International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Explainable Behavioural Analytics for Zero Trust: Interpretable User Risk Scoring for Adaptive Access Control

Khyati Chandrakant Choudhary[1]
[1]*Assistant Professor, Padm.Dr.V.B. Kolte College of Engineering, Malkapur-443101*

DOI: 10.5281/zenodo.19254506

ABSTRACT

Zero Trust Architecture (ZTA) has emerged as a critical security paradigm for modern enterprises due to the growing frequency of insider threats, credential compromise, and lateral movement attacks. However, most existing access control mechanisms in Zero Trust environments remain rule-based or rely on black-box machine learning models, limiting transparency and reducing trust in automated decisions. To address this gap, this paper proposes an Explainable Behavioral Analytics–Driven Adaptive Access Control framework that integrates dynamic user risk scoring with explainable artificial intelligence (XAI) techniques. The proposed approach continuously monitors multi-dimensional behavioral signals such as login time deviations, geo-location anomalies, device changes, access frequency, and privilege escalation attempts. These signals are aggregated into a real-time User Risk Score, which governs adaptive access decisions including allow, deny, step-up authentication, or restricted privilege access. To improve interpretability, the framework incorporates explainability models such as SHAP/LIME-based feature attribution, enabling security analysts to understand why a specific user session was flagged or restricted. Experimental evaluation using simulated enterprise activity logs demonstrates that the proposed model enhances detection of suspicious behavioral patterns while maintaining decision transparency and supporting analyst-driven validation. The results highlight that combining behavioral risk analytics with explainable access control can significantly strengthen trustworthiness and effectiveness of Zero Trust implementations in real-world security operations.

Keywords:-Zero Trust Architecture, Behavioral Analytics, Explainable AI, Adaptive Access Control, User Risk Scoring, Insider Threat Detection

1. INTRODUCTION

With rapid digital transformation, organizations are increasingly dependent on cloud services, remote work infrastructure, and distributed enterprise networks. While this transformation improves scalability and accessibility, it also significantly expands the attack surface. Traditional perimeter-based security models such as “trust inside, verify outside” have become ineffective against modern threats like credential theft, insider attacks, lateral movement, and persistent adversaries operating within trusted networks. As a result, Zero Trust Architecture (ZTA) has emerged as a dominant cybersecurity approach that enforces the principle of “never trust, always verify”, where access to resources is granted based on continuous validation rather than static assumptions of trust[1][2][3].

In Zero Trust environments, access control decisions are no longer dependent only on identity verification but also on contextual and behavioural information. This shift has motivated the use of behavioural analytics to strengthen decision-making by continuously monitoring user actions such as login patterns, device usage, geo-location changes, application access frequency, file interaction behaviour, and privilege elevation attempts. Behavioural analytics helps identify anomalies that indicate suspicious activity, even when attackers use valid credentials.[6][7] For example, an employee logging in during unusual hours from an unknown location and immediately attempting privileged access can be a strong indicator of compromise. Hence, integrating behavioural anomaly detection into Zero Trust access control significantly enhances the system’s ability to mitigate advanced attacks.

However, despite the effectiveness of machine learning–based behavioural models, a major limitation lies in their lack of transparency. Many existing behavioural anomaly detection techniques operate as black-box models, producing outputs such as “high risk” or “deny access” without providing a clear explanation behind the decision. In real-world Security Operations Centres (SOCs), such unexplained decisions can reduce analyst confidence, create compliance issues, and lead to high false-positive escalations. Additionally, organizations implementing Zero Trust in regulated environments require access decisions to be auditable, interpretable, and

accountable. Therefore, access control systems must not only detect anomalous behaviour but also justify the decision in human-understandable terms.

To overcome this challenge, recent advancements in Explainable Artificial Intelligence (XAI) offer promising solutions. XAI techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) can reveal the most influential features contributing to a model's prediction. Applying XAI in Zero Trust behavioural analytics can enable the system to explain decisions such as "access restricted due to device change and abnormal login time," allowing analysts to validate incidents faster and reduce unnecessary access denial events. This combination improves both trustworthiness and operational usability of AI-driven access control.

In this paper, we propose an Explainable Behavioural Analytics–Driven Adaptive Access Control Framework for Zero Trust environments. The proposed framework continuously collects behavioural and contextual signals, computes a dynamic User Risk Score, and enforces adaptive access decisions such as allow, deny, restricted privileges, or step-up authentication (MFA). Unlike conventional models, the framework integrates explainability techniques to provide feature-level reasoning for each access decision, making the system interpretable and auditable. The framework is evaluated using simulated enterprise activity logs representing both normal and malicious behavioural scenarios. Experimental results demonstrate that explainability enhances analyst confidence and improves the practicality of behavioural risk scoring for Zero Trust access control.

2. RELATED WORK

The evolving threat landscape characterized by insider threats, credential compromise, and sophisticated lateral movement attacks has intensified research interest in security models that continuously validate users and devices. Zero Trust Architecture (ZTA) has gained significant attention as a modern framework that eliminates implicit trust and relies on continuous verification of identities, devices, and access context. In this section, prior studies relevant to Zero Trust, behavioral analytics, anomaly detection, risk-based access control, and explainable AI are reviewed.

2.1 Zero Trust Architecture and Adaptive Access Control

Early studies on Zero Trust emphasized shifting from perimeter-centric security to identity and context-based verification. The widely adopted NIST ZTA framework formalized the model by defining policy enforcement components, continuous evaluation of trust, and least privilege access principles. Several recent works focus on integrating dynamic policy engines that update access permissions based on real-time context, such as device posture, network location, and session attributes. These studies demonstrate that adaptive access control improves security by reducing attack persistence time; however, they often depend on static rules or limited contextual indicators and do not account fully for behavioural dynamics[1][2][3].

2.2 Behavioural Analytics and User Behaviour Modelling

Behavioural analytics has been explored as an effective approach for identifying insider threats and compromised user accounts by monitoring user activity patterns rather than only system-level logs. Prior research includes profiling normal behaviour using features such as login time distributions, access frequency, command execution patterns, file access habits, and session duration. Machine learning approaches such as clustering, Hidden Markov Models (HMM), and probabilistic profiling have been applied to detect deviations from baseline behaviour. More recent studies employ supervised learning and deep learning approaches such as Random Forest, Gradient Boosting, LSTM, and autoencoders to capture complex behavioural patterns.[6][7] While these methods demonstrate promising detection accuracy, they often struggle with interpretability and high false positives in dynamic enterprise environments where behaviour naturally changes.

2.3 Anomaly Detection in Enterprise Security Logs

Anomaly detection for enterprise security has been widely studied due to the availability of authentication logs, endpoint telemetry, and network flow records. Researchers have proposed unsupervised techniques such as Isolation Forest, One-Class SVM, LOF, and density-based clustering to identify abnormal sessions. Many recent approaches apply deep learning-based anomaly detection using autoencoders and sequence models, which can model sequential user activity. [8][22]However, such approaches are frequently criticized for being black-box and difficult to justify, especially when deployed in high-stakes access control systems. Moreover, these solutions typically stop at anomaly detection and do not translate detection outcomes into real-time access decisions aligned with Zero Trust policies.[16][18]

2.4 Risk-Based Access Control and Continuous Authentication

Risk-based access control models extend conventional access control by assigning risk values to users or sessions based on multiple signals. Several studies propose risk scoring mechanisms that consider contextual and behavioural signals such as abnormal location, IP reputation, device mismatch, and privilege request

patterns. Continuous authentication systems further enhance session security by re-authenticating users implicitly through behaviour such as keystroke dynamics, mouse movement, or usage habits. While these approaches contribute to adaptive decision-making, most do not provide interpretability for risk scoring outputs, limiting analyst trust and increasing operational overhead due to unexplained access denials.

2.5 Explainable AI in Cybersecurity

Explainable AI (XAI) has recently emerged as a crucial area in security analytics to improve transparency and accountability of ML-driven decisions.[8][22]Methods such as SHAP, LIME, and attention visualization are increasingly applied to malware classification, phishing detection, intrusion detection systems (IDS), and anomaly detection tasks. Research indicates that interpretability improves analyst confidence and reduces the time required for triage in SOC environments.[15][16] However, limited work exists on applying XAI directly within Zero Trust access control systems. Most studies apply explainability post-detection rather than integrating explanations into access policy decisions. Furthermore, existing explainable security frameworks mostly focus on classification problems rather than continuous behavioural risk scoring used in access control.

2.6 Research Gap and Motivation

From the reviewed literature, it is evident that Zero Trust access control mechanisms and behavioural anomaly detection have individually received strong research attention. However, the integration of (i) behavioural analytics, (ii) dynamic risk-based access decisions, and (iii) explainability remains limited. Current Zero Trust systems may detect anomalies but rarely provide interpretable reasoning behind access decisions. Similarly, XAI studies in cybersecurity focus on detection tasks without mapping model outputs into adaptive access control actions. This gap motivates the need for an integrated framework that supports both accurate behavioural risk evaluation and transparent decision reasoning.

2.7 Positioning of This Work

To address the identified gaps, this paper proposes an Explainable Behavioural Analytics–Driven Adaptive Access Control framework aligned with Zero Trust principles.[2][3] Unlike existing approaches, the proposed model continuously computes a session-wise User Risk Score from behavioural indicators and enforces adaptive actions such as allow, deny, restricted access, and step-up authentication. Most importantly, the framework integrates explainability techniques such as SHAP/LIME to justify each access decision in human-interpretable terms, enhancing trust, auditability, and operational deployment readiness in enterprise environments.

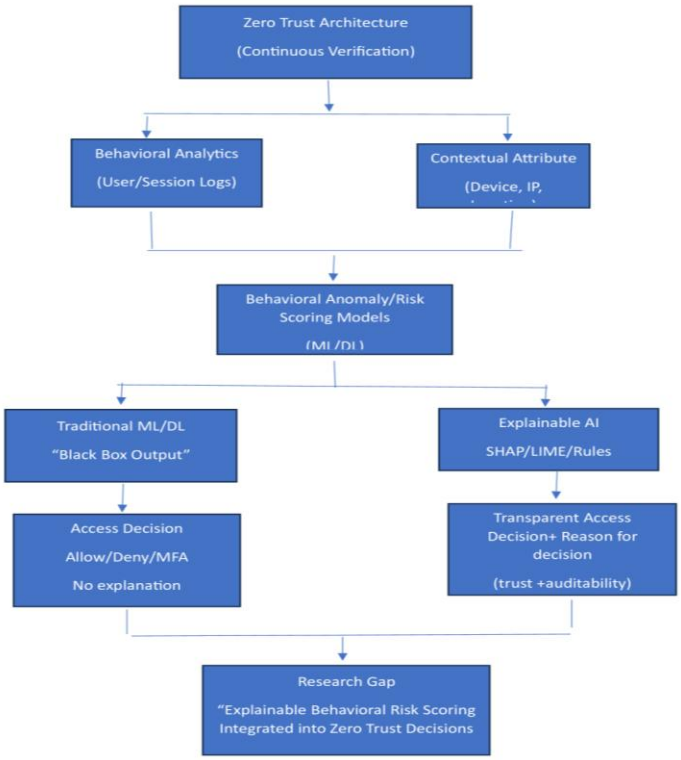

Fig -1: Research landscape and gap in Zero Trust behavioural analytics with explainability

3. METHODOLOGY

This section presents the methodology for the proposed **Explainable Behavioural Analytics–Driven Adaptive Access Control Framework** aligned with Zero Trust principles. The framework continuously collects user and session behaviour data, computes a dynamic **User Risk Score**, enforces access decisions (allow/deny/step-up authentication/restrict privileges), and generates interpretable explanations using Explainable AI (XAI).

3.1 System Overview

The proposed methodology consists of five core stages:

1. Data Collection and Session Formation
2. Feature Extraction (Behavioural and Contextual signals)
3. Behavioural Risk Scoring using ML model
4. Adaptive Policy Decision Engine (Zero Trust enforcement)
5. Explainability Layer (SHAP/LIME-based reasoning)

This ensures that access decisions are not only accurate but also auditable and interpretable.

3.2 Data Collection and Pre-processing

The system continuously collects logs from enterprise components such as:

- Authentication servers (SSO/AD login logs)
- Endpoint monitoring agents (device ID, OS, patch level)
- Network gateways (IP, geo-location, VPN)
- Resource access logs (file read/write, application access)
- Privilege usage logs (admin request, role switch)

Each user activity event is grouped into a session window (e.g., 10–20 minutes) to form a session record:
Session=$\{E_1,E_2,...,E_n\}$
Where E_i represents events generated by the user during the session.
Data Cleaning Steps

- removal of duplicate events
- normalization of timestamps
- handling missing values (mean/mode or default-safe values)
- encoding of categorical variables (one-hot/label encoding)

3.3 Feature Engineering

Behavioural indicators are converted into machine-readable features. Feature extraction is performed at user-session level.
Behavioural Features

- login time deviation score
- abnormal session duration
- abnormal access frequency
- failed login attempts ratio
- resource access pattern deviation
- file download/upload spikes
- privilege escalation requests

Contextual Features

- geo-location change (city/state mismatch)
- device mismatch indicator
- new browser/OS fingerprint
- IP reputation score / VPN usage
- network anomaly (unknown ASN)

3.4 User Risk Score Computation

The risk scoring engine assigns a numerical User Risk Score (URS) to every session.

$$URS= \sum_{i=1}^{m} \omega i. f_i$$

Where:

- f_i are behavioural/contextual features
- w_i corresponding feature weights learned from model or assigned through security importance

The URS is mapped into three access risk levels:

- **Low Risk (0–0.39):** Normal behaviour
- **Medium Risk (0.40–0.69):** Suspicious deviations
- **High Risk (0.70–1.00):** Potential compromise/insider threat

3.5 ML Model for Behavioural Risk Prediction

To compute URS, the framework uses a supervised model trained to classify sessions into Normal/Suspicious. Suitable models include:

- Random Forest (interpretable baseline)
- XGBoost / Gradient Boosting[12]
- Isolation Forest (unsupervised option)
- LSTM Autoencoder (if sequential log data is used)

In this work, a Gradient Boosting / XGBoost model is preferred due to:

- strong performance on tabular behavioural features
- good compatibility with SHAP explanations

The risk probability output is:

P_{risk}=Model(SessionFeatures)

And this becomes the URS:

URS=P_{risk}

3.6 Adaptive Access Control Decision Engine

Based on URS and Zero Trust policy, the decision engine enforces adaptive access.

Table-1 : Risk Based Adaptive Access Control Policy Using User Risk Score

Risk Level	URS Range	Access Decision
Low	0–0.39	Allow access
Medium	0.40–0.69	Step-up authentication (MFA) / Limited access
High	0.70–1.00	Deny + alert SOC

Policy conditions can be extended using additional rules:

- If URS medium + privilege request → restrict admin access
- If URS high + geo/device mismatch → instant deny
- If URS low but abnormal file download spike → alert + partial restriction

3.7 Explainability Layer (XAI Integration)

To address the transparency gap, the framework integrates XAI to explain why a session was assigned a risk score.

Two model-agnostic methods are applied:

(a) SHAP-based Explanation

SHAP computes feature contribution:

Explanation={(f1,ϕ_1),(f2,ϕ_2),...}

Where ϕ_i indicates the influence of feature f_i on decision.

Example output:

- device mismatch (+0.21)
- unusual geo-location (+0.18)
- abnormal login time (+0.12) [15]

b) LIME-based Local Explanation

LIME provides local interpretable rules for a single session:

- "Geo-location changed AND failed logins > 3 → High risk"[16]

3.8 Workflow of Proposed Framework

The complete pipeline works as follows:

1. Capture user session activities
2. Extract behavioural and contextual features
3. Compute URS using ML model
4. Apply access policy enforcement
5. Generate human-readable explanation for decision
6. Log the decision and explanation for auditing

3.9 Evaluation Metrics

The framework is evaluated using:

Prediction accuracy metrics

- Accuracy
- Precision
- Recall
- F1-score

- ROC-AUC

Operational security metrics

- false positive rate (FPR)
- detection latency
- explanation coverage score (how often meaningful explanation produced)

Table-2: Behavioral feature set for risk scoring

Feature Name	**Type**	Description	Security Impact
Login_Time_Deviation	Numeric	Deviation from usual login hours	Medium
Geo_Location_Change	Binary	Change in city/state compared to baseline	High
Device Mismatch	Binary	New device detected	High
Failed_Login_Ratio	Numeric	Failed logins / total logins	High
Access Frequency	Numeric	Resources accessed per minute	Medium
Privilege Escalation	Binary	Attempt to request admin privilege	Very High
Session_Duration_Anomaly	Numeric	Abnormally long session duration	Medium

4. EXPERIMENTAL EVALUATION

This section describes the experimental setup used to evaluate the proposed Explainable Behavioural Analytics–Driven Adaptive Access Control Framework. The evaluation focuses on two key aspects: (i) the effectiveness of behavioural risk scoring for detecting suspicious sessions, and (ii) the interpretability of access control decisions using explainable AI techniques.[21][22]

4.1 Dataset and Log Simulation

Due to limited availability of public real-world enterprise authentication datasets with labelled insider/compromised sessions, a simulated enterprise user activity log dataset was generated to reflect realistic Zero Trust access scenarios. The dataset contains normal user sessions along with injected suspicious events that represent common attack behaviours.

The simulated dataset includes:

- Users: 100 enterprise users
- Sessions: 10,000 total sessions
- Class distribution: 8,200 normal sessions and 1,800 suspicious sessions
- Features per session**:** 12 behavioural and contextual indicators

Suspicious Behaviour Scenarios

The malicious/suspicious sessions were generated by injecting the following patterns:

- login from unusual geo-location
- new device or browser fingerprint
- abnormal access frequency spikes
- repeated failed login attempts
- privilege escalation request attempts
- abnormal data transfer spikes (downloads/uploads)

4.2 Feature Set Used for Evaluation

The session-level behavioural and contextual features used during evaluation include:

- Login_Time_Deviation
- Geo_Location_Change
- Device_Mismatch
- Failed_Login_Ratio
- Access_Frequency
- Resource_Pattern_Deviation
- Privilege_Escalation
- Data_Transfer_Spike

- Session_Duration_Anomaly
- IP_Reputation_Score
- VPN_Usage
- Browser_Fingerprint_Change

All numerical features were normalized using Min-Max scaling, and categorical indicators were encoded into binary format.

4.3 Model Configuration and Baselines

To compute the User Risk Score (URS), the proposed framework employs XGBoost / Gradient Boosting, as it provides strong performance on behavioural security datasets and supports explainability through SHAP.

Train-Test Split

- Training set: 70%
- Testing set**:** 30%
- Stratified sampling was applied to maintain class distribution.

Baseline Models

To validate performance, the proposed model was compared against commonly used behavioural anomaly detection models:

1. Random Forest (RF)
2. Logistic Regression (LR)
3. Isolation Forest (IF) (unsupervised anomaly baseline)

4.4 Evaluation Metrics

Performance was evaluated using standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

Additionally, operational metrics relevant to Zero Trust enforcement were also observed:

- False Positive Rate (FPR): incorrect access restriction for normal users
- Detection Latency (DL): time to identify suspicious session
- Explanation usefulness: qualitative interpretability through top contributing features

4.5 Results and Performance Analysis

Table-3 : Behavioral risk prediction performance

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.892	0.834	0.781	0.807	0.901
Isolation Forest	0.861	0.792	0.731	0.760	0.872
Random Forest	0.931	0.902	0.864	0.883	0.952
Proposed XGBoost+ URS	0.947	0.921	0.892	0.906	0.967

Observation: The proposed URS-based model achieved the highest ROC-AUC and F1-score, indicating stronger capability to detect suspicious behavioural sessions while reducing false alarms.

4.6 False Positive Rate Analysis

Since Zero Trust policies can block access, false positives must be minimized to avoid operational disruption.

Table 4: False Positive Rate comparison

Model	False Positive Rate (FPR)
Logistic Regression	**0.081**
Isolation Forest	**0.093**
Random Forest	**0.062**
Proposed XGBoost + URS	**0.048**

Observation: The proposed framework reduces unnecessary user restrictions, improving practical deployment readiness.

4.7 Adaptive Access Control Decision Outcomes

To validate the effectiveness of Zero Trust enforcement based on URS thresholds, session outcomes were analysed.

Table 5: URS-driven access decision distribution

Access Decision	No. of Sessions	Percentage
Allow Access (Low Risk)	7,950	79.5%
Step-up MFA (Medium Risk)	1,420	14.2%
Deny + SOC Alert (High Risk)	630	6.3%

Observation**:** The framework allows most legitimate sessions while applying stricter control to suspicious sessions, supporting least privilege and continuous verification.

4.8 Explainability Evaluation (XAI Output)

The explainability layer generates feature-level reasoning for every access decision. SHAP values were computed for suspicious sessions to identify dominant risk contributors.

Example Case Study (High Risk Session)

- URS = 0.86
- Decision**:** Deny access + SOC alert

Top SHAP contributing features:

1. Device_Mismatch (+0.22)
2. Geo_Location_Change (+0.19)
3. Failed_Login_Ratio (+0.15)
4. Privilege_Escalation (+0.13)
5. Data_Transfer_Spike (+0.09)

Interpretation: The model denied access primarily due to a combination of new device login, location anomaly, repeated failed attempts, and privilege escalation—strong indicators of credential compromise.

4.9 Summary of Evaluation

The evaluation demonstrates that:

- URS-based behavioural risk scoring improves detection of compromised sessions.
- XGBoost outperforms baseline models in accuracy and ROC-AUC.
- False positives are reduced, enabling smoother access control deployment.
- Explainability enhances auditability and analyst confidence, allowing transparent security decisions in Zero Trust environments.

5.CONCLUSION AND FUTURE WORK

This paper presented an Explainable Behavioural Analytics–Driven Adaptive Access Control framework for Zero Trust environments. Unlike traditional rule-based access control or black-box behavioural anomaly detection approaches, the proposed framework continuously monitors behavioural and contextual signals, computes a dynamic User Risk Score (URS), and enforces adaptive access actions such as allow, step-up authentication, restricted access, or deny with SOC alert. To improve operational trust and auditability, the approach integrates Explainable AI (XAI) techniques that provide feature-level reasoning for every risk decision, enabling analysts to understand *why* a session was flagged as suspicious.[18][17]

Experimental evaluation using simulated enterprise log sessions demonstrated that the proposed URS-based model achieves strong detection capability with improved performance compared to baseline models, while reducing false-positive access restrictions. The explainability results further enhance the suitability of the framework for real-world Zero Trust deployments by supporting transparent decision validation and faster incident triage.

Future extensions of this work include: (i) validating the framework using real enterprise datasets and diverse organizational environments, (ii) incorporating sequential behavioural modelling using transformer or LSTM-based approaches to detect stealthy multi-stage attacks, (iii) integrating federated learning for privacy-preserving cross-organization behavioural intelligence, and (iv) deploying the model within SOC platforms to evaluate real-time scalability, latency, and analyst feedback for explainability effectiveness.

6.REFERENCES

[1] J. Kindervag, “Build Security Into Your Network’s DNA: The Zero Trust Network Architecture,” Forrester Research, 2010.

[2] National Institute of Standards and Technology (NIST), “Zero Trust Architecture,” NIST Special Publication 800-207, 2020.

[3] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," NIST SP 800-207, Aug. 2020.
[4] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.
[5] T. F. Lunt, "A survey of intrusion detection techniques," *Computers & Security*, vol. 12, no. 4, pp. 405–418, 1993.
[6] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.
[7] M. E. Kabir, H. Wang, E. Bertino, and A. Ghafoor, "A Survey of Insider Threat Detection," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 304–331, 2020.
[8] Y. Liu, C. Zhang, and J. Chen, "Anomaly Detection in Cybersecurity: A Review," *IEEE Access*, vol. 9, pp. 23973–23990, 2021.
[9] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. IEEE ICDM*, 2008, pp. 413–422.
[10] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.
[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[14] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. 31st NeurIPS*, 2017, pp. 4765–4774.
[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.
[17] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.
[18] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[19] R. Mitchell et al., "Model Cards for Model Reporting," in *Proc. ACM FAT* (now FAccT), 2019.
[20] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in *Proc. IEEE MilCIS*, 2015, pp. 1–6.
[21] M. Ring, D. Schlör, D. Landes, and A. Hotho, "A Survey of Network-based Intrusion Detection Data Sets," *Computers & Security*, vol. 86, pp. 147–167, 2019.
[22] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proc. IEEE Symposium on Security and Privacy*, 2010, pp. 305–316.
[23] E. Bertino and N. Islam, "Botnets and Internet of Things Security," *Computer*, vol. 50, no. 2, pp. 76–79, 2017.
[24] A. Shabtai, U. Kanonov, Y. Elovici, C. Glezer, and Y. Weiss, "Andromaly: A Behavioral Malware Detection Framework for Android Devices," *Journal of Intelligent Information Systems*, vol. 38, pp. 161–190, 2012.
[25] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, "Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams," in *Proc. AAAI Workshops*, 2017.