
MACHINE LEARNING-BASED IDENTITY PATTERN ANALYSIS FOR INTELLIGENT CYBERSECURITY THREAT DETECTION

Dr. Bodi Nidharshini

Lect.in Computer Science

Government Degree College (Autonomous), Medak.

Abstract

The increasing sophistication of cyber threats has created a need for intelligent security systems capable of identifying suspicious user activities, unauthorized access, compromised accounts, and insider threats. Traditional cybersecurity mechanisms that rely primarily on predefined rules and known attack signatures may struggle to detect previously unseen threats and subtle deviations in user behaviour. Machine learning provides opportunities to analyse identity-related activity patterns and distinguish normal behaviour from potentially malicious actions. This research proposes a machine learning-based framework for identity pattern analysis and intelligent cybersecurity threat detection. The framework examines authentication records, login frequency, access histories, device information, session characteristics, and deviations from established user behaviour. Data preprocessing and feature engineering are used to prepare identity-related records for analysis, while supervised and unsupervised learning techniques are considered for threat classification and anomaly detection. Candidate algorithms include Random Forest, Support Vector Machine, Isolation Forest, and Long Short-Term Memory networks for sequential behavioural data. The proposed framework will be evaluated using accuracy, precision, recall, F1-score, false-positive rate, and computational efficiency. Explainability and privacy considerations will also be incorporated to support responsible security decisions. The study aims to develop an adaptive approach to identity-based threat detection that can assist organizations in identifying suspicious activities and strengthening access security. The effectiveness of the framework must be established through empirical experimentation using appropriate cybersecurity datasets.

Keywords: Machine Learning, Identity Pattern Analysis, Cybersecurity, Threat Detection, User Behaviour Analytics, Anomaly Detection, Insider Threats, Authentication Security, Random Forest, Deep Learning.

1. Introduction

Cybersecurity has become a critical concern as individuals, businesses, educational institutions, and government organizations increasingly depend on digital platforms and interconnected information systems. The growing use of cloud computing, remote access, online services, and digital authentication has expanded the number of potential entry points for cyberattacks. Unauthorized account access, credential theft, privilege misuse, and insider threats can compromise confidential information and disrupt organizational operations. Consequently, cybersecurity systems must be capable of identifying suspicious activities accurately and responding to emerging threats in a timely manner.

Conventional cybersecurity approaches commonly rely on predefined rules, known attack signatures, and access-control policies. Although these mechanisms remain important, they may have limitations when attackers use legitimate credentials, change their behaviour, or employ previously unseen techniques. Machine learning offers an additional approach by learning patterns from historical and current activity data and identifying deviations that may indicate malicious behaviour. Research on machine learning and deep learning for authentication and authorization highlights their growing role in strengthening digital identity security.

Identity pattern analysis involves examining user-related information to understand typical authentication and access behaviour. Relevant characteristics may include login times, authentication failures, session duration, device changes, access frequency, and resource usage. When these characteristics are analysed over time, machine learning models can develop behavioural profiles and identify activities that differ significantly from established patterns. Such analysis is particularly useful for detecting compromised accounts and potential insider threats, where malicious activity may occur through otherwise legitimate access credentials.

Several machine learning techniques can support this process. Supervised algorithms, such as Random Forest and Support Vector Machine, can classify activities when reliable labelled examples are available. Unsupervised approaches, including Isolation Forest and clustering methods, can identify unusual patterns when labelled attack data are limited. Deep learning models, such as Long Short-Term Memory networks, can analyse sequential activity and potentially identify suspicious changes across multiple events. However, the performance of these methods depends on data quality, feature selection, class imbalance, and the ability to distinguish genuine threats from legitimate variations in user behaviour.

Yuan and Wu (2021) reviewed deep learning approaches for insider threat detection and identified challenges associated with complex, heterogeneous, sparse, and adaptive behavioural data. These challenges demonstrate why identity-based threat detection requires more than a simple classification model.

Against this background, the present study proposes a machine learning-based framework for identity pattern analysis and intelligent cybersecurity threat detection. The framework combines identity-related behavioural features, machine learning algorithms, anomaly detection, and performance evaluation to support the early identification of suspicious activities. Its contribution will be assessed by comparing candidate algorithms under consistent experimental conditions and examining their predictive performance, false alarms, interpretability, and computational requirements.

2. Objectives

The major objectives of the study are:

1. To analyse identity-related behavioural patterns associated with user authentication and system access.
2. To develop a machine learning framework for identifying suspicious identity activities and potential cybersecurity threats.

3. To examine the effectiveness of supervised and unsupervised learning algorithms for behavioural anomaly detection.
4. To identify deviations from normal user behaviour that may indicate compromised accounts, unauthorized access, or insider threats.
5. To evaluate the proposed approach using accuracy, precision, recall, F1-score, false-positive rate, and computational efficiency.
6. To compare the performance of selected machine learning algorithms under consistent experimental conditions.
7. To investigate the role of feature engineering and behavioural profiling in improving threat detection.
8. To identify practical limitations involving imbalanced data, evolving attack patterns, privacy, and model interpretability.

3. Research Methodology

3.1 Research design

The study adopts a quantitative, experimental research design. It proposes the development and evaluation of machine learning models for identifying identity-related behavioural anomalies and classifying potentially malicious activities. The evaluation will use an appropriate publicly available cybersecurity dataset containing user activity, authentication, access, or related security records.

The research should distinguish identity-based behavioural analysis from general network intrusion detection. If the selected dataset contains only network traffic features and no meaningful identity or user-behaviour attributes, the scope of the title and research objectives must be adjusted accordingly.

3.2 Data collection

The study will select a documented dataset suitable for the intended threat-detection task. Potential data sources include authentication logs, access-control records, user activity histories, and established insider-threat or intrusion-detection datasets.

Relevant attributes may include:

- Authentication success and failure events.
- Login frequency and time of access.
- Session duration and access sequences.
- Device or network context, where available.
- Access to files, applications, or protected resources.
- Labels indicating normal activity or known suspicious behaviour.

Only attributes necessary for the research task should be retained. Direct identifiers should be removed, pseudonymized, or otherwise protected where appropriate. The dataset, sampling strategy, class distribution, and limitations must be documented.

3.3 Data preprocessing

Raw security data may contain missing values, duplicate events, inconsistent formats, categorical variables, and highly imbalanced classes. The preprocessing stage will address these issues using suitable techniques.

The main steps include:

1. Removing duplicate records and correcting inconsistent data formats.
2. Handling missing values using documented rules.
3. Encoding categorical features and scaling numerical features when required by the selected algorithm.
4. Constructing behavioural features, such as failed-login counts, unusual access times, session frequency, and deviations from user baselines.
5. Separating training, validation, and test data to avoid information leakage.
6. Examining class imbalance and applying appropriate strategies, if necessary, using training data only.

Where activity is time-dependent, the data should be split chronologically or by user, as appropriate, to avoid unrealistic evaluation results caused by overlap between training and test records.

3.4 Feature extraction and identity pattern analysis

Feature extraction is an important stage because the quality of behavioural representations influences the performance of machine learning models. The study will construct features that capture both individual activity characteristics and deviations from established user patterns.

For example, a login at an unusual time may not be suspicious on its own. However, a combination of unusual access time, repeated authentication failures, an unfamiliar device, and access to sensitive resources may warrant further investigation. The framework should therefore assess relevant contextual features rather than automatically treating every deviation as a threat.

3.5 Machine learning algorithms

The study will consider the following candidate algorithms.

Algorithm	Proposed role
Random Forest	Classification of normal and suspicious activities using labelled data
Support Vector Machine	Classification of behavioural patterns with suitable feature representations
Isolation Forest	Detection of unusual activity in datasets with limited labels
Long Short-Term Memory (LSTM)	Analysis of sequential user activity when sufficient temporal data are available

A simple baseline model should be included to provide a meaningful comparison. Algorithm selection will depend on the dataset's size, feature structure, class balance, and computational requirements. Not all algorithms need to be used if the available data do not support them.

3.6 Proposed system architecture

The framework consists of five stages:

1. Cybersecurity data collection

Authentication logs, user activity and access records

2. Data preprocessing and feature engineering

Cleaning, encoding, behavioural features and data splitting

3. Machine learning and anomaly detection

Classification, behavioural profiling and anomaly scoring

4. Threat assessment

Risk scoring and identification of suspicious activity

5. Evaluation and security reporting

Performance metrics, false alarms and model limitations

3.7 Evaluation metrics

The models will be evaluated using standard classification and anomaly-detection measures.

- Accuracy: Proportion of all evaluated records classified correctly.
- Precision: Proportion of records classified as threats that are genuinely threats.
- Recall: Proportion of actual threats detected by the model.
- F1-score: Harmonic mean of precision and recall.
- False-positive rate: Proportion of legitimate activities incorrectly classified as threats.
- ROC-AUC: Ability to distinguish between threat and non-threat classes across classification thresholds, where appropriate.
- Computational efficiency: Training time, prediction latency, and resource consumption.

Because cybersecurity datasets often contain far fewer threat events than normal events, accuracy should not be used as the sole measure of success. The results should also report class-specific performance and, where suitable, precision–recall curves.

3.8 Comparative analysis and ethical considerations

The selected models will be trained and tested using consistent data partitions and evaluation procedures. Their performance will be compared to identify strengths and weaknesses across detection quality, false alarms, interpretability, and computational cost. Repeated runs or cross-validation may be used when appropriate for the dataset and experimental design.

The research will use public, synthetic, or appropriately authorized data. Sensitive personal information should be protected, and any proposed real-world deployment should include access controls, human review of high-impact alerts, and safeguards against unfair or inappropriate profiling.

4. Results and Discussion

4.1 Expected results

The proposed framework is intended to identify suspicious identity patterns by analysing deviations from normal authentication and access behaviour. Since the framework has not yet been

implemented and tested, the results presented here describe the expected evaluation outcomes rather than verified experimental findings.

The study will investigate whether machine learning models can distinguish legitimate user activities from suspicious behaviour and identify the most suitable algorithms for the selected dataset. The expected analytical benefits include improved recognition of unusual login patterns, identification of potentially compromised accounts, and prioritization of suspicious access events for further investigation.

4.2 Comparative performance analysis

The performance of Random Forest, Support Vector Machine, Isolation Forest, and LSTM will be compared using the same evaluation dataset wherever technically appropriate.

Table 1. Comparative Evaluation of Machine Learning Algorithms for Identity Pattern Analysis

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	To be measured	To be measured	To be measured	To be measured
Support Vector Machine	To be measured	To be measured	To be measured	To be measured
Isolation Forest	To be measured	To be measured	To be measured	To be measured
LSTM	To be measured	To be measured	To be measured	To be measured

The results will establish which algorithms perform most effectively under the selected experimental conditions. Random Forest may provide a useful baseline for structured behavioural data, while Isolation Forest may be suitable for detecting anomalies when reliable threat labels are limited. LSTM may be useful for sequential data if sufficient temporal information is available. These are methodological expectations, not conclusions about which algorithm will achieve the highest performance.

4.3 Detection of identity-based threats

The proposed framework will examine several categories of suspicious behaviour, including repeated failed authentication attempts, unusual access times, changes in device usage, abnormal session activity, and unexpected access to protected resources.

Table 2. Identity Behaviour Indicators and Potential Security Significance

Behavioural indicator	Potential security significance
Repeated failed login attempts	Possible credential guessing or unauthorized access
Login from an unfamiliar device	Possible account compromise or legitimate device change
Unusual access times	Possible deviation from established behaviour
Sudden increase in resource access	Possible misuse of privileges or unusual workload
Abnormal sequences of user actions	Possible compromised account or insider activity

These indicators should be interpreted contextually. A single unusual event does not establish that an attack has occurred. Combining multiple signals and considering legitimate changes in user behaviour can help reduce false alarms.

4.4 False-positive analysis and model reliability

False positives are a significant concern in identity-based threat detection. A system that repeatedly flags legitimate activity may increase the workload of security analysts and cause genuine alerts to be overlooked. The study will therefore examine false-positive rates alongside recall and precision.

A threshold-based risk-scoring mechanism could help categorize alerts according to their severity. However, thresholds must be selected using validation data and evaluated on a separate test set. Explainability techniques may also be used to identify the behavioural features that contributed to a prediction, helping analysts investigate alerts rather than relying solely on automated decisions.

4.5 Limitations and implications

The effectiveness of the proposed framework may be affected by imbalanced datasets, incomplete identity records, evolving attack strategies, and changes in legitimate user behaviour. Models trained on one organization or dataset may not generalize to other environments. In addition, an attacker who has valid credentials may deliberately imitate normal behaviour, making detection difficult.

Recent reviews of user authentication and authorization research identify continuing challenges in dataset selection, model evaluation, and the adaptation of machine learning techniques to real-world security settings.

Accordingly, the framework should be evaluated using clearly documented datasets, appropriate baselines, and consistent metrics. Its practical value will depend not only on predictive performance but also on its ability to operate reliably, explain suspicious activity, and minimize unnecessary alerts.

5. Conclusion

Machine learning-based identity pattern analysis offers a promising approach to strengthening cybersecurity through the identification of suspicious user behaviour, compromised accounts, unauthorized access, and potential insider threats. By examining authentication histories, access patterns, session characteristics, and behavioural deviations, machine learning models can support security monitoring beyond conventional rule-based mechanisms.

This study proposes a framework combining data preprocessing, behavioural feature extraction, supervised classification, and unsupervised anomaly detection. The performance of Random Forest, Support Vector Machine, Isolation Forest, and LSTM will be assessed using accuracy, precision, recall, F1-score, false-positive rate, and computational efficiency. The evaluation will help determine the suitability of different algorithms for identity-related threat detection.

The proposed approach also recognizes the importance of privacy, explainability, and responsible use of behavioural data. Although machine learning may improve the identification of suspicious activity, its effectiveness must be demonstrated through empirical testing rather than assumed.

Future research should examine real-world deployment, adaptation to evolving threats, interpretable risk scoring, and the integration of privacy-preserving techniques. A carefully validated framework could contribute to more adaptive and reliable cybersecurity monitoring.

6. References

1. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176. <https://doi.org/10.1109/COMST.2015.2494502>
2. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15. <https://doi.org/10.1145/1541880.1541882>
3. Akampurira, P., Edozie, E., Sadiq, B. O., & Buhari, M. D. (2026). A deep learning-based user behavior analytics model for proactive cyber threat detection and risk management: A review. *F1000Research*, 15, 674. <https://doi.org/10.12688/f1000research.178351.2>
4. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, 104, Article 102221. <https://doi.org/10.1016/j.cose.2021.102221>
5. Machine learning and deep learning for user authentication and authorization in cybersecurity: A state-of-the-art review. (2024). *Computers & Security*, 140, Article 103747. <https://doi.org/10.1016/j.cose.2024.103747>
6. Cui, J., Chen, Z., Tian, L., & Zhang, G. (2022). Overview of user and entity behavior analytics technology based on machine learning. *Computer Engineering*, 48(2), 10–20. <https://www.ecice06.com/EN/Y2022/V48/I2/10>
7. Zhang, H., Ye, J., Huang, W., Liu, X., & Gu, J. (2025). Survey of federated learning in intrusion detection. *Journal of Parallel and Distributed Computing*, 195, Article 104976. <https://doi.org/10.1016/j.jpdc.2024.104976>
8. Akampurira, P., Sadiq, B. O., Buhari, M. D., Venkateswarlu, M., Atuhe, A., & Mugisha, B. (2024). A review of neural networks for enhanced user entity behavior analytics in