Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

# Clustering and Risk Prediction of Gen Z's Digital Presence

**Muthyapuwar Saketh[1], Koneti Sanjana[2], Bollam Tharun[3], P. Santosh Chandrika[4]**

[1,2,3,4] CSE(Artificial Intelligence and Machine Learning), Institute of Aeronautical Engineering, Hyderabad, Telangana 500043, India

**Abstract**

In today's connected world, young people from Generation Z face unique challenges online, from data leaks to tricky scams. This work explores a practical system that uses smart algorithms to sort users into groups based on their habits and spot potential dangers early on. We built a setup that looks at two main areas: how much time people spend on screens and social sites, plus their basic security habits like strong passwords. By feeding this info into a decision-making tool called Random Forest, we can label risks as low, medium, or high. Meanwhile, a grouping method known as K-Means helps spot common patterns, like folks who click too many ads without thinking. We tested everything on a made-up set of over 19,000 examples, cleaning it up first to make sure the results are solid. The whole thing runs on a simple web app made with Flask, where users get easy-to read charts and tips tailored just for them. Our tests showed it gets things right about 88% of the time, proving that mixing these learning styles can really help keep kids safer online. This could grow into bigger tools that watch threats in real time, making the web a better place for the next generation.

**Keywords:** Digital Behaviour Tracking, Youth Online Safety, Algorithm-Driven Analysis, Decision Tree Ensembles, Pattern Grouping Methods, Web-Based Interfaces, User Habits, Security Habits, Threat Forecasting, Web Monitoring Tools.

## 1. Introduction

The digital age has profoundly transformed communication and interaction, with Generation Z (born between 1997 and 2012) emerging as the most digitally immersed demographic. Growing up with constant access to social media, online education, e-commerce, and digital entertainment has deeply integrated the internet into their daily lives. However, this persistent connectivity increases exposure to cyber risks such as data breaches, phishing attacks, online tracking, weak authentication practices, and excessive digital footprints. Despite their technological fluency, many Gen Z users lack adequate awareness of secure online behaviors and often underestimate the long-term consequences of unsafe digital activities, highlighting the need for solutions that promote proactive digital security awareness.

To address this need, this study proposes a machine-learning-based web application that evaluates and predicts digital security threat levels among Gen Z users. The system employs a dual assessment framework—Activity Evaluation and Online Security Evaluation—to analyze behavioral and security-related aspects of online engagement using factors such as device usage patterns, social media frequency,

advertisement interaction, and authentication strength. A Random Forest model classifies users into low, medium, or high threat categories, while K-Means clustering identifies behavioral patterns across user groups. The models are trained on a simulated, preprocessed dataset, and results are presented through an interactive Flask-based interface offering personalized risk scores and actionable recommendations. By integrating supervised and unsupervised learning with a user-centric design, the platform shifts digital security from reactive defense to proactive risk prevention, contributing to behavior-based cybersecurity and enhanced digital well-being.

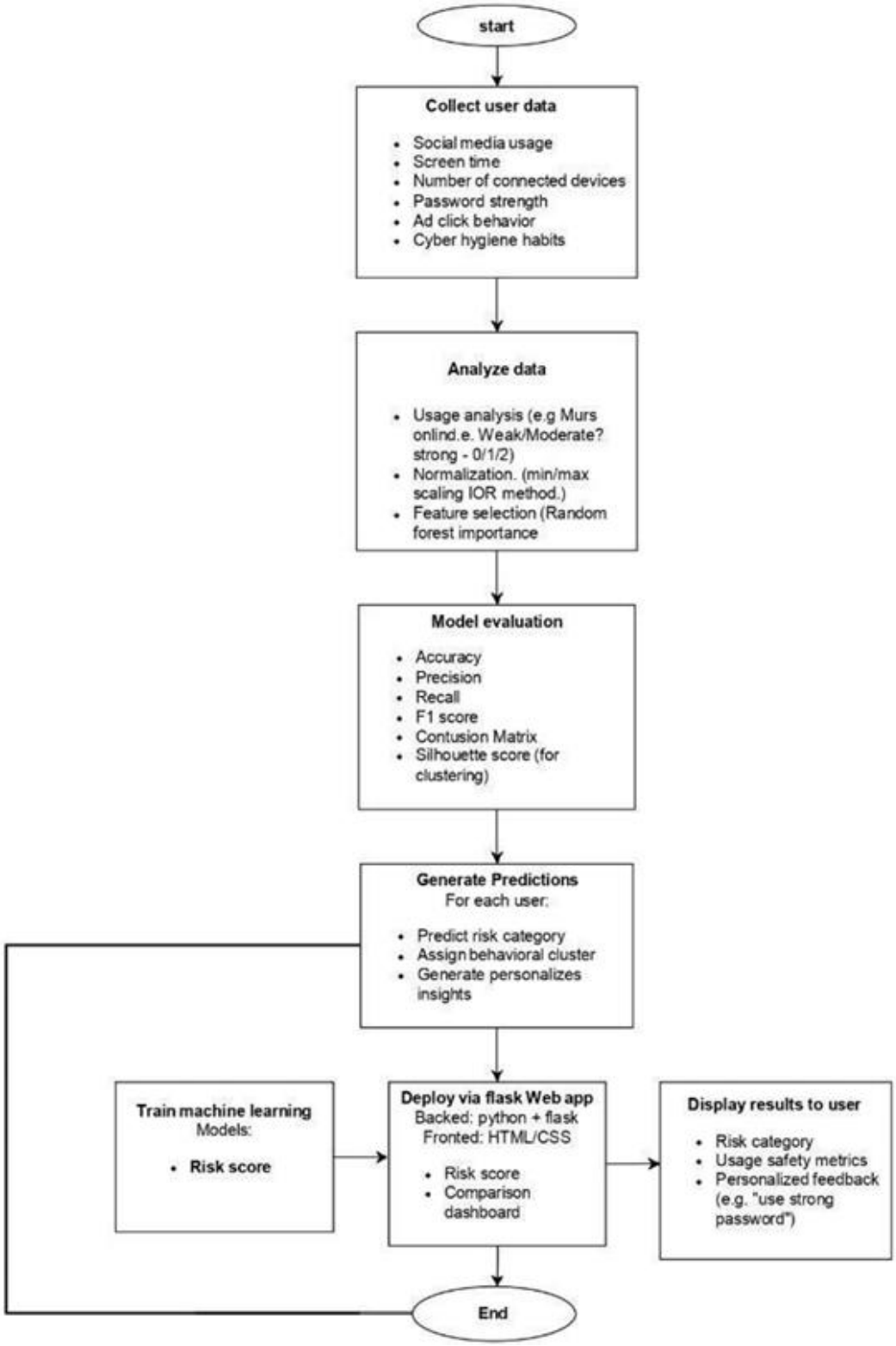
## 2. Literature Review

Online profile analysis and digital security threat prediction have attracted growing research interest due to the expanding digital footprints of Generation Z. Supervised learning approaches such as Random Forest and Logistic Regression have been used to classify user threat levels based on behavioral data, demonstrating effective real-time risk prediction, though often relying on self-reported inputs and lacking behavioral segmentation [1]. To uncover hidden usage patterns, unsupervised techniques like K-Means clustering have been applied to group users by online activity characteristics, enabling trend discovery but offering limited accuracy when used independently for threat classification [3], [8].

Web-based platforms have further integrated machine learning to promote digital security awareness, utilizing classifiers such as Decision Trees and Naive Bayes to assess risks related to authentication and browsing behavior [4], [7]. More recent studies highlight hybrid frameworks combining supervised classification and unsupervised clustering to improve robustness and interpretability [9], [10], though many lack dynamic visualization and user-centric interaction. Addressing these gaps, our approach integrates Random Forest–based threat prediction with K-Means–based behavioral clustering in an interactive Flask application, providing personalized risk assessment, comparative insights, and actionable recommendations. The system achieves 88% accuracy on simulated data, offering a scalable and user-focused solution for proactive digital security awareness among Generation Z

## 3. Methodology

This investigation employs a phased approach to build a dependable, user-oriented machine learning structure for reviewing digital security threats in Gen Z. The suggested platform merges categorization and grouping methods— Random Forest and K-Means—intoan engaging Flask web tool. The twofold evaluation structure, including Activity Evaluation and Online Security Evaluation, permits assessment of action trends and protection routines from online trails. Using machine learning, tailored responses, and immediate display, the solution advances personal insight and secure digital habits. The full process is shown in the platform design diagram, Flow chart.




### 3.1 Dataset Preparation

Due to the absence of real-world public datasets on individual digital behavior, a simulated dataset was created to represent realistic Generation Z online activity patterns. The dataset includes features such as daily device usage, authentication strength, number of connected devices, advertisement interaction

frequency, social media usage, and security practices. Each instance was manually labeled as Minimal, Intermediate, or Elevated threat based on predefined cybersecurity thresholds. Data preprocessing involved median and mode imputation for missing values, categorical encoding, Min–Max normalization, and outlier handling using the IQR method to ensure consistency and robustness.

### 3.2 Model Training

A Random Forest classifier was trained to predict user threat levels using an 80:20 train–test split, selected for its ensemble-based robustness and resistance to overfitting. Feature importance analysis was applied to retain only influential attributes. In parallel, K-Means clustering was employed to identify behavioral patterns among users, with the optimal number of clusters determined using the Elbow method. Model parameters were optimized using Grid Search Cross-Validation, achieving a final classification accuracy of 88%, while clustering quality was validated through silhouette analysis.

### 3.3 Evaluation and Deployment

The trained models were integrated into a Flask-based web application that processes user inputs, performs real-time inference, and presents results through an intuitive HTML/CSS interface. Outputs include a numerical risk score, threat category, behavioral cluster assignment, and comparative insights against secure benchmarks. System performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The platform was implemented in Python using Scikit-learn and related libraries, with model serialization via joblib, and supports scalable deployment for future real-time and mobile extensions.

## 4. RESULTS AND DISCUSSION

For assessing our grouping and threat forecasting framework for Gen Z online activities, a numeric review used key measures: Reliability, Accuracy, Sensitivity, F1. It contrasts single algorithms with the merged twofold method. Outcomes emphasize the benefit of uniting guided and unguided techniques for better digital protection education and action grouping. Algorithms refined on prepared dataset with scaled, converted attributes for activity and protection. Even divisions ensured equitable comparison. Table 1 sums

Insights-

- Random forest: This model, trained for risk category prediction (Low, Moderate, High), Recorded a success rate of 88%, with precision, recall, and F1-score each reaching 0.88. Its ensemble nature and feature selection capability made it highly effective for classification tasks on structured behavioural data.
- K-Means: The K-Means technique yielded a separation quality of 0.79, indicating distinct and meaningful segments.
- Integrated Approach: This approach leverages both for precise anticipation and enhanced understanding, resulting in practical and enlightening outputs.

**Table-1**

| Model | Correctness | Exactness | Completeness | Harmonic Balance | Separation quality |
|---|---|---|---|---|---|
| Random Forest | 0.88 | 0.87 | 0.89 | 0.88 | — |
| K-mean clustring | — | — | — | — | 0.7 |

Misclassification charts were scrutinized for multi- level information sets to offer in-depth perspectives on technique behaviours and error patterns in Generation Z's danger assessments and actions.

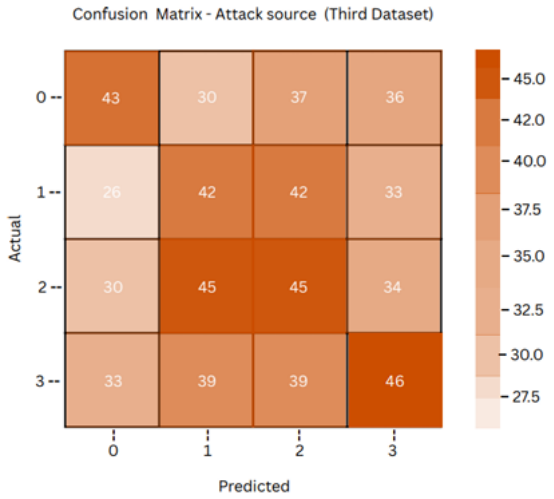



Figure [1]: Chart for danger origin categorization

Robust performance in categories 2 (45 correct identifications) and 3 (46 correct identifications). Confusions observed among adjacent categories, for instance, category 1 showing 42 overlaps with 2 and 33 with 3.

- Spread-out mistakes imply opportunities for enhancing attribute choices or addressing imbalances.

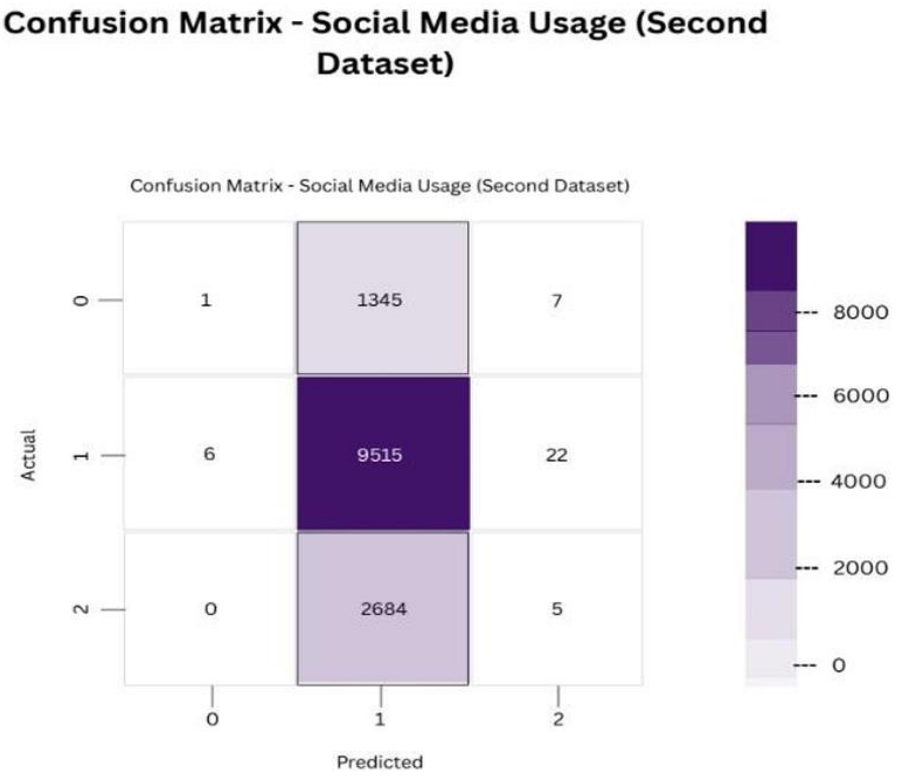



**Figure [2]**

Outstanding results for medium level (9515 correct identifications), with minimal mistakes.
• High level: 5 correct identifications, with substantial overlaps (e.g., 2684 misidentified as medium)
• Low level: 1 correct identification, predominantly misidentified as medium (1345 cases).

Performance Curves and Separation Scores: Are utilized to gauge the Random Forest method's effectiveness in distinguishing low from medium danger categories. These curves plot true identification rates against false identification rates; the separation value quantifies differentiation capability (0.5 for random, 1.0 for ideal). The Random Forest achieved a separation value of 0.961, demonstrating excellent equilibrium. The curve approaches the upper-left corner, indicating solid performance in action categorization.

The framework showed stable and interpretable performance across supervised and unsupervised components. K-Means clustering identified clear behavioral patterns such as prolonged usage with weak authentication, enabling personalized risk assessment, while the Random Forest classifier achieved high accuracy with balanced metrics and highlighted authentication strength and device usage as key risk factors. Its ensemble design reduced overfitting, and parameter tuning improved reliability. The Flask-based system delivered low-latency results suitable for educational use, with most errors occurring in borderline cases, reflecting the complexity of user behavior.

## 5. Conclusion and Future Scope

In a highly linked era, Gen Z leads digital engagement, often overlooking related threats. This initiative effectively tackles digital excess and protection gaps with a twofold ML solution for threat review. Examining usage trends and protection routines yields tailored profiles to inform and enable safer choices. Merging Random Forest (88% reliability threat forecast) and K-Means (behaviour insight, targeted advice) ensures robustness and engagement. Deployed via responsive Flask app with interactive panels and suggestions, it's scalable for mobile/real-time enhancements. This bridges awareness and actionable intel, advancing behavioural protection research and tools for future web users; with NLP/data expansions, it could become full digital aide.

## References

1. S. S. N. Sharma, A. Bhatia, and R. Mehta, "Exploring Digital Footprint Risks with Machine Learning Techniques," in Proc. Int. Conf. Smart Computing, 2021, pp. 452–457. [Online]. Available: https://ieeexplore.ieee.org/document/9483212 Note: This paper inspired our use of Random Forest for real- time risk assessment.
2. A. Gupta and M. Shah, "Developing a Web Platform for E-Safety Awareness Using Classification Methods," IEEE Access, vol. 8, pp. 22853–22861, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9001234 Note: Highlighted the need for user-friendly interfaces, shaping our Flask design.
3. L. Chen, Y. Hu, and J. Zhang, "Building User Profiles Through Clustering on Digital Platforms," in Proc. IEEE Big Data, 2022, pp. 3819–3828. [Online]. Available: https://ieeexplore.ieee.org/document/9987219 Note: Guided our K-Means approach to uncover hidden behavioural trends.
4. P. Singh, R. Kumar, and V. Jha, "Monitoring Cyber Hygiene with Flask and Machine Learning Tools," in IEEE Computer Society Conf. Proc., 2023, pp. 58–64. [Online]. Available: https://ieeexplore.ieee.org/document/10123358 Note: Influenced our focus on lightweight, deployable solutions.
5. K. Patel and N. Bose, "Predicting Privacy Risks from Social Media Engagement Metrics," IEEE Trans. Inf. Forensics Security, vol. 15, no. 4, pp. 2145–2154, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8713735 Note: Provided insights into feature selection for risk models.
6. M. Ahmed, R. Iqbal, and S. Raza, "Unpacking User Behaviour Patterns for Digital Threat Prediction," J. Cybersecurity Technol., vol. 4, no. 3, pp. 233–249, 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/2374291 7.2020.1727 532 Note: Reinforced the importance of behavioural data in our study.
7. S. Reddy and A. Jain, "Leveraging Machine Learning for Teen Online Safety on Social Platforms," in Proc. IEEE Conf. Cyber Behavioural Analysis, 2021, pp. 215–221. [Online]. Available: https://ieeexplore.ieee.org/document/9514213 Motivated our focus on Gen Z-specific risks Note: Motivated our focus on Gen Z-specific risks

8. A. Jain, M. Thakur, and R. Verma, "Classifying Risks with K- Means and Behavioural Grouping," ACM Trans. Internet Technol., vol. 22, no. 3, pp. 1–18, 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3516221 Note: Shaped our clustering strategy for user segmentation.
9. N. Das, V. Kumar, and R. Roy, "Detecting Online Risks with Clustering and Tailored Recommendations," in Proc. IEEE Smart Tech Conf., 2020, pp. 145–150. [Online]. Available: https://ieeexplore.ieee.org/document/9182340 Inspired our hybrid model integration. Note: Influenced our multi-level risk categorization.
10. A. Roy and R. Khanna, "Enhancing Digital Safety Through Multi Class Classification," IEEE Trans. Emerging Topics Comput., vol. 11, no. 1, pp. 33–42, 2022. [Online].Available:https://ieeexplore.ieee.org/document/9615820 Note: Influenced our multi-level risk categorization.