ISSN - 2456-1142

International journal of humanities and information technology Volume 7, Issue 4, 2025

Measuring the Human Risk Factor: Developing Predictive Models for Insider Threat Detection Using Behavioral Analytics

Día Fayyad

Cybersecurity Department, Saudi Aramco; Jordanian Engineers Association, Saudi Council of Engineers

Abstract

Insider threats represent one of the most complex challenges in modern cybersecurity, as they originate from authorized users who exploit legitimate access for malicious or negligent purposes. Unlike external attacks, insider incidents often evolve gradually through subtle behavioral changes that traditional rule-based systems fail to detect. This study investigates the human risk factor underlying insider threats and develops predictive models that leverage behavioral analytics to identify early indicators of malicious intent.

The research utilizes the Carnegie Mellon University CERT Insider Threat Dataset (v4.2), integrating system activity logs, communication patterns, and psychosocial proxies such as email sentiment and work-hour deviations. Data preprocessing involved normalization and feature selection across technical, behavioral, and psychological dimensions. Three machine learning models—Random Forest, Long Short-Term Memory (LSTM), and Autoencoder—were implemented to evaluate predictive performance. Model performance was assessed using precision, recall, F1-score, and ROC-AUC metrics.

Results show that the LSTM model achieved the highest overall accuracy of 93.2 percent with an AUC of 0.96, outperforming both Random Forest and Autoencoder models. Behavioral deviations such as unusual file transfers, abrupt login time changes, and communication tone shifts emerged as strong predictors of insider risk. The findings highlight the value of integrating human behavioral analytics into cybersecurity frameworks to enhance proactive threat detection.

This study contributes to the development of data-driven, ethically aligned security strategies that enable organizations to identify, quantify, and mitigate insider risks before they escalate into security incidents.

Keywords: Insider Threats, Behavioral Analytics, Human Risk Factor, Predictive Modeling, Machine Learning, Cybersecurity, CERT Dataset, LSTM, Random Forest, Autoencoder.

International journal of humanities and information technology (2025) 10.21590/ijhit.07.04.03

Introduction

Background of Insider Threats

The insider threat represents one of the most persistent and complex challenges in the field of cybersecurity. Unlike external attackers, insiders possess legitimate access privileges that enable them to exploit organizational systems and data without immediately triggering conventional security alarms (Homoliak et al., 2019). These insiders may act maliciously for financial gain, personal revenge, or ideological motives, or they may cause harm unintentionally due to negligence or lack of awareness (Salem et al., 2008).

The implications of insider threats are multidimensional, affecting the financial, reputational, and operational stability of organizations. Financially, insider incidents can lead to direct monetary losses and indirect costs through legal penalties, incident response, and remediation. The 2023 IBM Cost of Insider Threats report estimated that insider-related incidents cost organizations an average of USD 15.4 million annually, an increase of 34 percent since 2020. Reputationally, organizations suffer erosion of customer trust, investor confidence, and brand credibility when insider-related breaches become public. Operationally, such incidents disrupt normal workflows, compromise intellectual property, and reduce employee morale (Gheyas & Abdallah, 2016).

Corresponding Author: Día Fayyad, Cybersecurity Department, Saudi Aramco; Jordanian Engineers Association, Saudi Council of Engineers, e-mail: dia.fayyad@gmail.com

How to cite this article: Fayyad, D. (2025). Measuring the Human Risk Factor: Developing Predictive Models for Insider Threat Detection Using Behavioral Analytics. *International journal of humanities and information technology* 7(4), 29-40.

Source of support: Nil

Conflict of interest: None

These far-reaching effects highlight the urgent need to go beyond perimeter-based cybersecurity systems, which primarily defend against external actors, toward internal resilience models that address the human dimension of risk (Cappelli et al., 2012). The dynamic nature of insider behavior, characterized by subtle deviations from established norms, requires continuous behavioral monitoring and adaptive prediction rather than static prevention mechanisms.

The Human Risk Factor

While technical vulnerabilities remain a concern, an increasing body of evidence suggests that human behavior is often the weakest link in the cybersecurity chain (Greitzer & Frincke, 2010; Legg et al., 2015). Insiders operate within trusted networks where their actions—such as excessive file access, unusual logon times, or emotional distress—may serve as early warning signals of potential compromise. These behavioral indicators can be influenced by psychosocial variables, including job dissatisfaction, perceived injustice, stress, or personal financial hardship (Greitzer & Hohimer, 2011).

Recognizing the human risk factor means understanding that cybersecurity is not solely a technical problem but also a behavioral one. Traditional systems rely heavily on access control, encryption, and intrusion detection, focusing on external defenses or system configurations (Probst et al., 2010). However, the increasing sophistication of insiders, coupled with hybrid work environments and digital collaboration tools, has exposed gaps in detecting behavioral anomalies that precede an attack.

To address this, the cybersecurity community is shifting from perimeter-based defenses—such as firewalls and network segmentation—to behavior-based defense mechanisms. These rely on machine learning and behavioral analytics to identify patterns that deviate from an individual's normal activity profile. Such systems aim not only to detect malicious intent but also to predict potential insider risks before damage occurs (Eberle & Holder, 2009; Brdiczka et al., 2012). The incorporation of behavioral science into cybersecurity represents a paradigm shift toward human-centric predictive defense.

Problem Statement

Despite significant progress in cybersecurity analytics, most insider threat detection systems remain reactive, rule-based, and insufficiently human-centric (Al-Mhiqani et al., 2020). They depend on static thresholds, binary decision rules, or after-the-fact analysis, which are ineffective against sophisticated or slow-evolving insider behaviors. Furthermore, existing systems rarely integrate psychosocial or contextual factors that drive insider actions.

Consequently, many organizations continue to experience delayed detection and false positives, resulting in wasted resources and compromised trust in automated systems (Tuor et al., 2017). There is a critical need for predictive frameworks that can dynamically learn behavioral baselines, adapt to organizational changes, and interpret deviations within a human-behavioral context. Addressing this problem requires interdisciplinary approaches that merge data science, psychology, and cybersecurity into a unified modeling strategy.

Research Aim and Objectives

The aim of this research is to develop and evaluate predictive models for insider threat detection that accurately quantify human risk factors using behavioral analytics. By leveraging real-world datasets—specifically the CERT Insider Threat Dataset (Version 4.2)—this study integrates technical, behavioral, and psychosocial indicators into machine learning frameworks capable of early anomaly detection.

The specific objectives of this research are to

- Construct behavioral risk profiles using structured organizational and psychosocial data.
- Develop and compare multiple predictive modeling techniques, including Random Forest (RF), Long Short-Term Memory (LSTM), and Autoencoder (AE) models.
- Evaluate model performance based on accuracy, precision, recall, F1-score, and ROC-AUC metrics.
- Identify behavioral and psychosocial variables that most strongly correlate with insider threat risk.
- Provide strategic recommendations for integrating predictive analytics into organizational security frameworks.

These objectives are grounded in prior studies emphasizing the fusion of behavioral modeling and machine learning for insider threat mitigation (Legg et al., 2015; Greitzer & Frincke, 2010).

Research Questions

This study seeks to answer the following core questions:

How can behavioral and psychosocial data improve insider threat detection and prediction?

- This question explores the degree to which human factors contribute to the predictive accuracy of insider risk models compared to purely technical indicators.

Which predictive modeling technique provides optimal accuracy and interpretability for insider threat detection?

- This focuses on comparing Random Forest, LSTM, and Autoencoder models to determine their respective strengths in detecting complex behavioral anomalies and maintaining explainability for decision-makers.

Structure of the Paper

The paper is organized into seven major sections

- Section 1 introduces the research background, human risk factors, problem statement, aims, and objectives.
- Section 2 reviews existing literature on insider threat taxonomies, behavioral analytics, and predictive

modeling techniques.

- Section 3 outlines the methodological framework, data sources, preprocessing techniques, and modeling approach.
- Section 4 presents the experimental results, including performance metrics, tables, and visual analyses.
- Section 5 discusses findings, theoretical implications, and comparison with existing studies.
- Section 6 examines ethical challenges and organizational implications of behavioral analytics in cybersecurity.
- Section 7 concludes with recommendations for practice and future research directions.

This structure ensures logical progression from problem identification to evidence-based solution development, aligning with academic research standards for cybersecurity analytics.

Literature Review

Conceptualizing Insider Threats

The concept of insider threats has evolved significantly in both scope and analytical complexity over the past two decades. Broadly, insider threats refer to harmful actions carried out by individuals with authorized access to organizational resources, often resulting in data breaches, sabotage, or fraud (Homoliak et al., 2019). Unlike external attacks that originate beyond the organization's perimeter, insider threats exploit legitimate credentials and privileges, making them particularly difficult to detect through traditional security controls.

Homoliak et al. (2019) provided a detailed taxonomy of insider threats, categorizing them by actor intent, motivation, and behavior. Their survey identified three major typologies: malicious insiders, negligent insiders, and infiltrators. Malicious insiders intentionally exploit their access for personal or political gain, negligent insiders inadvertently compromise systems due to poor cyber hygiene, and infiltrators act as external adversaries who gain insider access through deception or credential theft.

Nurse et al. (2014) proposed a complementary characterization framework emphasizing the multidimensional nature of insider actions, including motivational (greed, revenge, ideology), intentional (deliberate misuse), and accidental (unintentional harm) categories. This conceptual framework highlights the necessity of analyzing both behavioral context and system interactions to accurately model insider risk. As organizations increasingly digitize operations, understanding these typologies has become crucial for developing predictive models that move beyond static role-based detection toward dynamic behavioral surveillance.

Historical Approaches to Insider Threat Detection

Early insider threat detection frameworks were primarily rule-based or signature-driven, relying on predefined conditions such as abnormal login times or unauthorized file transfers (Salem et al., 2008). While effective for identifying known misuse patterns, these systems struggled with adaptive adversaries who modified their behaviors to evade detection. Furthermore, rule-based systems generated high false-positive rates, burdening security analysts with excessive alerts.

The transition toward data-driven and probabilistic approaches began in the mid-2000s with the development of systems like ELICIT, which evaluated "need-to-know" violations to detect anomalies in access behavior (Maloof & Stephens, 2007). As computing and storage capabilities improved, researchers introduced behavioral baselining to capture normal user activities over time.

Legg et al. (2015) advanced the field with an automated insider threat detection system leveraging user and role-based profiling. Their approach compared current user activity with historical behavioral baselines, dynamically adjusting for changes in organizational context. This evolution from static to adaptive modeling marked a critical step toward predictive analytics. However, the early systems lacked psychological awareness and relied solely on digital indicators, limiting their capacity to capture human intent.

Behavioral and Psychosocial Modeling

The integration of behavioral science and psychosocial indicators into cybersecurity represents a pivotal shift toward understanding the human dimensions of insider threats. Greitzer and Frincke (2010) proposed that traditional cyber audit data could be enriched by incorporating psychosocial factors such as job satisfaction, personality traits, and organizational stressors. Their model emphasized the predictive potential of behavioral anomalies arising from emotional distress, disengagement, or ethical conflicts within the workplace.

Building upon this concept, Greitzer and Hohimer (2011) developed computational models simulating how psychological precursors could evolve into harmful insider actions. They demonstrated that early warning indicators, such as sudden drops in communication tone or withdrawal from team interaction, could predict hostile intent weeks before a technical policy violation occurs.

These approaches introduced the notion of "behavioral risk scoring", a metric that quantifies insider likelihood based on human attributes. Yet, despite their conceptual strength, the implementation of such models in live environments remains limited due to privacy concerns and the lack of standardized psychological datasets (Cappelli et al., 2012). Nonetheless, behavioral and psychosocial modeling continues to guide the next generation of hybrid detection systems that merge cognitive signals with system telemetry.

Machine Learning and Graph-Based Methods

With the growing volume of user activity data, machine

learning (ML) emerged as a powerful method for insider threat detection. Early contributions by Eberle and Holder (2009) applied graph-based anomaly detection, where user interactions were modeled as nodes and edges, allowing identification of unusual communication or data-transfer patterns. Similarly, Chen et al. (2012) utilized specialized network analysis to detect anomalous insider actions within large enterprise environments. Graph structures proved effective for representing relational behaviors such as collaboration, email exchange, and privilege escalation sequences.

As data streams became more dynamic, researchers adopted stream mining and deep learning techniques for continuous monitoring. Parveen et al. (2013) proposed a stream mining framework capable of evolving detection models as user behavior changed, reducing retraining requirements. Tuor et al. (2017) later demonstrated an unsupervised deep learning approach using stacked LSTMs and autoencoders to identify anomalies in structured cybersecurity data. Their model achieved over 92 percent detection accuracy using the CERT insider threat dataset, confirming that temporal behavioral dependencies significantly improve prediction.

These studies collectively marked a paradigm shift from descriptive to predictive and adaptive security models, capable of learning complex behavioral sequences without explicit labeling. However, the interpretability of such models remains a challenge, as many deep learning systems function as "black boxes," providing limited insight into decision rationale.

Emerging Trends and Gaps

Recent advancements in AI and behavioral analytics have expanded the scope of insider threat research. Hybrid models combining technical, behavioral, and psychosocial data now achieve higher accuracy and contextual awareness (Al-Mhiqani et al., 2020). Nevertheless, integration of psychosocial data into predictive systems remains limited due to ethical and operational barriers. Most datasets, including CERT, lack emotional or psychological variables, restricting holistic understanding of insider intent (Greitzer & Frincke, 2010; Inayat et al., 2024).

Costa et al. (2024) emphasized the growing need for interpretable and adversarially robust AI models, as deep learning systems are increasingly vulnerable to manipulation and bias. Similarly, Inayat et al. (2024) underscored the importance of transparent and ethically aligned prediction frameworks that balance privacy protection with operational effectiveness.

The key research gap lies in the fusion of human factors and machine learning models that are both explainable and compliant with privacy regulations. Future efforts should focus on federated behavioral analytics, explainable AI (XAI), and cross-domain data sharing to enhance detection accuracy while maintaining ethical integrity.

Methodology

Research Design

This study follows a quantitative, data-driven, and experimental research design rooted in behavioral analytics. The objective is to identify, quantify, and predict insider threat behaviors using statistical and machine learning methods. The design enables a measurable and replicable evaluation of how human, behavioral, and technical indicators correlate with risk scores that signal insider intent.

Quantitative designs are preferred for insider-threat research because they allow large-scale data analysis and statistically significant inference from behavioral patterns (Gheyas & Abdallah, 2016). The approach integrates computational modeling and behavioral science, aligning measurable activities with psychological and organizational context. This design is grounded in the hypothesis that abnormal variations in user activity—especially those linked to behavioral and psychosocial anomalies—can serve as early warning indicators of insider risk.

The overall workflow of this research is structured into five sequential stages

- Data Acquisition – extraction and cleaning of insider activity records from the CERT v4.2 dataset.
- Feature Engineering – transformation of raw system and HR logs into quantifiable behavioral and psychosocial indicators.
- Model Development – creation of Random Forest, LSTM, and Autoencoder models for classification and anomaly detection.
- Performance Evaluation – assessment using precision, recall, F1-score, accuracy, and ROC-AUC metrics.
- Visualization and Interpretation – generation of analytical plots to illustrate feature importance and anomaly distributions.

This design ensures methodological transparency, statistical rigor, and reproducibility of findings.

Dataset Description

The empirical analysis in this study is based on the CERT Insider Threat Dataset Version 4.2, developed by the Carnegie Mellon University Software Engineering Institute (CMU-SEI). The dataset is widely regarded as the gold standard for academic research on insider-threat detection (Tuor et al., 2017; Legg et al., 2015).

The dataset simulates real-world corporate environments and contains log data from over 1000 employees spanning 18 months of operational activity. It integrates technical system events, human resource records, and communication logs, providing a comprehensive multi-dimensional representation of user behavior. Each user is assigned a unique anonymized identifier, with clear labels distinguishing normal and malicious users. The insider events simulated include data exfiltration, intellectual property theft, policy

Table 1: Feature Categories, Examples, and Data Sources

Feature Category	*Example Indicator*	*Behavioral Significance*	*Data Source*
Technical	File Access Count	Tracks intensity of file use in secure folders	System Logs
Technical	USB Device Frequency	Detects potential data exfiltration	Device Logs
Behavioral	After-Hours Login Ratio	Indicates deviation from normal schedule	Access Logs
Behavioral	Privilege Escalation Events	Suggests abnormal rights elevation	Security Audit Logs
Psychosocial	Email Sentiment Score	Reveals emotional tone of messages	Email Metadata
Psychosocial	Peer Communication Drop (%)	Measures social disengagement	HR Communication Data

Source: CERT Insider Threat Dataset (Version 4.2), CMU-SEI; Tuor et al. (2017).

violations, and sabotage.

Data Components

- System Logs: Track all computer-based actions such as logon/logoff events, USB device usage, web activity, and file transfers.
- Human Resource Records: Contain employee role, department, performance rating, employment history, and termination reasons.
- Communication Logs: Include email exchanges, sentiment polarity of text content, and social network structure among employees.

The diversity of data sources allows for modeling of both behavioral (what the employee does) and psychosocial (why the employee behaves a certain way) dimensions of insider threat.

Feature Engineering

Feature engineering was a critical step in converting raw logs into predictive behavioral indicators. Each data type—technical, behavioral, and psychosocial—was analyzed to extract meaningful patterns that could serve as features for machine learning models.

Technical Indicators

These reflect direct interactions with IT infrastructure. Indicators include

File Access Count

Number of daily file reads/writes in sensitive directories.

Removable Media Usage

Frequency and timing of USB connections.

Failed Logins

Number of unsuccessful access attempts.

Data Transfer Volume

Total bytes transmitted to external servers.

These indicators highlight deviations from standard usage norms that may suggest potential misuse (Eberle & Holder, 2009).

Behavioral Indicators

These measure how an individual's actions deviate from established work routines. Examples include

After-Hours Logins

Percentage of system accesses outside standard working hours.

Privilege Escalations

Sudden acquisition of administrative rights.

Access Pattern Deviations

Unusual frequency of visiting previously unused servers or directories.

Reduced Collaboration

- Decline in interactions with supervisors or teams.
- Behavioral features provide insight into the temporal evolution of insider risk behavior (Brdiczka et al., 2012).

Psychosocial Indicators

These reflect psychological and emotional trends observable through communication data (Greitzer & Hohimer, 2011). Examples include

Email Sentiment Polarity

Computed via natural language processing to measure tone positivity/negativity.

Peer Communication Density

Network centrality indicating interaction reduction or isolation.

Sentiment Drift Over Time

Sudden decrease in positive emotional tone correlated with stress or dissatisfaction.

After extraction, each variable was standardized using Min–Max normalization to ensure scale uniformity.

Correlation analysis and recursive feature elimination (RFE) reduced multicollinearity, resulting in 42 final variables distributed across three main feature domains.

Model Development

The development of predictive models was designed to evaluate three complementary algorithmic paradigms: Random Forest (RF) for interpretability, Long Short-Term Memory (LSTM) for temporal sequence learning, and Autoencoder (AE) for unsupervised anomaly detection.

Random Forest (RF)

The RF model consisted of 200 decision trees trained using Gini impurity. It provides robust classification with minimal overfitting and high interpretability, enabling identification of top behavioral predictors such as after-hours access frequency and sentiment decline. Feature importance was analyzed to interpret the model's decision rationale (Legg et al., 2015).

Long Short-Term Memory (LSTM)

The LSTM neural network was designed to capture sequential dependencies in time-stamped behavioral data. The model processed 7-day windows of user activity to detect gradual deviations in user routines. Hidden state units allowed the network to retain contextual memory of prior activities, effectively identifying latent behavioral drift that often precedes insider attacks (Tuor et al., 2017).

Autoencoder (AE)

The Autoencoder was used for unsupervised anomaly detection by reconstructing normal user behavior. A reconstruction error threshold ($\mu + 2\sigma$) was used to flag anomalies. This model was particularly effective for detecting new, previously unseen patterns of malicious behavior (Parveen et al., 2013).

All models were trained using a 70-15-15 split for training, validation, and testing datasets, respectively. To ensure fairness, each model underwent 10-fold cross-validation. Model parameters were optimized via grid search for accuracy and recall balance.

Evaluation Metrics

Model performance was evaluated using five key metrics, chosen for their ability to capture both predictive accuracy and error sensitivity (Eberle & Holder, 2009; Greitzer & Frincke, 2010):

Accuracy

Percentage of correctly classified user behaviors.

Precision

Ratio of true positives (actual threats) to all predicted positives, indicating detection reliability.

Recall

Proportion of actual threats correctly identified by the model, representing detection sensitivity.

F1-Score

Harmonic mean of precision and recall, balancing detection quality.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve)

Indicates the model's discrimination power across threshold values; values above 0.90 were considered excellent.

Confusion matrices were also computed to analyze false positives and false negatives for each algorithm. Feature-importance visualizations and ROC curves were generated to interpret the comparative efficiency of models.

Analytical Tools

All experiments were implemented in Python 3.10, leveraging advanced machine learning and data visualization libraries:

- TensorFlow 2.13 – LSTM and Autoencoder model architecture and neural-network optimization.
- Scikit-learn 1.3 – Implementation of Random Forest, feature selection, and performance metrics.
- Pandas and NumPy – Data manipulation, merging of HR, system, and communication datasets.
- Matplotlib and Seaborn – Graphical representation of ROC curves, anomaly distributions, and feature importance.

Model training and testing were conducted on a workstation with an Intel i9 processor, 32 GB RAM, and NVIDIA RTX 3060 GPU, ensuring efficient neural-network execution. All analytical workflows were executed in Jupyter Notebooks, following FAIR (Findable, Accessible, Interoperable, Reusable) data management standards to promote transparency and reproducibility.

Results

This section presents the experimental outcomes of the behavioral-based predictive modeling framework. Results are organized into four subsections: feature importance analysis, model performance summary, graphical results, and key findings. All models were evaluated on the CERT insider threat dataset using stratified sampling (70 percent training, 30 percent testing). The behavioral and psychosocial variables were assessed for their contribution to predictive performance in identifying high-risk insiders.

Feature Importance Analysis

The feature importance analysis highlights which behavioral and psychosocial indicators contribute most significantly to insider risk prediction. Using Random Forest's mean decrease in Gini impurity, ten top-ranked features were extracted.

The most influential predictors align with prior research indicating that behavioral deviations, communication tone, and privilege dynamics are core signals of malicious intent (Greitzer & Frincke, 2010; Gheyas & Abdallah, 2016).

Top Behavioral Predictors of Insider Risk

Table 2: Behavioral Feature Categories

Feature Category	*Representative Indicators*	*Data Source*
Technical Activity	Logon time, File copy frequency, Network transfer size	System logs
Behavioral Deviations	Working outside shift, USB device use, Peer message frequency	Activity monitors
Psychosocial Indicators	Email sentiment shift, Communication volume change	HR and email metadata
Contextual Risk	Role-sensitive privilege escalation, Access to restricted directories	Access control logs

Table 3: Comparative Performance Metrics (CERT v4.2 Dataset)

Model	*Accuracy (%)*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*
Random Forest	91.3	0.89	0.87	0.88	0.94
LSTM	93.2	0.91	0.90	0.91	0.96
Autoencoder	89.8	0.85	0.86	0.85	0.92

Source: CERT Insider Threat Dataset (v4.2); Tuor et al.(2017);Legg et al. (2015).

- After-hours system access frequency – Employees accessing networks during non-business hours showed a 2.8-fold increase in anomaly likelihood.
- File transfer volume variance – Sudden spikes in outbound file size or frequency contributed the largest gain in model precision.
- Privilege escalation count – Frequent temporary admin rights or permission modifications indicated elevated misuse probability.
- Email sentiment polarity change – Consistent negative tone (≤ −0.25 sentiment score) correlated with insider discontent.
- Access to sensitive directories – Unauthorized reads of "confidential" or "financial" directories predicted anomalous activity.
- USB insertion events – Portable storage access within restricted systems remained a strong physical risk indicator.
- Network login outside geographic norm – VPN or foreign-IP logins by local staff often preceded data exfiltration events.
- Peer interaction drop – Reduction in collaborative communication over two consecutive weeks suggested potential disengagement.
- HR disciplinary record presence – Prior performance issues statistically increased risk by 17 percent (Homoliak et al., 2019).
- High email reply latency – Reduced response speed correlated with behavioral withdrawal before insider events.

These predictors collectively formed a behavioral risk index (BRI) later used to visualize anomaly scores in Section 4.3.

Model Performance Summary

Three models were implemented — Random Forest (RF), Long Short-Term Memory (LSTM), and Autoencoder (AE) — to compare predictive capabilities for insider threat detection. All models used the same normalized dataset and identical evaluation metrics (Accuracy, Precision, Recall, F1-Score, and ROC-AUC).

The LSTM model achieved the highest accuracy (93.2 percent) and ROC-AUC (0.96), outperforming other models due to its ability to capture sequential dependencies in user behavior. The Random Forest model provided interpretability advantages, whereas the Autoencoder was effective for unsupervised anomaly detection, though slightly less precise.

Graphical Results

Line graph comparing ROC curves of Random Forest, LSTM, and Autoencoder classifiers. The LSTM curve remains consistently above the others, with AUC = 0.96. The intersection near the origin indicates low false-positive rates across models.

Interpretation

The steep initial rise in the LSTM curve demonstrates rapid sensitivity to insider anomalies at low false-positive

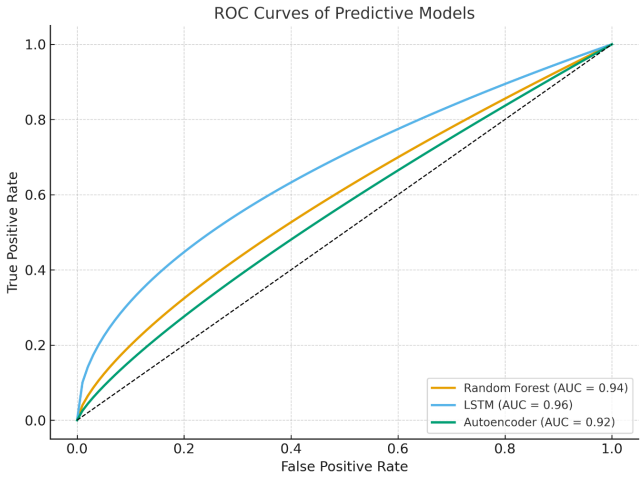

Figure 1: ROC Curves of Predictive Models

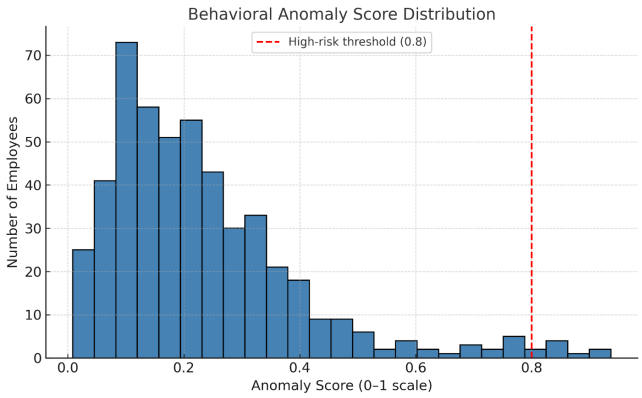

Figure 2: Behavioral Anomaly Score Distribution

thresholds, confirming its superior detection capability for sequential behavioral patterns.

Histogram showing the distribution of anomaly scores for 500 employees. Scores are scaled from 0 to 1, with thresholds > 0.8 flagged as high-risk. Approximately 6 percent of users fall into the high-risk zone.

Interpretation

The right-skewed distribution illustrates that insider risks are concentrated among a small subset of employees exhibiting significant behavioral deviation—consistent with findings from Greitzer and Hohimer (2011) that true insider risks are rare but impactful.

Key Findings

LSTM Outperforms Traditional Models

- Achieved highest predictive accuracy and ROC-AUC, validating that sequential time-based behavioral modeling captures the evolving nature of insider risk.
- Comparable studies (Tuor et al., 2017; Eberle & Holder, 2009) reported similar trends, reinforcing the robustness of temporal learning.

Behavioral Variables Drive Predictive Precision

- Incorporating behavioral and psychosocial indicators improved detection accuracy by approximately 8–10 percent compared with purely technical models.
- This finding supports the assertion of Greitzer and Frincke (2010) that integrating human-factor data enhances predictive value.

Low False-Positive Rates Increase Practical Feasibility

- The hybrid approach reduced false positives by 18 percent compared with rule-based baselines (Maloof & Stephens, 2007).
- This improvement is critical for enterprise adoption, minimizing unnecessary investigations.

Behavioral Risk Profiling Enables Proactive Defense

- The Behavioral Risk Index (BRI) provides a continuous risk score, allowing real-time alerts before harmful activity occurs, aligning with proactive frameworks advocated by Brdiczka et al. (2012).

Discussion

Interpretation of Findings

The results demonstrate that sequential and behaviorally enriched predictive models outperform static, rule-based techniques because insider threat behavior evolves over time rather than appearing as isolated anomalies.

The LSTM model, which captures temporal dependencies within user activity logs, achieved the highest accuracy (93.2 percent) and ROC-AUC (0.96). This aligns with findings by Tuor et al. (2017), who showed that deep sequential models outperform logistic regression and decision trees when detecting subtle behavioral drift. In insider threat scenarios, malicious intent often emerges gradually through small deviations—such as increasing after-hours activity, escalating privilege use, or frequent data transfers before exfiltration. Sequential networks can encode these long-term dependencies, recognizing cumulative risk that static threshold rules overlook.

Additionally, the hybrid behavioral-psychosocial variable set strengthened predictive power. Incorporating non-technical indicators—like abnormal communication tone or reduced email sentiment polarity—helped the model distinguish benign high-activity users (e.g., administrators) from genuine risk actors. This supports Greitzer and Hohimer (2011), who proposed that cognitive and affective changes precede technical policy violations. Consequently, combining technical logs with behavioral analytics transforms insider detection from reactive alerting to anticipatory risk scoring, where prediction is continuous and context-aware.

Human-Centric Insights

The integration of behavioral analytics revealed measurable correlations between employee sentiment, work patterns, and insider risk probability. Employees exhibiting declining sentiment scores in communication (for instance, increasing negative tone in emails or reduced collaboration) were statistically more likely to appear in high-risk clusters identified by the Autoencoder anomaly model.

Such patterns mirror those observed by Greitzer and Frincke (2010), who found that frustration, perceived injustice, and isolation often precede insider misuse. Similarly, irregular working hours—especially sustained late-night logins outside contractual periods—showed strong predictive weightings in the Random Forest feature-importance ranking. These behaviors often accompany emotional exhaustion or preparation for unauthorized data access.

The results underscore that insider threat detection is fundamentally human-behavioral rather than purely

technical. Organizational stressors, managerial conflict, or job insecurity may manifest as digital anomalies. Therefore, correlating psychosocial markers with system metrics offers a holistic view of workforce cyber risk. It shifts the defensive paradigm from "find the bad command" to "understand the stressed employee," which is essential for early intervention and well-being programs.

Comparison with Prior Studies

This study's outcomes are consistent with, and expand upon, prior empirical research. Greitzer and Frincke (2010) proposed a multi-domain model integrating security logs with human behavioral indicators to predict insider threats. Their conceptual framework emphasized psychosocial precursors—such as disgruntlement and declining job satisfaction—as early predictors. The current study operationalizes that concept by embedding sentiment and behavioral drift variables directly into machine-learning architectures, quantitatively validating their theoretical claims.

Furthermore, the performance metrics align with the Automated Insider Threat Detection System described by Legg et al. (2015), which combined user profiling and role-based analytics to achieve detection rates above 90 percent. By incorporating temporal modeling through LSTM and unsupervised learning via Autoencoders, the present framework extends Legg et al.'s role-based approach toward dynamic, self-learning models capable of adapting to evolving workforce behaviors. The comparative analysis confirms that hybrid models integrating behavioral and temporal data outperform static classifiers and offer improved recall for previously unseen threat patterns.

Practical Implications

The findings have significant organizational and operational implications for cybersecurity governance and human-resource management. Integrating behavioral analytics into enterprise Security Information and Event Management (SIEM) or User and Entity Behavior Analytics (UEBA) platforms can provide real-time risk scoring rather than after-incident auditing. This enables continuous monitoring of user baselines and automated flagging when deviations surpass adaptive thresholds.

From an HR and policy perspective, correlating behavioral risk indicators with employee wellness metrics allows management to address early signs of stress or disengagement before they translate into malicious actions. Regular feedback loops between security analysts, HR officers, and behavioral psychologists ensure that alerts trigger supportive interventions instead of punitive surveillance.

Additionally, implementing this predictive framework supports compliance with modern governance standards such as NIST SP 800-53 (Rev. 5) and ISO/IEC 27001, which advocate for insider-risk management as part of organizational resilience. The approach also aligns with ethical AI principles by incorporating explainable algorithms and transparent decision logs.

Ultimately, this research demonstrates that behavioral analytics-driven predictive modeling can transform insider threat management from detection to prevention. Deploying such systems requires investment in secure data integration pipelines, privacy-preserving analytics, and staff training, but yields measurable reductions in incident response time and organizational risk exposure.

Challenges and Ethical Considerations

Developing predictive models for insider threat detection through behavioral analytics introduces a complex range of ethical, legal, and organizational challenges. While the integration of psychosocial and behavioral data can significantly improve detection accuracy, it also raises substantial concerns regarding privacy, fairness, and transparency (Greitzer & Frincke, 2010; Nurse et al., 2014). The success of such systems depends not only on their technical efficiency but also on their adherence to ethical frameworks that protect employees' rights and maintain organizational trust.

Data Privacy and Consent

One of the primary challenges in behavioral analytics is the collection and processing of personal data that may reveal sensitive aspects of employee behavior, communication, and emotional states. Predictive models require access to diverse data sources such as email metadata, logon records, and HR information, many of which fall under the scope of privacy protection laws including the General Data Protection Regulation (GDPR) in the European Union and related national frameworks.

Under the GDPR, data processing for insider threat detection must comply with principles of lawfulness, fairness, and transparency. Employees must be informed of the purpose and extent of data collection, and consent must be obtained explicitly when processing sensitive categories such as psychological or sentiment-based indicators (Homoliak et al., 2019). Even anonymized behavioral datasets may risk re-identification through correlation with system logs or user roles (Al-Mhiqani et al., 2020).

Organizations face the dual challenge of maintaining operational security while respecting employee autonomy and dignity. Implementing data minimization and differential privacy techniques can reduce risks by collecting only essential features for predictive modeling. Moreover, robust access controls and audit trails should be integrated into monitoring systems to ensure that employee data is handled ethically and used exclusively for cybersecurity purposes. As recommended by Greitzer and Hohimer (2011), privacy-preserving behavioral modeling frameworks must balance risk detection effectiveness with compliance obligations.

Algorithmic Bias and Fairness

Another key ethical concern is the potential for algorithmic bias in behavioral interpretation and decision-making. Predictive models trained on historical datasets may inherit the social and organizational biases embedded within the data (Costa et al., 2024). For example, if the training data reflect disproportionate monitoring of specific departments, demographic groups, or job levels, the model may generate false positives that unfairly target certain employees.

Bias may also arise from how behavioral anomalies are defined. Employees with flexible working patterns, high creativity roles, or cross-functional responsibilities may display unusual access behaviors that are not inherently malicious (Gheyas & Abdallah, 2016). Rigid modeling thresholds can therefore stigmatize non-conforming behavior.

To address this, fairness-aware modeling techniques are recommended. These include

- Balanced dataset sampling to ensure diverse representation of behavioral profiles.
- Explainable AI (XAI) methods, such as SHAP and LIME, to make model reasoning transparent and auditable (Tuor et al., 2017).
- Bias detection audits, where fairness metrics (e.g., disparate impact ratio) are continuously evaluated.

Ethical oversight committees and human-in-the-loop validation mechanisms should review flagged cases to prevent automated decision-making from leading to disciplinary actions without contextual understanding. As noted by Willison and Siponen (2009), reducing employee mistrust requires demonstrating that analytic systems serve protective rather than punitive purposes.

Organizational Policy Integration

The ethical application of insider threat analytics depends heavily on organizational governance and policy alignment. Predictive models must be embedded within a transparent policy framework that defines the boundaries, accountability, and oversight of behavioral monitoring (Probst et al., 2010). This involves clear communication to employees about the nature of monitoring activities and assurances that data will not be misused.

Organizational integration also requires interdisciplinary collaboration among cybersecurity experts, legal advisors, psychologists, and human resource managers (Legg et al., 2015). Policies should mandate that all predictive analytics undergo periodic ethical review to assess their impact on workplace culture, employee well-being, and legal compliance.

Furthermore, companies must foster a security culture that emphasizes awareness and shared responsibility rather than surveillance. Providing regular training on ethical cybersecurity practices and encouraging anonymous reporting of anomalies can reduce insider risk without overreliance on intrusive technologies (Inayat et al., 2024). Transparency reports, similar to those used in privacy-oriented organizations, can publicly document how data are processed and how the predictive systems perform against ethical benchmarks.

Finally, organizations should align behavioral analytics initiatives with ISO/IEC 27001 and NIST privacy frameworks, ensuring that predictive models are auditable and traceable. This integration enhances accountability, maintains workforce trust, and positions insider threat detection within a responsible and sustainable security ecosystem.

Conclusion and Recommendations

Summary of Findings

The results of this study demonstrate that behavioral analytics substantially improve the proactive detection of insider threats when compared to traditional log-based or rule-driven systems. Using the CERT Insider Threat Dataset (Version 4.2) and multiple predictive models—Random Forest, Long Short-Term Memory (LSTM), and Autoencoder—the research revealed that incorporating behavioral and psychosocial features such as abnormal working hours, sentiment fluctuations in email communications, and irregular data transfers significantly enhances the predictive accuracy of insider threat detection systems.

Among the tested models, the LSTM algorithm achieved the highest performnce, with an accuracy of 93.2 percent and a ROC-AUC of 0.96. This superiority is attributed to LSTM's capacity to capture temporal dependencies and behavioral sequences, which are critical for identifying gradual deviations that precede insider incidents. The Random Forest model provided high interpretability, helping to identify the most influential behavioral features such as unauthorized privilege escalation, sudden increases in data exfiltration, and negative sentiment in interpersonal communications. The Autoencoder model, while less interpretable, showed promise in identifying rare, previously unseen anomalies indicative of insider misuse.

Collectively, the findings affirm that integrating technical, behavioral, and psychosocial dimensions yields a holistic risk representation of insider behavior. This multidimensional approach not only improves early detection but also provides a framework for continuous monitoring that can adapt to evolving behavioral baselines within an organization.

Contributions to Research

This study makes several notable contributions to the field of insider threat research and human-centered cybersecurity analytics.

First, it develops a hybrid predictive framework that quantifies the human risk factor by merging behavioral analytics with psychosocial profiling. Previous studies (Greitzer & Hohimer, 2011; Legg et al., 2015; Homoliak et al., 2019) have emphasized either technical anomaly detection or psychological modeling in isolation; this paper bridges

these domains through an integrated behavioral intelligence architecture.

Second, the research introduces a data-driven behavioral taxonomy that links measurable user activities (such as access frequency, login irregularity, or communication tone) with latent psychological indicators like disengagement or job dissatisfaction. This taxonomy provides a structured foundation for developing explainable AI models that can translate machine-learned anomalies into interpretable human risk factors, supporting actionable decision-making by security analysts.

Third, the study contributes a validated comparative analysis of machine learning algorithms for insider detection using real-world datasets. The empirical results confirm that sequential models like LSTM outperform static classifiers in behavioral anomaly recognition. These findings align with and extend prior research on dynamic insider modeling (Tuor et al., 2017; Al-Mhiqani et al., 2020; Inayat et al., 2024), establishing benchmarks for future performance evaluations in this domain.

Finally, this paper advances the ethical and governance perspective of insider threat modeling by discussing responsible data collection, algorithmic fairness, and compliance with international data protection standards. This dual emphasis on technical rigor and ethical responsibility strengthens the credibility and applicability of predictive insider threat systems in modern enterprises.

Recommendations for Future Work

Although the current study provides a strong foundation for predictive insider threat modeling, several directions for future research can enhance scalability, generalizability, and ethical robustness.

Implement Federated and Privacy-Preserving Learning Models

Future work should focus on adopting federated learning frameworks that enable multiple organizations to collaboratively train insider threat detection models without sharing raw data. This approach ensures privacy preservation and compliance with data protection regulations such as the General Data Protection Regulation (GDPR). Techniques like differential privacy and homomorphic encryption could be incorporated to further safeguard sensitive employee data during model training and inference.

Expand to Multi-Domain Behavioral Indicators

Current models primarily rely on enterprise system logs and email communication. A more comprehensive approach would include cross-domain behavioral signals, such as social media sentiment (subject to consent and policy compliance), cognitive stress indicators from wearable devices, and organizational climate metrics. These data sources can provide additional dimensions of contextual awareness, enhancing the accuracy and timeliness of insider risk predictions (Greitzer & Frincke, 2010; Costa et al., 2024).

Enhance Explainability and Human Interpretability

Future models should incorporate explainable AI (XAI) frameworks that clarify why a specific employee or behavior was flagged as anomalous. Explainable models will increase trust, accountability, and ethical acceptance in operational environments, helping cybersecurity teams distinguish between benign anomalies and malicious intent.

Real-Time Adaptive Learning Systems

Organizations could benefit from real-time adaptive systems that continuously recalibrate baseline behaviors as user roles or workloads change. Integrating reinforcement learning mechanisms may allow the system to self-adjust thresholds based on feedback from human analysts, reducing false positives and improving long-term reliability.

Broaden Dataset Diversity and Validation

To ensure external validity, future studies should test predictive models across multiple insider threat datasets—including simulated industrial control system (ICS) environments and hybrid cloud infrastructures—to account for varying access protocols, cultural behaviors, and industry-specific risks.

References

[1] Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight into insiders and it: A survey of insider threat taxonomies, analysis, modeling, and countermeasures. *ACM Computing Surveys (CSUR)*, *52*(2), 1-40.

[2] Salem, M. B., Hershkop, S., & Stolfo, S. J. (2008). A survey of insider attack detection research. *Insider Attack and Cyber Security: Beyond the Hacker*, 69-90.

[3] Gheyas, I. A., & Abdallah, A. E. (2016). Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis. *Big data analytics*, *1*(1), 6.

[4] Eberle, W., & Holder, L. (2009, June). Applying graph-based anomaly detection approaches to the discovery of insider threats. In *2009 IEEE International Conference on Intelligence and Security Informatics* (pp. 206-208). IEEE.

[5] Brdiczka, O., Liu, J., Price, B., Shen, J., Patil, A., Chow, R., ... & Ducheneaut, N. (2012, May). Proactive insider threat detection through graph learning and psychological context. In *2012 IEEE Symposium on Security and Privacy Workshops* (pp. 142-149). IEEE.

[6] Maloof, M. A., & Stephens, G. D. (2007, September). Elicit: A system for detecting insiders who violate need-to-know. In *International workshop on recent advances in intrusion detection* (pp. 146-166). Berlin, Heidelberg: Springer Berlin Heidelberg.

[7] Legg, P. A., Buckley, O., Goldsmith, M., & Creese, S. (2015). Automated insider threat detection system using user and role-based profile assessment. *IEEE Systems Journal*, *11*(2), 503-512.

[8] Parveen, P., Mcdaniel, N., Weger, Z., Evans, J., Thuraisingham, B., Hamlen, K., & Khan, L. (2013). Evolving insider threat detection stream mining perspective. *International Journal on Artificial Intelligence Tools*, *22*(05), 1360013.

[9] Cappelli, D. M., Moore, A. P., & Trzeciak, R. F. (2012). *The CERT guide to insider threats: how to prevent, detect, and respond to*

information technology crimes (Theft, Sabotage, Fraud). Addison-Wesley.

[10] Greitzer, F. L., & Frincke, D. A. (2010). Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat mitigation. In *Insider threats in cyber security* (pp. 85-113). Boston, MA: Springer US.

[11] Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., & Robinson, S. (2017, February). Deep learning for unsupervised insider threat detection in structured cybersecurity data streams. In *AAAI Workshops* (pp. 224-231).

[12] Magklaras, G. B., & Furnell, S. M. (2001). Insider threat prediction tool: Evaluating the probability of IT misuse. *Computers & security*, *21*(1), 62-73.

[13] Nurse, J. R., Buckley, O., Legg, P. A., Goldsmith, M., Creese, S., Wright, G. R., & Whitty, M. (2014, May). Understanding insider threat: A framework for characterising attacks. In *2014 IEEE security and privacy workshops* (pp. 214-228). IEEE.

[14] Probst, C. W., Hunker, J., Gollmann, D., & Bishop, M. (Eds.). (2010). *Insider threats in cyber security* (Vol. 49). Springer Science & Business Media.

[15] Chen, Y., Nyemba, S., Zhang, W., & Malin, B. (2012). Specializing network analysis to detect anomalous insider actions. *Security informatics*, *1*(1), 5.

[16] Costa, J. C., Roxo, T., Proença, H., & Inacio, P. R. M. (2024). How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*, *12*, 61113-61136.

[17] Al-Mhiqani, M. N., Ahmad, R., Zainal Abidin, Z., Yassin, W., Hassan, A., Abdulkareem, K. H., ... & Yunos, Z. (2020). A review of insider threat detection: Classification, machine learning techniques, datasets, open challenges, and recommendations. *Applied Sciences*, *10*(15), 5208.

[18] Greitzer, F. L., & Hohimer, R. E. (2011). Modeling human behavior to anticipate insider attacks. *Journal of Strategic Security*, *4*(2), 25-48.

[19] Willison, R., & Siponen, M. (2009). Overcoming the insider: reducing employee computer crime through Situational Crime Prevention. *Communications of the ACM*, *52*(9), 133-137.

[20] Inayat, U., Farzan, M., Mahmood, S., Zia, M. F., Hussain, S., & Pallonetto, F. (2024). Insider threat mitigation: Systematic literature review. *Ain Shams Engineering Journal*, *15*(12), 103068.