CYBER SECURITY

CyberSentinel: Network Intrusions, Insider Threats, and Ransomware Attacks

Aditya Kalbhor, Sanika Kalbhor, Shrinath Kalbhor, Priya Shimpi

School of Computing, MIT Art, Design and Technology University (MIT ADT), Pune, Maharashtra, India

Published November 2026

Abstract

The increasing complexity of cyber threats has highlighted the necessity of intelligent and adaptive intrusion detection systems designed to detect malicious activities across diverse computing environments. Traditional signature-based detection approaches face limitations in recognizing previously unseen attacks, while conventional machine learning and deep learning methods may encounter challenges related to data heterogeneity, temporal attack patterns, computational efficiency, and model interpretability. Recent research explores artificial intelligence techniques, multimodal behavioural analysis, temporal deep learning, dimensionality reduction, and explainable artificial intelligence (XAI) to overcome these challenges. Through the examination of network traffic, system logs, Internet of Things (IoT) telemetry, and user behavioural patterns, AI-driven approaches can enhance the detection and categorization of suspicious activities, including advanced persistent threats and complex network intrusions. Furthermore, explainability techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and attention-based interpretation offer explanations regarding the attributes influencing model predictions, thereby improving interpretability and assisting threat investigations. Despite these advancements, several unresolved issues persist in achieving reliable detection across heterogeneous environments, mitigating false positives, handling evolving attack strategies, and maintaining an appropriate trade-off between detection effectiveness and model transparency and computational requirements. Continued research into robust, explainable, and adaptive AI-based intrusion detection is consequently vital to enhancing cybersecurity resilience across enterprise networks, cloud infrastructures, and IoT environments.

Keywords: cyber sentinel; SHAP; artificial intelligence; intrusion detection; insider threat

1 Introduction

As cyberattacks become more sophisticated, organizations face growing difficulties in protecting their digital infrastructure. Traditional access-control mechanisms verify user permissions but do not necessarily assess whether an action is consistent with normal behaviour. However, authorization alone cannot establish whether an action is consistent with the user's normal behaviour. Consequently, Insiders misusing legitimate privileges and external attackers using compromised accounts may perform harmful activities while appearing to have legitimate access.

Artificial Intelligence (AI) and Machine Learning (ML) techniques offer potential solutions for identifying suspicious patterns in security data. Recent research has investigated machine-learning classifiers, deep-learning architectures, multimodal behavioural analysis, and temporal modelling to improve intrusion detection [1]. Furthermore, Explainable Artificial Intelligence (XAI) techniques have gained importance because security analysts require understandable evidence to interpret automated predictions and investigate potential threats.[2, 3].

Nevertheless, important issues persist in detecting insider misuse, reducing false alarms, and interpreting automated predictions, including the identification of insider misuse, the analysis of heterogeneous security data, the reduction of false-positive alerts, and the interpretation of model predictions [4]. Independent detection systems may generate multiple alerts for related activities, making it difficult for analysts to identify the underlying incident and prioritize their response [1, 5].

In response to these limitations, this study introduces the preferring Cyber-Sentinel framework**, an Explainable AI System for Intelligent Threat Detection and Response** [3, 6]. The architecture put forward introduces a Behavioural Context Engine that complements existing authorization mechanisms by evaluating user activities in relation to contextual information and historical behavioural patterns. The system is designed to investigate three threat categories: network intrusion, insider threats, and ransomware. Threat-specific machine-learning and deep-learning models are filed to analyse relevant security data, while a common event schema enables their outputs to be represented in a consistent format.

Moreover, Cyber-Sentinel aims to correspond parallel security events into combined occurences and provide an transparent event-specific risk evaluation, using SHapley Additive exPlanations (SHAP). An analyst dashboard is planned to present incident information and supporting evidence, enabling security analysts to make informed decisions while retaining human control over critical response actions.

This research aims to examine whether how behavioural analysis, threat-specific detection, event correlation, and explainability can be integrated into a unified architecture to support more transparent and effective cybersecurity investigations [1, 7–9].

2 Literature survey

Current study has explored several artificial intelligence techniques for improving intrusion detection, identifying behavioural anomalies, and explaining security-related predictions. The selected research papers collectively provide a foundation for understanding the opportunities and limitations of AI-driven cybersecurity systems [1, 9].

2.1 AI-based intrusion detection

Machine-learning and deep-learning approaches have been investigated for recognizing malicious network activities and discriminating attacks from legitimate traffic [1, 2, 5]. Conventional machine-learning classifiers can learn patterns from labelled network datasets, while deep-learning architectures can pinpoint more complex relationships in high-dimensional data [2, 5].

These approaches provide a substructure for automated threat classification. However, their effectiveness depends on the quality and comprehensiveness of the training data, feature selection, and evaluation procedures. Models trained on particular attack distributions may also face challenges when encountering earlier unseen threats or changing network conditions [2, 5].

For Cyber-Sentinel, these findings motivate the investigation of established machine-learning classifiers, including Random Forest and XGBoost, for network intrusion detection [1–3]. Their relevance will need to be evaluated experimentally using an apposite dataset, such as CIC-IDS2017.

2.2 Temporal and multimodal threat detection

Exploration into temporal deep learning and multimodal behavioural analysis call attention to the importance of examining relationships between events and analysing different types of security information. Network traffic, system logs, and user activities can provide complementary evidence about potentially spiteful behaviour [4, 8, 10].

Fleeting models, including Long Short-Term Memory (LSTM) networks, are designed to learn sequential dependencies. Other neural architectures can learn multiplex representations from behavioural or network data [2, 3].

Nevertheless, merging heterogeneous information sources presents challenges because different data types may have different structures, time intervals, and feature representations [1, 2, 5]. Analysing individual sources independently may also limit the ability to recognize relationships between seemingly unrelated security events.

Cyber-Sentinel labels this research direction through its suggested multi-source architecture. Network flow logs, user activity logs, and endpoint or system-call logs will be processed through appropriate detection pipelines. A common event schema will then standardize model outputs to hold up subsequent event correlation.

2.3 Insider-threat detection and behavioural analysis

Insider threats present a distinctive cybersecurity challenge because authorized users may embezzle legitimate privileges [9]. Permission-based security mechanisms can determine whether ingress is allowed but may not associate whether the resulting activity is suspicious in its behavioural context [1].

Behavioural analysis provides an alternative perspective by determining factors such as access frequency, login times, device usage, and interactions with sensitive resources. Deviations from established patterns may specify activities requiring further investigation [1].

However, an unusual activity does not necessarily stipulate vicious intent. Legitimate changes in employee responsibilities or working patterns may also cause anomalies. Therefore, behavioural detection requires applicable contextual information and careful evaluation to control false-positive alerts [1].

Cyber-Sentinel addresses a Behavioural Context Engine that supplements authorization information with contextual attributes, including time, device, location, resource sensitivity, and historical behavioural baselines. The CERT Insider Threat Dataset is proposed for building and evaluating the insider-threat detection component.

2.4 Explainable artificial intelligence in cybersecurity

Explainable Artificial Intelligence has become an important research direction because complex detection models may give rise to predictions that are difficult for analysts to interpret. Methods such as SHAP and LIME can provide explanations of model predictions by connecting features that influence individual outputs [5, 11].

SHAP estimates feature enhancements to a model prediction based on Shapley values from cooperative game theory. These explanations can help analysts understand why an activity was classified as dubious and link the attributes that contributed to a particular prediction [6, 12–14].

However, describing an individual model prediction does not automatically explain an entire security incident. When several alerts originate from related activities, analysts may need both feature-level explanations and situational information about the relationship between events [5, 11].

Cyber-Sentinel proposes using SHAP to support the elucidation of threat-specific model predictions and provide evidence for exposure potential assessments. The practical usefulness and reliability of these explanations will require assesment during execution [6, 13, 14].

2.5 Research synthesis and relevance to CyberSentinel

The selected papers inclusively highlight the importance of intelligent detection, behavioural analysis, temporal and multimodal information processing, feature representation, and interpretability in cybersecurity [2]–[5].

These research directions stimulate an integrated architecture that stretches isolated threat classification. Cyber-Sentinel proposes bridging three threat-specific detection components through a familiar event schema and a correlation layer, with the objective of consolidating related alerts into incidents that analysts can investigate.

The core investigation opportunity is to merge behavioural context, multiple threat-detection components, cross-event correlation, and explainable risk assessment within a unified architecture. The effectiveness of this unification remains to be established through implementation and experimental evaluation.

Table 1. Summary of selected studies on AI-based intrusion and threat detection.

Sr.no.	Author	Title	Limitation	Future Scope
1.	Adel Alshamrani / 2026	Advanced persistent threat detection through multi-modal behavioral analysis	Performance drops on external real-world APT traces; CERT is synthetic; limited organizational diversity/adversarial evaluation	More real-world data, adversarial robustness, federated learning, real-time adaptation and automated response
2.	Divine Ezeagwuna / 2026	AI-Driven Intrusion Detection Systems for Cybersecurity	Adversarial attacks, privacy, computational complexity and model interpretability	Explainable AI, federated learning, edge intelligence and predictive threat detection
3.	Anurag Jain; Abhirup Khanna; Amit Pimpalkar; Bhupesh Kumar Dewangan; Nilesh Shelke / 2026	An explainable multimodal temporal deep learning framework for intrusion detection (EMT-IDNet) in IoT environments	Focused on IoT benchmark data; explanation evaluation is mainly visualization/attention-based; real-time edge validation remains open	Asynchronous multimodal data, adversarial robustness, lightweight edge deployment and stronger quantitative XAI evaluation
4.	Adel Alabbadi; Fuad Bajaber / 2025	An Intrusion Detection System over the IoT Data Streams Using eXplainable Artificial Intelligence (XAI)	System-specific attack coverage and extremely high benchmark scores may not generalize; real-time deployment remains to be validated	Broader datasets, online/streaming adaptation, stronger cross-domain validation and operational response integration

Sr.no.	Author	Title	Limitation	Future Scope
5.	Bokang Motala; Swelihle Nosipho Ngubane; Likhona Yedwa / 2025	Artificial Intelligence in Cybersecurity: AI-Driven Threat Detection and Response Systems	Methodological inconsistency, dataset availability, computational/ deployment cost and skills gaps	Broader cross-sector validation; standardized datasets/metrics; practical human-AI response studies

3 Methodology

The conveyed Cyber-Sentinel methodology consists of data collection, pre-processing, feature engineering, threat-specific detection, contextual analysis, event normalization, alert correlation, risk assessment, explainability, and dashboard presentation.

3.1 Data collection

The system is designed to execute three categories of security information:

- **Network flow logs:** Network communication features used to acknowledge doubtful traffic and potential intrusions.
- **User activity logs:** Records of employee activities used to concede potential insider misuse.
- **Endpoint and system-call logs:** System-level events that may help pinpoint ransomware behaviour and other sceptical endpoint activities.

Public datasets are put forward for initial development and evaluation. CIC-IDS2017 will be investigated for network intrusion detection, the CERT Insider Threat Dataset for insider-threat detection, and an particular ransomware behavioural dataset for ransomware detection[[4, 10, 12, 15].

Because these datasets may distinguish in format, labelling, and collection conditions, each dataset will be scrutinize and exercised according to the requirements of its corresponding detection task.

3.2 Data preprocessing and feature engineering

The gathered datasets will undergo pre-processing to make them appropriate for machine-learning and deep-learning models. Bound to the dataset, this stage may include:

- Managing missing, illegitimate, or incorrect values.
- Recognizing and managing repealed records.
- Encoding categorical article.
- Scaling numerical attributes where required.
- Selecting relevant characteristics.
- Examining class variance.
- Drawing up labels for supervised learning.
- Organizing records into sequences when transitory modelling is required.

Training and testing data will be distinguished carefully to minimize data leakage. Where the data contains fleeting dependencies or related records, the splinter strategy will be selected to avoid overly optimistic evaluation.

3.3 Threat-specific detection models

Cyber-Sentinel proposes different detection pipelines for its three target threat categories.

1. **Network intrusion detection:** Random Forest and XGBoost are put forward as candidate classifiers for identifying normal and vicious network traffic. Their performance will be determined using suitable evaluation metrics before picking a model [11].
2. **Insider-threat detection:** Behavioural characteristics will be derived from employee activity records. A suitable classification or anomaly-detection approach will be evaluated to point out activities that fall apart significantly from established behavioural patterns [5].
3. **Ransomware detection:** Endpoint and system-level behavioural information will be examined to recognize potentially hostile activity. Candidate approaches may include a Multilayer Perceptron (MLP), an Autoencoder, or an LSTM network, hinging on the available dataset, feature structure, and provisional characteristics [5].

The suggested models are candidates rather than finalized selections. Their suitability will be determined through implementation and experimental evaluation.

3.4 Behavioural context engine

The Behavioural Context Engine is designed to accompaniment existing authorization mechanisms. Instead of addressing only whether a user is permitted to perform an action, it considers whether the activity is appropriate with the surrounding context.

The presenting contextual attributes include:

- Time of activity.
- Device information.
- Location information, where available.
- Authorization status.
- Sensitivity of the accessed resource.
- Historical behavioural baseline.

For example, an employee acquiring a sensitive resource at an unusual time may require further investigation if the activity also vary significantly from their established pattern. However, the system must consider the existing facts rather than treating any single documentation as proof of an attack.

3.5 Common event schema

The outputs of the threat-specific detection components may vary in structure and meaning. Cyber-Sentinel therefore suggests a common event schema to constitute their outputs consistently.

Hinging on data availability, the schema may contain:

- Event identifier and timestamp.
- Threat category.
- Relevant source or entity identifier.
- Predicted class.
- Model confidence or anomaly score, where supported.
- Contextual attributes.
- Relevant evidence or feature information.

This standardized rendition is intended to support downstream event correlation. Since model scores may have different statistical meanings, they must not be treated as directly distinguishable without pertinent calibration or normalization.

3.6 Cross-event correlation and incident risk scoring

The correlation component is designed to pinpoint related security events using proper contextual relationships, such as timestamps, shared entities, event sequences, and applicable system or user information.

Related events may then be grouped into a consolidated incident, lowering the need for analysts to investigate each alert separately.

A severity score will be investigated using apposite information from the correlated events and their contextual attributes. The scoring method will require explicit design and validation before its outputs can be labelled as reliable risk assessments.

The objective is to systematize incidents for investigation rather than to make unsubstantiated claims about an attack's cause or severity.

3.7 Explainability using SHAP

SHAP will be investigated to brief the predictions of compatible trained models. It estimates how single input attribute contribute to a model's prediction relative to an appropriate groundline [6, 12–14].

For example, the clarification for a network alert may point out traffic-related features that contributed to a malicious classification. An insider-threat prediction may instead be influenced by behavioural characteristics derived from employee activity [5].

The resulting simplifications can be presented alongside the relevant incident information to help analysts understand the evidence behind an alert.

SHAP explanations will not, by themselves, establish causation or guarantee that a prediction is correct. Their usefulness will depend on the quality of the trained model and the data supplied to it [6, 12–14].

3.8 Evaluation

After execution, each detection component will be evaluated individually before the unsegregated architecture is assessed.

The suggested evaluation metrics include:

- Precision.
- Recall.
- F1-score.
- False-positive rate.
- Confusion matrix.
- Processing time, where relevant.

The incident-correlation component may additionally be assessed using the correctness of event grouping, while the explainability component will require appropriate criteria for examining clarification usefulness and consistency [4].

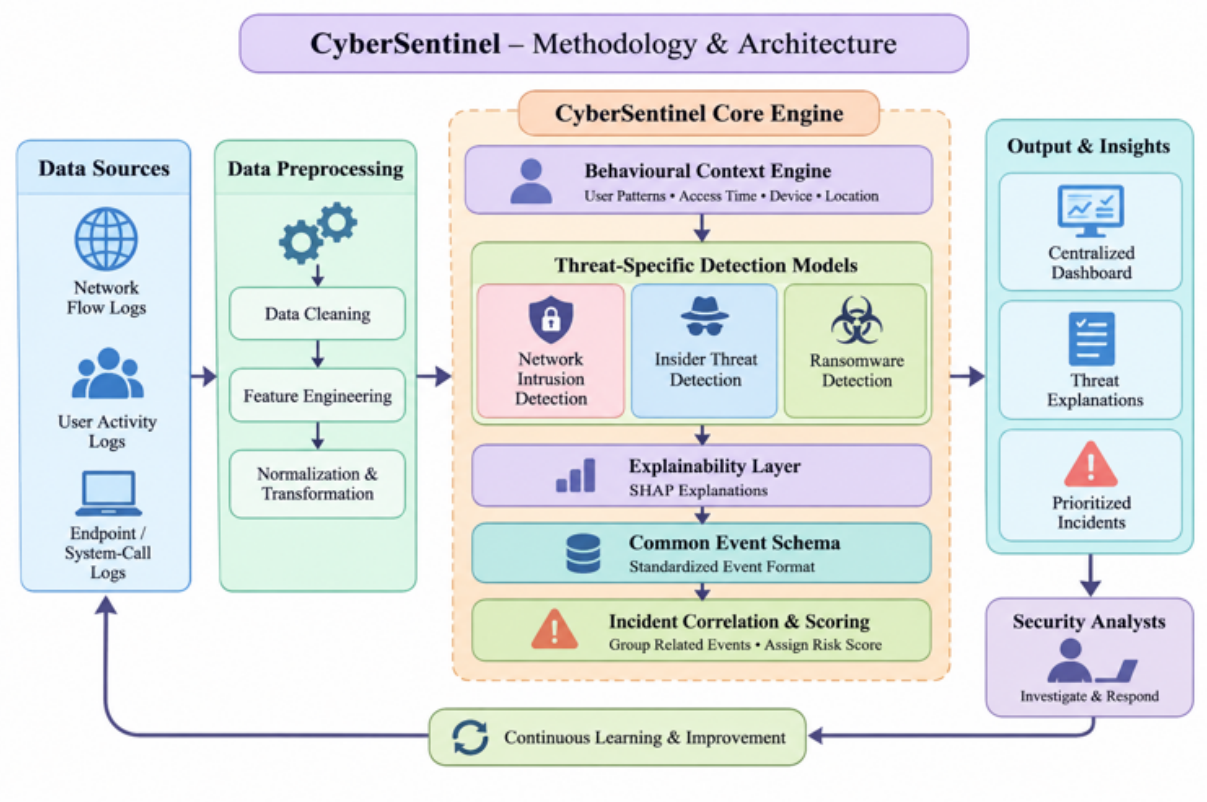

Figure 1. Proposed architecture and methodology of CyberSentinel.

4 Proposed solution

Cyber-Sentinel put forwards an AI-based intrusion detection and response architecture that combines behavioural context, threat-specific detection, cross-event correlation, and accountable risk assessment [4, 12, 13]. The system is intended to complement existing IAM and RBAC mechanisms without requiring changes to the organization's authorization infrastructure.

4.1 Architectural overview

The suggested architecture consists of the following stages:

1. **Data acquisition:** The system receives network flow logs, user activity records, and endpoint or system-call logs from appropriate data sources.
2. **Pre-processing and feature engineering:** The incoming records are consistent, modified, and converted into articles suitable for their correlated detection models.
3. **Behavioural Context Engine:** Contextual attributes, as well as time, device, location, authorization status, resource sensitivity, and historical behavioural patterns, are reviewed when assessing potentially suspicious activities.
4. **Threat-specific detection:** Individual detection components analyse the obtainable data for network intrusion, insider threats, and ransomware [12].

5. **Common event schema:** The outputs are transformed into a standardized event representation to ease further processing.
6. **Correlation and incident risk scoring:** Related events are unified into incidents, and an accountable risk assessment is generated using a method that must be implemented and validated.
7. **Explainability and analyst dashboard:** Relevant model explanations and incident evidence are hand over to the Security Operations Centre (SOC) analyst to support investigation and response decisions.

4.2 Detection beyond authorization

Conventional access-control systems determine whether an action satisfies established permissions. Cyber-Sentinel proposes extending this assessment by considering behavioural and contextual evidence.

For example, an employee may possess legitimate access to confidential files but suddenly begin accessing an unusually large number of sensitive resources. The proposed system would assess this deviation alongside other available contextual information to determine whether the activity warrants investigation.

This approach aims to recognize potentially wary behaviour that permission-based mechanisms alone may not recognize.

4.3 Multiple threat-detection components

Cyber-Sentinel is designed to address three complementary threat categories.

Table 2. Threat categories, data sources and candidate detection approaches.

Threat category	Proposed data source	Candidate approach
Network intrusion	Network flow records	Random Forest or XGBoost
Insider threat	Employee activity records	Behavioural anomaly detection and suitable classification models
Ransomware	Endpoint and system-behaviour records	MLP, Autoencoder, or LSTM, depending on the data

The candidate models will be evaluated separately. The final model selection will depend on the characteristics of the available datasets and the observed experimental results.

4.4 Unified incident correlation

Independent detection components may produce numerous alerts for activities associated with the similar underlying incident. Cyber-Sentinel proposes a correlation mechanism that inspects relationships between events and consolidates relevant alerts into an individual incident record.

For example, unusual employee file access and unsure outgoing network traffic might be unified for investigation when their timestamps, entities, and other contextual features provide sufficient evidence of a relationship.

This process is intended to offer analysts with a more coherent incident view and reduce fragmented alert handling. Correlation rules must be validated to avoid integrating unrelated events or overlooking relevant relationships.

4.5 Explainable incident assessment

Cyber-Sentinel proposes incorporating SHAP explanations into its investigation workflow. These explanations can show which input features influenced a model's prediction, providing supporting evidence for an alert [6, 12–14].

The proposed incident assessment will combine relevant contextual information with the outputs of the detection components. A suitable risk-scoring method must be developed to determine how these signals contribute to incident prioritization.

The system will distinguish between model predictions, explanatory evidence, and the final incident assessment so that analysts can interpret the results appropriately [13, 14].

4.6 SOC analyst dashboard and response

A unified analyst interface will present incident details, contextual evidence, and model explanations in one location for reviewing detected threats and their supporting evidence. Depending on implementation, it may display:

- Incident identifiers and timestamps.
- Threat categories.
- Risk-priority levels.
- Related events and contextual information.
- Relevant feature-level explanations.
- Suggested investigation or response steps.

The SOC analyst will review the evidence and determine the appropriate response. Critical actions will remain under human control rather than being automatically executed solely on the basis of a model prediction.

4.7 Scope and feasibility

The proposed architecture can be developed using established open-source machine-learning frameworks, publicly available datasets, and standard dashboard technologies.

However, feasibility depends on the availability and suitability of the datasets, computational requirements, pre-processing effort, and the complexity of integrating the detection components. In particular, the three threat categories require different features and labels, so their integration must be designed carefully.

At the current stage, Cyber-Sentinel is a prototype and proposed architecture. The full detection pipeline, model training, explainability, incident correlation, and dashboard integration remain subject to implementation and validation.

5 Expected outcomes

The framework aims to establish a broader and context-sensitive approach to cybersecurity threat detection by integrating network traffic analysis, user activity monitoring, and endpoint-level information. By examining multiple sources of security data, the framework aims to improve the visibility of suspicious activities that may not be identified through conventional permission-based security mechanisms alone.

The consolidation of threat-specific machine learning models is intended to support the recognition of network intrusions, insider threats, and ransomware-related activities [3, 7, 16]. The Behavioural Context Engine is presumed to provide supplemental context by examining factors such as user activity patterns, access time, device information, location, and resource sensitivity. This may help differentiate legal operations from potentially unusual behaviour, including activities performed through authorized accounts.

To make threat predictions easier to understand, the framework is designed to use SHAP to recognize the factors that influence the output of its detection models [6, 12–14]. The Common Event Schema and incident linking process are intended to organize related security alerts into consolidated incidents, while the proposed severity rating framework aims to assist in prioritizing events according to their assessed risk levels.

A consolidated security dashboard is expected to provide analysts with a unified view of detected events, associated explanations, and incident priorities. This may hold up more efficient security investigations, improve threat visibility, and help lower the effort required to survey multiple independent alerts.

Altogether, Cyber-Sentinel aims to set up a footing for explainable, behaviour-aware, and integrated cybersecurity monitoring. The verified detection performance, reliability, scalability, effectiveness of incident prioritization, and reduction in false-positive alerts will be set on through future execution and experimental evaluation using satisfactory cybersecurity datasets [4].

6 Advantages

The proposed Cyber-Sentinel structure offers several potential lead for intelligent and explainable cybersecurity monitoring:

- **Multi-Source Threat Analysis:** The structure is designed to evaluate network flow logs, user activity logs, and endpoint or system-call logs, assuming a broader perspective on security-related activities [1].
- **Behaviour-Based Anomaly Detection:** By encompassing a Behavioural Context Engine, Cyber-Sentinel targets to recognize deviations from established user activity patterns, including suspicious actions performed through authorized accounts [1].
- **Detection of Multiple Threat Categories:** The proposed architecture includes separate perception pipelines for network intrusions, insider threats, and ransomware-related activities, carry the analysis of different categories of cybersecurity threats [5].
- **Improved Model Interpretability:** The arranged addition of SHAP explanations is intended to point out the characteristics influencing individual model predictions, making security analysts understand why an activity may have been classified as vicious [5].

- **Centralized Event Representation:** The Common Event Schema is designed to normalize the outputs of various detection models into a accordant format, making security events easier to organize and analyse.
- **Cross-Event Correlation:** The postulated correlation mechanism targets to associate related security alerts and unite them into blended incidents, rather than requiring analysts to investigate every alert independently.
- **Risk-Based Incident Prioritization:** The incident-scoring process is intended to assign risk-based priority levels to corresponding events, helping analysts focus their attention on likely reproving incidents.
- **Centralized Security Monitoring:** The arranged panel will provide a consolidated interface for observing detected threats, model explanations, and incident priorities, supporting security investigation and monitoring activities.
- **Compatibility with Existing Access-Control Infrastructure:** CyberSentinel is made to correlate enduring Identity and Access Management (IAM) and Role-Based Access Control (RBAC) process rather than substitute them, allowing behavioural analysis to be introduced as an additional security layer [1, 7–9].
- **Support for Human-Led Security Decisions:** The model intends to provide security analysts with situational information, descriptions, and categorized incidents while keeping critical response decisions under manual human intervention [3, 4, 6, 12].

7 Future scope

Future development of Cyber-Sentinel could traverse the combination of modern threat intelligence sources to provide additional insights into unfolding attack techniques and developing cyber threats. The structure may also be extended to incorporate amalgamate learning, allowing organizations to associatively improve threat detection framework without directly linking sensitive security data [6, 11, 12]. Moreover, investigating privacy-preserving techniques and adversarial robustness could reinforce the reliability of the system in sophisticated security infrastructure. Another potential direction is the enlargement of interactive investigation capabilities that allow security analysts to explore relationships between incidents, understand attack progression, and examine the ratiocinate behind model predictions. These ameliorations could help evolve Cyber-Sentinel into a more versatile, collaborative, and authentic cybersecurity monitoring framework [2, 3].

8 Conclusion

This paper presents Cyber-Sentinel, a conceptual interpretable AI-based system for intelligent intrusion detection and mitigation. Motivated by limitations in permission-based security and isolated threat-detection approaches, the proposed architecture combines behavioural context, threat-specific detection, standardized event representation, cross-event correlation, and explainable incident assessment [5, 6, 10, 13].

The platform is designed to investigate network breaches, potential insider threats, and ransomware utilizing suitable public datasets and candidate machine-learning or deep-learning approaches. A Situational Behavioural Engine is outlined to assess activities beyond their authorization status, while a common event schema and correlation component aim to organize related alerts into consolidated incidents. SHAP is include to deliver feature-level explanations that assist analysts understand model

predictions, and an analyst dashboard will support evidence-based investigation while preserving human oversight regarding vital response actions [6, 12–14].

The suggested methodology is motivated by studies into AI-based threat detection, source evaluation and multimodal analysis, behavioural anomaly detection, and explainable AI [11]. Its primary goal is the consolidation of these capabilities into a cohesive incident-oriented framework instead of depending on standalone detection alerts [5, 6].

Because Cyber-Sentinel is presently at a stage, its detection performance, scalability, false-positive reduction, incident-correlation accuracy, and explanation quality have been established experimentally but needs verification through more testing [4]. Future work will focus on dataset preparation, model development and assessment, creation of the unified event layout, execution of incident linking and risk qualification, and incorporation of SHAP clarifications into the interface. The finished platform will subsequently be tested utilizing suitable identification benchmarks and incident-level validation procedures.

Ultimately, Cyber-Sentinel seeks to examine whether combining behavioural context, multiple detection components, event correlation, and transparency can deliver a more interpretable and useful approach to cybersecurity threat analysis.

References

[1] A. Alshamrani, "Advanced persistent threat detection through multi-modal behavioral analysis," *PLOS ONE*, vol. 21, no. 6, Art. no. e0349607, 2026, doi: 10.1371/journal.pone.0349607.

[2] D. Ezeagwuna, "AI-driven intrusion detection systems for cybersecurity," *Journal of Science*, Technology and Social Transformation, vol. 2, no. 2, pp. 37–51, 2026, doi: 10.64235/ntebwf71.

[3] B. Motala, S. N. Ngubane, and L. Yedwa, "Artificial intelligence in cybersecurity: AI-driven threat detection and response systems," *SSRN*, Nov. 2025, doi: 10.2139/ssrn.5647571.

[4] S. J. Moore, F. Cruciani, C. D. Nugent, S. Zhang, I. Cleland, and S. Sani, "Deep learning for network intrusion: A hierarchical approach to reduce false alarms," *Intelligent Systems with Applications*, vol. 18, Art. no. 200215, 2023, doi: 10.1016/j.iswa.2023.200215.

[5] H. Alamro et al., "Enhanced intrusion detection in cybersecurity through dimensionality reduction and explainable artificial intelligence," *Scientific Reports*, vol. 15, Art. no. 33848, 2025, doi: 10.1038/s41598-025-06761-9.

[6] G. Nalinipriya, S. Rama Sree, K. Radhika, E. Laxmi Lydia, F. K. Karim, M. K. Ishak, and S. M. Mostafa, "Leveraging explainable artificial intelligence for early detection and mitigation of cyber threat in large-scale network environments," *Scientific Reports*, vol. 15, Art. no. 24662, 2025, doi: 10.1038/s41598-025-08597-9.

[7] F. R. Alzaabi and A. Mehmood, "A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods," *IEEE Access*, vol. 12, pp. 30907–30927, 2024, doi: 10.1109/ACCESS.2024.3369906.

[8] M. Du, F. Li, G. Zheng, and V. Srikumar, "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," in *Proc*. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2017, pp. 1285–1298, doi: 10.1145/3133956.3134015.

[9] CERT Division, Software Engineering Institute, Carnegie Mellon University, "Insider threat test dataset," *Carnegie Mellon University*, 2016, doi: 10.1184/R1/12841247.v1.

[10] N. Al-Nomasy, A. Alamri, A. Aljuhani, and P. Kumar, "Transformer-based knowledge distillation for explainable intrusion detection system," *Computers & Security*, vol. 154, Art. no. 104417, 2025, doi: 10.1016/j.cose.2025.104417.

[11] A. Alabbadi and F. Bajaber, "An intrusion detection system over the IoT data streams using eXplainable Artificial Intelligence (XAI)," *Sensors*, vol. 25, no. 3, Art. no. 847, 2025, doi: 10.3390/s25030847.

[12] O. Arreche, T. Guntur, and M. Abdallah, "XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," *Applied Sciences*, vol. 14, no. 10, Art. no. 4170, 2024, doi: 10.3390/app14104170.

[13] S. Patil, V. Varadarajan, S. M. Mazhar, A. Sahibzada, N. Ahmed, O. Sinha, S. Kumar, K. Shaw, and K. Kotecha, "Explainable artificial intelligence for intrusion detection system," *Electronics*, vol. 11, no. 19, Art. no. 3079, 2022, doi: 10.3390/electronics11193079.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017, doi: 10.48550/arXiv.1705.07874.

[15] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc*. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116, doi: 10.5220/0006639801080116.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324

[17] M. Pawlicki, A. Pawlicka, R. Kozik, and M. Choraś, "The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation," *Artificial Intelligence Review*, 2024, doi: 10.1007/s10462-024-10972-3.

[18] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112392–112415, 2022, doi: 10.1109/ACCESS.2022.3216617.

[19] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, Art. no. 15, 2009, doi: 10.1145/1541880.1541882.