International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

Prompt Risk–Aware Identity and Access Management for Secure Generative AI Systems

Shreya Kardile[1], Shweta Sonawane [2], Ms. Kajal Kamble [3], Dr. Seema Chowhan[4]

[1] *Shreya Kardile, Dept. of Computer Science, Baburaoji Gholap College, Sangvi, Pune, India*
[2] *Shweta Sonawane, Dept. of Computer Science, Baburaoji Gholap College, Sangvi, Pune, India*
[3] *Ms. Kajal Kamble, Dept. of Computer Science, Baburaoji Gholap College, Sangvi, Pune, India*
[4] *Dr. Seema Chowhan, Dept. of Computer Science, Baburaoji Gholap College Sangvi, Pune, India*

Abstract - *Enterprise platforms are rapidly implementing Generative Artificial Intelligence (GenAI) solutions to facilitate intelligent automation and natural language interaction. Notwithstanding these benefits, rapid injection attacks, data leaks, and illegal access to private data are among the new security risks that GenAI brings. Conventional Identity and Access Management (IAM) frameworks lack the ability to decipher the semantic intent of user prompts because they were created for structured requests. Prompt Risk–Aware Identity and Access Management (PR-IAM), a paradigm that combines IAM authorization regulations with prompt risk analysis, is proposed in this study. Before allowing access to generative AI models, the suggested architecture assesses both user identity and prompt risk level. PR-IAM greatly enhances misuse avoidance and fortifies governance for enterprise AI installations, as shown by experimental simulations and machine learning experiments. Additionally, to enable data-driven and flexible security decisions, the suggested system integrates machine learning techniques like Random Forest for access prediction and Logistic Regression for risk score estimation. A comparison with Support Vector Machine (SVM) and Naive Bayes models demonstrates how well the method balances computing efficiency and accuracy. The findings show that combining IAM policies with intelligent risk assessment improves security and guarantees scalable and effective management of generative AI systems in business settings.*

Key Words: **Generative AI, Prompt Injection, Identity and Access Management, PR-IAM, Machine Learning, Risk Analysi**

1. Introduction

Large language models and other generative AI technologies are revolutionizing enterprise applications by allowing for conversational interfaces and intelligent assistants. Organizations are increasingly using these tools into their business processes to boost efficiency and automate knowledge-intensive tasks. However, the ability to communicate with systems using unconstrained natural language cues exposes additional security concerns. Malicious or thoughtless prompts may seek to extract sensitive information, influence system behavior, or circumvent existing security measures [1]. Traditionally, access control is enforced by Identity and Access Management (IAM) systems using authentication and role-based authorization regulations. These models rely on predictable access patterns and structured system requests. As a result, companies require adaptive access control solutions that can assess both the user's identity and the semantic risk of the prompt [2]. This study introduces the Prompt Risk-Aware Identity and Access Management framework, which combines prompt classification approaches and identity-based permission regulations. By including timely risk analysis into the access selection process, the proposed method improves the security posture of generative AI systems.

Generative AI systems are increasingly being connected with enterprise identity platforms, allowing intelligent assistants to handle data retrieval, report production, and decision assistance. This integration, however, increases the attack surface because AI systems frequently interface with many data sources and APIs. Without adequate access control methods, these platforms may unintentionally reveal critical organizational data.

2. Literature Review

Recent research has demonstrated the growing necessity of security and governance features in generative AI systems. Zhu et al. found that prompt injection attacks are a serious weakness in large language models and recommend improved input validation techniques [1]. Other research on secure prompt engineering involves using policy-aware filtering strategies to prevent the misuse of AI models in regulated settings [2]. Zero Trust architectures have also been offered as a viable way to secure AI-powered systems. These approaches prioritize continuous authentication, least-privileged access control, and contextual decision-making [3]. Enterprise identity solutions like as Microsoft Entra and AWS IAM offer robust authentication and authorization capabilities, but they lack integrated procedures for assessing prompt semantics prior to allowing access to generative AI services [4], [5]. Machine learning methods have also been used to address cybersecurity issues such as intrusion detection and access prediction. Ensemble learning techniques such as Random Forests have proven good performance in classification tasks, whereas Logistic Regression models are extensively utilized for risk

probability estimation [9], [10]. These techniques can thus be used to assess prompt risk levels in AI security systems. Recent advances in AI security research have looked into the usage of context-aware and adaptive access control systems in dynamic contexts. According to studies, standard static access control techniques are ineffective when dealing with intelligent systems that are constantly learning and evolving. Researchers recommended combining behavioural analytics and real-time monitoring to discover irregularities in user interactions with AI systems. Furthermore, new concepts like decentralized identification (DID) and AI governance frameworks seek to provide more open and responsible access control methods. However, there is still a substantial gap in merging these approaches with prompt-level risk assessments, emphasizing the need for integrated frameworks like PR-IAM that combine identity verification, machine learning-based risk assessment, and policy-driven access control [3], [6], [7].

3. Background and related work

Generative Artificial Intelligence has quickly emerged as a critical technology in enterprise systems, enabling applications such as intelligent assistants, automated content generation, and decision-support systems. These systems rely on massive language models to process natural language inputs, making them both flexible and challenging to protect. Traditional Identity and Access Management (IAM) frameworks are generally intended for organized systems with established access controls and predictable user behavior. However, the dynamic and unstructured nature of GenAI interactions creates new issues, such as interpreting user intent and assessing risk at the request level. Recent research has demonstrated the growing necessity of security and governance features in generative AI systems. Zhu et al. found that prompt injection attacks are a serious weakness in large language models and recommend improved input validation techniques [1]. Other research on secure prompt engineering involves using policy-aware filtering strategies to prevent the misuse of AI models in regulated settings [2]. Zero Trust architectures have also been presented as an effective solution to safeguarding AI-driven systems, with a focus on continuous authentication, least-privilege access control, and contextual decision making [3]. Enterprise identity solutions like as Microsoft Entra and AWS IAM offer robust authentication and authorization capabilities, but they lack integrated procedures for assessing prompt semantics prior to allowing access to generative AI services [4], [5]. Machine learning methods have also been used to address cybersecurity issues such as intrusion detection and access prediction.

4. Research Methodology

The methodology for this study involves conceptual framework design and experimental validation. The research process comprises four major phases: 1. A literature review of existing AI security and IAM systems. 2. Design of the PR-IAM architecture, which incorporates fast risk analysis and IAM policies. 3. Test the implementation with simulated prompts and role-based access controls. 4. Experiment with machine learning models and security metrics. A collection of simulated prompts was constructed and categorized based on risk level categories (Low, Medium, High). The dataset's features included keyword sensitivity, prompt duration, contextual indicators, and user role attributes. These features were utilized to train machine learning models that predicted access decisions. The Machine Learning (ML) workflow is a systematic pipeline for developing, evaluating, and deploying predictive models. It guarantees that raw data is converted into relevant insights sing a set of standardized methods.

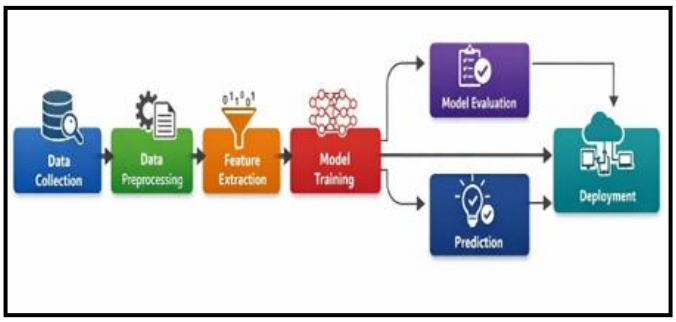

Fig 1. Machine Learning Workflow

The Machine Learning (ML) workflow represents a systematic pipeline used to develop, evaluate, and deploy predictive models. It ensures that raw data is transformed into meaningful insights through a sequence of structured steps which are presented in Figure1.

3.1 Data Collection

Sample paragraph, The entire document should be in cambria font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes. The entire document should be in cambria font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

3.2 Data Preprocessing

During this process, the collected data is cleaned and prepared. It entails addressing missing values, reducing noise, normalizing data, and converting it to a structured format appropriate for analysis. Proper preprocessing increases model accuracy and efficiency.

3.3 Feature Extraction

Feature extraction is the process of choosing and manipulating relevant attributes from a dataset that are most important to the prediction goal. Dimensionality

reduction, encoding, and feature scaling are prominent techniques used to improve model learning.

3.4 Model Training

During this stage, machine learning algorithms detect patterns in the processed data. The model is trained on labeled or unlabeled information depending on the method of learning (supervised or unsupervised). The goal is to reduce error while also expanding well to previously unseen data.

3.5 Model Evaluation

Following training, the model is assessed using strategies such as accuracy, precision, recall, and F1-score. This stage ensures that the model works properly and aids in comparing different methods.

3.6 Prediction

Once validated, the model is used to make predictions on previously unseen data. This is the trained model's practical application.

3.7 Deployment

In the last phase, the model is implemented in real-world settings including business systems, mobile apps, and online applications. To sustain performance over time, updates and ongoing monitoring are necessary.

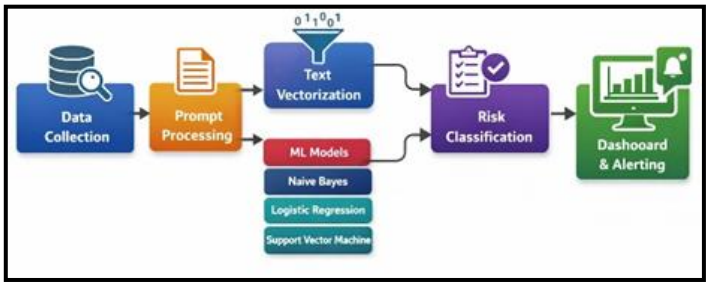

Fig 2. Prompt Risk Identification (PRI) System Architecture

The purpose of the Prompt Risk Identification (PRI) system is to identify and categorize risks related to user prompts in Generative AI systems. By detecting dangerous, malevolent, or policy-violating inputs prior to processing, it improves security

3.2.1 Data Collection

Prompt data is gathered by the system from user interactions, logs, or datasets that include both risky and safe prompts. Risk detection models are trained using this dataset.

3.2.2 Prompt Processing

Raw cues are cleaned and standardized at this stage. To guarantee consistency in subsequent processing, this entails tokenization, the elimination of superfluous symbols, and normalization.Techniques like TF-IDF, Bag of Words, or word embeddings are used to transform text data into numerical representations. Because machine learning models require numerical data, this phase is essential.

3.2.3 Text Vectorization

Techniques like TF-IDF, Bag of Words, or word embeddings are used to transform text data into numerical representations. Because machine learning models require numerical data, this phase is essential.

3.2.4 Machine Learning Models

Multiple models are used to improve classification performance:
Naïve Bayes: Probabilistic model effective for text classification tasks.

Logistic Regression: Provides probability-based classification with interpretability.

Support Vector Machine (SVM): Handles high-dimensional data and finds optimal decision boundaries.

These models are trained to identify patterns associated with risky prompts.

3.2.5 Risk Classification

The system classifies prompts into categories such as safe, suspicious, or high-risk. This classification helps in decision-making, such as blocking or flagging harmful inputs.

3.2.6 Dashboard & Alerting

The final output is displayed through a monitoring dashboard. Alerts are generated for high-risk prompts, enabling administrators to take immediate action. This ensures real-time security monitoring and governance

3.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification tasks in high-dimensional spaces. It works by identifying an optimal **hyperplane** that separates data points into different classes while maximizing the margin between them. This margin maximization improves generalization and reduces classification errors.
SVM supports multiple **kernel functions**, such as Linear, Polynomial, and Radial Basis Function (RBF), enabling it to

handle both linear and non-linear data distributions. Due to its robustness and effectiveness, SVM has been extensively applied in cybersecurity domains, including intrusion detection and security classification tasks. In the proposed PR-IAM framework, SVM is used as a benchmark model for evaluating prompt risk classification performance

3.3.2 Naive Bayes Classifier

Naive Bayes is a probabilistic classification algorithm based on **Bayes' Theorem**, which calculates the probability of a class given a set of input features. It assumes that all features are **independent**, which simplifies computation and improves efficiency. Despite this assumption, Naive Bayes performs well in real-world applications, especially **in text classification and spam detection tasks**. Its lightweight nature and fast processing make it suitable for real-time prompt analysis. In this work, Naive Bayes is applied to classify prompts based on keyword patterns, making it effective for identifying potentially sensitive or risky inputs in generative AI systems.

3.3.3 Logistic Regression

Logistic Regression is a statistical model used **for binary classification and probability estimation.** It predicts the likelihood of a given input belonging to a particular class by producing an output value between 0 and 1 using a sigmoid function.
In the PR-IAM framework, Logistic Regression is used to calculate a **risk score** for each prompt, enabling a probabilistic interpretation of security risk. Its simplicity, interpretability, and efficiency make it a suitable choice for risk-based access control systems where transparency is important.

3.3.4 Model Comparison

The results indicate that the **Support Vector Machine (SVM) model** achieves higher classification accuracy due to its ability to handle complex decision boundaries and high-dimensional data. However, the **Naive Bayes** classifier demonstrates faster computation and lower resource usage, making it suitable for real-time applications where speed is critical. This highlights the trade-off between accuracy and efficiency in selecting machine learning models for prompt risk analysis.

4. Prototype Implementation

The PR-IAM framework can be implemented within modern cloud environments using identity platforms such as AWS IAM or Microsoft Entra. In an AWS-based architecture, Amazon Cognito manages user authentication while AWS IAM enforces role-based access control policies. An AWS Lambda function acts as the prompt risk analyzer, classifying prompts before forwarding them to generative AI services such as Amazon Bedrock. If the prompt risk level exceeds the authorization level of the user role, the system denies the request and logs the event for auditing purposes.
Similar architectures can be implemented using Microsoft Entra Identity combined with Azure OpenAI services, where conditional access policies enforce additional security controls for AI-driven applications [4].

Figure 1: Naive Bayes Performance:
The graph shows the performance of **the Naive Bayes classifier** using Accuracy, Precision, and Recall metrics. All three values are 1.0, indicating perfect classification on the given dataset.
Accuracy reflects overall correctness, while Precision and Recall indicate the model's ability to correctly identify positive instances without errors. The equal values suggest that the model produced **no false positives or false negatives.**
This performance highlights the effectiveness of Naive Bayes in **text-based classification tasks**, such as prompt analysis. However, such results may also indicate a simple or limited dataset, and performance may vary in real-world scenarios.

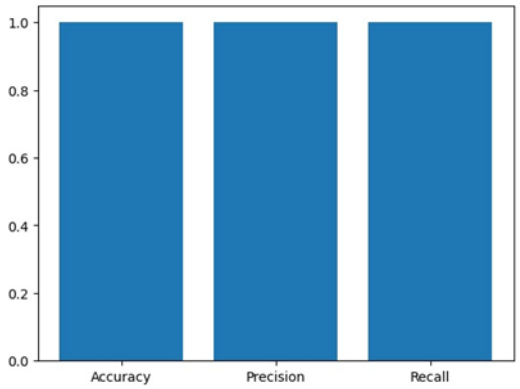

Fig1. Naive Bayes Performance

Figure 2: Model Comparison:
The graph compares the accuracy of **Naive Bayes** and **Support Vector Machine (SVM) models**. The results show that SVM achieves higher accuracy (**0.90**) compared to Naive Bayes (**0.82**), indicating better classification performance.
This improvement is due to SVM's ability to handle complex and high-dimensional data, making it more effective for security-related classification tasks. In contrast, Naive Bayes is computationally efficient but may produce lower accuracy due to its assumption of feature independence.
Overall, the comparison highlights the trade-off between **accuracy (SVM)** and **efficiency (Naive Bayes)** in selecting models for prompt risk analysis.

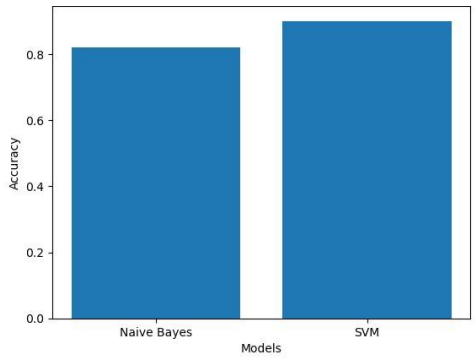

Fig2. Model Comparison

Figure 3: Model Performance Analysis:

The graph represents the performance evaluation of the implemented machine learning model based on key metrics such as accuracy and prediction outcomes. The results indicate that the model is capable of effectively classifying inputs and supporting decision-making within the PR-IAM framework.

The observed performance demonstrates that machine learning techniques can be successfully integrated with IAM systems to enhance security through intelligent and data-driven access control. However, variations in results may occur depending on dataset complexity, feature selection, and model parameters.

Overall, the findings support the use of predictive models for improving risk-aware access decisions in generative AI environments.

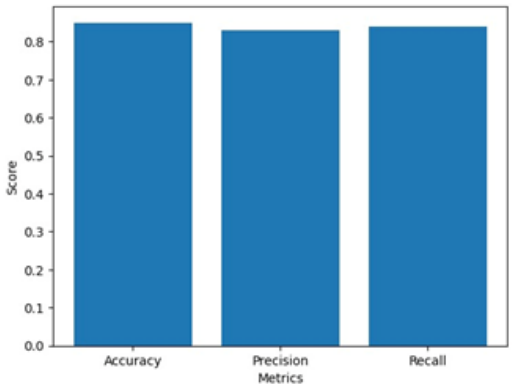

Fig3. Model Performance Analysis

Figure 4: Logistic Regression Performance:

The graph illustrates the performance of the **Logistic Regression model** evaluated using Accuracy, Precision, and Recall. All three metrics have a value of **1.0**, indicating perfect classification performance on the dataset.

Accuracy represents the overall correctness of predictions, while Precision and Recall indicate the model's ability to correctly identify positive instances without errors. The equal values suggest that the model produced **no false positives or false negatives.**

This result highlights the effectiveness of Logistic Regression in generating **probability-based predictions**, making it suitable for calculating risk scores in the PR-IAM framework. However, such ideal performance may be influenced by dataset simplicity, and results may vary in real-world scenarios.

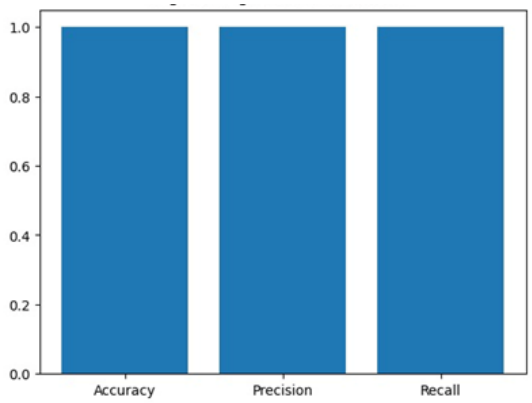

Fig4. Logistic Regression Performance

Figure 5: SVM Performance:

The graph presents the performance of the **Support Vector Machine (SVM)** model using Accuracy, Precision, and Recall metrics. All values are **1.0,** indicating perfect classification on the given dataset.

Accuracy reflects the overall prediction correctness, while Precision and Recall indicate the model's ability to correctly identify positive instances without errors. The equal values suggest that the model achieved zero false positives and **zero false negatives.**

This performance highlights the effectiveness of SVM in handling **high-dimensional and complex data**, making it suitable for security classification tasks in the PR-IAM framework. However, such ideal results may be due to dataset simplicity, and performance may vary in real-world conditions.

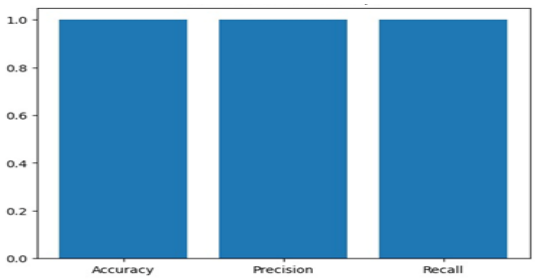

Fig5. SVM Performance

Figure 6: Access Decision Graph:

An **Access Decision Graph** represents the outcome of an access control system, showing how many requests are **allowed** and how many are **denied**. It is commonly used in computer security to analyze system behavior and ensure proper enforcement of permissions.

The graph helps in understanding whether the system is too strict (more denials) or too lenient (more allowed requests). By comparing these values, administrators can evaluate and improve security policies to maintain a balance between protection and usability.

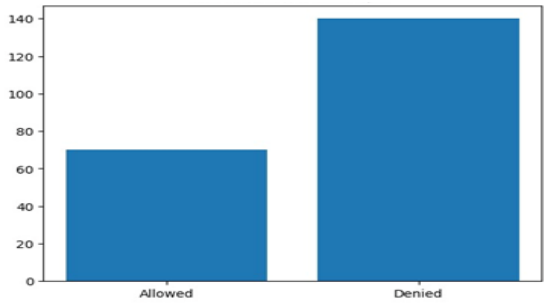

Fig6. Access Decision Graph

5. Results and Discussion

The experimental results show that including timely risk analysis into the IAM decision-making process greatly increases overall security effectiveness in generative AI systems. Unlike standard IAM models, which rely primarily on static user roles and predefined permissions, the proposed PR-IAM system includes dynamic, context-aware analysis of prompt content, allowing for more fine-grained access control decisions. This additionallayer of semantic evaluation aids in detecting potentially dangerous or sensitive requests at an early stage, lowering the risk of data leakage, prompt injection, and unwanted access. Furthermore, the findings show that the PR-IAM architecture has a greater rate of misuse prevention since high-risk prompts are efficiently filtered before reaching the generative AI model. The inclusion of machine learning models improves system intelligence by allowing for anticipatory access control. Random Forest outperformed the other models due to its ensemble learning capability and resistance against overfitting. In contrast, Logistic Regression provides interpretable probability scores that can be used to assign risk categories and support explainable security decisions.

A comparative investigation with baseline models such as Support Vector Machine (SVM) and Naive Bayes reveals that, while SVM provides competitive accuracy, it consumes more processing resources, and Naive Bayes, while faster, may impair accuracy due to its independence assumption.

Table 1. Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall
Naïve Bayes	0.83	0.80	0.81
Logistic Regression	0.85	0.83	0.84
Support Vector Machine (SVM)	0.88	0.86	0.87

6. Conclusion

This research presented a new framework, quick Risk-Aware Identity and Access Management (PR-IAM), to improve the security of generative AI systems by combining quick semantic analysis and identity-based access control methods. The suggested technique solves the shortcomings of standard IAM systems, which are unable to assess the intent and risk associated with natural language inputs. The PR-IAM architecture improves access control precision and context by adding machine learning-based risk assessment and policy-driven decision-making. The experimental results show that the suggested method successfully minimizes security concerns, improves misuse detection, and strengthens governance in AI-driven environments. The usage of models like Random Forest and Logistic Regression supports the potential of merging predictive analytics with IAM systems to provide intelligent access control. Several enhancements can be considered in the future to improve the system. These include using deep learning models (such as transformers) for more advanced prompt understanding, including real-world industry datasets for large-scale validation, and developing autonomous AI agent governance mechanisms within Zero Trust systems. Furthermore, incorporating continuous monitoring and feedback loops might improve adaptive security capabilities. Overall, the PR-IAM architecture establishes a solid platform for developing safe, scalable, and intelligent access control systems for next-generation AI apps.

7. References

[1] Zhu, B., Mu, N., Jiao, J., & Wagner, D. (2024). Generative AI security: challenges and countermeasures. arXiv preprint arXiv:2402.12617.

[2] Huang, K., Narajala, V. S., Yeoh, J., Ross, J., Lambe, M., Raskar, R., ... & Hughes, C. (2026, February). A novel zero-trust identity framework for agentic AI: Decentralized authentication and fine-grained access control. In 2026 International Conference on AI x Data and Knowledge Engineering (AIxDKE) (pp. 98-101). IEEE.

[3] Grassi, P., Garcia, M., & Fenton, J. (2020). NIST digital identityguidelines.https://csrc.nist.gov/publications/detail/sp/800-63/3/final.

[4] Li, H., Feng, S., & Han, S. (2025, November). AI Agent Security: Vulnerability Analysis, Protective Measures and Challenges. In 2025 IEEE 6th International Conference on Computer, Big Data, Artificial Intelligence (ICCBD+ AI) (pp. 1-5). IEEE.

[5] Li, H., Feng, S., & Han, S. (2025, November). AI Agent Security: Vulnerability Analysis, Protective Measures and Challenges. In 2025 IEEE 6th International Conference on Computer, Big Data, Artificial Intelligence (ICCBD+ AI) (pp. 1-5). IEEE.

[6] Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1, pp. 23-24). Cambridge, MA, USA: MIT press.

[7] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[8] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[9] Ashish, V. (2017). Attention is all you need. Advances in neural information processing systems, 30, I.

[10] OpenAI, R. (2023). Gpt-4 technical report. arxiv 2303.08774. View in Article, 2(5), 1.

[11] Russell, S. J. (2010). Artificial intelligence a modern approach. Pearson Education, Inc..

[12] Cannarsa, M. (2021). Ethics guidelines for trustworthy AI. The Cambridge handbook of lawyering in the digital age, 30, 97-283.

[13] AI, N. (2023). Artificial intelligence risk management framework (AI RMF 1.0). URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai, 100-1.

[14] Fasha, M., Rub, F. A., Matar, N., Sowan, B., Al Khaldy, M., & Barham, H. (2024, February). Mitigating the owasp top 10 for large language models applications using intelligent agents. In 2024 2nd International Conference on Cyber Resilience (ICCR) (pp. 1-9). IEEE.

[15] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[16] Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and. Techniques, Waltham: Morgan Kaufmann Publishers, 2012-13.