User Behavior Modeling

Sachit Desai   
*CS Department  
SIT PUNE*  
[sachit.desai.btech2021@sitpune.edu.in](mailto:sachit.desai.btech2021@sitpune.edu.in)

Sarothi Adhikary   
*CS Department  
SIT PUNE*  
sarothi.adhikary.btech2021@sitpune.edu.in

Tejas Ulabhaje   
*CS Department  
SIT PUNE*  
tejas.ulabhajebtech2021@sitpune.edu.in

*Abstract*— User behavior modelling is an important area of research in fields such as marketing, cyber security as well as psychology. In today’s world understanding and predicting user behavior is essential for designing effective systems which provides market strategies and helps in providing personalized user experiences. The following report delves into how user behavior is used in the field of cyber security including key factors such as the role of data collection and analysis as well as the application of machine learning techniques. Traditional security measures such as firewalls and antivirus software are no longer sufficient approach to defend against cyber threats. Therefore, User Behavior Modelling as emerged as a promising approach. By using such techniques professionals can identify deviations from a pattern, which may indicate a security breach or unauthorized access. In this study, we gather a collection of biometrics data such as Keyboard, Mouse and Touchscreen (KMT) dynamics. The dataset consists legitimate and illegitimate data entries of users, collected during each user entry session. We then use Decision Tree Classifier, to classify whether a user is authentic or not.

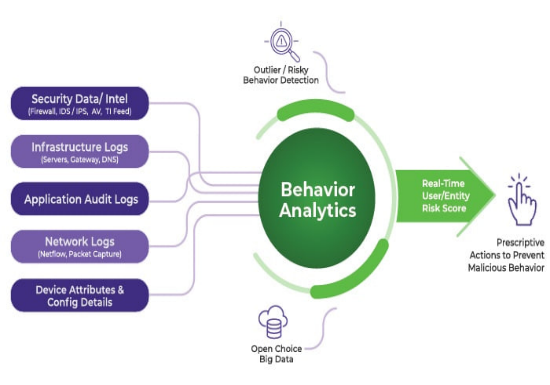
# INTRODUCTION

In today's technologically advanced landscape, organizations face an increasingly sophisticated array of cyber threats, including those that target user login credentials, jeopardizing sensitive data without detection. The potential breach of login credentials raises concerns about unauthorized data access and compromise, demanding the implementation of effective cybersecurity measures. One of the most important vulnerabilities is the human element. Most attacks carried out in recent times include human elements through social engineering or the blatant use of restricted physical systems by an external threat actor. If an unauthorized individual enters a facility through social engineering methods and accesses digital systems it poses a huge potential for data privacy to be violated which can result in the loss of millions of dollars’ worth of intellectual property and other assets. To address this issue, there is a need for a proactive solution that leverages machine learning algorithms to analyze and comprehend user behavior patterns.

Based on these established patterns of user behavior, it is possible to detect whether an unauthorized user is operating a system locally or remotely. Based on these algorithms silent alarms may be triggered prompting the security team to assess the situation and prevent breaches of both types; Social Engineering and Unauthorized remote access.

User Behavior Modelling can also help to potentially prevent zero-day attacks from doing exceeding amounts of damage by detecting abnormalities in how the system is being used although that is not the scope of this project.

The main scope of this project focuses on insider threats i.e. a user of the same organization accessing restricted resources in a non-flat business model by remotely or locally accessing systems with sensitive data.



[1] Figure Role of Behavior Analytics

# Literature Review

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Reference No. | Method Used | Dataset Used | Performance Parameters | Findings/Limitation/Future Scope |
| [2] | Using Supervised Learning to develop models to predict user behavior | True/fake content containing tweets containing at least one link to an article which can either be true or false | Silhouette, Calinski-Harabaz, Davies-Bouldin | Limitation in data sampling and extracting of features |
| [4] | The proposed approach involves utilizing the User Experience Scale (UES), which comprises six distinct dimensions: Aesthetic Appeal, Felt Involvement, Focused Attention, Novelty, Perceived Usability, and Endurability | Three sets of data obtained from prior research, consisting of behavioral information extracted from system log files and perception data. | F1Score, Accuracy | Interpreting user behavior and its relationship with the UES and search behavior patterns can be difficult. Finding accurate conclusions requires careful analysis and consideration of various factors that can influence user behavior. |
| [5] | LSTM based autoencoder to identify anomalous data points | Log data from users systems | Accuracy, true positives, false positives | Accuracy of 90.17%, TP 91.03%, FP 9.84%. These results concluded that this methodology could be used in the future for anomaly detection. |
| [6] | One class learning, unary classification, GRU | CERT insider threat dataset | True positives, True negatives, precision, f1 score | TP and TN rates at 79.81%, AUC of 0.87, precision of 80.1%, f1 score of 79.4% |
| [7] | This paper uses Machine learning and Big data analytics to figure out the biometric differences between legitimate and malicious user in logs generated by user browser history and subsequent patterns found in them. | This methodology uses the current users browser history and patterns as a a base dataset for applying its machine learning algorithms.  This data is collected in real time by checking the browser history of the user. | Accurac | The model was able to achieve 70% accuracy in determining whether the user was malicious or non-malicious. |
| [8] | This papers approach was to extract real time users preference to predict user intentions to recommend to the user accurately. The process was divided into two phases, user preference learning and user intention prediction. The model used to make the prediction was long short-term memory (LSTM), CNN, MCM, FPM. | dataset used is two thousand from the scopus repository of research papers. | Accuracy, f1, score, precision, recall | it was found that using LSTM produced the highest accuracy in suggesting the right paper to a user. MCM was lowest, then came FPM, followed by CNN.  LSTM was successful in consistently recommending the right papers to users.  Although the project is in the domain of user modelling, it does not address the issue of user behaviour for detecting malicious intent. |

# Problem Statement

In an increasingly interconnected digital landscape, ensuring the security and integrity of online platforms has become a paramount concern. [2]The rapid proliferation of malicious activities, such as fraud, cyberattacks, and data breaches, poses significant risks to individuals, organizations, and society at large. To mitigate these threats, there is an urgent need to develop advanced user behavior modeling techniques that can accurately detect and preemptively identify malicious behavior within online systems.

The primary challenge lies in the complex and evolving nature of malicious activities. Traditional rule-based approaches and static authentication methods often fall short in detecting sophisticated and novel forms of malicious behavior. As malicious actors continuously adapt their tactics, a dynamic and adaptive solution is required to stay ahead of the ever-changing threat landscape.

Furthermore, [3] user privacy and ethical considerations must be carefully balanced while developing such models. Striking the right equilibrium between safeguarding user information and enabling accurate detection of malicious behavior is of utmost importance.

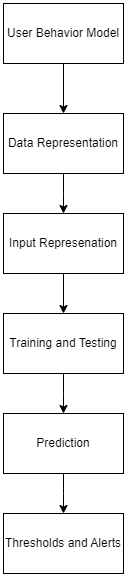
# Objectives

* Design and implement a user behavior modeling system that can accurately distinguish between legitimate and malicious behavior based on real-time data analysis.
* Develop machine learning algorithms capable of detecting emerging forms of malicious behavior through continuous learning and adaptation.
* Conduct experimentation and performance evaluation to measure the accuracy and efficiency of the developed model.

# Proposed Methadology

In order to achieve the problem above, a systematic approach is employed. It begins with data representation where the KMT dynamics of each user is recorded and the converted to a .json file. Each event bears a timestamp linked to the user. Machine Learning models like Support Vector Machines (SVM) or Decision Tree Classifiers are used to identify temporal dependencies. The models are then trained, to predict the subsequent events based on past actions. After training, the model predicts the patterns and anticipates the next event. Deviations from these prediction serves as anomalies. A threshold of the same is defined and if the observed variance between predicted and actual events.

surpasses a predefined threshold, alerts are generated which prompts cybersecurity team to investigate potential security breaches.



* Data Representation: Represent user login and logout activities as a sequence of events, with each event being a timestamped action (log in or logout) associated with a user. This sequence of events becomes your input data.
* Input Representation: Encode each event with appropriate features, such as the time of day, day of the week, location, device used, and so on. These features help the model understand the context of each event.
* Training and Testing: For Testing, various machine learning models, like Support Vector Machine (SVM) or Decision Tree Classifier, can be used. Users tend to have consistent patterns when it comes to logging in and out, and anomalies might break these patterns. By training the model to predict the next event in the sequence based on past events, you create a model that learns the expected behavior. Deviations from the predicted behavior can then be flagged as anomalies.
* Prediction: Train the model on historical data of user behavior. The model learns the usual patterns and can predict the next event based on the sequence. During prediction, if the actual event deviates significantly from the predicted event, it could indicate an anomaly.
* Thresholds and Alerts: Determine thresholds for anomaly detection. If the difference between the predicted and actual events exceeds a certain threshold, generate an alert for the cybersecurity team to investigate.

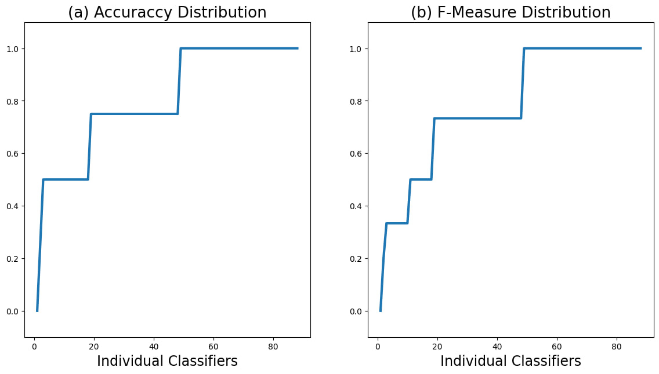
# Results and Discussions

To derive the results from our comparative analysis of multiple machine learning classifier algorithms for user authentication based on user behavior modeling from monitoring data of keyboard strokes and mouse movement we used the following algorithms given with the respective accuracies achieved

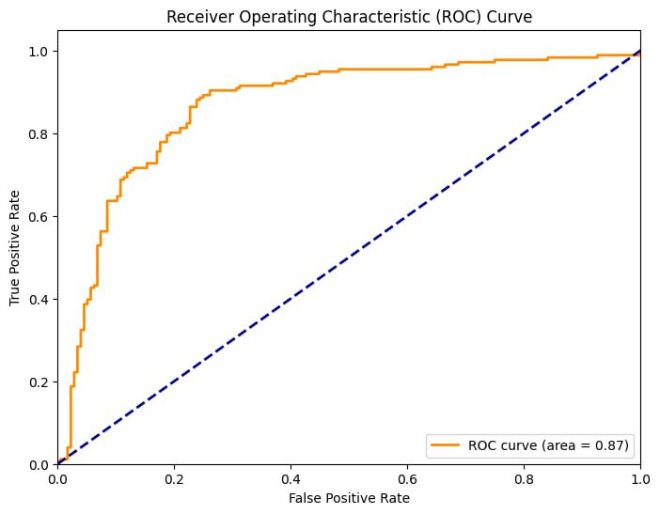
* **Naive Bayes - 80.39**
* Support Vector Machine - 78.12
* Decision Tree – 74.43
* Random Forest – 80.11

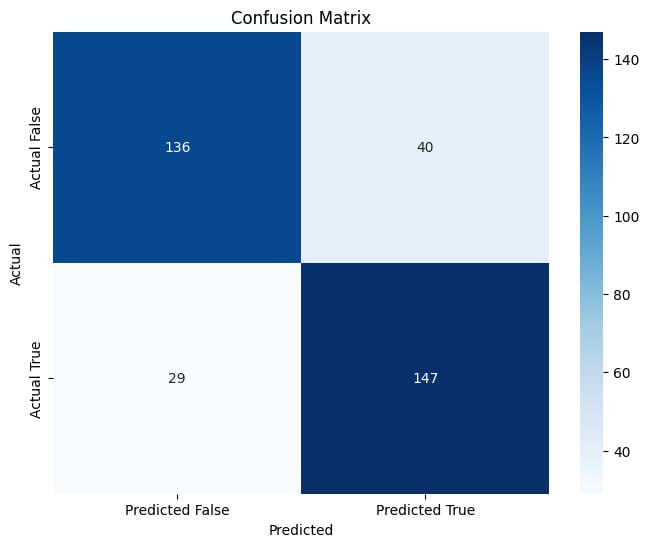
|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F1-score** |
| Support Vector Machine | 63.63 | 58.59 |
| Decision Tree | 74.43 | 72.23 |
| Random Forest | 80.11 | 78.10 |
| **Naive Bayes** | **80.39** | **78.25** |

From the above results, it is glaringly obvious that Naive Bayes shows the highest accuracy for our dataset. With an accuracy of 80.39, it confirms our hypothesis about the potential of drawing accurate predictions about user authentication to bolster organizational security and generate alarms for unauthorized users.



To confirm the performance of our classification model at all classification thresholds we employed a ROC Curve method by plotting true positive rates and false positive rates at different classification thresholds. Further, we measured the entire two-dimensional area under the ROC curve. We were able to achieve an area under ROC value of 0.87 which is considered to be a strong positive indicator of

the models' strength.  


To further solidify our findings we decided to implement a confusion matrix to visualize the performance of our classification algorithm.  
Through the above results, we were able to prove the performance of our model using various evaluation metrics and comparative analysis.

# Comparitive Analysis

A number of similar examinations were done for various multi-class prediction models using K-means and agglomerative clustering. The dataset mostly consisted of a number of sampled users on which clustering is performed to obtain labels subjected to around 10-fold cross-validation. A 60-40 split strategy was used for train-test sets. Overall, the best performance was for MLP and SVM. However, this evidence suggested that user behavior predictive models can be further improved with larger datasets or more features by exploiting real-time data.

# Conclusion

In this paper we have presented a simple machine-learning approach to user behavior modeling for differentiating between fake and authentic users. We identified diverse features from the user behavior categories. Super Learning models depict that user behavior can be predicted from such features. Although our approach has many limitations, mainly the accuracy of our model. Given such limitations, this works as a foundation work which can be improved upon in the upcoming years. This research provides groundwork for advanced user analysis methods focused on cyber security research.

##### References

[1] https://gurucul.com/blog/how-behavior-analytics-improves-cybersecurity

[2]Zahra Rajabi, Amardu Shehu, Hemant Purohit: User Behavior Modelling for Fake Information Mitigation on Social Web (May 2019)

[3] Nnamoko, Nonso; Barrowclough, Joe; Liptrott, Mark; Korkontzelos, Ioannis (2022), “Behaviour Biometrics Dataset”, Mendeley Data, V1, doi: 10.17632/fnf8b85kr6.1

[4] July 9-12, 2017 Modelling User Behavior based on Process Mengdie Zhuang University of Shefield Management School Shefield, UK.

[5] IAIT2020: Proceedings of the 11th International Conference on Advances in Information TechnologyJuly 2020 [5] Neely-Barnes, S.L., Hall, H.R., Roberts, R.J. and Graff, J.C., 2011.

[6] Nepal, Santosh & Joshi, Basanta. (2022). User Behavior Analytics for Insider Threat Detection using Deep Learning.

[7] March 2022Rohit Ranjan, Shashi Shekhar Kumar,

User behaviour analysis using data analytics and machine learning to predict malicious user versus legitimate user,High-Confidence Computing,Volume 2, Issue 1,2022, 100034,ISSN 2667-2952

[8] Modeling User Behaviour in Research Paper Recommendation System Arpita Chaudhuri1 · Debasis Samanta1 · Monalisa Sarma