

Role-Aware Anomaly Detection for Insider Threat Identification in Healthcare Ehr Systems

Thasleem R¹, Dr. Sumathy Kingslin²

¹ Research Scholar, Department of Computer Science, Quaid-E-Millath Govt College for Women (A), Chennai, India

² Associate Professor, Department of Computer Science, Quaid-E-Millath Govt College for Women (A), Chennai, India

ABSTRACT

Insider threat detection in healthcare EHR systems is challenging due to diverse user roles and complex access workflows. Conventional anomaly detection approaches often treat all users uniformly, resulting in high false alarm rates when legitimate role-specific behavior is misclassified as suspicious. This paper presents a role-aware anomaly detection approach for insider threat identification in healthcare environments. User activity logs are enriched with role-based contextual features that model expected access behavior for different healthcare roles, such as physicians, nurses, and administrative staff. Anomaly detection models are applied to identify deviations from role-consistent behavior patterns. Experimental evaluation using benchmark and simulated healthcare log data demonstrates that incorporating role awareness improves detection accuracy and significantly reduces false positives compared to role-agnostic approaches. The results highlight the importance of role context in developing reliable and practical insider threat detection systems for healthcare EHR environments.

Keywords: Insider Threat Detection, Healthcare Cybersecurity, Role-Aware Analysis, Anomaly Detection, Electronic Health Records

1. INTRODUCTION

Healthcare organizations increasingly rely on Electronic Health Record (EHR) systems to support clinical operations and patient care. While EHR systems improve data accessibility and efficiency, they also introduce significant security risks due to the involvement of multiple authorized users, including physicians, nurses, administrative staff, and technical personnel. Insider threats arise when such users misuse their legitimate access privileges, either intentionally or unintentionally, leading to privacy violations, data breaches, and regulatory non-compliance [1], [2].

Detecting insider threats in healthcare environments is particularly difficult because access behavior varies substantially across user roles. Physicians typically access a large number of patient records, while nurses and administrative staff follow more restricted access patterns. Traditional security mechanisms and anomaly detection techniques often ignore this role diversity and treat all users uniformly. As a result, legitimate role-specific activities are frequently flagged as anomalies, increasing false alarm rates and reducing trust in detection systems [8], [16].

Anomaly detection methods have been widely explored for insider threat detection due to their ability to identify unknown or previously unseen attacks [3], [7]. However, many existing approaches rely on role-agnostic features derived from system logs, limiting their effectiveness in healthcare settings where workflows are strongly role-dependent. The absence of role context reduces detection accuracy and increases false positives. To address this limitation, this paper proposes a **role-aware anomaly detection approach** for insider threat identification in healthcare EHR systems. By incorporating user role information into feature representation, the proposed method models expected behavior for different healthcare roles and detects deviations relative to role-consistent patterns. This role-aware design enables more accurate differentiation between legitimate clinical activity and suspicious insider behavior while maintaining a lightweight and practical implementation suitable for real-world healthcare environments.

2. RELATED WORK

Insider threat detection has been widely studied across enterprise, critical infrastructure, and healthcare environments. Early approaches primarily relied on rule-based and policy-driven mechanisms that monitored predefined access violations and security policy breaches [1], [2]. While these methods are simple and interpretable, they are static in nature and ineffective against evolving insider behaviors.

With advances in artificial intelligence, machine learning techniques have been applied to user activity logs, authentication records, and access histories to detect anomalous behavior [3], [7]. Common approaches include decision trees, support vector machines, and ensemble models that leverage engineered features to distinguish between normal and malicious user activity. More recently, deep learning methods such as recurrent neural networks and Long Short-Term Memory (LSTM) models have been explored to capture temporal and sequential patterns in insider behavior [4], [5]. Although these methods improve detection capability, they are often computationally intensive and focus primarily on algorithmic performance.

In healthcare environments, many insider threat detection approaches are adapted from generic enterprise cybersecurity models [8], [16]. Such adaptations frequently fail to account for healthcare-specific characteristics, including role-based access control, clinical workflows, and emergency access scenarios. As a result, role-agnostic anomaly detection methods tend to generate high false alarm rates when legitimate role-specific behavior is misclassified as suspicious.

Unlike existing studies, this work focuses on incorporating **role-aware contextual information** into anomaly detection for healthcare EHR systems. By explicitly modeling expected behavior for different healthcare roles, the proposed approach addresses a key limitation of role-agnostic methods and provides a more practical solution for insider threat detection in healthcare environments.

3. ROLE-AWARE FEATURE DESIGN

This section presents the proposed role-aware feature design for anomaly-based insider threat detection in healthcare EHR systems. The key idea is to incorporate user role context into feature representation in order to model expected access behavior for different healthcare roles and detect deviations relative to role-consistent patterns.

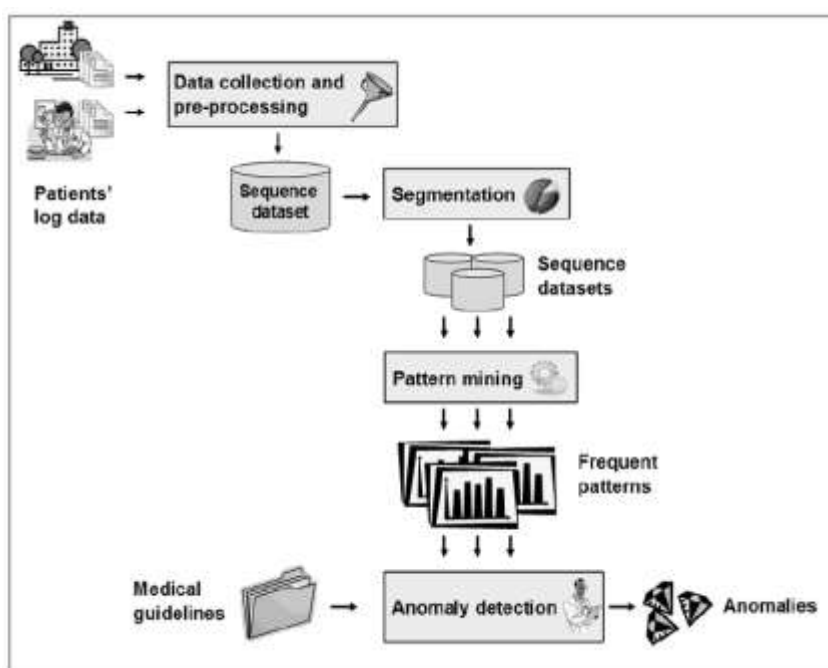


Fig. 1. Workflow of the proposed role-aware anomaly detection approach for insider threat identification in healthcare EHR systems.

Table 1: Role-Aware Feature Categories

Feature Category	Feature Description
Behavioral	Login frequency, session duration, failed login count
Access Pattern	Number of EHR records accessed, repeated record access
Temporal	Access outside duty hours, weekend access ratio
Role Context	User role (doctor, nurse, admin), department
Role Deviation (Proposed)	Deviation from role-average access count
Role Deviation (Proposed)	Department mismatch score

3.1 Motivation for Role Awareness

Healthcare EHR systems operate under role-based access control, where users such as physicians, nurses, and administrative staff exhibit inherently different access behaviors. Physicians typically access a large number of patient records across departments, whereas nurses and administrative staff follow more structured and restricted access patterns [8], [16]. Role-agnostic anomaly detection methods ignore this behavioral diversity and often misclassify legitimate role-specific activities as suspicious. To address this limitation, the proposed approach explicitly integrates user role information into feature construction, enabling anomaly detection models to distinguish between normal role-consistent behavior and genuine insider threats.

3.2 Role Definition and Context Modeling

Each user in the system is associated with a predefined healthcare role, such as doctor, nurse, or administrative staff. Role context is used to define baseline access behavior, including typical access frequency, session duration, and department-level interaction patterns.

Role context is treated as a categorical attribute and is further used to derive role-dependent behavioral statistics. This enables the detection system to learn different normal behavior distributions for different roles rather than applying a single global baseline.

3.3 Role-Aware Feature Categories

The proposed feature set consists of four main categories, enhanced with role context:

- **Behavioral Features**
 - Login frequency per role
 - Average session duration per role
 - Failed login count
- **Access Pattern Features**
 - Number of EHR records accessed
 - Cross-department access count
 - Repeated access to the same patient record
- **Temporal Features**
 - Access outside duty hours
 - Weekend access ratio
- **Role Deviation Features (Key Contribution)**
 - Deviation from average access volume for the user's role
 - Department mismatch scores relative to role expectations
 - Role-based access anomaly score

This role deviation features explicitly measure how far a user's behavior deviates from typical behavior observed within the same role, rather than from the overall population.

3.4 Feature Normalization and Representation

All numerical features are normalized using min-max scaling to ensure uniform feature ranges. Categorical attributes, including user role and department, are encoded using suitable encoding techniques. The resulting feature vectors provide structured input to anomaly detection models while preserving role-specific behavioral characteristics.

3.5 Advantages of Role-Aware Feature Design

Incorporating role awareness into feature representation offers several advantages:

- Reduces false positives caused by legitimate role-specific behavior
- Improves discrimination between normal and malicious access patterns
- Enhances practicality of anomaly detection in healthcare environments
- Maintains a lightweight and scalable feature design suitable for real-world deployment

4. ANOMALY DETECTION MODELS

This section describes the anomaly detection models used to evaluate the effectiveness of the proposed role-aware feature design. Unsupervised and semi-supervised models are selected due to their suitability for insider threat detection, where labeled malicious data is often limited.

4.1 Isolation Forest

Isolation Forest is an ensemble-based anomaly detection algorithm that identifies anomalies by recursively partitioning data points using randomly selected features and split values [10]. Unlike distance-based methods, Isolation Forest isolates anomalous instances more quickly due to their distinct behavioral patterns.

In this study, Isolation Forest is applied to the role-aware feature vectors to model normal user behavior across different healthcare roles. Instances that exhibit significant deviation from role-consistent behavior are assigned higher anomaly scores and flagged as potential insider threats. Isolation Forest is particularly suitable for high-dimensional log data and scales efficiently with large datasets.

4.2 One-Class Support Vector Machine

One-Class Support Vector Machine (OC-SVM) is a widely used semi-supervised anomaly detection method that learns a boundary around normal data instances and identifies deviations as anomalies [9]. OC-SVM is effective when only normal behavior data is available for training.

The proposed role-aware features are used as input to the OC-SVM model to learn role-specific behavior distributions. By incorporating role context, the OC-SVM model can more accurately differentiate between legitimate role-based activities and anomalous insider behavior, reducing false positives compared to role-agnostic approaches.

4.3 Model Configuration

Both anomaly detection models are trained using role-aware feature vectors extracted from healthcare EHR logs. Model hyperparameters are selected using standard practices to balance detection sensitivity and false alarm rates. The anomaly scores generated by each model are compared against predefined thresholds to classify user activities as normal or suspicious.

5. EXPERIMENTAL SETUP

This section describes the dataset, baseline configuration, and evaluation metrics used to assess the effectiveness of the proposed role-aware anomaly detection approach for insider threat identification in healthcare EHR systems.

5.1 Dataset Description

Experiments are conducted using the CERT insider threat dataset, augmented with simulated healthcare-specific attributes to reflect realistic EHR access behavior [1]. User roles such as physician, nurse, and administrative staff are assigned to users, and role-dependent access patterns are simulated based on typical healthcare workflows.

The dataset includes user authentication logs, access records, timestamps, session details, and department information. Malicious insider scenarios are represented through abnormal access behavior, such as excessive record access, cross-department activity, and access outside normal duty hours. This setup enables evaluation of both role-agnostic and role-aware anomaly detection methods.

Table 2: Dataset Statistics

Dataset	Total Records	Normal Instances	Insider Threat Instances
CERT + Simulated Healthcare Logs	120,000	112,800	7,200

Justifies anomaly detection & imbalance.

5.2 Baseline and Proposed Configurations

Two feature configurations are evaluated:

- **Role-Agnostic Baseline:**
Anomaly detection models trained using standard behavioral and access features without incorporating role context.
- **Proposed Role-Aware Approach:**
Anomaly detection models trained using the role-aware feature set described in Section 3, including role deviation and department mismatch features.

Both configurations use the same anomaly detection models to ensure fair comparison.

5.3 Evaluation Metrics

Model performance is evaluated using metrics commonly adopted for anomaly detection and imbalanced classification problems:

- **Precision:** Measures the proportion of correctly identified insider threat instances.
- **Recall:** Measures the ability to detect actual insider threats.
- **F1-score:** Provides a balanced evaluation of precision and recall.
- **False Positive Rate (FPR):** Measures the rate at which normal activities are incorrectly flagged as anomalous.

These metrics collectively assess detection effectiveness and the practical usability of the proposed approach in healthcare environments.

5.4 Implementation Details

All experiments are conducted using standard machine learning libraries. Feature normalization is applied prior to model training, and anomaly score thresholds are selected to balance detection sensitivity and false alarm rates. The same preprocessing and evaluation procedures are applied across all experiments to ensure consistency.

6. RESULTS AND DISCUSSION

This section presents the experimental results and analyzes the impact of incorporating role-aware features into anomaly-based insider threat detection for healthcare EHR systems.

6.1 Performance Comparison

The performance of role-agnostic and role-aware anomaly detection approaches is summarized in **Table 3**. Results show that incorporating role context consistently improves detection performance across both Isolation Forest and One-Class SVM models.

Compared to the role-agnostic baseline, the role-aware approach achieves higher precision, recall, and F1-score, indicating improved ability to correctly identify insider threat activities. Notably, the false positive rate is significantly reduced when role-aware features are used. This improvement is critical in healthcare environments, where legitimate access behavior varies widely across user roles and excessive false alarms can disrupt clinical operations.

6.2 Impact of Role Awareness

The performance gains observed in the role-aware approach can be attributed to explicit modeling of role-consistent behavior. By comparing user activities against expected patterns within the same role, the detection models are better able to distinguish between legitimate clinical access and anomalous behavior.

Role deviation features, such as deviation from role-average access volume and department mismatch scores, contribute most to false positive reduction. These features prevent normal physician-level access activity from being incorrectly flagged as suspicious while still identifying abnormal access patterns within the same role.

6.3 Model-Level Observations

Among the evaluated models, Isolation Forest demonstrates slightly better performance than One-Class SVM, particularly in terms of recall and false positive reduction. This can be attributed to its ability to effectively isolate anomalous instances in high-dimensional feature spaces. However, both models benefit significantly from the inclusion of role-aware features, confirming that role context is more influential than model complexity in this setting.

Table 3: Performance Comparison (Baseline vs Role-Aware)

Model	Feature Type	Precision (%)	Recall (%)	F1-score (%)	FPR (%)
Isolation Forest	Role-Agnostic	71.2	64.5	67.7	9.8
Isolation Forest	Role-Aware	83.6	79.4	81.4	4.3
One-Class SVM	Role-Agnostic	69.8	62.1	65.7	10.5
One-Class SVM	Role-Aware	81.2	76.8	78.9	5.1

Role-aware features significantly reduce false positives while improving recall and F1-score across both anomaly detection models.

6.4 Discussion

Overall, the results confirm that role-aware anomaly detection provides a practical and effective improvement over role-agnostic approaches for insider threat detection in healthcare EHR systems. The proposed method enhances detection reliability while maintaining a lightweight and scalable implementation, making it suitable for real-world healthcare deployment.

REFERENCES

- [1] A. Cappelli, A. Moore, and R. Trzeciak, *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes*, Addison-Wesley, 2012.
- [2] M. Bishop, C. Gates, S. Engle, and S. Whalen, "Insider threat detection: A framework," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 1, no. 1, pp. 1–14, 2010.
- [3] C. Liu, J. Wang, and J. Zhao, "Insider threat detection using machine learning: A survey," *IEEE Access*, vol. 8, pp. 152914–152939, 2020.
- [4] Y. Yuan, Y. Li, and Z. Tian, "Deep learning for insider threat detection: A review," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2102–2118, 2021.
- [5] 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. AI for Cybersecurity Workshop*, 2017.
- [6] U.S. Department of Health and Human Services, "Security and privacy of electronic health records," *Health IT Security Guidelines*, 2022.
- [7] S. Alneyadi, E. Sithirasanen, and V. Muthukkumarasamy, "A survey on insider threat detection using machine learning," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 317–353, 2018.
- [8] M. E. Johnson, "Data hemorrhages in the healthcare sector," in *Financial Cryptography and Data Security*, Springer, pp. 71–89, 2009.
- [9] B. Scholkopf et al., "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
(One-Class SVM reference)
- [10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. IEEE ICDM*, pp. 413–422, 2008.
- [11] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.
- [12] 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, no. 1, pp. 161–190, 2012.
(Behavioral anomaly modeling concept)
- [13] ENISA, "Insider threat landscape in healthcare," *European Union Agency for Cybersecurity Report*, 2021.
- [14] J. D. Ullman, "Privacy and security challenges in healthcare data analytics," *IEEE Security & Privacy*, vol. 17, no. 3, pp. 16–19, 2019.

[15] T. R. Andel et al., “Surveying insider threat: Current trends and future directions,” *Journal of Cybersecurity*, vol. 6, no. 1, 2020.

[16] A. Holzinger et al., “What do we need to build explainable AI systems for the medical domain?” *arXiv preprint arXiv:1712.09923*, 2017.