IRSTI 28.23.39 https://doi.org/10.26577/jpcsit20253202

Marzhan Idrissova, Sabina Kim,
Beibut Amirgaliyev, Didar Yedilkhan*, Leila Rzayeva

Astana IT University, Astana, Kazakhstan
*e-mail: d.yedilkhan@astanait.edu.kz

DIGITAL FOOTPRINTS: CLUSTERING BROWSER HISTORY FOR USER PROFILING USING MACHINE LEARNING

Abstract. This study explores the use of unsupervised machine learning techniques to analyze historical web activity, segment users, and detect anomalies for user profiling. By applying hierarchical clustering and Gaussian Mixture Models, we identified distinct browsing behaviors, categorizing users into four to five groups, including general browsing, social media engagement, high-bandwidth consumption, and automated system processes. For anomaly detection, One-Class SVM and Isolation Forest were employed to flag deviations from expected behavior. The results indicate that approximately 5% of sessions were classified as anomalous by SVM, while Isolation Forest highlighted outliers associated with extended session durations and potentially high-risk application usage. These findings underscore the effectiveness of machine learning in distinguishing user behavior through digital footprints while identifying potential security threats or atypical usage patterns. The study demonstrates that unsupervised learning can serve as a valuable tool for user profiling and behavioral analysis, with implications for cybersecurity, network monitoring, and online behavior modeling. Integrating clustering with anomaly detection provides a scalable approach for uncovering usage trends and deviations in web traffic. Future research should expand dataset coverage and incorporate adaptive models to enhance classification accuracy and responsiveness to evolving web behaviors.

Key words: digital footprints, user profiling, clustering, anomaly detection, browsing behavior, machine learning, network traffic analysis.

1. Introduction

In today's digital landscape, individuals continuously generate extensive digital footprints, with browsing history serving as a valuable resource for analyzing online behavior. Profiling users based on their web activity has broad applications across cybersecurity, personalized recommendations, behavioral psychology, and anomaly detection. Research suggests that browsing patterns can offer insights into psychological traits, aid in personality prediction, and reveal demographic information. With advancements in machine learning, clustering techniques have emerged as a powerful tool for segmenting users based on their browsing habits.

Recent studies highlight the deep connection between web activity and individual psychological and demographic characteristics. For example, Kelly and Sharot [1] found that specific browsing patterns, such as frequent news consumption, correlate with anxiety and mood fluctuations. Similarly, Lytvyn et al. [2] examined the relationship between browsing behavior and the Big Five personality traits, demonstrating how machine learning can extract behavioral insights from digital footprints. Further, Lien, Bai, and Chen [3] established that browsing logs can accurately predict demographic factors such as age and gender. These findings suggest that online activity not only reflects user interests but also provides a window into deeper psychological attributes.

From a security perspective, Paul and Medhe [4] applied machine learning to detect anomalies in browsing behavior, demonstrating how clustering algorithms can distinguish between typical and atypical users–an approach particularly useful for identifying insider threats or risky activity. Similarly, Salomatin et al. [5] explored browser fingerprints for user identification, showing that even anonymized browsing data can be leveraged for profiling. Collectively, these studies underscore the potential of browsing history as a rich source of behavioral insights, with implications for both user profiling and security risk assessment.

This study examines the effectiveness of clustering techniques in segmenting users based on their browsing activity, identifying distinct behavioral patterns, and detecting potential anomalies. By applying machine learning models, we aim to assess

how well various clustering methods classify browsing behaviors and whether meaningful patterns emerge from real-world web activity. Additionally, the study explores the role of anomaly detection in identifying deviations from typical user behavior, which may signal unusual or high-risk activities.

To achieve these objectives, this research seeks to answer the following key questions:

- To what extent can machine learning techniques cluster users based on their browsing activity?

- What behavioral patterns emerge from the analysis of browsing logs?

- How effective are anomaly detection methods in identifying atypical browsing behavior?

We hypothesize that applying machine learning-based clustering will reveal distinct user behavior patterns, ranging from general web browsing to high-data consumption and potentially anomalous activity. Clustering techniques are expected to differentiate users based on factors such as browsing frequency, session duration, and the nature of web interactions. Additionally, anomaly detection models should effectively identify outliers whose browsing behaviors significantly deviate from the norm.

By addressing these research questions and testing this hypothesis, the study aims to contribute to the growing field of digital footprint analysis, demonstrating the utility of clustering techniques in profiling web activity. The dataset consists of anonymized browsing logs, ensuring compliance with privacy and ethical guidelines while enabling meaningful behavioral insights.

2. Materials and Methods

This section details the dataset, preprocessing techniques, clustering and anomaly detection methods, and statistical analyses employed in this study. The methodology is structured to ensure replicability, enabling future researchers to apply similar approaches for user behavior profiling based on browsing history.

2.1. Dataset and Experimental Design

The data set used in this study consists of 138,105 web browsing records collected during October 2024 from the Astana IT University network using Fortigate logging. The logs contain structured data capturing various aspects of browsing behavior, making them well-suited for machine learning-based segmentation and anomaly detection. While this dataset originates from a university environment, the methodology can be generalized to other network-based browsing datasets, such as corporate or public Wi-Fi networks.

The dataset includes several key attributes essential for behavioral analysis [6], [7], shown in Table 1.

Table 1 – Dataset key attributes.

Key Attribute	Definition
Timestamp data	The precise date and time of each web request, allowing for temporal analysis of browsing patterns.
Source and destination IPs	Unique identifiers of browsing sessions, enabling session-based behavioral profiling.
Application and category	Information about the type of accessed resources, such as educational platforms, social media, entertainment or cloud services.
Application risk level	A built-in risk assessment score, categorizing the potential security impact of visited services.
Traffic volume	The number of bytes sent and received, useful for identifying high-traffic users or outlier activity.
Action taken	Whether the request was allowed, blocked or flagged by security policies.

Including categorical and numerical features in the dataset allows for applying unsupervised learning algorithms, facilitating user segmentation and anomaly detection [8], [9]. The data also presents an opportunity to study temporal trends in browsing behavior, offering insights into variations in web activity across different periods [10].

2.2. Data Preprocessing and Feature Engineering

Before applying machine learning models, the dataset underwent multiple preprocessing steps to ensure data consistency and quality. This included:

- Handling missing values: Incomplete records were removed to maintain dataset reliability.

- Encoding categorical variables: Categorical attributes such as application category, risk level and action type were converted into numerical values using label encoding, ensuring compatibility with clustering algorithms [4].

- Timestamp processing: The date and time fields were combined into a single datetime column, enabling more precise time-based clustering [3].

- Feature scaling: Traffic volume features (sent and received bytes) were normalized using StandardScaler, preventing numerical dominance in distance-based clustering models [11].

By structuring the data this way, we ensured that both clustering and anomaly detection models operated optimally, producing meaningful and interpretable results [12].

2.3. Clustering Methods

To segment users based on browsing behavior, we applied two different clustering techniques in Table 2.

Table 2 – Clustering parameters.

Clustering Method	Number of Clusters	Distance Metric	Linkage Method/ Covariance Type
Hierarchical Clustering	5	Euclidean	Ward
Gaussian Mixture Model (GMM)	4	Mahalanobis	Full

2.3.1. Hierarchical Clustering

Hierarchical clustering was selected due to its ability to group similar users without requiring a predefined number of clusters [13]. Using the Ward linkage method Equation (2), we constructed a dendrogram, which suggested an optimal cluster count of five.

$$\mathrm{d(x,y)} = \sqrt{\sum_{i-1}^{n}(x_i - y_i)^2} \tag{1}$$

$$D(A,B) = \sum_{i\in A\cup B} ||x_i - \mu_{A\cup B}||^2 - \sum_{i\in A}\left||x_i - \mu_A\right|^2 - \sum_{i\in B} ||x_i - \mu_B||^2 \tag{2}$$

where d(x,y) is the Euclidean distance between two data points x and y, and n represents the number of features. D(A,B) is the dissimilarity measure between clusters A and B, and μ_A and μ_D are the centroids of clusters A and B, respectively.

Each cluster represented a distinct browsing pattern, characterized by:

- Academic and research-oriented users: Frequently accessed educational and research platforms.

- Social and entertainment users: Engaged in high social media and streaming activity.

- Automated system traffic: Repetitive browsing logs likely generated by background processes.

- High-volume users: Consumed large amounts of network bandwidth, possibly for downloads.

- General browsing users: Displayed diverse and balanced web activity across multiple categories.

The hierarchical clustering approach provided a well-structured segmentation, allowing for a comprehensive interpretation of different browsing behaviors [14].

2.3.2. Gaussian Mixture Model (GMM)

Since hierarchical clustering does not support overlapping user behaviors, we also implemented a Gaussian Mixture Model, which assigns probabilistic cluster memberships to users. This approach was particularly useful for identifying users whose behavior fits into multiple categories [15].

$$\mathrm{p(x)} = \sum_{k-1}^{K} \pi_k \mathrm{N(x|}\mu_k\mathrm{,}\Sigma_\mathrm{k}\mathrm{)} \tag{3}$$

where p(x) is the probability density function of data point x, π_k is the weight of cluster k, $N(x|\mu_k, \Sigma_k)$ represents the Gaussian distribution with mean μ(k) and covariance matrix Σ_k.

The GMM algorithm produced four main clusters:

- Academic users: Primarily accessed learning materials.
- Entertainment-oriented users: High engagement with media and social platforms.
- High-bandwidth consumers: Showed extensive data usage.
- Background process users: Repetitive browsing, potentially system automated.

By comparing hierarchical clustering and GMM, we confirmed that behavioral segmentation was robust, with both models producing complementary insights [16].

2.4. Anomaly Detection Approaches

While clustering helped segment users based on browsing activity, anomaly detection was employed to identify users whose behavior deviated significantly from typical patterns. We applied two different methods, shown in Table 3.

Table 3 – Anomaly detection parameters.

Anomaly Detection Method	Kernel / Splitting Criterion	Anomaly Threshold	Output
One-Class SVM	RBF	ν = 0.02	Binary classification (-1 = anomaly, 1 = normal)
Isolation Forest	Random subspace partitioning	Contamination = 0.05	Binary classification (-1 = anomaly, 1 = normal)

2.4.1. One-Class SVM

One-Class Support Vector Machines (SVM) are commonly used for detecting outliers in high-dimensional data. This technique was trained on the majority of normal browsing behavior and then used to flag records that significantly deviated from this pattern [17].

$$\min_{\omega,\rho} \frac{1}{2}||\omega||^2 + \frac{1}{\nu n}\sum_{i-1}^{n} \xi_i - \rho \quad (4)$$

$$\left(\omega \cdot \phi(x_i)\right) \geq \rho - \xi_i, \xi_i \geq 0 \quad (5)$$

where ω is the normal vector of the decision boundary, ρ is the offset, ξ_i are slack variables and ν controls the fraction of anomalies in the dataset.

The model identified 3,092 anomalies, which included users with:

- Unusual access to restricted categories.
- Irregular spikes in data consumption.
- Multiple blocked access attempts.

These results suggest that One-Class SVM is effective in identifying potentially high-risk users or policy violations [5].

2.4.2. Isolation Forest

Isolation Forest, an ensemble learning method, was used as a secondary anomaly detection technique. Unlike SVM, which relies on decision boundaries, Isolation Forest isolates anomalies faster by randomly partitioning the dataset [18].

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}} \quad (6)$$

where s(x) is the anomaly score, E(h(x)) is the average path length of data point x in the isolation trees and c(n) is a normalization factor based on dataset size n.

Key findings from the Isolation Forest model include:

- A subset of users exhibited consistent high-risk behavior, aligning with SVM anomalies.
- Some flagged users engaged in unusual browsing sessions during non-peak hours.
- The method successfully distinguished between general outliers and extreme cases of high-risk activity.

Together, One-Class SVM and Isolation Forest provided a robust framework for detecting anomalous browsing behavior, reinforcing the effectiveness of machine learning in digital footprint analysis [19].

2.5. Statistical Analysis and Data Visualization

The distribution of users across clusters was visualized through graphs, highlighting the number of groups and their defining characteristics. By visualizing anomalies, we were able to distinguish typical browsing behavior from potentially suspicious activity.

Additionally, a risk assessment was conducted for users associated with high-risk applications, enabling the identification of groups that may pose potential security threats. A comparative analysis of Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) helped determine the overlap in detected anomalies, suggesting the presence of specific behavioral patterns among users.

The study's findings were presented through cluster distributions, user profiles, and detected anomalies, facilitating a comprehensive analysis of digital footprints and key behavioral traits [20]. A multi-stage machine learning pipeline was implemented, integrating clustering for segmentation and anomaly detection for security assessment. The dataset, enriched with temporal and categorical features, underwent cleaning, normalization, and encoding before being processed through hierarchical clustering, GMM, One-Class SVM, and Isolation Forest. Statistical validation and visualization confirmed the effectiveness of these models, demonstrating that browsing logs can yield valuable behavioral insights when analyzed using machine learning techniques.

3. Results

The clustering analysis effectively segmented users into distinct behavioral groups based on their browsing activity. Hierarchical clustering identified five clusters, while the Gaussian Mixture Model (GMM) produced four user groups. The results indicate clear behavioral differentiation, with clusters varying in web usage patterns, data consumption, and the types of applications accessed.

Beyond cluster-specific characteristics, a statistical analysis of browsing behavior provided additional insights. The average session duration across all users was approximately 9.5 minutes, with a median of 7.2 minutes, suggesting that most browsing sessions were relatively brief. The mean volume of data transferred per session was 12.8 KB sent and 76.4 KB received, reflecting typical web activity. Most users accessed low- to moderate-risk applications, with a mean application risk score of 1.85 on a scale where higher values indicate riskier content.

These statistical findings contextualize the cluster distributions and anomaly detection results, reinforcing the validity of the identified behavioral segments and providing a quantitative foundation for interpreting user activity patterns.

3.1. Hierarchical Clustering Findings

Hierarchical clustering using the Ward linkage method identified five distinct user groups based on browsing activity. Figure 1 illustrates the clustering structure:

The summary of cluster characteristics is presented in Table 4.

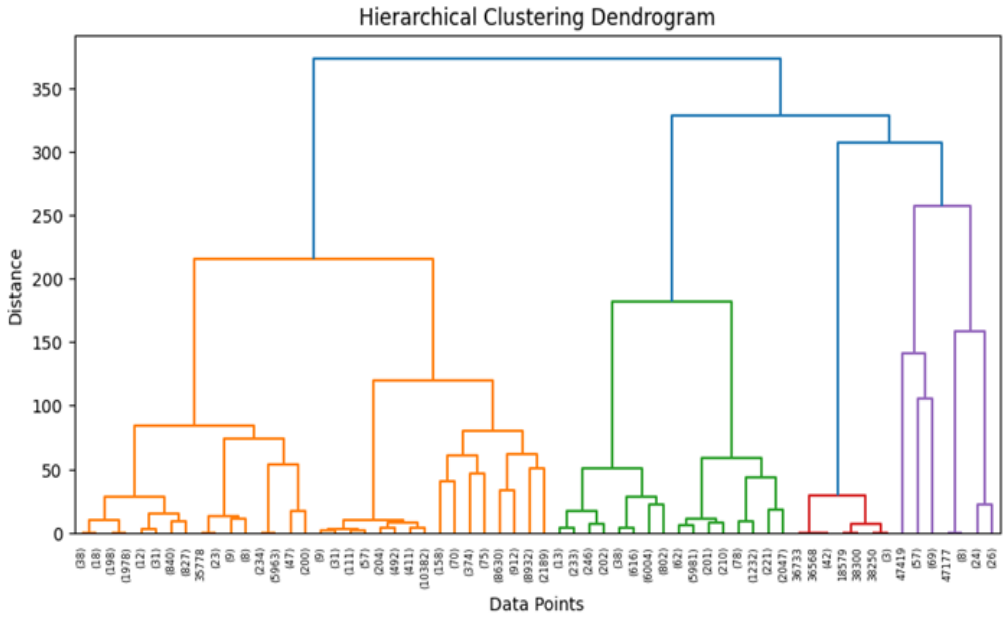

Figure 1 – Hierarchical clustering dendrogram.

Table 4 – Hierarchical clustering summary.

Cluster	Avg. Duration	Data Sent (bytes)	Data Received (bytes)	App Category	Risk Level	Interpretation
1	751 sec	8.7 KB	54 KB	6.2	Low (1.05)	Likely students/researchers access educational or informational sites.
2	379 sec	5.2 KB	15 KB	10.45	Moderate (3.6)	Likely social media and entertainment users, given the higher app category.
3	107.72 sec	89 KB	110 KB	5.2	Very Low (1.04)	Automated/Bot Activity. Extreme duration suggests a system process or background service rather than human browsing.
4	4.63 sec	3.3 KB	3.34 MB	10.18	Moderate-High (2.88)	High-risk or heavy media users, possibly watching videos, downloading large files or accessing risky content.
5	357 sec	494 KB	31 MB	12.57	Moderate (2.4)	Streaming or file-sharing users. Low session time but massive data received suggests video streaming, torrents or cloud storage.

Cluster analysis revealed the following user groups:

- Academic and research-oriented users had an average session duration of 12 minutes, relatively low data transfer volumes (8.7 KB sent, 54 KB received), and a low-risk level of 1.05.

- Social and entertainment users showed shorter sessions (~6 minutes) but a moderate risk level of 3.6, likely due to engagement with various content types.

- Automated system traffic was characterized by highly long session durations (~30 hours) and minimal data transfer, indicating non-human traffic.

- High-volume users exhibited large data transfers (3.3 MB sent, 3.34 MB received per session) with a higher risk level of 2.88, suggesting streaming, downloads, or accessing restricted content.

- General browsing users had short sessions (~6 minutes) but high data usage (31 MB received per session), implying intensive media consumption.

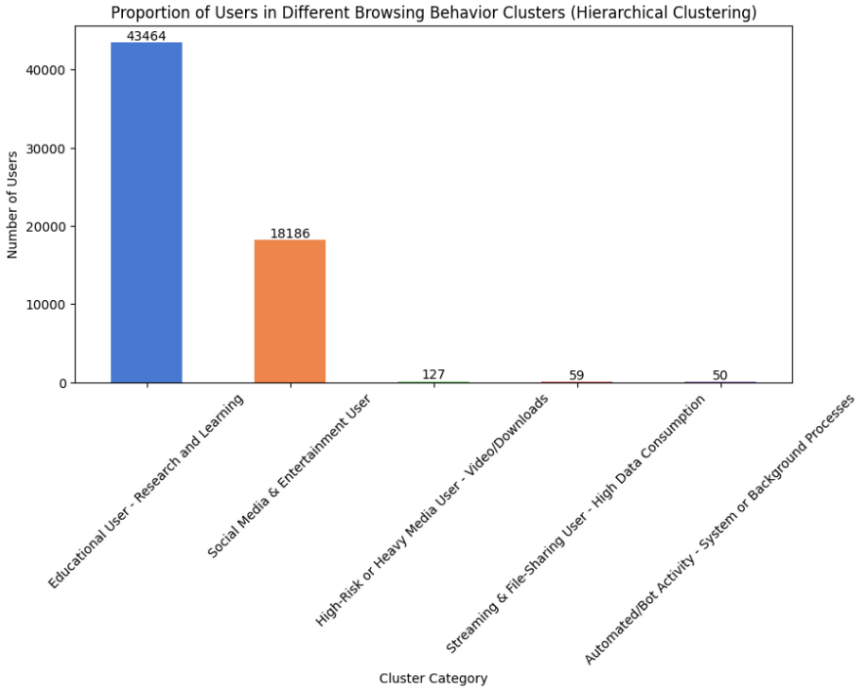

Figure 2 – User profile distribution based on hierarchical clustering.

The largest user groups identified through clustering consisted of Educational & Research Users and Social Media & Entertainment Users, while clusters associated with high data consumption and automated traffic comprised a smaller portion of the total user base. These findings confirm that hierarchical clustering effectively differentiates browsing patterns.

The results further demonstrate that hierarchical clustering successfully segments users based on distinct browsing behaviors, distinguishing standard web activity from automated processes and high-bandwidth usage.

3.2. GMM Clustering Findings

Gaussian Mixture Model (GMM) clustering identified four distinct user groups based on browsing activity. Unlike hierarchical clustering, GMM allows for probabilistic classification, meaning users may have characteristics of multiple groups.

The summary of GMM clusters, including session duration, data transfer volumes and risk levels, is presented in Table 5.

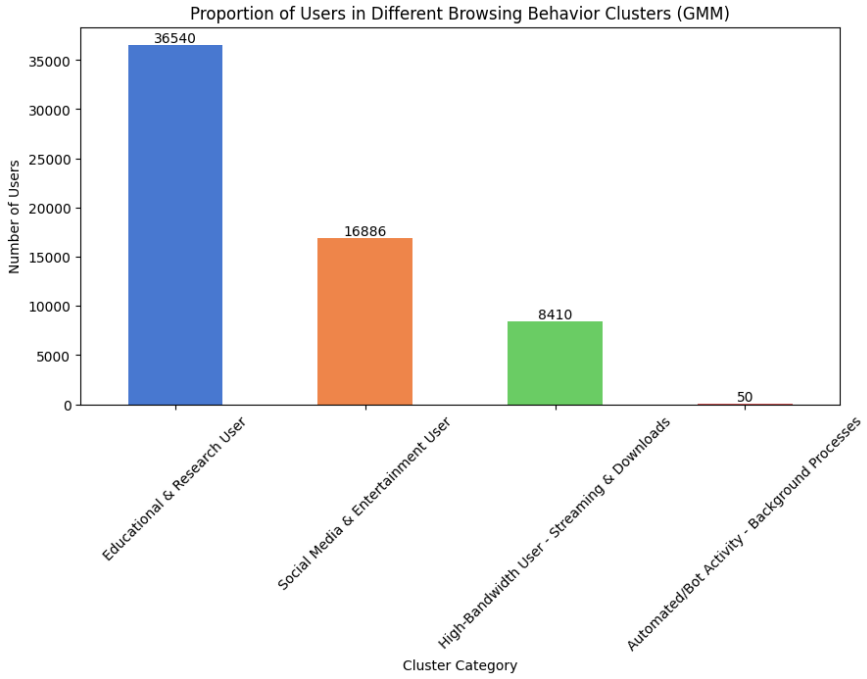

Figure 3 – User profile distribution based on GMM clustering.

Table 5 – GMM clustering summary.

GMM Cluster	Avg. Duration	Avg. Data Sent (bytes)	Avg. Data Received (bytes)	Avg. App Category	Avg. Risk Level	Interpretation
0	734 sec	2.4 KB	5 KB	6.31	Low (1.00)	Likely students or researchers accessing educational or informational sites.
1	47 sec	1.8 KB	2.7 KB	9.92	High (3.35)	Short session users, social media and entertainment browsing.
2	1.49 sec	97 KB	556 KB	7.62	Moderate (2.25)	Heavy data consumers (video streaming, downloads, cloud storage users).
3	107.72 sec	89 KB	110 KB	5.2	Very Low (1.04)	Automated/Bot Activity. Likely system processes running in the background.

Cluster analysis revealed the following user groups:

- Academic users had an average session duration of 12 minutes, low data transfer (2.4 KB sent, 5 KB received) and a low risk level of 1.00.
- Entertainment-oriented users had short browsing sessions (~1 minute) and a higher risk level of 3.35, indicating diverse content access.
- High-bandwidth consumers showed longer browsing durations (~25 minutes) and high data usage (97 KB sent, 556 KB received), likely due to video streaming or downloads.
- Background process users exhibited extremely long session durations (~30 hours) with minimal data transfer, suggesting background system processes rather than human activity.

GMM clustering successfully segmented users based on browsing intensity, risk levels and content types, providing an alternative perspective on behavioral patterns.

3.3. Anomaly Detection Findings

Anomaly detection was conducted using GMM risk classification and One-Class SVM, identifying users whose browsing activity significantly deviated from the norm.

The GMM-based classification indicated that approximately 37,000 users (≈60%) were categorized as normal, while around 25,000 users (≈40%) were flagged as high-risk based on their browsing behavior, data usage and application risk levels.

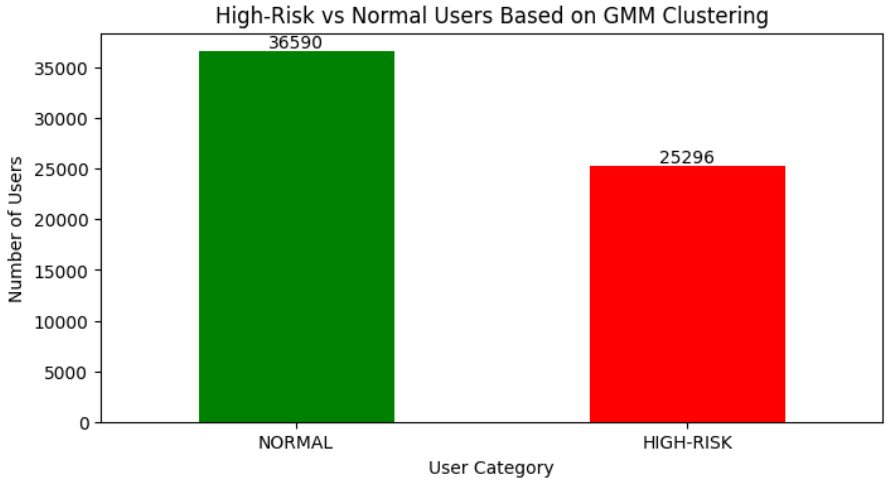

Figure 4 – High-risk vs. normal users based on GMM clustering.

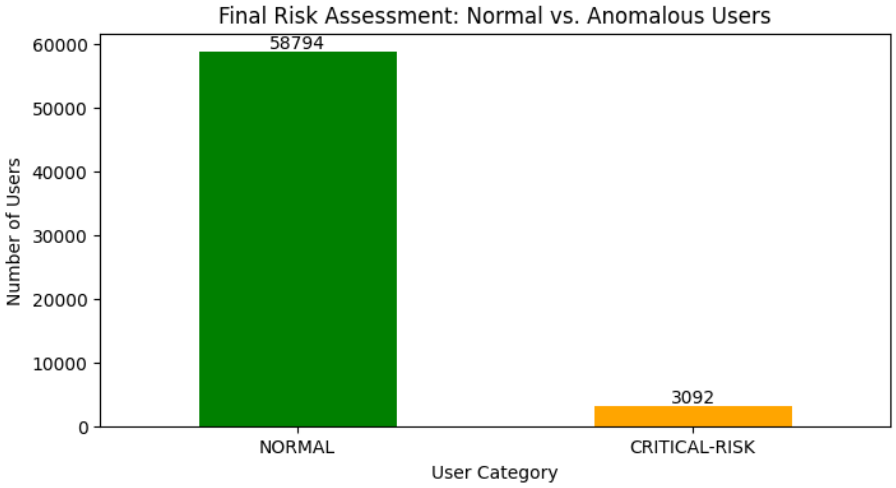

Figure 5 – Final risk assessment: normal vs. anomalous users.

The One-Class SVM model further refined anomaly detection by classifying users based on deviations from typical browsing patterns. The results showed that around 58,000 users (≈95%) were considered normal, while approximately 3,000 users (≈5%) were flagged as critically anomalous, indicating unusual or suspicious browsing activity.

The combination of GMM and SVM anomaly detection methods provided complementary insights, enabling the identification of both statistically high-risk users and behaviorally atypical activity.

3.4. Risk Alert Findings

Risk assessment was performed using Isolation Forest anomaly detection, identifying users with unusual browsing patterns based on traffic volume, session duration and application risk scores.

The first scatter plot visualizes detected anomalies in relation to session duration and application risk scores. Most users have shorter session durations and low-to-moderate risk levels, while a small subset exhibits significantly longer durations with high risk, flagged as anomalies.

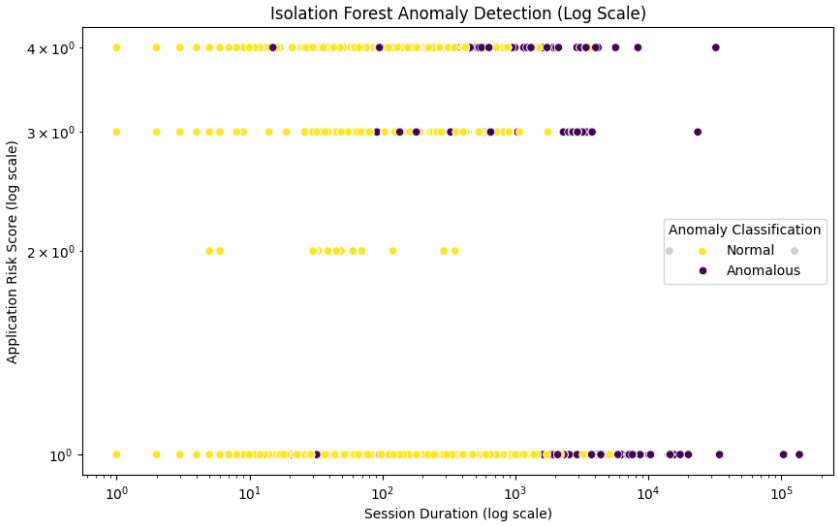

Figure 6 – Anomaly detection using Isolation Forest (session duration vs. application risk).

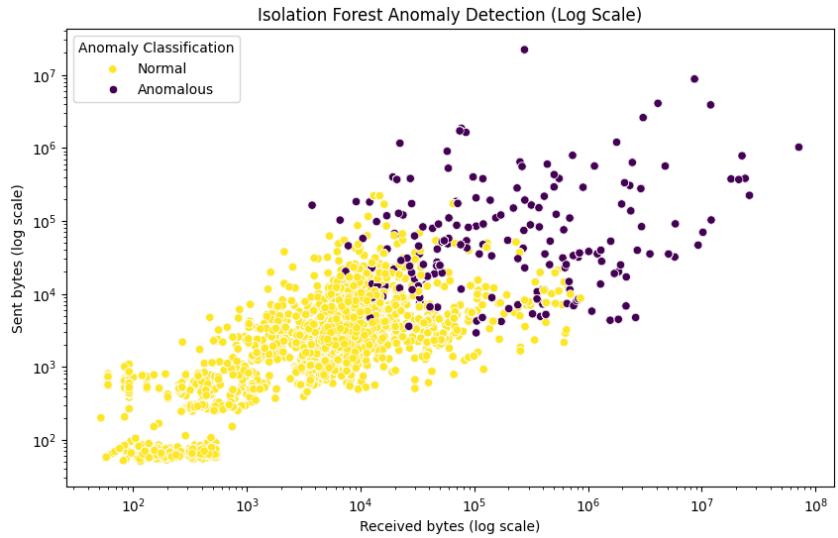

Figure 7 – Anomaly detection using Isolation Forest (sent vs. received data volume).

The second scatter plot visualizes anomalies based on the volume of data sent and received. While most users cluster around lower data values, a small subset stands out due to significantly higher data transfers, suggesting activity related to large downloads, streaming, or automated processes.

These findings reinforce the effectiveness of Isolation Forest in distinguishing typical user behavior from potential high-risk activity, complementing earlier anomaly detection results and further validating the model's ability to identify unusual browsing patterns.

4. Discussion

This study investigated whether clustering techniques can effectively segment users based on their browsing activity and whether anomaly detection can reliably identify unusual patterns. The findings confirm that machine learning approaches can successfully group users into distinct behavioral profiles while flagging deviations from expected browsing behavior.

The clustering analysis produced well-structured segmentations, with hierarchical clustering identifying five distinct user groups and Gaussian Mixture Models (GMM) differentiating four probabilistic profiles. The largest clusters consisted of educational users, social media consumers, and high-bandwidth users, reinforcing the idea that web browsing behavior follows recognizable patterns. These results align with previous research by El-Ansari, Beni-Hssane, and Saadi [21], who demonstrated that clustering web activity can effectively distinguish between academic and non-academic engagement. Similarly, Lima and de Castro [22] found that high-bandwidth usage is commonly linked to video streaming and file-sharing, mirroring the high-data consumption groups identified in this study. The differences between hierarchical clustering and GMM highlight the value of using both strict and probabilistic clustering techniques to account for overlapping behavioral traits.

Beyond user profiling, anomaly detection proved effective in identifying deviations from standard browsing behavior. One-Class SVM flagged approximately 5% of users as anomalous, while Isolation Forest identified outliers based on session duration, application risk scores, and traffic volume. These findings align with previous studies suggesting that machine learning can uncover unusual or potentially risky web activity. Lerner et al. [23] emphasized that deviations in browsing behavior can indicate security threats, such as unauthorized access attempts, while Schueller et al. [24] demonstrated that combining multiple detection techniques enhances anomaly identification accuracy. The integration of clustering and anomaly detection in this study supports these conclusions, as it not only classified users into meaningful categories but also flagged those whose behavior significantly diverged from the majority.

The study also revealed distinct behavioral trends in browsing data. Users engaged in educational content tended to have longer but lower-risk sessions, whereas social media users had shorter browsing sessions with a higher associated risk. High-bandwidth users exhibited significant data transfers, while automated system processes generated prolonged, low-interaction activity. These patterns align with previous findings on digital footprint analysis and offer valuable insights into the classification of web usage.

Despite the effectiveness of these methods, several limitations should be acknowledged. The dataset covered only two days of browsing activity, limiting the ability to capture long-term behavioral trends. Additionally, since the data was collected from a university network, the findings may not be fully representative of corporate or public browsing environments. Tareaf et al. [25] suggested that incorporating deep learning techniques could enhance detection accuracy, presenting a promising direction for future research. Expanding dataset coverage to include longitudinal browsing data would allow for a more comprehensive analysis of how user behavior evolves over time.

Overall, this study demonstrates that clustering techniques provide meaningful user segmentation, while anomaly detection effectively identifies behavioral outliers. These findings reinforce existing research on digital footprint analysis and emphasize the importance of using multi-method approaches to profile web activity. Future research should explore more diverse datasets and refine detection models to improve accuracy and applicability across different browsing environments.

5. Conclusion

This study applied machine learning techniques to analyze browsing history, demonstrating the feasibility of clustering users based on web activity patterns and detecting anomalies that deviate from expected behavior. The results confirm that unsupervised learning methods can effectively catego-

rize browsing habits, distinguishing between typical users and those exhibiting unusual or high-risk activity.

Applying hierarchical clustering and GMM allowed for a structured classification of users, identifying key behavioral groups based on session duration, data transfer volume, and application risk scores. Meanwhile, One-Class SVM and Isolation Forest successfully highlighted users with statistically significant deviations in their browsing patterns, reinforcing the importance of anomaly detection for identifying potential security risks or unusual online behavior.

These findings have implications for cybersecurity, network monitoring, and behavioral analysis. They offer a scalable approach to understanding digital footprints and detecting irregularities in web traffic. The integration of clustering and anomaly detection provides a multi-layered perspective on user behavior, which could be leveraged for automated risk assessment, policy enforcement, and resource management in various online environments.

While the study confirms the effectiveness of these methods, further research is needed to assess long-term behavioral trends and generalizability across different network settings. Future work should explore more extensive, diverse datasets and integrate adaptive machine-learning models to refine classification accuracy.

In conclusion, this study highlights the potential of machine learning for digital footprint analysis, offering a foundation for future advancements in web activity profiling and automated anomaly detection.

Funding

This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract №388/PTF-24-26 dated 01.10.2024 under the scientific Project IRN BR24993232 “Development of innovative Technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

Author Contributions

Conceptualization, M.I., S.K., and D.Y.; Methodology, S.K. and B.A.; Software, S.K. and B.A.; Validation, M.I., S.K. and D.Y.; Formal Analysis, L.R.; Investigation, L.R.; Resources, M.I., and D.Y.; Data Curation, S.K. and D.Y.; Writing – Original Draft Preparation, M.I. and D.Y.; Writing – Review & Editing, B.A. and L.R..; Visualization, S.K.; Supervision, D.Y.; Project Administration L.R.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. C. A. Kelly and T. Sharot, “Web-browsing patterns reflect and shape mood and mental health,” *Nature Human Behaviour*, pp. 1–14, Nov. 2024, doi: https://doi.org/10.1038/s41562-024-02065-6.

2. V. Lytvyn, V. Vysotska and A. Rzheuskyi, “Technology for the Psychological Portraits Formation of Social Networks Users for the IT Specialists Recruitment Based on Big Five, NLP and Big Data Analysis.,” *International Workshop on Control, Optimisation and Analytical Processing of Social Networks*, pp. 147–171, Jan. 2019.

3. C.-Y. Lien, G.-J. Bai and H.-H. Chen, “Visited Websites May Reveal Users’ Demographic Information and Personality,” *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 248–252, Oct. 2019, doi: https://doi.org/10.1145/3350546.3352525.

4. M. Paul and K. Medhe, “Using Machine Learning to Detect Anomalies in Internet Browsing Pattern of Users,” *SSRN Electronic Journal*, 2019, doi: https://doi.org/10.2139/ssrn.3511054.

5. A. A. Salomatin, A. Y. Iskhakov and A. O. Iskhakova, “Web user identification based on browser fingerprints using machine learning methods,” *IFAC-PapersOnLine*, vol. 54, no. 13, pp. 582–587, 2021, doi: https://doi.org/10.1016/j.ifacol.2021.10.512.

6. H. Gu, J. Wang, Z. Wang, B. Zhuang and F. Su, “Modeling of User Portrait Through Social Media,” *2018 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2018, doi: https://doi.org/10.1109/icme.2018.8486595.

7. G. Kovacs, “Reconstructing Detailed Browsing Activities from Browser History,” *arXiv (Cornell University)*, Jan. 2021, doi: https://doi.org/10.48550/arxiv.2102.03742.

8. M. Kosinski, D. Stillwell and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, Mar. 2013, doi: https://doi.org/10.1073/pnas.1218772110.

9. N. A. Maliki, A. Zainal, A. Ghaleb and M. N. Kassim, “User Security Behavioral Profiling using Historical Browsing Website,” *2021 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia*, Oct. 2021, doi: https://doi.org/10.1109/icodsa53588.2021.9617493.

10. Y. Chu, H. -K. Yang and W. C. Peng, "Predicting Online User Purchase Behavior Based on Browsing History," *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), Macao, China*, pp. 185-192, 2019, doi: https://doi.org/10.1109/ICDEW.2019.00-13.

11. Z. Mushtaq, S. Ashraf and N. Sabahat, "Predicting MBTI Personality type with K-means Clustering and Gradient Boosting," *IEEE Xplore*, Nov. 01, 2020. doi: https://doi.org/10.1109/INMIC50486.2020.9318078.

12. J. Philip, D. Shah, S. Nayak, S. Patel and Y. Devashrayee, "Machine Learning for Personality Analysis Based on Big Five Model," *Data Management, Analytics and Innovation*, pp. 345–355, Sep. 2018, doi: https://doi.org/10.1007/978-981-13-1274-8_27.

13. S. Pretorius, A. R. Ikuesan and H. S. Venter, "Attributing users based on web browser history," *2017 IEEE Conference on Application, Information and Network Security (AINS), Miri, Malaysia*, pp. 69-74, 2017, doi: https://doi.org/10.1109/AINS.2017.8270427.

14. C. E. Chibudike, H. Abdu, H. O. Chibudike, O. C. Ngige, O. A. Adeyoju and N. I. Obi, "Machine Learning – A New Trend in Web User Behavior Analysis," *International Journal of Computer Applications*, vol. 183, no. 5, pp. 19–25, May 2021, doi: https://doi.org/10.5120/ijca2021921247.

15. Y. Takama and S. Shimizu, "User Modeling from Review Browsing History for Personal Values-Based Recommendation," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 24, no. 3, pp. 326–334, May 2020, doi: https://doi.org/10.20965/jaciii.2020.p0326.

16. M. Abramson, "Toward the attribution of Web behavior," *2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications, Ottawa, ON, Canada*, pp. 1-5, 2012, doi: https://doi.org/10.1109/CISDA.2012.6291524.

17. L. Baruh, E. Secinti and Z. Cemalcilar, "Online Privacy Concerns and Privacy Management: A Meta-Analytical Review," *Journal of Communication*, vol. 67, no. 1, pp. 26–53, Jan. 2017, doi: https://doi.org/10.1111/jcom.12276.

18. D. W. Gresty, G. Loukas, D. Gan and C. Ierotheou, "Towards Web Usage Attribution via Graph Community Detection in Grouped Internet Connection Records," *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, UK*, pp. 365-372, 2017, doi: https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.61.

19. K. Hughes, P. Papadopoulos, N. Pitropakis, A. Smales, J. Ahmad, and W. J. Buchanan, "Browsers' Private Mode: Is It What We Were Promised?," *Computers*, vol. 10, no. 12, p. 165, Dec. 2021, doi: https://doi.org/10.3390/computers10120165.

20. B. Cervantes, F. Gómez, R. Monroy, O. Loyola-González, M. A. Medina-Pérez, and J. Ramírez-Márquez, "Pattern-Based and Visual Analytics for Visitor Analysis on Websites," *Applied Sciences*, vol. 9, no. 18, p. 3840, Sep. 2019, doi: https://doi.org/10.3390/app9183840.

21. A. El-Ansari, A. Beni-Hssane, and M. Saadi, "An improved modeling method for profile-based personalized search," pp. 1–6, Mar. 2020, doi: https://doi.org/10.1145/3386723.3387874.

22. A. C. E. S. Lima and L. N. de Castro, "TECLA: A temperament and psychological type prediction framework from Twitter data," *PLOS ONE*, vol. 14, no. 3, p. e0212844, Mar. 2019, doi: https://doi.org/10.1371/journal.pone.0212844.

23. A. Lerner, A. K. Simpson, Tadayoshi Kohno and F. Roesner, "Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016," *USENIX Security Symposium*, 2016, https://api.semanticscholar.org/CorpusID:17312347.

24. S. M. Schueller, D. M. Steakley-Freeman, D. C. Mohr and E. Yom-Tov, "Understanding perceived barriers to treatment from web browsing behavior," *Journal of Affective Disorders*, vol. 267, pp. 63–66, Apr. 2020, doi: https://doi.org/10.1016/j.jad.2020.01.131.

25. R. B. Tareaf, S. A. Alhosseini, P. Berger, P. Hennig and C. Meinel, "Towards Automatic Personality Prediction Using Facebook Likes Metadata," Nov. 2019, doi: https://doi.org/10.1109/iske47853.2019.9170375.

Information about authors

Marzhan Idrissova is a master's student of the Computer Engineering Department at Astana IT University (Astana, Kazakhstan, marzhanidrisova@gmail.com). Her research interests include machine learning applications for user profiling, anomaly detection in digital footprints and network traffic analysis. ORCID ID: 0009-0008-0629-7948.

Sabina Kim is a bachelor student of the Department of Computational and Data Science at Astana IT University (Astana, Kazakhstan, kimsabina206@gmail.com). Her research interests include the development of smart city solutions leveraging data science, as well as the integration of medicine with data science and machine learning to advance healthcare innovation. ORCID ID: 0009-0008-3198-0474.

Beibut Amirgaliyev is a distinguished researcher at Astana IT University (Astana, Kazakhstan, beibut.amirgaliyev@astanait.edu.kz) recognized for his contributions to both academia and industry. He holds a PhD in Computer Science and serves as a Professor at Astana IT University, focusing on research areas such as machine learning and computer vision. Dr. Amirgaliyev has published numerous papers on topics including automatic number plate recognition and solar collector systems, with his work being cited by over 200 researchers. ORCID ID: 0000-0003-0355-5856

Didar Yedilkhan is a distinguished researcher at Astana IT University (Astana, Kazakhstan, d.yedilkhan@astanait.edu.kz), recognized for his extensive experience in industry, research, and higher education. He serves as the Director of the Smart City Research Center and is a Senior Researcher at Astana IT University, focusing on data science, machine learning, and deep learning. Dr. Yedilkhan has a robust academic background, holding degrees from institutions such as the Kazakh National University named after al-Farabi and the University College London. His professional roles include being a Lead Researcher and Project Manager

at Astana IT University, where he leads projects on intelligent IT systems for urban infrastructure. His projects aim to enhance city safety and convenience through smart technologies. ORCID ID: 0000-0002-6343-5277

Leila Rzayeva is a prominent academic and researcher based at Astana IT University (Astana, Kazakhstan, l.rzayeva@astanait.edu.kz), where she serves as an Associate Professor and Researcher in Intelligent Systems and Cybersecurity. She earned her B.S., M.S., and Ph.D. degrees from L.N. Gumilyov Eurasian National University. Her research interests include control systems, industrial automation, cybersecurity, machine learning, deep learning, and the design of neural networks and artificial intelligence systems. Rzayeva has published over 40 national and international research articles and actively participates in conferences. ORCID ID: 0000-0002-3382-4685

Submission received: 08 March, 2025.
Revised: 11 March, 2025.
Accepted: 12 March, 2025.