are elaborated below [37, 38].

- **Asset Context** lays out the foundation defining the assets by describing their nature and values for the organization. The proposed approach follows the NIST Common Platform Enumeration (CPE) catalogue to identify and category the asset including hardware and software in a structured manner. However, in case of customised application or open source software package Package URL (pURL) naming convention is used. The asset context including different types such as asset name, type, CPE(if available), pURL, etc.
- **Threat Intelligence** enhances any security posture by analysing and compiling current and historical data concerning threat landscape. This encompasses gathering data regarding known attack vectors, adversary tactics, techniques, and procedures (TTPs) and even Indicators of Compromise (IoCs). Such information helps determine which vulnerabilities may be exploited by potential attackers based on how they orchestrate breach attempts, thus improving risk predictions and enabling proactive countermeasures.
- **Security Telemetry** encompasses a stream of data that is collected from various infrastructures of the organization, such as system logs and security alerts or events from endpoints. Gathering this data reflects ongoing processes and suspicious activities within the organization's infrastructure. It is crucial for real-time threat detection, enabling both AI-based analysis and human-based risk assessments. It provides evidence of attempted breaches or completed attacks, thereby enhancing the ability to set up dynamic defenses that respond actively and providing continuous monitoring alongside feedback loops.
- **Vulnerability** outlines some weaknesses or defects that lie in the given assets. Each vulnerability has an identifier (such as a CVE ID), an extended description, the product and vendor it affects, and predictive metrics such as the Exploit Prediction Scoring System (EPSS), which estimates the likelihood of exploitation. Prioritization of vulnerabilities also exists, both based on technical severity and relevant context. This component is the one that provides threat detection analytics and is informed by asset characteristics and AI-based risk models.
- **Risk** is defined as the probability of specific vulnerability exploitation and potential impact. It is a measure of how likely a vulnerability can be exploited and severity due to the exploitation. Generally, a risk quantified by the potential damage such as loss of revenue, increased operating cost, reputational damage, legal penalty, and many more that can be caused by vulnerabilities, both in terms of the likelihood of exploitation and the resulting effect. Risk, by this proposed approach, is prioritized into five different levels, i.e., very high, high, medium, low, and very low. It is necessary to choose the right control based on their priority. Depending on specific risk context, any residual risk that remains after implementing the control, can be transferred using cyber insurance.
- **AI Model** is the computational mind of the system, using inputs from vulnerabilities, threat intelligence, and telemetry to identify patterns, classify risks, and rank responses. It consists of the model architecture (decision trees or neural networks), fine tuned hyper-parameters,

and performance metrics like recall, precision, and accuracy. The model allows rapid processing of intricate data which would take an unreasonable amount of time to analyse manually.

- **XAI** applies interpretable techniques such as feature contribution or correlation to explain how certain vulnerabilities are assigned a specific score. Consider a model, for example, that indicates which asset attributes contributed most to a high-risk rating. These clear explanations build trust among security, development, and business groups, which leads to collaborative remediation.
- **Control** is defined as a security measure that aims to prevent, reduce, transfer the risk so that the organization can continue operations. These are the technical, management, and operational course of actions for tackling the identified risks. The control selection follows the guidance provided by the NIST SP 800-53 framework based on the twenty control families [53]. Moreover, the controls are also categorised into three distinct function types including corrective, detective and preventive.
- **Cyber Insurance** refers to a form of coverage that aims to alleviate the impact of residual risk, allowing the organization to continue its operation. It is a risk mitigation strategy that transfers the financial impact to the insures, ensuring operational and financial stability of the organization. Cyber insurance includes several properties including first and third party coverage, protection scope, mandated controls, and exclusion.

The conceptual model presented in Figure 1 links these concepts. Asset context defines an inventory of Common Platform Enumerations (CPEs) and Package URLs(pURLs) related to the organization. This asset inventory is connected to vulnerability and threat data sourced from threat intelligence. For vulnerabilities prioritization and risk quantification, Explainable AI (XAI) is used to correlate the vulnerability data and inform decision-making. The resulting risk level determines the selection of security controls. The control justification, using XAI, is then used to support cyber insurability decisions.

5.2 Process

A process is a collection of actions, steps, or operations that are performed in a sequential manner to achieve a desired result or goal [39]. Our process includes AI and explainable AI based decision-making approach along with a risk level calculation for cyber insurance procurement. This process contains six interdependent steps which have distinct tasks guiding the process from gaining insight on an organization's assets to obtaining cyber insurance as illustrated in Figure 2. It initiates with key services and assets identification then

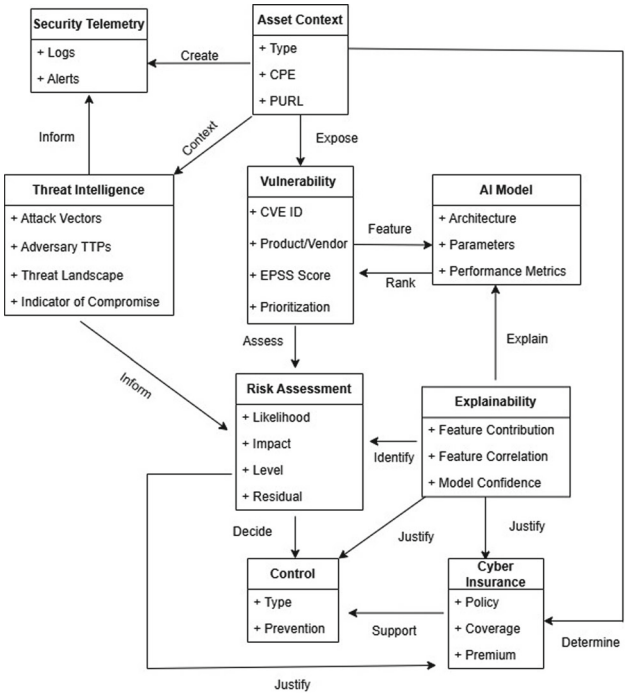

Fig. 1 Conceptual Model

moves to vulnerability detection and prioritization through AI model application. Risk is assessed dynamically in a graded manner translating into better understanding. These decisions facilitate selection and validation of security controls which support explanation able decision-making during assessment of residual risks in conjunction with appropriate cyber insurance frameworks.

5.2.1 Phase 1: Organizational Asset Context

The process of the proposed approach triggers this phase which aims to identify the services and related asset context within these services. This allows to identify the potential vulnerabilities that are linked with the asset so that risk posed from these vulnerabilities can be assessed and managed. At the beginning, the phase reviews the existing services of the organization that run the key business functions. Therefore, service identification requires to consider core business functions and their dependencies among business areas of the organizations. The identified services can be categorised based on their criticality. For instance, some services may deal with recurring business functions which are less critical than the services that are business and mission critical. organizations generally rely on business-critical services for managing day to day operation and mission critical ones with highest impact on the key business functions and potential for recovery.

After defining the services, next step is to identify the assets linked to these services. The assets include exist-

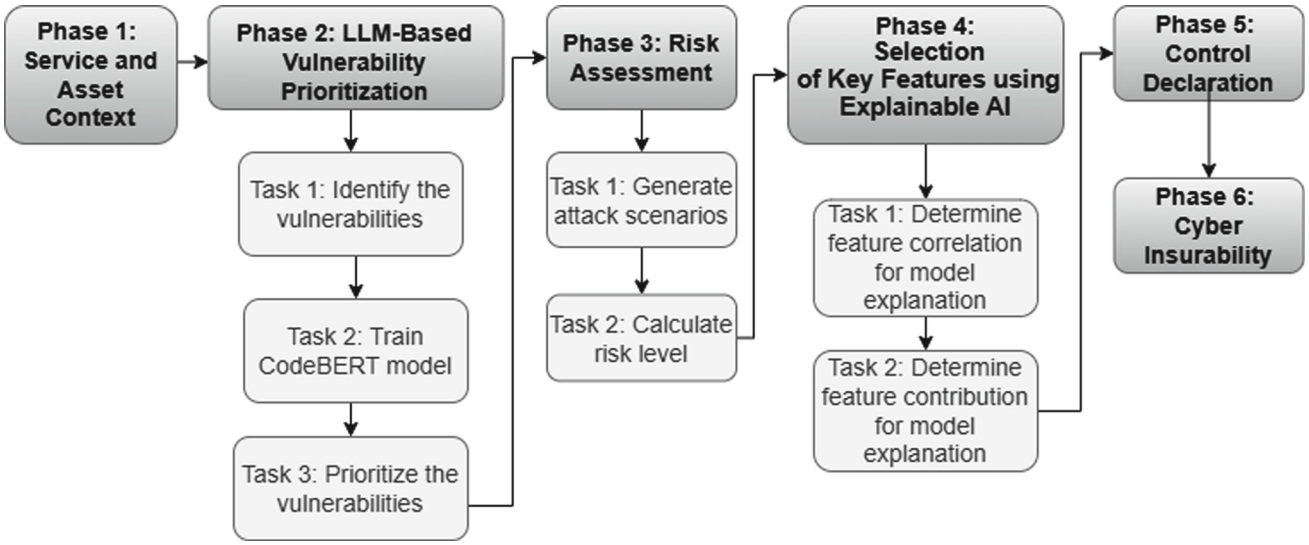

Fig. 2 Process of the proposed approach

ing ICT products such as hardware, software, and operating systems. This phase advocates to use the NIST Common Platform Enumeration (CPE) catalogue and its specific URIs to systematically categorize these known components [40]. It provides a structured naming scheme for IT products, enabling their linkage to common weaknesses and vulnerabilities. However, note that CPE doesn't cover all assets, particularly custom applications and software components within organizations built with open-source packages or represented in a Software Bill of Materials (SBOM). In this context, the Package URL (pURL) naming scheme is used to track the underlying libraries and packages. The asset discovery mainly follows the existing inventory of assets within the organization and maps the listed assets with the services. The CPE catalogue can also be used for a structured representation of the asset. Moreover, regarding custom applications, it is necessary to create a complete list of open source packages and other libraries as third party components. Each entry of the list needs a pURL for precise structured definition. Generally, there are several platforms available for this mapping such as synk or syft. Finally, this phase also focuses on defining security objectives and functional requirements for the identified assets so that assets can be mapped with necessary security controls. It follows EU common criteria based cybersecurity certification scheme (EUCC) for this purpose the outcome of this phase is also considered for the vulnerability prioritization [67].

5.2.2 Phase 2: LLM-Based Vulnerability Prioritisation

Once the services and related asset based CPEs and pURLs are identified, it is necessary to identify and prioritize the associated vulnerabilities. This phase focuses on using the capabilities of advanced language models for vulnerability detection, and prioritization against the identified assets. The primary objective is to improve the efficiency of vulnerability detection by incorporating CodeBERT, a transformer model trained on both programming and natural languages. The integration of AI in this phase significantly reduces manual overhead while enhancing scalability and precision in finding complex security vulnerabilities. The phase includes three tasks covering vulnerability detection and prioritization using LLM model capabilities.

Task 2.1: Identify the Vulnerability This task entails extracting confirmed vulnerability related information from the identified assets, specifically the CPEs and pURLs from phase 1. It investigates the existing security knowledge bases such as CVE and National Vulnerability Databse (NVD) for identifying vulnerability related information. However, as stated before, pURLs provide a standardized way to uniquely identify a software component. By capturing key aspects such as the package type, name, and version, a pURL enables accurate mapping between the component and publicly available vulnerabilities. Therefore, this task also investigates the existing vulnerability knowledge base such as GitHub Security Advisories and Open Source Vulnerabilities (OSV), to identify the documented vulnerabilities associate with specific software package. Although a pURL does not directly reference security weaknesses, it enables an indirect connection to the MITRE Common Weakness Enumeration (CWE) framework through its link with known vulnerabilities. Most publicly available vulnerabilities include one or more CWE classifications, which describe the underlying software weakness. Therefore, when a component is represented using a pURL and is matched with specific CVEs,

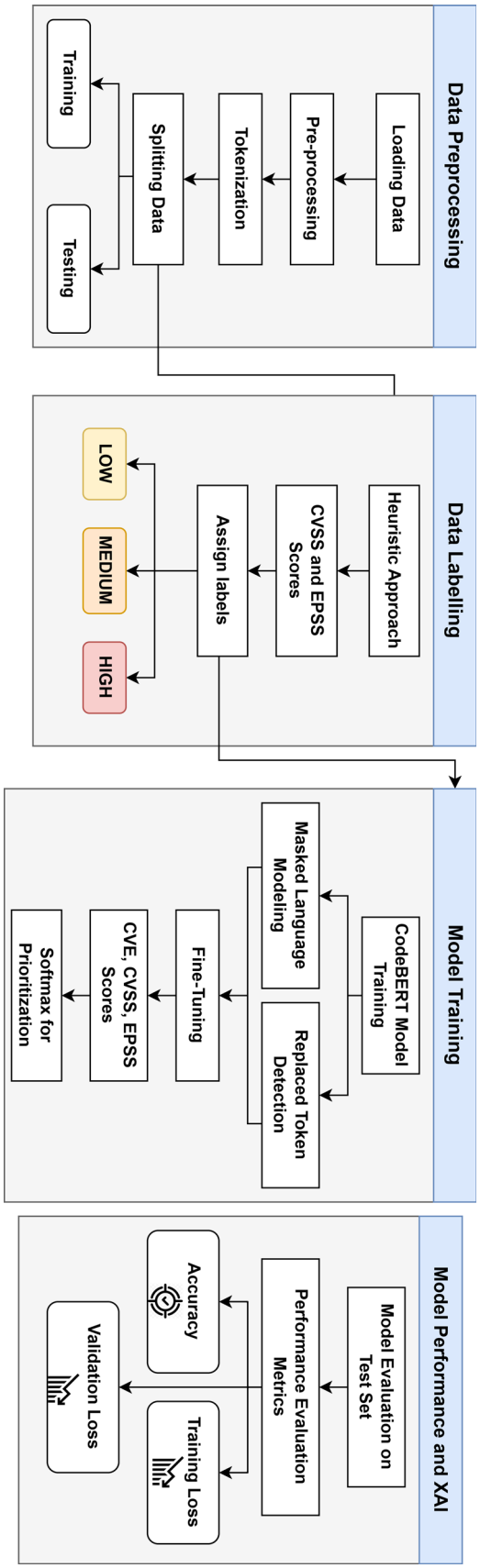

Fig. 3 LLM-Based Vulnerability Prioritization

OSV entries, or GitHub advisories, the related CWE categories can also be retrieved.

Task 2.2: Train the CodeBERT Model This task focuses on the training the chosen model by following key steps to clean, structure, and enrich the data for meaningful analysis. The dataset is used by this task to fine-tune CodeBERT, enabling it to learn domain-specific patterns and produce accurate vulnerability predictions. This task involves a series of steps, as illustrated in Figure 3, which are discussed below.

Step 2.2.1: Data Preprocessing The initial step of this task is data preprocessing which involves cleaning, formatting, and optimizing raw data with respect to machine learning tasks. This involves loading information from different sources like vulnerability databases, threat intel feeds, or internal enterprise logs. These sources often contain heterogeneous datasets which vary in structure and completeness. Also, a high-quality and diverse dataset can be selected for this step. After loading the data, proper pre-processing steps are taken to resolve issues such as missing values class imbalance, and out-of-structure text. Missing values pose a risk in distorting model performance and are addressed by mean imputation techniques [41]. Tokenization is performed on text data to convert the text into numerical representations suitable for transformer models like CodeBERT. Finally, the preprocessed data is separated into training and test sets by techniques like stratified sampling in order to have the same class distributions. This separation is crucial for evaluating the model's generalization ability during later steps.

Step 2.2.2: Data Labelling Following the preprocessing, next step is to label each instance of the refined dataset for effective supervised learning. A Heuristic approach is utilized to create labels that are based on rule-based logic and expert-defined thresholds, in contrast to solely relying on already annotated datasets which might be outdated or sparse [42]. This real-world and experience-based approach provides an efficient solution when the ground truth label is absence or incomplete. In this context, heuristics are used to generate labels by examining available features such as scores of vulnerabilities and exploitability with respect to well-defined rules based on knowledge of the cybersecurity domain. In our case, we have considered CVSS which provides an industry standard for the severity score, and EPSS which measures the likelihood of exploitation in the near timescale. These scores are combined to generate a composite risk score for each occurrence of the dataset. Based on this composite score, the vulnerability instances are ranked into LOW, MEDIUM, or HIGH priority levels.

Step 2.2.3: Model Training The focus of this step is to train the model that can effectively understand and prioritize vulnerabilities using the representation power of transformer-based architectures. As stated before, Code-

BERT is employed as a transformer-based model developed by Microsoft Research to handle tasks with both source code and natural language [43]. It improves upon BERT's architecture by learning from an extensive corpus of paired data consisting of natural language descriptions and their corresponding Python, Java, or JavaScript code snippets. It starts with byte pair encoding which splits text into tokens. Then integrates a question posed by a human and the corresponding code answer as one sequence, where the split is marked by [SEP] and starting with [CLS] as shown in Figure 4. During this process, the model generates rich embeddings for all tokens, thus enabling retrieval of the [CLS] token to serve as a key feature for simple classification tasks. The primary steps of pre-training CodeBERT are by using two self-supervised steps. Initially, the model performs tasks involving natural language and programming inputs through Masked Language Modelling (MLM) where random words are concealed within the text, and the model has to predict what they are [43]. This improves the effectiveness of the model in grasping grammar, structure, and meaning in both languages. The second task, inspired by ELECTRA, is Replaced Token Detection (RTD), where plausible stand-ins for genuine tokens are distributed throughout text, and the model learns to identify them [44]. RTD fortifies a model's capability of pinpointing minute shifts in semantics which is crucial when determining flaws or weaknesses in code snippets. When CodeBERT is taken out of pretraining context into a practical scenario like predicting vulnerabilities, then the model needs to be fine-tuning. Therefore, the model is trained on a labelled dataset containing CVE, CVSS, EPSS scores along with other frequently occurring attributes. It learns from these evidence-based labelled data and anticipated potential attacks. During this fine-tuning process, the outcome corresponding to a special token known as [CLS] is passed through a dense layer resulting in human readable probabilities after applying softmax function which converts raw score into clear probability to demonstrate how likely the specific vulnerability is to be exploited based raw data and exploitability features.

Moreover, the primary reason for preferring CodeBERT over other models stems from its distinct feature of processing natural language alongside source code at the same time. Unlike other language models which focus exclusively on text, CodeBERT learns from parallel datasets of code and natural language descriptions to address code-related tasks [43]. This allows it to grasp deeper meanings and grammatical structures, essential for evaluating potential security risks. For prediction tasks such as exploitation prediction, its architecture is particularly suited as it can read both structured code patterns and unstructured vulnerability descriptions with greater precision.

Step 2.2.4: Model Performance This final step involves evaluating the model performance after training is completed. This starts with model evaluation on the test set, where unseen data is used to establish the model's ability to generalise from the training data. Several performance measurements are utilized, including accuracy, precision, recall, F1-score, training loss, and validation loss. Accuracy represents the proportion of correct predictions by the model in terms of true positives and true negatives relative to the total predictions [45]. It gives an overall idea regarding the model performance but can be deceptive if the dataset is imbalanced. Precision, on the other hand, represents the proportion of true positives to all predicted positives and thus represents the quality of positive predictions [46]. A higher precision score indicates good detection and a low rate of false alarms. Recall is capacity to detect all positive instances pertinent and higher value means fewer missed detections [47]. Finally, the F1-score is a harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives [48]. The integration of these metrics provides a complete picture of model performance and success and justifying its deployment in the given domains. These checks not only determine model efficacy but also serve as pointers for further optimization and tuning.

Task 2.3: Prioritize the Vulnerabilities This final task of Phase 2 is vulnerability prioritization, where risks associated with specific assets are methodically ranked to aid in effective and informed risk mitigation. After evaluating the trained LLM model, the model outputs guide the prioritization process. Vulnerability instances are then filtered based on their relevance to the assets within a given system or service. The selected vulnerabilities are subsequently assessed and ranked using a key dataset features describing technical severity and probability of exploitation. This prioritization focuses mitigation and response efforts on the most exploitable and high priority vulnerabilities, ensuring appropriate resource utilization and overall security posture as shown in Table 2.

5.2.3 Phase 3: Dynamic Risk Assessment

This phase identifies and assesses the risks associated with assets and services. The prioritized vulnerabilities from the previous phase and overall impact are primarily used to calculate the risk level. This phase encompasses two tasks for risk assessment.

Task 3.1: Generate attack scenarios This task generates the possible attack scenarios based on the identified vulnerabilities that can be exploited against specific asset. These attack scenarios demonstrate how threat actors can exploit the identified vulnerabilities on related asset, leading to potential risk. In this context, the EPSS score from the previous phase and CVSS metrics are considered for formulating the

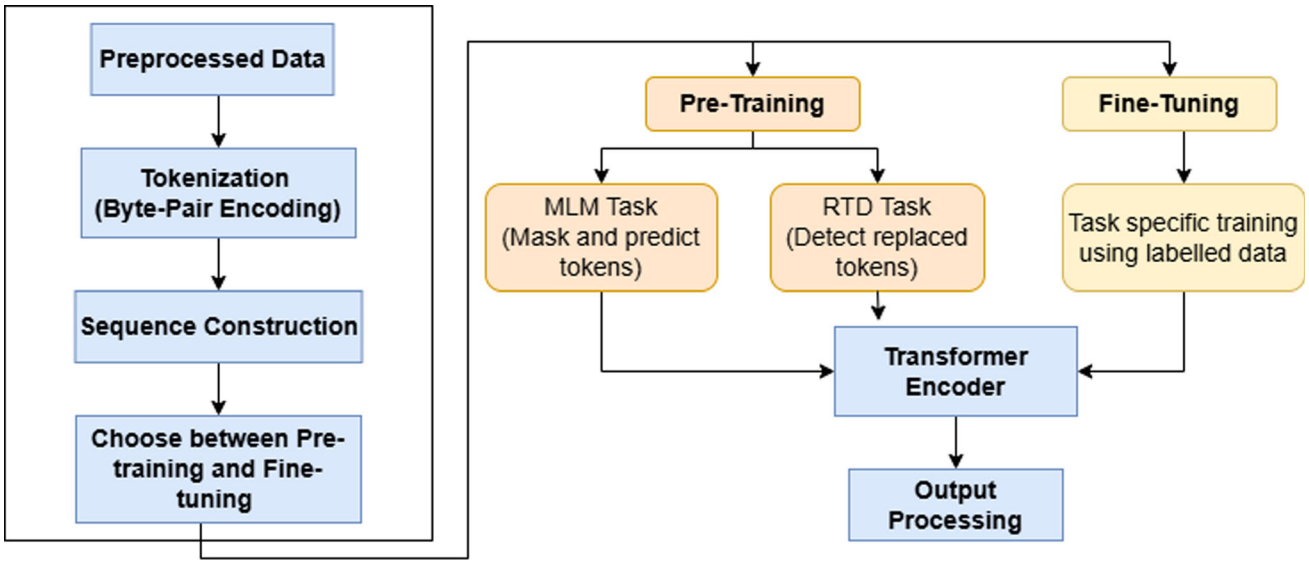

Fig. 4 Inner working of the CodeBERT model

Table 2 Vulnerability prioritization based on EPSS Score

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

attack scenario[49]. Each attack scenario may involve a single or multiple assets and related vulnerabilities. We follow the steps provided by [50] for the attack path generation. Specifically, the task initiates by considering the possible entry points threat actors use to exploit vulnerabilities related to specific assets. If dependencies exist among the assets, the attack can be propagated from the entry point asset to the target asset. This allows us to link vulnerabilities to the assets and determine the potential risk exposure that can impact both the assets and services.

Task 3.2: Calculate the Risk Level Once the attack scenarios is generated, it is necessary to calculate the associated risk level. The risk level is calculated based on the likelihood of the exploitation of the vulnerabilities and the potential impact. he likelihood of vulnerability exploitation is primarily determined by the identified vulnerability priority level. The impact measures the effect on the single or multiple assets that can be expected as the result of the successful exploitation of a vulnerability. The recorded CVSS CIA impact attributes of vulnerabilities are taken into consideration for the impact measurement. This is because CVSS is a standard method for calculating the global severity impact of an exploited vulnerability. These impact attributes relate to other organizational aspects; for instance, high data confidentiality impact from exploitation may result in organizational financial loss due to a data breach, legal fines, customer loss, and other factors. Similarly high impact on service availability can lead to service outage, ransomware or revenue loss. We define five impact levels, ranging from minimal to very significant to classify the impact as shown below:

- **Very Significant**: Major widespread damage to asset and/or infrastructure, complete disruption of operational continuity.
- **Significant**: Significant damage to critical asset and /or infrastructure, disruption of operational continuity.

– **Average**: Moderate damage to asset and/or infrastructure preventing, partial disruption of operational continuity.
– **Minor**: Considerable damage to asset and/or infrastructure preventing, minimal disruption of operational continuity.
– **Minimal**: Negligible or no significant damage to asset and/or infrastructure , minimum impact on operational continuity.

The risk level is calculated as follows:

$$\text{Risk} = V_L \times I_L \tag{4}$$

Where

– V_L is the likelihood of the exploitation of vulnerability that exist in the asset
– I_L is the impact level due to the exploitation (Table 3)

5.2.4 Phase 4:Selection of Security Control using Explainable AI

This phase aims to explain the model's decision-making for choosing security controls by employing XAI techniques. This ensures that the model predictions are transparent and interpretable, which is essential for sensitive domains such as vulnerability management where trust in system output is imperative for decision makers. In this phase, the relevant XAI components were selected based on the model architecture and nature of input features. These components facilitate understanding how the model processes inputs to reach conclusions, which helps in building confidence for its results. In this step, we consider two tasks for the XAI implementation for the chosen AI model.

Task 4.1: Determine Feature Correlation for Model Explanation This task focuses on explaining how input features cooperate with one another to influence the prediction. The goal is to uncover the interdependencies and contextual relationships between features that collectively contribute to the model's output. By understanding how features complement or counteract one another, we can identify patterns that may be difficult to spot with individual feature importance. For instance, an insignificantly predictive feature may become incredibly powerful when paired with another. This information is vital to the subsequent security control mapping stage, where features highly correlated may mean similar risk dimensions or similar mitigants. To achieve this, we apply the correlation heatmap, as explained in Section 4. While useful, it is important to note that correlation does not imply causation, thus, these results need to be supplemented with model-specific interpretability tools such as SHAP interaction values or integrated gradients to test assumptions.

Task 4.2: Determine Feature Contribution for Model Explanation This task seeks to determine the independent contribution of each input feature to the model prediction. Identifying which features contribute most to the output is essential for building transparency, explaining model behavior, and achieving insightful results from complex machine learning models. To achieve this, we utilize SHAP, an explanation technique based on cooperative game theory. SHAP calculates Shapley values for all features by assessing their marginal contribution across different sets of inputs, delivering a robust and fair measure of feature influence [51]. These values can be assigned locally–for one prediction–and globally–by summing over the data set. With the help of SHAP analysis, we can visualize feature importance, assess their positive or negative impact on predictions, as well as detect patterns indicating potential bias or over-fitting. This analysis allows us to examine between how and why the model comes to its decision while confirming that appropriate reasoning informs those decisions.

5.2.5 Phase 5: Control Declaration

Control declaration is a foundational part in cybersecurity risk management as it ensures the implementation of suitable and precise protective measures for the identified risks and enhances system resilience [52]. Our approach relies on XAI techniques aimed at justifying control selection to achieve evidence-based strategies. Tye feature contribution implements SHAP to determine individual feature significance in the model prediction. While, feature correlation uses attention mechanisms and layer-wise analysis to explain the relationships among features or tokens in the model input. The outcomes from both components are integrated to determine common influential features and tokens, which are then used to map the relevant security controls. Control selection follows the guidance provided by the NIST SP 800-53 framework, which ensures organizational compliance with relevant cybersecurity norms and industry standards [53]. The framework covers security and privacy area holistically using eighteen distinct families. Security and privacy controls are classified according to these families, encompassing operational, technical, and management perspectives.

5.2.6 Phase 6: Cyber Insurability

The decision to purchase cyber insurance stems from the need to cover the risk exposure gap remaining after implementing of security controls. While both preventive and detective controls mitigate cyber threats and their impacts, certain vulnerabilities, particularly those with high exploitability

Table 3 Risk Assessment Matrix

Likelihood	Impact				
	Minimal	**Minor**	**Average**	**Significant**	**Very Significant**
Risk Level					
Very High	Medium	High	Very High	Very High	Very High
High	Low	Medium	High	High	Very High
Medium	Low	Medium	Medium	High	Very High
Low	Very Low	Low	Medium	High	High
Very Low	Very Low	Very Low	Low	Medium	Medium

characteristics, continue to threaten invaluable assets [54]. The specific, remaining level of risk is referred as residual risk. For business organizations, this residual risk creates the demand for more layers of security to safeguard their assets through the use of insurance. Therefore, cyber insurance is a risk transfer strategy that provides financial security and operational assurance against a cyber attack. The following strategies are followed by this final phase for the cyber insurance decision making:

- **Asset and Vulnerability Profiling:** Insurance decisions take into account high-value assets (such as operating system, software, custom application). They also consider the history of attacks and published vulnerabilities (such as CVEs) against those assets. By understanding which assets have been impacted previously and identifying the most common security vulnerabilities, organizations and insurers can better determine the risk potential and consequently the appropriate coverage or controls required to protect those systems.
- **Explainability and Control Justification:** In providing justification for model predictions concerning selected controls, XAI effectively supports to justify the controls. This enables insurers together with other relevant decision-makers to appreciate why certain risks categorized as vulnerable were chosen and whether appropriate actions have been taken.
- **Residual Risk:** Once the technical, administrative, and physical controls have been implemented, it is necessary to investigate the existence of residual risks. Despite the implementation of security control, the existence of residual risk is often validated by the presence of key exploitation parameters associated with the confirmed vulnerability. Our approach advocates to consider these parameters as valid evidence that can pose exploitation of vulnerability to compromise an asset.
- **Premium Adjustment and Policy Structuring:** As previously stated, the extent of residual risk plays a significant role in determining the cost and design of a cyber insurance policy. The residual risk is determined by insurers upon analysis of an organization's existing security posture and the effectiveness of its preventive, detective, and response controls. A higher residual risk, particularly involving key systems or highly exploitable vulnerabilities, tends to result in higher premiums, more exclusions, and restricted coverage [55]. Organizations that can demonstrate a mature cybersecurity framework, are often categorized as lower risk. They are likely to have more favourable policy terms, e.g., lower premiums, broader coverage, and fewer exclusions [56]. In some cases, insurers can even offer incentives or discounts for adhering to best practices or for taking further risk mitigation steps during policy term.

6 Evaluation: Industrial Use Case and Experiment

We have considered an industrial use case to evaluate the proposed approach with insurability decision as shown in Figure 5. The evaluation also includes an experiment using widely used dataset to determine the vulnerability exploitation predication and link it with the use case scenario. As shown in Figure 5, the use case is initially investigated to understand the context, including the possible assets along with related CPE and pURLs. Once the assets are identified from the pilot case, then the identifiers (CPEs and pURLs) are mapped with the prioritized vulnerabilities based on the result from the experiment part. The experiment part utilizes the CVEjoin dataset [57] to identify the prioritized vulnerabilities, maps them further with the identified assets from the use case context, and calculates the risk level. The main aims of the evaluation are:

- To demonstrate the applicability of the proposed approach on a real industrial use case scenario in terms of understanding vulnerabilities and risks associated with the assets and their mitigations with cyber insurability decision.
- To prioritize the vulnerability based on the level of exploitation using CodeBERT-based LLM model, trained

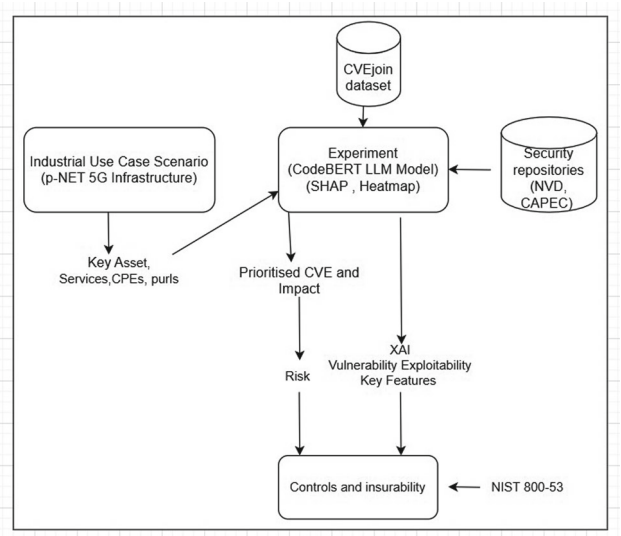

Fig. 5 Implementation of Use Case Scenario

using a dataset that contains relevant features and instances of real-world vulnerabilities.
- To evaluate the predictive performance and accuracy of the CodeBERT-based model in classifying the likelihood of vulnerability exploitation, thereby ensuring its reliability for supporting risk assessment and insurability decisions.
- To operationalise the XAI practice for informed decision making regarding choosing the right security control and cyber insurability.

6.1 Industrial use case scenario

As stated before, the industrial use case scenario is considered as a part of the evaluation to demonstrate the applicability of the approach. It facilitates to identify the assets and pURLs of the organization which further links with the vulnerabilities and risks using the proposed approach so that suitable control action can be selected along with insurability decision. The scenario is based on p-NET which is a competence centre providing services related to digitalization consulting, testing and integration, and digital skill development.

Network as a Service (NaaS) model and Infrastructure: p-NET provides advanced 5G infrastructure by following NaaS model with dynamic testing environment that facilitated testing of next generation network capabilities with customised deployment. It is a cloud-based infrastructure with distributed test beds to provide capabilities for 5G/6G Cloud-to-Edge deployments, non-public network configuration along with various integration and scalability related testing services using end users' devices. The two key services are added below as a part of scenario:

- *Performance Monitoring and Network Data Analytics:* This service is based on real time analytics and optimisation using NITRO a VIAVI's Network Integrated test environment. Products are tested in various points of the 5G network (core network, transmission layer, end user devices) and compared with selected network Key Performance Indicators (KPIs) being commercially available using this service. The platform collects metrics via the use of HW and SW Viavi network probes placed in the 5G network.
- *Network Optimisation:* This service collects information from user equipment end-devices and the radio access network and identifies actions that can improve the network performance through radio resource allocation. Specifically, radio resource management algorithms account for key performance indicators of the radio access network, e.g., coverage level and data rate perceived at end-users. Radio resource allocation optimisation is then performed, leveraging time, frequency, power, and/or spatial degrees of freedom available at the network cells (gNBs). Examples may include time-frequency resource block allocations to different users, multi-antenna beamforming and spatial multiplexing configuration, as well as dynamic power control.

The advanced 5G infrastructure of p-NET includes various software and hardware products inside the core network to support these capabilities and some of these products are manufactured and managed by third parties. In this context, vulnerabilities that arise from any of these products can pose any potential threats to the entire 5G infrastructure. It is critical for the p-NET to be able to ensure the highest level of cybersecurity for any kind of external services used inside their infrastructure. Therefore, it is necessary to identify the possible vulnerabilities and their probability of exploiting relevant with the assets of advanced 5G infrastructure so that suitable mitigation strategy can be recommended by using the proposed approach.

Assets and corresponding CPEs and pURLs for the advanced 5G infrastructure: We have identified several key assets that are relevant with the 5G infrastructure and their services. The identified assets are mainly individual hardware and software components rather than custom applications utilized by the NaaS (Network-as-a-Service). Therefore, no pURL (Package URL) identifiers were included in the evaluation. The assets and related CPEs are shown in Table 4.

6.2 Experiment

6.2.1 Dataset Description

The CVEJoin dataset [57] combines various vulnerability records from public repositories, with a special focus

Table 4 Identified key asset of the 5G Network

Asset Name	Product & Description	CPEs	Security Objectives	Security Functional Requirements (SFRs)
Operating System	**Oracle Linux Enterprise**; Server-grade Linux distribution	cpe:2.3:o:oracle: linux:8:*:*:*:*:*:*:*; cpe:2.3:o:oracle: linux:7:*:*:*:*:*:*:*	O.IDENTIFICATION_ AUTHENTICATION; O.ACCESS_CONTROL; O.AUDIT; O.INTEGRITY	FIA_UID.2 (User identification before action); FIA_UAU.2 (User authentication before action); FDP_ACF.1 (Security attribute based access control); FAU_GEN.1 (Audit data generation)
Application Server	**Apache Tomcat**; Java-based web application server	cpe:2.3:a:apache: tomcat:9.0: *:*:*:*:*:*:*; cpe:2.3:a:apache: tomcat:8.5: *:*:*:*:*:*:*	O.SECURE_ COMMUNICATIONS; O.ACCESS_CONTROL; O.AUTHENTICATION; O.SESSION_ MANAGEMENT	FCS_HTTPS.1 (HTTPS protocol); FDP_ACC.2 (Complete access control); FIA_UAU.1 (Timing of authentication); FTP_ITC.1 (Inter-TSF trusted channel)
DBMS Application	**MySQL**; Relational database management system	cpe:2.3:a:oracle: mysql:8.0:*:*:*:*:*:*:*; cpe:2.3:a:mysql: mysql:5.7:*:*:*:*:*:*:*	O.DATABASE_ACCESS_ CONTROL; O.DATA_INTEGRITY; O.AUTHENTICATION; O.AUDIT	FDP_ACC.1 (Subset access control); FDP_ACF.1 (Security attribute based access control); FIA_ATD.1 (User attribute definition); FAU_SAR.1 (Audit review)
Time Scale Database	**PostgreSQL on Analytics VM**; Advanced analytical database	cpe:2.3:a:postgresql: postgresql:13:*:*:*:*:*:*:*; cpe:2.3:a:postgresql: postgresql:12:*:*:*:*:*:*:*	O.DATA_INTEGRITY; O.BACKUP_RECOVERY; O.ACCESS_CONTROL; O.CRYPTOGRAPHIC_ SUPPORT	FDP_RIP.1 (Subset residual information protection); FCS_CKM.1 (Cryptographic key management); FDP_ACC.2 (Complete access control)
Content Service Switch	**CISCO 11000**; High-performance network hardware device	cpe:2.3:h:cisco: content_services_ switch_11000: *:*:*:*:*:*:*:*	O.NETWORK_ACCESS_ CONTROL; O.TRAFFIC_ FILTERING; O.SECURE_ MANAGEMENT; O.INTEGRITY	FFW_RUL.1 (Simple firewall rules); FPF_RUL.1 (Packet filtering rules); FMT_SMF.1 (Specification of management functions); FPT_STM.1 (Reliable time stamps)

on Common Vulnerabilities and Exposures (CVEs). This in turn helps researchers studying or enhancing vulnerability detection, risk assessment, exploit prediction, and AI-based cybersecurity solutions. By integrating data from the National Vulnerability Database, EPSS system, threat intelligence feeds, and social indicators, CVEJoin provides a comprehensive contextual insight for each vulnerability. This aggregation allows researchers to not only understand the technical aspects of vulnerabilities, but also the frequency of their exploitation in the wild and their relevance in contemporary attacks. Every instance of the dataset aligns with one CVE and contains numerous traits ranging from severity scores to social metrics. By connecting static details like CVSS scores to the dynamic signals (such as CTI counts or mentions on social media) the dataset supports richer model training. This blend can enhance the interpretability when implementing explainable AI models that require both technical and contextual information to justify predictions.

The dataset contains 39 features and approximately over 200,000 and is considered due to its diversity. It covers operating systems and software, handles multiple severity scoring models, and incorporates contextual threat intelligence data. This makes it a useful source for the analysis of the lifecycle of the vulnerabilities and building good predictive models that consider technical severity as well as real-world exploitability. The key features of the CVEjoin dataset are:

- **EPSS Scores:** Standardized metrics that assess technical severity and likelihood of exploitation.
- **CWE Categories:** Helps classify vulnerabilities by weakness type (e.g., CWE-20 for input validation).
- **Exploit and Threat Intelligence Data:** Includes exploit counts and CTI signals like social media audience and Google interest.
- **Vendor and Product Diversity:** Represents vulnerabilities from a wide spectrum of vendors and product families.
- **Temporal and Update Signals:** Tracks publication dates, updates, and whether patches are available (update_available column).

6.2.2 Experimental Setup

We designed the experiment to efficiently train and test a transformer model on software vulnerability records. The experiment was carried out in a Google Colab notebook powered by an NVIDIA A100 GPU, imparting plenty of raw compute and parallel power. The operating system utilized for this setup was Ubuntu 20.04, which supports a wide range of cutting-edge software libraries and tools. The code was built using Python 3.10, a widely used programming language in machine learning and data science operations. The deep-learning tasks relied on PyTorch 2.0, prized for its adaptive graph-building and solid support for NVIDIA GPUs.

Hardware configuration included an NVIDIA A100 GPU, which facilitated high-performance parallel computation, ideal for cost-effective training of transformer models. The GPU configuration significantly optimized the model training and validation processes, especially when working with long sequence input and increased batch sizes. The system further provided sufficient memory, typically around 32 GB RAM to seamlessly process data loading and preprocessing. We chose CodeBERT-a pretrained code-focused transformer-and loaded it directly from the Hugging Face hub [58]. Following the standard practice in AI, we split the full dataset into an 80-20 training-test pair [59]. This allows the model to be trained on a major part while testing on a reserved set for unbiased evaluation, thereby preventing over-fitting.

6.2.3 Data Preprocessing

The first step of the experiment is data preprocessing, which involves a series of pre-processing steps to convert the raw JSON entries into a structured format that is suitable for efficient and accurate model performance. The CVEjoin dataset, being descriptive in nature, contained null values for several significant features such as EPSS and CVSS scores. The null or missing values need to be addressed prior to training to avoid compromising the model performance. To ensure data quality without adding noise, rows containing missing values for these columns were removed using the dropna() function. Moreover, we utilized the CodeBertTokenizer from the Hugging Face's Transformer library. Built to match the BERT-style architecture, this tokenizer converts plain text into the numbered tokens the model expects. Each vulnerability summary was tokenized, then truncated and padded to a cap of 128 tokens. This ensures uniform input size and prevents lengthy entries from slowing training or overflowing memory. The cleaned tokens and their labels are then packed into a custom VulnDataset class that is compatible with PyTorch's DataLoader. This setup handles batches of tokens and label tensors smoothly, during training and testing.

6.2.4 Data Labelling

After the preprocessing, data is then heuristically labelled, allowing the model to learn appropriately and make precise predictions in downstream tasks. This process supports efficient classification where there are no public or manually provided labels. For the CVEjoin dataset, heuristic rules were formulated based on the distribution and severity of EPSS values, yielding in the following classification criteria:

– The Low level is made up of vulnerabilities that have an EPSS less than 0.3, i.e., not likely to be exploited.
– The Medium level is made up of vulnerabilities with an EPSS of 0.4-0.7, i.e., moderate likelihood of being exploited.
– The Highest level is made up of those whose EPSS is more than 0.8, i.e., highest level of vulnerabilities that need immediate attention.

These thresholds were carefully selected using exploratory data analysis to ensure agreement with real risk levels and leverage the importance of probabilities. These labelled instances are the foundation for training supervised models, allowing the model to learn explicit patterns between risk levels and making informed predictions on unseen data during testing.

6.2.5 Model training

After data labelling, the model was trained using a transformer-based architecture, i.e., the CodeBERT sequence classification model. The pre-trained model was fine-tuned using the vulnerability dataset to output a list of prioritized vulnerabilities including their CVE and EPSS. The dataset was first converted into tokenized input forms and then added into the training pipeline in 16-sample mini-batches. While training, the model was shifted to an environment based on a GPU for utilizing the strength of quick computation and parallelism. The model was trained for four epochs, wherein in each epoch the model learned from the entire training dataset. The parameters of the model were updated using an AdamW optimizer, and the cross-entropy loss function measured how much the predictions varied from the actual labels, as shown in Table 5. The selected parameters conformed to best practice when tuning transformer models like CodeBERT on binary classification tasks. The configuration optimizes training performance, model stability, and generalizability in order to learn effectively without overfitting [66]. A balanced learning rate, sufficient batch size, and optimal number of epochs were some of the parameters that were selected to accommodate the computational budget and the nature of the vulnerability dataset. Additionally, at each training step, the model received batches of input IDs, attention masks, and labels. It predicted, computed the loss, and updated the weights via back propagation. The mean training loss across each epoch was monitored to measure convergence and confirm the model's effective learning.

Table 5 Parameters for model training

Parameters	Value
Pre-trained Model	CodeBERT (microsoft/codebert-base)
Number of Output Labels	2
Tokenizer	CodeBERT Tokenizer (BERT-based)
Max Token Length	128
Batch Size	16
Number of Epochs	4
Optimizer	AdamW
Learning Rate	2e-5
Loss Function	CrossEntropyLoss

6.2.6 Performance Evaluation

The trained model is evaluated using widely accepted performance metrics. To ensure an equal and unbiased evaluation to each vulnerability class, the test data contains 500 samples from each class, least probable, medium probable, and highly probable vulnerabilities. The training results show that the model consistently achieved high accuracy on both the training and validation datasets with 96.9% and 96.7% respectively, showing strong generalization and not overfitting, as shown in Figure 6. The loss values consistently decreased across epochs, showing that the model was still learning and refining its understanding of the data. Precision also rose significantly from 0.59 in the first epoch to over 0.75 by the final epoch, illustrating the model's growing ability to correctly identify positive cases. Although recall was generally very low, it followed a gradual increasing trend and finished at 0.0889 in epoch 5. The increase in F1-score, from 0.11 to 0.16, suggests ongoing but consistent improvement in precision-recall trade-off, especially in detecting the minority class.

6.2.7 Vulnerability Prioritization

After the model's performance, we subsequently ranked the identified vulnerabilities in order of their respective EPSS scores. EPSS is a value-added contribution to the model output by including a pragmatic, risk-based perspective for vulnerability ranking based on exploit potential in the real world. From the results in Table 6, the top of the list is CVE-2014-0160 (Heartbleed) with an EPSS score of 0.96076, which showcases its severity, ease of exploitation, and its prevalence among all users of OpenSSL. Similarly, CVE-2018-7600 (Drupalgeddon2) also tops the list with a score of 0.96053 due to the possibility of unauthenticated remote code execution on widely used Drupal websites. CVE-2014-0160 and CVE-2018-7600 were among the most critical vulnerabilities with scores greater than 0.96, indicating a nearly hundred percent rate of exploitability. Following behind were the Apache Tomcat vulnerabilities CVE-2017-12617 and CVE-2019-0232, both with more than a score

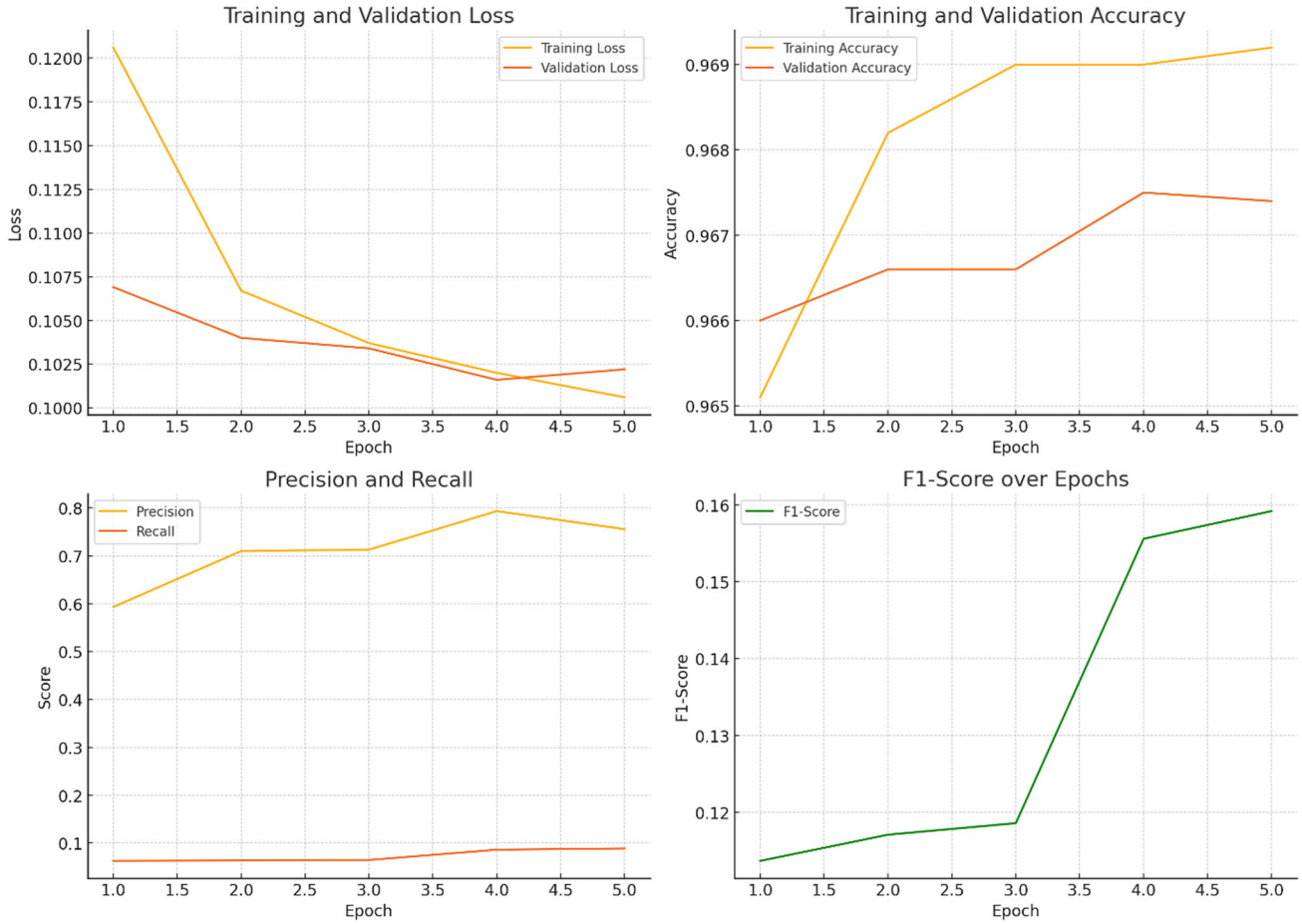

Fig. 6 Performance evaluation of the model over 5 epochs

of 0.94, reiterating their severity as exposed application servers. Additionally, some of the MySQL-related vulnerabilities (such as CVE-2022-22963, CVE-2018-10933, and CVE-2012-5613) reflected high EPSS scores ranging from 0.88 to 0.94, which necessitated their priority because they impacted widely deployed database systems. On the lower end, CVE-2011-3310, which was related to Cisco devices, reflected an EPSS score of only 0.21, which indicated a much lower likelihood of exploitation and, therefore, a lower priority during remediation planning.

6.2.8 XAI results

The XAI phase enhances model explainability, allowing all stakeholders and end-users to understand how the decisions are derived. To accomplish this, we utilized two complementary methods: feature contribution and feature correlation. Specifically, SHAP is implemented to assess the effect of each feature on the model's output. This was beneficial in identifying key features that shaped the model's predictions globally and locally. Moreover, for feature correlation, we used heatmap visualizations which are powerful tools for interpreting relationships among features. These visual explanations enable stakeholders to get insights into relationships and dependencies among various inputs, which enhances trust in model's logic.

– **Feature Contributions:** The SHAP feature importance plot shown in Figure 7 provides a global explanation of how each feature contributes to the model's predictions on the entire dataset. The x-axis indicates the mean absolute SHAP value, which reflects each feature's average impact on the model's output in terms of its effect (either positively or negatively). The feature with the highest contribution is exploitability_score. In a cybersecurity context, this aligns with intuition because vulnerabilities that are easier to exploit need to be prioritized for implementing appropriate mitigations. Following closely is epss which shows that the model leans heavily on signals of exploitability in predicting outcomes. confidentiality_impact also stands out as a leading contributor, demonstrating the value of

Table 6 CVE vulnerabilities with EPSS scores

CVE	CPE	Product/ Vendor	EPSS score
CVE-2014- 0160	cpe:2.3:o:linux: linux_kernel: *:*:*:*:*:*:*:* cpe:2.3:o:redhat: linux:6.2:*:alpha: *:*:*:*:*	Linux	0.96076
CVE-2018- 7600	cpe:2.3:o:redhat: linux:6.2:*:sparc: *:*:*:*:*	Linux	0.96053
CVE-2017- 12617	cpe:2.3:a:apache: tomcat:4.1.20: *:*:*:*:*:*:* cpe:2.3:a:apache_ tomcat:apache_ tomcat:5.0.16: *:*:*:*	Tomcat	0.94973
CVE-2019- 0232	cpe:2.3:a:apache: tomcat:4.1.20: *:*:*:*:*:*:* cpe:2.3:a:apache_ tomcat:apache_ tomcat:5.0.16: *:*:*:*	Tomcat	0.94954
CVE-2022- 22963	cpe:2.3:a:oracle: mysql:8.0: *:*:*:*:*:*:*	MySQL	0.94581
CVE-2018- 10933	cpe:2.3:a:mysql: maxdb:7.5.00.16: *:*:*:*:*:*:*	MySQL	0.92171
CVE-2019- 0230	cpe:2.3:a:apache: struts:*:*:*:*:*:*:*:* cpe:2.3:a:oracle: mysql_enterprise_ monitor:*:*:*:*:*:*:*:*	MySQL	0.89451
CVE-2019- 9193	cpe:2.3:a:postgresql: postgresql:8.3.2: *:*:*:*:*:*:* cpe:2.3:a:postgresql: postgresql:8.4.1: *:*:*:*:*:*:*	PostgreSQL	0.88202
CVE-2012- 5613	cpe:2.3:a:oracle: mysql:5.5.19: *:*:*:*:*:*:*	MySQL	0.88189
CVE-2008- 3257	cpe:2.3:o:oracle: linux:8:-: *:*:*:*:*:*	Apache	0.78095
CVE-2011- 3310	cpe:2.3:h:cisco: content_services_ switch_11000: *:*:*:*:*:*:*:*	Cisco	0.61344

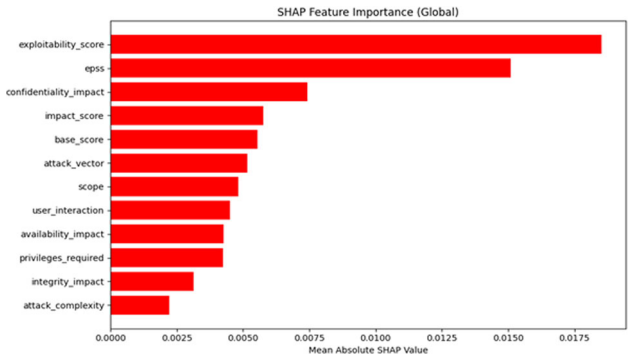

Fig. 7 Global SHAP Result

data sensitivity and the severity of a breach. Other key features are impact_score, base_score, and attack_ vector, all of which are traditional CVSS elements that describe the technical severity and exploitable complexity of the vulnerability. Lower-priority features such as attack_complexity, integrity_impact, and privileges _required had less overall impact on the model in a global setting. While still relevant, their overall contribution to the model's findings was relatively smaller. In conclusion, this global SHAP analysis indicates that the model places highest significance on real-world exploitability and technical severity when determining whether a vulnerability deserves immediate attention.

The SHAP result in Figure 8 graphically displays each feature contribution towards the model's prediction for a particular instance of the dataset. Exploitability_score is the most significant feature, with the highest positive SHAP value. It indicates that the model positively associated this feature with a higher likelihood of the vulnerability being exploitable. This suggests that the more readily the vulnerability can be exploited, the greater the likelihood the model will flag it as critical. In contrast, EPSS contributes strongly negatively to prediction. That is, in this specific instance, the EPSS rating diminished the model's ability in flagging the vulnerability as one that can be exploited. This might imply that though technically the vulnerability is an easy one to exploit, it is not widely exploited in the real world, hence the model is downgrading it slightly. Other features like attack_complexity were also negatively affecting the model, showing that the greater complexity in performing the attack, diminishes the urgency within the model's decision-making. The opposite is true for confidentiality_impact, attack_vector, and impact_score, although their contributions were not as significant. These features show that if the vulnerability is impacting confidentiality, remotely accessible, or has a significant overall impact, the model will be more inclined to flag it as critical. The SHAP explanation in this instance provides a comprehensible view of how the model is balancing technical severity with real-world exploitability signals. It portrays the model's ability to move beyound raw severity scores in isolation to consider many factors that play a role in vulnerability prioritization within a real-world context.

Feature Correlation Figure 9 demonstrates the correlation heatmap, illustrating how specific features used in model training relate to each other. To distinguish between positive and negative correlations, the heatmap uses gradient colours, where positive contributions are shown in warmer tones and negative contributions in cooler shades. One of the strongest

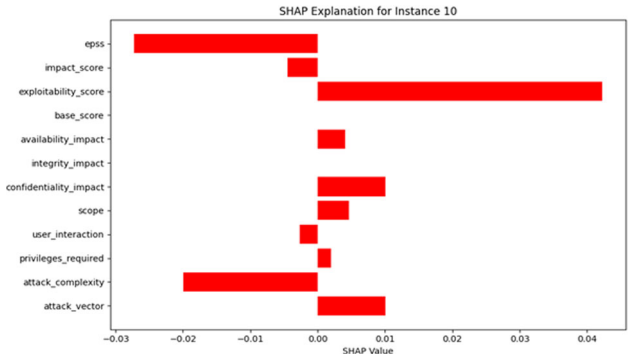

Fig. 8 Local SHAP Result

patterns observed is the strong positive correlation between impact_score and base_score (0.69), which could be due to the fact that they are derived from the CVSS scoring system. Similarly, exploitability_score has a moderate positive relationship with both impact_score and epss, suggesting that these variables are likely to co-vary in vulnerability data and perhaps prove useful together as inputs for the estimation of exploit potential. Moreover, there is a strong relationship among the features that are associated with the target update_available. The strongest association is seen with privileges_required _LOW (0.09), indicating that no single feature is highly predictive of the existence of an update. This suggests that much more complex interaction rules might determine the target, underlining an important need to use machine learning for non-linear dependencies for accurate prediction. The one-hot encoded variables mirror expected inverse relationships. For instance, attack_ vector _LOCAL and attack_vector _NETWORK strongly negatively correlate at -0.91 because these values are exclusive of each other. The same is seen across the different privilege classes, where privileges_required_LOW and privileges_required _NONE correlate at -0.79. The result confirms proper encoding and shows how differential categorization is mirrored in the correlation matrix.

6.2.9 Risk Calculation

At this stage, the risk level is calculated for each prioritized vulnerability, based on the likelihood of exploitation and the associated impact. The impact is estimated by analysing the consequences on single or multiple assets resulting from a successful exploitation of specific vulnerability. It is measured with the aid of recorded CVSS CIA impact attributes of recorded vulnerabilities. For instance, CVE-2018-7600 associated with the Linux product is recorded as critical with a high value for confidentiality, integrity and availability. Therefore, the impact is ranked as very significant and with a very high likelihood of exploitation, the risk is calculated as very high. Similarly, CVE-2019-9193 is ranked as significant impact associated with PostgreSQL asset and resulting in an overall risk level of High. A high impact on confidentiality may reflect the potential for a data breach or legal fines. The Table 7 presents the risk level for the identified assets.

6.2.10 Control declaration

This control declaration phase identifies the potential controls to mitigate the risks associated with the identified asset from the use case scenario. It follows the experiment results from the previous phases so that prioritized vulnerabilities associated with the risks can be taken into consideration for the control declaration. In this context, the integration of XAI enables a deeper understanding of the model's decision-making process by identifying key features that strongly correlate with the potential vulnerability exploitation. Our approach considers these features as risk factors, which are the main causes allowing the vulnerabilities to be exploited for a specific risk. The most influential features uncovered through XAI analysis included epss, impact score, exploitability _score, attack _vector,user_ interaction, privileges required and impact on confidentiality, integrity and availability. Note that impact score as a part of base metrics measures the possible impact due to potential consequence of successful exploitation. This score considers impact on vulnerable and subsequent system's confidentiality, integrity and availability. We considered the impact score as a part of one of the key features rather than considering individual CIA impact. These features provided insights into the risk posture of each vulnerability, supporting not only prioritization but also strategic mitigation planning. To translate these insights into actionable cybersecurity practices, we mapped the identified features to relevant NIST SP 800-53 security controls as shown in Table 8. This mapping allowed us to determine the appropriate technical and organizational controls required to reduce the risk associated with each vulnerability, ensuring that mitigation efforts align with established security frameworks.

6.2.11 Insurability

This final phase aims to support informed insurability decision making as part of a risk mitigation strategy. The proposed approach identifies and justifies the necessary controls to mitigate the exploitation of vulnerabilities associated with the risks and assets. The adoption of XAI further provides the necessary justification for security control selection. This reduces information asymmetry,supporting in the understanding of the organizational security posture and insurability decisions. Table 8 identifies key controls which p-NET requires to be implemented to tackle the exploitability features associated with the vulnerabilities and assets. However, despite the implemented security controls and complex

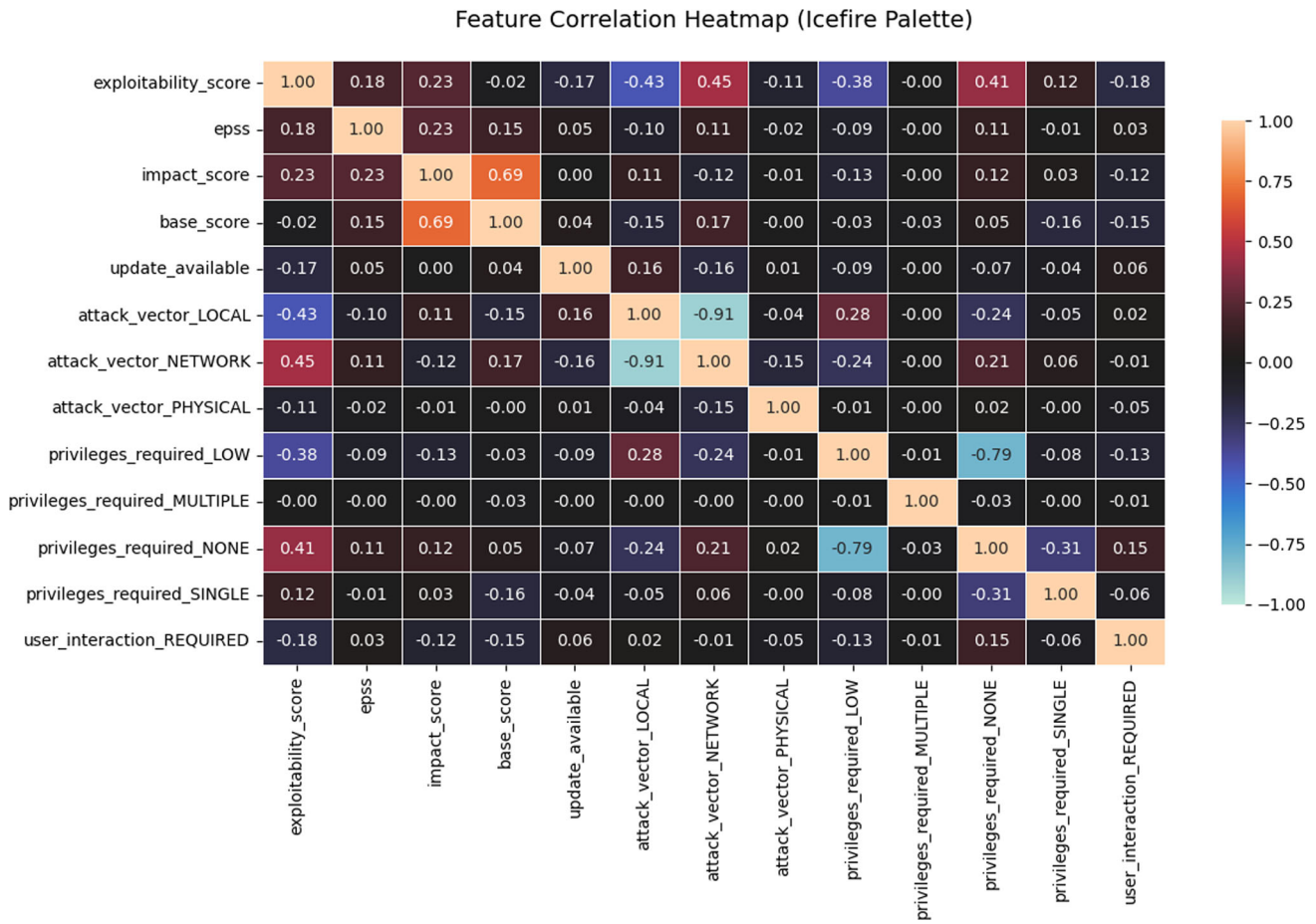

Fig. 9 Feature Correlation Heatmap

nature of the vulnerability exploitation features, the residual risk level could be higher for some features and their associated assets. This brings the necessity of additional mitigation strategy for ensuring the security and resilience of p-NET. The following vulnerability exploitation features are taken into consideration for determining the additional security mitigation strategy. Note that we do not focus on exploitability and base score due to their dependencies with other features.

- **EPSS:** This feature defines the potential likelihood of successful exploitation of a vulnerability and relies on comprehensive risk and vulnerability management practice to reduce the probability of exploitation. The calculation of EPSS is data driven and dynamic based on data available from various security repositories and threat feeds. Generally, AI and hybrid models are utilized to determine the EPSS. There are several assets of p-NET exhibits very high EPSS score such as Linux with 96%, Tomcat with 94%, and MySql with 94%, which are the key assets for the p-NET 5G infrastructure. Therefore, despite the list of control declared previous phase, additional mitigation strategy is required to reduce the risk related to the high EPSS score
- **Attack vector:** This feature relies on four distinct means, i.e., network, adjacent, local and physical for successful exploitation. The declared controls are mainly focused on Access control (AC), Physical and Environmental Protection (PE), System and Communications Protection (SC), Assessment, Authorization, and Monitoring (CA) classes. The related p-NET products for attack vector are linux OS and with CVE-2018-7600 and PostgreSQL with CVE-2019-9193 and both include network attack vector. Therefore, residual risk level could be higher for possible exploitation for these assets regardless of implementation of the control.
- **User interaction:** This feature determines the necessity of the user interaction for exploitation of a vulnerability with two possible options, i.e., none and required. Therefore, no user interaction is required for exploitation in case of none option and vice versa. The two related p-NET products are MySQL with CVE-2022-22963 and tomcat with CVE-2019-0232 and both require NONE option for user interaction. Therefore, vulnerabil-

Table 7 Assigning risk level to the prioritized vulnerabilities

CVE ID	Product/ Vendor	Likelihood of Exploitation	Impact	Risk Level
CVE-2014-0160	Linux	Very High (0.96076)	Significant	Very High
CVE-2018-7600	Linux	Very High (0.96053)	Very Significant	Very High
CVE-2017-12617	Tomcat	Very High (0.94973)	Significant	Very High
CVE-2019-0232	Tomcat	Very High (0.94954)	Significant	Very High
CVE-2022-22963	MySQL	Very High (0.94581)	Very Significant	Very High
CVE-2018-10933	MySQL	Very High (0.92171)	Very Significant	Very High
CVE-2019-0230	MySQL	High (0.89451)	Very Significant	Very High
CVE-2019-9193	PostgreSQL	High (0.88202)	Significant	High
CVE-2012-5613	MySQL	High (0.88189)	Average	High
CVE-2008-3257	Apache	High (0.78095)	Very Significant	Very High

ity exploitation for these products could be higher due to the no user interaction and this implies higher residual risk regardless of controls.

– **Privileges required:** This feature indicates the level of privilege, i.e,, none, low and high, attackers required to exploit specific vulnerabilities. With regard to p-NET, MySql with CVE-2018-10933 includes no required privilege, which implies exploitation of the vulnerability without any privilege or as a general user.

The key features and their potential causes for residual risk are discussed above. The other inherent causes include such as correct implementation and regular update of controls and attack history of p-NET. We have recommended that p-NET should acquire cyber insurance besides existing implemented controls. The justification of controls and residual risk allows for the adjustment of premiums and the determination of exclusions of the insurance policy.

7 Discussion

This section examines the contributions and implications of the proposed approach across four key dimensions. We first discuss how the operationalisation of explainable AI enhances security decision-making and supports cyber insurance decisions. We then position our work within the existing vulnerability detection landscape, comparing our approach with recent studies and acknowledging implementation challenges. We explore the broader practical implications for organizations, including operational resilience, governance alignment, and knowledge democratization. Finally, limitations and future directions for this work is presented.

7.1 XAI-Informed Security Decision-Making

This study operationalises XAI practices to ensure that decisions made by the CodeBERT-based LLM model are transparent, interpretable, and understandable to diverse user groups, including security analysts, auditors, and decision-makers. Accurate decision-making in mitigating cybersecurity threats is critical to guarantee organizational security and resilience. Our solution incorporates both feature correlation (heatmaps) and feature contribution (SHAP) techniques, providing insights into how individual features contribute to model predictions while exposing interactions among different features that contribute to vulnerability exploitation in meaningful and systematic ways. The evaluation of the proposed model integrates a real-world industrial use case–the p-NET 5G infrastructure–with experimental analysis using a large vulnerability dataset. The results demonstrated accurate identification of high-risk, highly exploitable vulnerabilities for assets in the use case scenario. For instance, CVE-2014-0160 (Heartbleed) and CVE-2018-7600 (Dru-

Table 8 Control declaration mapping XAI features to NIST security controls

Key Features from XAI	CVE and Products	Declared Security Requirements	Key Controls
exploitability_score (Measure the ease of vulnerability exploitation)	CVE-2014-0160 (Linux) CVE-2018-7600 (Linux) CVE-2022-22963 (MySQL)	FIA_UID.2 (User identification before action); FIA_UAU.2 (User authentication before action); FDP_SDI.2 (Stored data integrity monitoring); FCS_CKM.1 (Cryptographic key management)	System and Information Integration (SI) SI-2 Flaw Remediation SI-3 Malicious code protection Risk Assessment (RA) RA-5 Vulnerability monitoring and scanning Incident Response (IR) IR-4 Incident Handling
epss (Estimates the probability to exploit a vulnerability)	CVE-2019-9193 (PostgreSQL) CVE-2022-22963 (MySQL) CVE-2018-10933 (MySQL)	FDP_RIP.1 (Subset residual information protection); FDP_SDI.1 (Stored data integrity monitoring); FPT_RCV.3 (Automated recovery without undue loss); FDP_ACC.2 (Complete access control)	Risk Assessment (RA) RA-3 Risk Assessment RA-5 Vulnerability Monitoring and Scanning System and Information Integration (SI) SI-2 Flaw Remediation SI-4 Information System Monitoring Assessment, Authorization, and Monitoring (CA) CA-7 Continuous Monitoring System and Communications Protection (SC) SC-7 boundary protection
impact_score (Measures the effects of successful exploitation)	CVE-2019-0230 (MySQL) CVE-2014-0160 (Linux) CVE-2017-12617 (Tomcat)	FDP_DAU.1 (Basic data authentication); FCS_HTTPS.1 (HTTPS protocol); FTP_ITC.1 (Inter-TSF trusted channel); FAU_SAR.1 (Audit review)	Access Control (AC) AC-3 Access Enforcement System and Communications Protection (SC) SC-16 Transmission of Security and Privacy Attributes SC-28 Protection of Information at Rest SC-28(1) Cryptographic Protection System and Information Integrity (SI) SI-7 Software, Firmware, and Information Integrity Contingency Planning (CP) CP-10 System Recovery and Restoration Audit and Accountability (AU) AU-2 Audit Events
base_score (Provides an overall rating of the vulnerability)	CVE-2019-0232 (Tomcat) CVE-2008-3257 (Apache) CVE-2012-5613 (MySQL)	FIA_UAU.1 (Timing of authentication); FDP_ACC.1 (Subset access control); FIA_ATD.1 (User attribute definition); FAU_GEN.1 (Audit data generation)	Risk Assessment (RA) RA-5 Vulnerability Monitoring and Scanning Access Control (AC) AC-2 Account Management AC-6 Least Privilege Configuration Management (CM) CM-2 Baseline Configuration
attack_vector (Describes the possible means such as network, adjacent, local and physical used by the attacker to execute an attack)	CVE-2014-0160 (Linux) CVE-2018-7600 (Linux) CVE-2011-3310 (Cisco) CVE-2019-9193 (PostgreSQL)	FFW_RUL.1 (Simple firewall rules); FPF_RUL.1 (Packet filtering rules); FMT_SMF.1 (Specification of management functions); FDP_ITC.2 (Import of user data with security attributes)	Access control (AC) AC-3 Access enforcement AC-6 Least privilege AC-17 Remote access Physical and Environmental Protection (PE) PE-3 physical access control System and Communications Protection (SC) SC-7 boundary protection SC-12 cryptographic key establishment Assessment, Authorization, and Monitoring (CA) CA-3 Information exchange

Table 8 continued

Key Features from XAI	CVE and Products	Declared Security Requirements	Key Controls
user_interaction (Indicates the requirement of human user interaction, i.e, passive and active to compromise a system by an attacker)	CVE-2018-7600 (Linux) CVE-2022-22963 (MySQL) CVE-2019-0232 (Tomcat)	FTA_SSL.3 (TSF-initiated session locking); FDP_ACF.1 (Security attribute based access control); FPT_STM.1 (Reliable time stamps); FIA_UAU.2 (User authentication before action)	Awareness and Training (AT) AT-2 Security Awareness Training Access control (AC) AC-6 Least Privilege System and Communications Protection (SC) SC-18 Mobile Code System and Information Integration (SI) SI-4 Information System Incident Response (IR) IR-4 Incident Handling
privileges_required (Level of privilege, i.e., none, low, high, that attackers required to execute an attack)	CVE-2018-10933 (MySQL) CVE-2012-5613 (MySQL) CVE-2008-3257 (Apache)	FDP_ACF.1 (Security attribute based access control); FDP_ACC.2 (Complete access control); FIA_ATD.1 (User attribute definition); FCS_CKM.1 (Cryptographic key management)	Access Control (AC) AC-2 Account Management AC-5 Separation of Duties AC-6 Least Privilege AC-17 Remote Access Identification and Authentication (IA) IA-5 Authenticator Management Audit and Accountability (AU) AU-6 Audit Review AU-12 Audit Generation

palgeddon2), both concerning Linux-based systems, have EPSS scores of 0.96076 and 0.96053 respectively, indicating severe exploitability and high-risk materialization potential. The CodeBERT model achieved 96.9% accuracy on the training dataset and 96.7% on the validation dataset, confirming its effectiveness in identifying and prioritizing such vulnerabilities.

The experimental results identified the most significant features driving vulnerability exploitation, thereby operationalising XAI in a concrete, actionable manner. SHAP analysis revealed that EPSS, confidentiality_impact, attack_vector, and user_interaction serve as key drivers of model predictions. These features have direct implications for critical assets such as MySQL, Tomcat, and PostgreSQL in the p-NET system. This explainability enables security analysts not only to identify which vulnerabilities require priority mitigation but also to understand the underlying causes linked with specific assets, facilitating informed and evidence-based risk mitigation actions. Building upon these insights, the key features obtained from XAI were mapped to relevant NIST 800-53 security controls, enabling the definition of technical and organizational measures to appropriately reduce exploitation risks. For instance, vulnerabilities influenced by exploitability_score and attack_vector were mapped to Access Control (AC), System and Communications Protection (SC), and Incident Response (IR) controls. This transparent and auditable rationale for model predictions supports informed decision-making in cyber insurance. By understanding which vulnerabilities are associated with high residual risk–even after control deployment–organizations are better positioned to negotiate insurance coverage, premium adjustments, and exclusions. The integration of XAI into the risk management and insurability decision bridges the gap between AI-generated insights and real-world security governance, supporting both operational resilience and insurability.

7.2 Positioning Within the Vulnerability Detection Landscape

Having established the effectiveness of XAI in our approach, it is essential to position these results within the broader landscape of vulnerability detection research to understand our contribution's relative strengths and limitations. Recent studies on vulnerability detection have increasingly employed transformer-based models to enhance prediction accuracy and facilitate cybersecurity workflows. However, these models face challenges concerning sophistication, scalability, and practical deployment beyond academic settings. Our proposed approach, utilizing a fine-tuned CodeBERT model, achieved high accuracy on both training (96.9%) and validation (96.7%) sets while integrating EPSS-based scoring to facilitate dynamic risk evaluation. In comparison, the study by Tiwari [60] employed a standard BERT model for binary classification with 85.2% performance. While this demonstrates good application of pretrained models, it lacked integration with risk-based scoring or multi-class support. The work by Li et al. [61] proposed an ensemble pipeline incorporating 13 large language models (LLMs), but the heightened integration complexity and computational cost limited performance to only 67% accuracy. Our single-model design minimizes deployment complexity, maximizes real-time usability, and provides substantially improved accuracy. Li et al. [62] proposed the VDMAF model using Word2Vec,

GGNN, and BiLSTM layers, achieving a high F1-score of 98.9% on the Devign dataset but suffering from accuracy loss on other datasets, introducing generalizability concerns. Additionally, its complex architecture presents significant deployment challenges. Our model addresses these gaps by presenting a scalable, structured, and interpretable solution with multi-class classification capabilities and real-time vulnerability prioritization supported by an efficient transformer-based architecture. Furthermore, Aghaei et al. [63] conducted contextual vulnerability detection analysis with lightweight transformers, emphasizing the need for models to predict contextual information aligned with CVEs for better interpretation–a requirement that motivated our XAI-enabled, EPSS-integrated framework for actionable cybersecurity insights.

Despite these comparative advantages in accuracy and efficiency, transitioning from experimental validation to operational deployment presents several challenges warranting careful consideration. While our proposed CodeBERT-based architecture has improved vulnerability detection and risk estimation, real-world implementation poses problems that must be addressed for large-scale and sustained deployment. Chief among these is dependency on training data representativeness and quality. Although the CVEjoin dataset is extensive, it primarily consists of publicly disclosed vulnerabilities and lacks zero-day exploits or proprietary threats. Deletion of incomplete records during preprocessing also reduces data diversity and may limit model generalizability. Moreover, heuristic-based thresholding on EPSS scoring, while useful for laboratory settings, may not effectively capture real-world risk environment nuances.

Operationally, the computational intensity of transformer models like CodeBERT can prove a barrier to adoption, especially for organizations with limited infrastructure. While CodeBERT optimizes processes by consolidating multiple detection and scoring functionalities, initial setup costs–including cloud provisioning, employee training, and infrastructure upgrades–represent substantial investments. This challenge can be mitigated through cloud-hosted pre-trained models to facilitate rapid prototyping, combined with phased development of in-house capability for customized deployment and maintenance.

7.3 Practical Implications for Organizations and the Cybersecurity Industry

This research offers practical implications that extend beyond theoretical AI-enabled dynamic risk management, impacting cybersecurity operations, governance frameworks, and financial risk strategies. The primary benefit is to enable organizations to transition from reactive measures to proactive cybersecurity postures. By incorporating Explainable AI (XAI) into AI-powered vulnerability assessment, organizations can identify critical risks based on level of vulnerability exploitation and associated impact. This enhanced operational resilience not only improve performance but also minimizes high risk exposure such as 5G networks and cloud infrastructures. Beyond this operational advantage, another important implication relates to contextualized with organizational security policies and compliance requirements. Integrating vulnerability attributes with well-defined control objectives such as NIST 800-53, establishes guiding mechanisms for converting model outcomes into actionable security responses [64].

Complementing these governance benefits, the operationalisation of explainable AI democratizes security knowledge across diverse user roles [65]. Since model predictions can be interpreted and explained at appropriate levels, all relevant stakeholders,i.e., risk or business manager, system administrator, and management can readily understand the rationale behind security decisions. This bridges gap between technical and non-technical teams, enhancing decision-making and security governance across the organization. Furthermore, this transparency directly supports cyber insurance acquisition decisions by providing insurers with interpretable evidence of organizational security posture, control effectiveness, and residual risk. The ability to justify control selection through XAI techniques reduces information asymmetry in the cyber insurance market, addressing one of the key challenges preventing market maturity identified in our introduction.

7.4 Limitations and Future Research

The proposed approach demonstrates significant potential in advancing cyber insurability through feature-based risk characterization, calculating the risk level using dynamic security properties and providing justification for chosen security controls and informed cyber insurance decision making. The experimental results also show achieving 96.9% accuracy for the vulnerability exploitability prediction. However, we observe several limitations in the proposed approach. Firstly, the impact estimation only focuses on the published CVE impact level. In the future, we aim to consider the direct impact on other parameters, such as financial loss and reputational damage, for the insurability decision. Secondly, the XAI enhanced control selection and residual risk characterization only considers SHAP and heat map techniques. Although, SHAP and heatmaps provide valuable insights for feature importance and interdependencies, they lack a focus on capturing non-linear feature interactions and temporal dependencies among features. In this context, we aim to integrate advanced XAI techniques such as LIME (Local Interpretable Model-agnostic Explanations) for local interpretability validation and counterfactual explanations to demonstrate how changes in specific features would alter

feature importance leading to comprehensive and actionable insights for decision making. Thirdly, the proposed approach only considers vulnerability exploitation as a dynamic security parameter for the risk assessment. In future, we would like to extend these dynamic parameters to include asset dependencies and zero-day exploitation for risk assessment. Finally, a single pilot case based evaluation limits the generalisation of our finding and we are aiming to evaluate the approach using other sector specific pilot cases.

8 Conclusion

The adoption of cyber insurance has significantly increased in recent years to protect businesses from potential losses and to ensure overall business continuity. However, the security context is continuously evolving and organizations need to understand their overall security posture to make the right decisions. This research addresses the challenge of dynamic cybersecurity risk management, enabling informed cyber insurability decision within evolving security and system contexts. The proposed approach introduces AI based dynamic risk management methodology that incorporates temporal security parameters using CodeBERT based LLM to identify exploitation of vulnerability which then used to calculate the risk level. The adoption of XAI for feature correlation and contribution justifies the chosen security control. The comprehensive evaluation through the p-NET 5G infrastructure use case validates the practical applicability of the approach, demonstrating how XAI insights directly support control selection and identification of residual risk for actionable cyber insurance decisions. We believe that such an approach will advocate wider adoption of cyber insurance and minimize the variation in risk perception for a given context.

Acknowledgements The authors acknowledge P-NET to support with the pilot case study evaluation.

Author Contributions Spyridon Papastergiou: Conceptualization, Writing - original draft, Writing - review and editing, Methodology, Project administration, Supervision, Funding acquisition, Resources, Investigation. Nihala Basheer: Conceptualization, Writing - original draft, Writing - review and editing, Methodology, Investigation, Validation, Software (CodeBERT model implementation, XAI techniques development), Data curation (CVEJoin dataset processing, vulnerability labeling). Kostas Lampropoulos: Resources, Project administration, Validation, Investigation (P-NET infrastructure access and use case implementation), Writing - review and editing. Panayiotis Verrios: Software (industrial use case development, system integration), Investigation, Resources, Writing - review and editing. Shareeful Islam: Conceptualization, Methodology, Supervision, Writing - original draft, Writing - review and editing, Formal analysis (XAI results interpretation, control mapping validation).

Funding This work was partially supported by the European Union's Horizon Europe Project, 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, funded under grant agreement No. 101120779
and CURIUM-Cra sUppoRt continuum, funded under the Grant Agreement No. 101190372
and CUSTODES - A Certification approach for dynamic, agile and reUSable assessmenT fOr composite systems of ICT proDucts, servicEs, and processeS, funded under grant agreement number 101120684.

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 creatures performed by any of the others.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Jennings-Trace, E.: Three massive UK retailers have been hit by cyber attacks this week – so what's going on? TechRadar (2025). https://www.techradar.com/pro/security/three-massive-uk-retailers-have-been-hit-by-cyber-attacks-this-week-so-whats-going-on
2. Kumar, R., Singh, S.: Cyber insurance in India: An overview. Int. J. Res. Finance Manag. **6**, 373–378 (2023). https://doi.org/10.33545/26175754.2023.v6.i1d.230
3. Islam, S., Basheer, N., Papastergiou, S., Ciampi, M., Silvestri, S.: Intelligent dynamic cybersecurity risk management framework with explainability and interpretability of AI models for enhancing security and resilience of digital infrastructure. Journal of Reliable Intelligent Environments 11, Article 12 (2025). https://doi.org/10.1007/s40860-025-00253-3
4. Eling, M., Schnell, W.: What do we know about cyber risk and cyber risk insurance? The Journal of Risk Finance **17**, 474–491 (2016). https://doi.org/10.1108/JRF-09-2016-0122
5. Zeller, G., Scherer, M.: Risk mitigation services in cyber insurance: optimal contract design and price structure. The Geneva Papers on Risk and Insurance - Issues and Practice **48**, 502–547 (2023). https://doi.org/10.1057/s41288-023-00289-7
6. Cremer, F., Sheehan, B., Fortmann, M., Kia, A.N., Mullins, M., Murphy, F., Materne, S.: Cyber risk and cybersecurity: a system-

atic review of data availability. The Geneva Papers on Risk and Insurance - Issues and Practice **47**, 698–736 (2022). https://doi.org/10.1057/s41288-022-00266-6
7. Chong, W.F., Linders, D., Quan, Z., Zhang, L.: Incident-specific cyber insurance. ASTIN. Bulletin **55**, 395–425 (2025). https://doi.org/10.1017/asb.2025.9
8. Gonzalez-Granadillo, G., Dubus, S., Motzek, A., Garcia-Alfaro, J., Alvarez, E., Merialdo, M., Papillon, S., Debar, H.: Dynamic risk management response system to handle cyber threats. Future Gener. Comput. Syst. **83**, 535–552 (2018). https://doi.org/10.1016/j.future.2017.05.043
9. Schneider, D., Reich, J., Adler, R., Liggesmeyer, P.: Dynamic risk management in cyber physical systems. arXiv:2401.13539 (2024). arXiv:2401.13539
10. Cheimonidis, P., Rantos, K.: Dynamic risk assessment in cybersecurity: A systematic literature review. Future Internet 15 (2023). https://doi.org/10.3390/fi15100324
11. Naumov, S., Kabanov, I.: Dynamic framework for assessing cyber security risks in a changing environment. In: 2016 International Conference on Information Science and Communications Technologies (ICISCT), pp. 1–4 (2016). https://doi.org/10.1109/ICISCT.2016.7777406
12. Panou, A., Ntantogian, C., Xenakis, C.: RiSKi. In: 21st Pan-Hellenic Conference on Informatics, pp. 1–6 (2017). https://doi.org/10.1145/3139367.3139426
13. Wang, S.S.: Integrated framework for information security investment and cyber insurance. Pac.-Basin Finance J. 57, 101173 (2019). https://doi.org/10.1016/j.pacfin.2019.101173
14. Zhang, R., Zhu, Q.: Optimal Cyber-Insurance contract design for dynamic risk management and mitigation. IEEE Trans. Comput. Soc. Syst. **9**, 1087–1100 (2021). https://doi.org/10.1109/tcss.2021.3117905
15. Thlon, M., Strupczewski, G.: Assessing the impact of cyber risk perception on cyber insurance purchase decisions. Sci. Pap. Silesian Univ. Technol. Organ. Manag. Ser. 179 (2023). https://doi.org/10.29119/1641-3466.2023.179.32
16. Jawhar, S., Kimble, C.E., Miller, J.R., Bitar, Z.: Enhancing Cyber Resilience with AI-Powered Cyber Insurance Risk Assessment. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0435–0438 (2024). https://doi.org/10.1109/ccwc60891.2024.10427965
17. Shaikh, M.U.R., Ullah, R., Akbar, R., Savita, K.S., Mandala, S.: Fortifying Against Ransomware: Navigating Cybersecurity Risk Management with a Focus on Ransomware Insurance Strategies. Int. J. Acad. Res. Bus. Soc. Sci. 14 (2024). https://doi.org/10.6007/ijarbss/v14-i1/20566
18. Biswas, B., Mukhopadhyay, A., Kumar, A., Delen, D.: A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks. Decis. Support Syst. **177**, 114102 (2024). https://doi.org/10.1016/j.dss.2023.114102
19. Mukhopadhyay, A., Jain, S.: A framework for cyber-risk insurance against ransomware: A mixed-method approach. Int. J. Inf. Manage. **74**, 102724 (2024). https://doi.org/10.1016/j.ijinfomgt.2023.102724
20. Pavlík, L., Ficek, M., Rak, J.: Dynamic assessment of cyber threats in the field of insurance. Risks **10**(12), 222 (2022). https://doi.org/10.3390/risks10120222
21. Uuganbayar, G., Yautsiukhin, A., Martinelli, F., Massacci, F.: Optimisation of cyber insurance coverage with selection of cost effective security controls. Computers & Security **101**, 102121 (2021). https://doi.org/10.1016/j.cose.2020.102121
22. Boonen, T.J., Feng, Y., Tong, Z.: Cybersecurity investments and cyber insurance purchases in a non-cooperative game. ASTIN Bulletin: The Journal of the IAA **55**(2), 426–448 (2025). https://doi.org/10.1017/asb.2024.31
23. Awiszus, K., Knispel, T., Penner, I., Svindland, G., Voß, A., Weber, S.: Modeling and pricing cyber insurance: Idiosyncratic, systematic, and systemic risks. Eur. Actuar. J. **13**(1), 1–53 (2023). https://doi.org/10.1007/s13385-023-00342-9
24. French, C.C.: Five approaches to insuring cyber risks. Maryland Law Review **81**, 103 (2021)
25. Gordon, L.A., Loeb, M.P., Sohail, T.: A framework for using insurance for cyber-risk management. Commun. ACM **46**(3), 81–85 (2003). https://doi.org/10.1145/636772.636774
26. Eling, M., Schnell, W.: What do we know about cyber risk and cyber risk insurance? The Journal of Risk Finance **21**(1), 1–14 (2020). https://doi.org/10.1108/JRF-03-2019-0045
27. Romanosky, S., Ablon, L., Kuehn, A., Jones, T.: Content analysis of cyber insurance policies: How do carriers price cyber risk? Journal of Cybersecurity 5(1), tyz002 (2019). https://doi.org/10.1093/cybsec/tyz002
28. Jacobs, J., Romanosky, S., Edwards, B., Roytman, M., Adjerid, I.: Exploit prediction scoring system (EPSS). Digital Threats: Research and Practice **2**(3), 1–17 (2021). https://doi.org/10.1145/3436242
29. Mohitkar, C., Lakshmi, D.: Explainable AI for transparent Cyber-Risk assessment and Decision-Making. In: Advances in computational intelligence and robotics book series, pp. 219–246 (2024). https://doi.org/10.4018/979-8-3693-7540-2.ch010
30. Mishra, S., Pradhan, R.K.: Analyzing the Impact of Feature Correlation on Classification Accuracy of Machine Learning Model. In: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), pp. 879–883. IEEE, Greater Noida (2023). https://doi.org/10.1109/AISC56616.2023.10085542
31. Macdonald, R.R.: Correlation and covariance matrices. Wiley StatsRef: Statistics Reference Online (2014). https://doi.org/10.1002/9781118445112.stat06481
32. Weisburd, D., Britt, C., Wilson, D.B., Wooditch, A.: Measuring Association for Scaled Data: Pearson's correlation coefficient. In: Springer eBooks, pp. 479–530 (2020). https://doi.org/10.1007/978-3-030-47967-1_14
33. Zhong, J., Negre, E.: Context-aware feature attribution through argumentation. arXiv (Cornell University) (2023)https://doi.org/10.48550/arxiv.2310.16157
34. Zhang, Y., Tan, X., Xi, H., Zhao, X.: Real-time risk management based on time series analysis. In: 2008 7th World Congress on Intelligent Control and Automation, pp. 2518–2523. IEEE, Chongqing (2008). https://doi.org/10.1109/WCICA.2008.4593320
35. Cheng, X., Che, C.: Interpretable Machine Learning: Explainability in Algorithm Design. J. Ind. Eng. Appl. Sci. 2, 65–70 (2024). https://doi.org/10.70393/6a69656173.323337
36. Mouratidis, H., Islam, S., Santos-Olmo, A., Sanchez, L.E., Ismail, U.M.: Modelling language for cyber security incident handling for critical infrastructures. Computers & Security **128**, 103139 (2023). https://doi.org/10.1016/j.cose.2023.103139
37. Kure, H.I., Islam, S., Mouratidis, H.: An integrated cyber security risk management framework and risk predication for the critical infrastructure protection. Neural Comput. Appl. **34**(18), 15241–15271 (2022). https://doi.org/10.1007/s00521-022-07059-6
38. 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. Integrated Computer-Aided Engineering 10692509251368663 (2025). https://doi.org/10.3233/ICA-251368663
39. Van Aartsengel, A., Kurtoglu, S.: Analyze process steps and tasks. In: Springer eBooks, pp. 483–517 (2013). https://doi.org/10.1007/978-3-642-35901-9_27
40. NVD - Common Platform Enumeration (CPE). https://nvd.nist.gov/products/cpe

41. Zhang, Z.: Missing data imputation: focusing on single imputation. PubMed **4**, 9 (2016). https://doi.org/10.3978/j.issn.2305-5839.2015.12.38
42. Azar, D., Harmanani, H.: Heuristic approaches for optimizing the performance of rule-based classifiers. In: 2011 IEEE International Conference on Information Reuse & Integration, pp. 25–31. IEEE, Las Vegas (2011). https://doi.org/10.1109/IRI.2011.6009515
43. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., Zhou, M.: CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020 (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.139
44. Khanfir, A., Jimenez, M., Papadakis, M., Traon, Y.L.: CodeBERT-nt: Code Naturalness via CodeBERT. In: 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS), pp. 936–947. IEEE, Guangzhou (2022). https://doi.org/10.1109/QRS57517.2022.00098
45. Cabot, J.H., Ross, E.G.: Evaluating prediction model performance. Surgery **174**, 723–726 (2023). https://doi.org/10.1016/j.surg.2023.05.023
46. Naidu, G., Zuva, T., Sibanda, E.M.: A review of evaluation metrics in Machine learning Algorithms. In: Lecture Notes in Networks and Systems, pp. 15–25 (2023). https://doi.org/10.1007/978-3-031-35314-7_2
47. Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Sci. Rep. 14 (2024)https://doi.org/10.1038/s41598-024-56706-x
48. Fourure, D., Javaid, M.U., Posocco, N., Tihon, S.: Anomaly Detection: How to Artificially Increase Your F1-Score with a Biased Evaluation Protocol. In: Lecture notes in computer science, pp. 3–18 (2021). https://doi.org/10.1007/978-3-030-86514-6_1
49. FIRST - Common Vulnerability Scoring System Specification Document. https://www.first.org/cvss/specification-document
50. Applied Sciences: Special Issue on Cybersecurity Risk Assessment and Management. https://www.mdpi.com/2076-3417/12/9/4443
51. Li, M., Sun, H., Huang, Y., Chen, H.: Shapley value: from cooperative game to explainable artificial intelligence. Auton. Intell. Syst. 4 (2024)https://doi.org/10.1007/s43684-023-00060-8
52. Khan, M.M.: Cyber Security Risk Management. Int. J. Multidiscip. Res. 6 (2024). https://doi.org/10.36948/ijfmr.2024.v06i04.23754
53. NIST: Security and privacy controls for information systems and organizations (2020). https://doi.org/10.6028/nist.sp.800-53r5
54. Tsegaye, T., Flowerday, S.: Controls for protecting critical information infrastructure from cyberattacks. In: World Congress on Internet Security (WorldCIS-2014), pp. 24–29. IEEE, London (2014). https://doi.org/10.1109/worldcis.2014.7028160
55. Cremer, F., Sheehan, B., Fortmann, M., Mullins, M., Murphy, F.: Cyber exclusions: An investigation into the cyber insurance coverage gap. In: 2022 Cyber Research Conference - Ireland (Cyber-RCI), pp. 1–10. IEEE, Galway (2022). https://doi.org/10.1109/cyber-rci55324.2022.10032678
56. P-NET: Cybersecurity Research Network. https://p-net.gr/
57. CVEjoin: An Information Security Vulnerability and Threat Intelligence Dataset. Figshare (2022). https://doi.org/10.6084/m9.figshare.21586923
58. Microsoft CodeBERT-base Model. https://huggingface.co/microsoft/codebert-base
59. Kholiev, V., Barkovska, O.: Analysis of the of training and test data distribution for audio series classification. ІнформаціЙно–керуючі Системи На Залізничному Транспорті 28, 38–43 (2023). https://doi.org/10.18664/ikszt.v28i1.27634
60. Tiwari, H.: Advancing Vulnerability Classification with BERT: A Multi-Objective Learning Model. arXiv preprint arXiv:2503.20831 (2025). https://arxiv.org/abs/2503.20831
61. Li, Y., Li, X., Wu, H., Xu, M., Zhang, Y., Cheng, X., Zhong, S.: Everything You Wanted to Know About LLM-based Vulnerability Detection But Were Afraid to Ask. arXiv preprint arXiv:2504.13474 (2025)
62. Li, Y., Luo, Q., Wu, P., Zheng, H.: VDMAF: Cross-language source code vulnerability detection using multi-head attention fusion. Inf. Softw. Technol. 107739 (2025). https://doi.org/10.1016/j.infsof.2025.107739
63. Aghaei, E., Al-Shaer, E., Shadid, W., Niu, X.: Automated CVE analysis for threat prioritization and impact prediction. arXiv (Cornell University) (2023). https://doi.org/10.48550/arxiv.2309.03040
64. Rahman, M.M., Kshetri, N., Sayeed, S.A., Rana, M.M.: AssessITS: Integrating procedural guidelines and practical evaluation metrics for organizational IT and Cybersecurity risk assessment. arXiv (Cornell University) (2024). https://doi.org/10.48550/arxiv.2410.01750
65. Umm-E-Habiba, N., Habibullah, K.M.: Explainable AI: A Diverse Stakeholder Perspective. In: 2024 IEEE 32nd International Requirements Engineering Conference (RE), pp. 494–495. IEEE, Reykjavik (2024). https://doi.org/10.1109/re59067.2024.00060
66. Mosbach, M., Andriushchenko, M., Klakow, D.: On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. arXiv:2006.04884 (2020)
67. European Commission, European Common Criteria-based cybersecurity certification scheme (EUCC), Commission Implementing Regulation (EU) 2024/482, 2024. https://eur-lex.europa.eu/eli/reg_impl/2024/482/oj

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.