JURNAL EMACS e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.7 No.3 September 2025: 269-274 DOI: 10.21512/emacsjournal.v7i3.13445

Gender Classification Using Keystroke Dynamics: Enhancing Performance with Feature Selection and Random Forest

Ayu Maulina[1*], Rifqi Alfinnur Charisma[2]
[1,2]Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
ayu.maulina001@binus.ac.id; rifqi.charisma@binus.ac.id

*Correspondence: ayu.maulina001@binus.ac.id

Abstract – *The purpose of this study is to improve gender categorization by examining the usage of keyboard dynamics, with enhanced model performance through data standardization and appropriate feature selection. Features including gender, age, handedness, language, education, and metrics measuring typing behavior like mean_latency, std_latency, and frequency are all included in the dataset. Correlation analysis served as the foundation for the feature selection procedure, which is essential for effective model training, and data normalization was performed to guarantee consistency among the characteristics that were chosen. Because of its stability and capacity to handle complicated data, the Random Forest classifier was selected. The findings demonstrate that the Random Forest model achieved an accuracy of 95% and an F1-score of 95% when using all features, and 82% accuracy with an F1-score of 82.5% when using only the selected features. The results emphasize how important it is to choose the appropriate characteristics and standardize the data in order to increase predictive accuracy. By showcasing keystroke dynamics' capacity for gender categorization, this study advances the area and creates opportunities for further research in user experience improvement, digital service customization, and online behavioral analysis. Overall, the study emphasizes the importance of feature engineering, normalization, and model tuning for achieving accurate and reliable classification outcomes.*

Keywords: *Classification; Gender Classification; Keystroke Dynamics; Random Forest*

I. INTRODUCTION

Biometric technologies are becoming an essential feature of many applications in the modern digital world, especially in the areas of user authentication and security (Sriman et al., 2024). Facial recognition, fingerprint analysis, and iris scanning are just a few of the biometric modalities that have been developed to improve identification accuracy and system resilience (Thakare et al., 2021). One of the more appealing of these is keystroke dynamics, which focuses on analyzing individual typing patterns, particularly the rhythm and timing of key presses and releases, in order to identify or categorize users (Raul et al., 2020)(Shekhawat & Bhatt, 2022). The term "keystroke dynamics" describes each person's distinct typing habits, such as dwell time (the length of time between keystrokes) and flight time (the interval between keystrokes). Similar to fingerprints, these patterns are specific to each individual and can be used as a biometric identifier (Tsvetkova & Bakhteev, 2024). This approach has several benefits, including the ability to be smoothly integrated into current keyboard-based interfaces, being non-intrusive, and being hardware-independent.

Beyond the realm of authentication, keystroke dynamics has been used more and more to categorize personal traits like age, gender, emotional state, and stress levels (Tsimperidis et al., 2021)(Cascone et al., 2022). Gender categorization is one area of this field that shows great promise for applications,

including user experience optimization, digital service customization, and behavioral analysis on online platforms. Keystroke dynamics have been shown to be effective in a variety of settings in earlier studies. For example, Buker et al.'s research(Buker et al., 2019) created a gender classification model based on keyboard patterns in live chat settings, and it achieved accuracy levels of above 95%. Similarly, Pentel(Pentel, 2019) investigated keystroke patterns for age classification, with encouraging outcomes. Kołakowska and Lndowska also examined keyboard dynamics when participants wrote both positive and negative judgments(Kołakowska & Landowska, 2021). Based on typing activity, their study's support vector machine (SVM) model demonstrated a moderate amount of efficacy in determining the text's emotional tone, as seen by the highest F1-score of 0.76. Furthermore, Marrone and Sansone (Marrone & Sansone, 2022) explored the application of keystroke dynamics for continuously predicting users' emotional states during message writing sessions. Their study showed that the processing method significantly influenced the outcomes, with the multiple-instance learning–support vector machine (MIL-SVM) model yielding the highest accuracy when trained on bags of variable sizes. To address the limitations of existing datasets in this field, which remain relatively rare in the literature, the IKDD dataset was introduced in 2024 (Tsimperidis et al., 2024). Such datasets, derived from user keystroke recordings, are essential for advancing research in keystroke dynamics, yet they are scarce. The IKDD dataset aims to fill this gap by providing valuable data for further research. It comprises input from 164 volunteers and includes 533 log files, recording approximately 1.85 million keystrokes in total. Its ecological validity, which gathers keystroke data from users' regular computer interactions, is one of its main benefits. Five important demographic characteristics are also included in the dataset: gender, age group, handedness, native language, and educational attainment. Machine learning models can attain up to 80% accuracy in gender classification tasks, according to earlier research using the IKDD dataset. Device dependency—the propensity for keystroke patterns to alter across various hardware configurations, such as keyboard kinds, key sensitivities, and device dimensions—remains a significant obstacle, nevertheless. Even though this field has seen tremendous advancements, issues with hardware variability still exist. Typing behavior can be significantly changed by variations in keyboard responsiveness and design, underscoring the necessity of normalization procedures to increase model generalizability and robustness.

The purpose of this research is to investigate the use of keystroke dynamics for machine learning-based gender classification. Because of its better stability over other models and ability to handle complicated, high-dimensional data, the Random Forest algorithm has been chosen as the main model for classification. Combining the results of several weak learners (decision trees) to produce more reliable and accurate predictions is its predictive strength (Hu & Szymczak, 2023).

II. METHODS

2.1 Dataset

The data utilized in this study is derived from the IKDD Keystroke Dataset(Tsimperidis et al., 2024), comprising a total of 276,782 data entries in .txt file format, with each file corresponding to a single user. Initially, the dataset contains seven primary features: user id, gender, age group, handedness, mother tongue, education level, and device.

Each record within the dataset represents keystroke activity and is structured as follows. For example, a segment of the dataset appears as:

48–0,62,65,74,64,60,45
49–0,95,91,82,108
50-0,98,88,87,103,104,59,87,65,60,48, 83
69–82,272,316,671,391,96,928,550,74
69–83,125,193,170,142,235,168,310

In the above representation, the first value in each record denotes the keystroke feature; for instance, 50–0 indicates the keystroke duration for the key 2, while 69–84 signifies the digram latency between the characters E–T.

To facilitate the integration and analysis of the dataset, all individual text files were consolidated into a single .xlsx file. Moreover, additional features were constructed from the raw keystroke data to enhance the dataset's

analytical potential. Specifically, two new features were introduced: keystroke duration, which captures the time taken to press each individual key, and digram latency, which measures the time interval between consecutive key presses.

Furthermore, an additional feature, frequency, was created to quantify how often each character is pressed, derived from the count of keystroke duration data corresponding to each character within the dataset.

Following these preprocessing steps, the dataset was augmented to include a total of 12 features. In the final data cleaning phase, the user id feature was excluded from the dataset, as it did not contribute substantively to the analysis and posed a risk of data leakage during model development.

2.2 Random Forest

Random Forest is an ensemble learning technique that aggregates the predictions of several decision trees to enhance the overall accuracy compared to using a single decision tree(Breiman, 2001). Random Forest is trained by constructing multiple decision trees, each built from a bootstrapped version of the training dataset(Hu & Szymczak, 2023). To generate a prediction using Random Forest, an observation is passed through all the decision trees in the forest. The final output is determined by either taking the majority vote (for classification) or calculating the average (for regression) from the predictions of all trees. Some observations are not included in the training of individual trees since each tree is trained using a bootstrap sampling of the data. The prediction error of the entire forest can be estimated using these excluded observations, sometimes referred to as out-of-bag (OOB) samples, as test data (Breiman, 2001; Hu & Szymczak, 2023; Talekar & Agrawal, 2020). Furthermore, Random Forest is thought to be very noise-resistant because it usually performs better than a single decision tree when combining forecasts from several trees (Hengbo et al., 2020). It is one of the most popular machine learning algorithms because of its resilience, adaptability, and capacity to function effectively with both small and large datasets. Furthermore, Random Forest may be applied to a variety of real-world issues more easily because it doesn't require a lot of hyperparameter adjustment (Geetha Vadav et al., 2024).

Lastly, because each tree in the forest can be trained separately, the algorithm's parallel design enables efficient processing. Because of this characteristic, Random Forest can be used in applications that need a lot of computing and have big datasets. All things considered, Random Forest is an effective tool that blends accuracy, adaptability, and simplicity in predictive modeling (Salman et al., 2024) (Vázquez-Novoa et al., 2023).

2.3 Evaluation

Several common classification measures, including accuracy, precision, recall, and F1-score, are used to assess the performance of the suggested model. These metrics provide a comprehensive insight of the predicted efficacy of the model and are frequently employed in binary classification assignments.

- *Accuracy* assesses the overall performance of the model by calculating the ratio of correctly predicted instances to the total number of predictions.
- *Precision* calculates the percentage of accurate positive predictions, which aids in lowering false positives.
- *Recall (or sensitivity),* assesses how well the model can detect every true positive instance.
- *F1-Score* represents the harmonic mean of precision and recall, providing a balanced evaluation metric, particularly when dealing with imbalanced classes.

The "IKDD: A Keystroke DynamicDataset for User Classification" benchmark will be used to compare the experimental outcomes of the suggested approach (Tsimperidis et al., 2024). According to the study, the gender categorization accuracy was 81.2%.

III. RESULTS AND DISCUSSION

In the initial phase of this study, the dataset employed consists of 11 features: gender, age, handedness, language, education, device, session, combination, mean_latency, std_latency, and frequency, with gender serving as the target variable. A correlation analysis was subsequently conducted to assess the strength and direction of the relationship between each feature and the target variable. The correlation

coefficients were computed using the Pearson method, which yields values ranging from -1 to 1, where:

- Values approaching 1 indicate a strong positive correlation,
- Values approaching -1 indicate a strong negative correlation,
- Values near 0 suggest no significant correlation.

The results of this correlation analysis are visualized through a heatmap, as shown in Figure 1. In the heatmap, shades of red indicate strong positive correlations, whereas shades of blue represent strong negative correlations. This visualization facilitates the identification of features that are potentially informative for predicting the target variable.

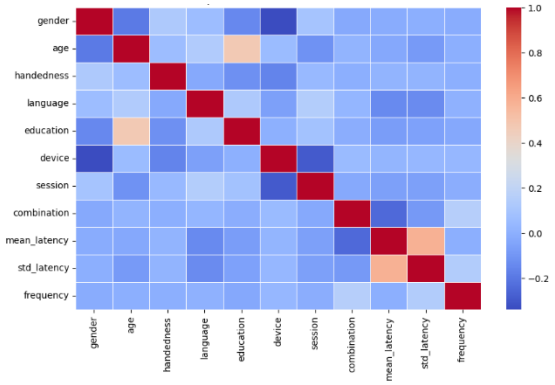

Figure 1. Heatmap correlation between features

Based on the results of the correlation analysis, the eight most significant features with the strongest associations to the target variable gender were selected. These features include: education, age, language, handedness, mean_latency, std_latency, frequency, and device. To ensure that all selected features are on a comparable scale, normalization was performed using the StandardScaler. This preprocessing step is crucial to facilitate more effective learning by the Random Forest model.

Subsequently, the Random Forest classifier was trained using the selected features with the following parameters: n_estimators = 100 and random_state = 42 to ensure reproducibility. The dataset was split into training and testing subsets with a ratio of 80:20, allowing for robust model evaluation. Evaluation metrics—accuracy, precision, recall, and F1-score—are presented in Table 1, showing that the model with selected features achieved 82% accuracy and 82.5% F1-score, indicating solid performance but limited by feature selection.

Table 1. Model Result

Class	Accuracy	Precision	Recall	F1-Score
0	82%	83%	84%	84%
1		82%	80%	81%

In the second experiment, the entire set of features available in the dataset was utilized without applying any feature selection techniques. Employing all features resulted in a notable enhancement in the performance of the Random Forest model compared to the initial experiment. The evaluation metrics indicated a substantial improvement, with the model achieving an accuracy approaching 95%. Furthermore, the F1-score for both classes reached 95%, demonstrating balanced and robust predictive capability, as detailed in Table 2.

Table 2. Model Result with All Features

Class	Accuracy	Precision	Recall	F1-Score
0	95%	96%	95%	95%
1		94%	95%	95%

This improvement implies that certain features that were not included in the selection process may still provide supplementary data that improves model prediction. Although feature selection lowers computing costs and complexity, it may unintentionally exclude minor patterns that are essential for classifying gender based on keystroke dynamics.

To further evaluate the performance of the model developed in this study, the experimental results were compared with benchmark studies employing Multi-Layer Perceptron (MLP) and Radial Basis Function Network (RBFN). Table 3 presents a comparative analysis between the current experimental findings and the benchmark results.

Table 3. Comparison of model performance

Model	Accuracy	F1-Score
MLP	77.1%	77.1%
RBFN	81.2%	81.2%
RF (With Feature Selection)	**82%**	**82.5%**
RF (With All Features)	**95%**	**95%**

The Random Forest model with chosen characteristics obtained an accuracy of 82% and an F1-score of 82.5%, according to the

comparison findings. In contrast, the model with all features showed a notable increase, achieving 95% for both accuracy and F1-score. Both setups fared better than the benchmark models, Radial Basis Function Network (RBFN) and Multi-Layer Perceptron (MLP), which only managed 81.2% and 77.1%, respectively. This comparison demonstrates that the Random Forest model developed in this study, utilizing all features, delivers markedly superior performance over the existing benchmark models across all evaluation metrics, including accuracy, precision, recall, and F1-score. This improvement indicates that proper feature selection, combined with data normalization applied in this experiment, has a significant impact on enhancing the model's performance in the classification task.

IV. CONCLUSION

This study shows that a relatively good gender categorization model may be generated utilizing the IKDD dataset with keyboard dynamics. For both classes, the experiments with all characteristics demonstrated good accuracy and F1-scores that were close to 95%. In benchmark investigations, these outcomes perform better than the MLP and RBFN models. But there is still a lot of space for research and enhancement, especially when it comes to resolving issues with device dependency and other elements that might affect the model's functionality.The conclusion of this study indicates that proper feature selection, followed by data normalization, can significantly impact the performance of classification models. The application of the Random Forest model achieved superior results compared to benchmark models, particularly in terms of F1-score, precision, and recall. These findings emphasize the importance of feature selection and normalization techniques in prediction accuracy, while also opening opportunities for further development in classification applications using keystroke dynamics data.

REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Buker, A. A. N., Roffo, G., Vinciarelli, A., & Cambria, E. (2019). Type like a man! Inferring gender from keystroke dynamics in live-chats. IEEE Intelligent Systems, 34(6), 53–59. https://doi.org/10.1109/MIS.2019.2948514

Cascone, L., Nappi, M., Narducci, F., & Pero, C. (2022). Touch keystroke dynamics for demographic classification. Pattern Recognition Letters, 158, 63–70. https://doi.org/10.1016/J.PATREC.2022.04.023

Geetha Vadav, M., Rajasekhar, N., Reddy, E. S., Vishal, M. S., & Vishal, G. (2024). The Role of Machine Learning in Crime Analysis and Prediction. Proceedings - 2024 International Conference on Expert Clouds and Applications, ICOECA 2024, 885–890. https://doi.org/10.1109/ICOECA62351.2024.00157

Hengbo, X., Fengjun, L., Xuan, D., & Zhu, T. (2020). Analysis on the Applicability of the Random Forest. Journal of Physics: Conference Series, 1607(1), 012123. https://doi.org/10.1088/1742-6596/1607/1/012123

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 24(2), 1–11. https://doi.org/10.1093/BIB/BBAD002

Kołakowska, A., & Landowska, A. (2021). Keystroke Dynamics Patterns While Writing Positive and Negative Opinions. Sensors 2021, Vol. 21, Page 5963, 21(17), 5963. https://doi.org/10.3390/S21175963

Marrone, S., & Sansone, C. (n.d.). Identifying Users' Emotional States through Keystroke Dynamics. https://doi.org/10.5220/0011367300003277

Pentel, A. (2019). Predicting User Age by Keystroke Dynamics. Advances in Intelligent Systems and Computing, 764, 336–343. https://doi.org/10.1007/978-3-319-91189-2_33

Raul, N., Shankarmani, R., & Joshi, P. (2020). A Comprehensive Review of Keystroke Dynamics-Based Authentication

Mechanism. Advances in Intelligent Systems and Computing, 1059, 149–162. https://doi.org/10.1007/978-981-15-0324-5_13

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, 69–79. https://doi.org/10.58496/BJML/2024/007

Shekhawat, K., & Bhatt, D. P. (2022). A novel approach for user authentication using keystroke dynamics. Journal of Discrete Mathematical Sciences and Cryptography, 25(7), 2015–2027. https://doi.org/10.1080/09720529.2022.2133241

Sriman, J., Thapar, P., Alyas, A. A., & Singh, U. (2024). Unlocking Security: A Comprehensive Exploration of Biometric Authentication Techniques. Proceedings of the 14th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2024, 136–141. https://doi.org/10.1109/CONFLUENCE60223.2024.10463322

Talekar, B., & Agrawal, S. (2020). A Detailed Review on Decision Tree and Random Forest. Biosc.Biotech.Res.Comm. Special Issue, 13, 245–248. https://doi.org/10.21786/bbrc/13.14/57

Thakare, A., Gondane, S., Prasad, N., & Chigale, S. (2021). A Machine Learning-Based Approach to Password Authentication Using Keystroke Biometrics. Lecture Notes in Electrical Engineering, 749 LNEE, 395–406. https://doi.org/10.1007/978-981-16-0289-4_30

Tsimperidis, I., Asvesta, O.-D., Vrochidou, E., & Papakostas, G. A. (2024). IKDD: A Keystroke Dynamics Dataset for User Classification. Information, 15(9), 511. https://doi.org/10.3390/INFO15090511

Tsimperidis, I., Yucel, C., & Katos, V. (2021). Age and Gender as Cyber Attribution Features in Keystroke Dynamic-Based User Classification Processes. Electronics, 10(7). https://doi.org/10.3390/ELECTRONICS10070835

Tsvetkova, A. D., & Bakhteev, D. V. (2024). KEYSTROKE DYNAMICS FEATURES IN FORENSIC IDENTIFICATION:: theoretical and experimental approaches. Revista EJEF, 5, 2024. https://doi.org/10.70982/REJEF.V1I5.66

Vázquez-Novoa, F., Conejero, J., Tatu, C., & Badia, R. M. (2023). Scalable Random Forest with Data-Parallel Computing. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14100 LNCS, 397–410. https://doi.org/10.1007/978-3-031-39698-4_27