

Journal Scientific and Applied Research, Vol. 31, 2026
International Journal

ISSN 1314-6289 (Print)
ISSN 2815-4622 (Online)

Original Contribution

CONCEPTUAL MODEL FOR GENDER PREDICTION BASED ON USER INPUT

Valentin T. Atanasov

Department of Communication and Computer Engineering and Security Technologies, Faculty of Technical Sciences, Konstantin Preslavsky University of Shumen, Shumen 9712, 115 Universitetska St., e-mail: v.atanasov@shu.bg

Abstract: *The present study examines an approach to the synthesis of a conceptual model for gender prediction based on user input. The predictive capability of the model is grounded in statistical differences observed within specific datasets and in the applicability of a machine-learning classification model. The predictive analysis is governed by a set of parameters, including language, the specifications of the peripheral input device, sample size, and biological indicators. It should be noted that reliable classification of a user's gender is a function of predefined key indicators.*

Key words: *Keystroke dynamics, Gender classification, Prognosis, Modelling.*

1. Introduction

User identification is one of the key factors influencing the subsequent behavior of modern systems. There is a gap not in the concept of gender recognition itself, but in the definition of an efficient set of predefined statistical metrics.

There is a theoretical predictive possibility of determining biological characteristics, such as the user's gender, based on a specific set of statistical indicators directly or indirectly related to the analysis of structured texts. These indicators include the time interval between two consecutive keystrokes, key press duration, intervals within disyllabic, trisyllabic, or multisyllabic strings, typing speed, frequency of deletion keystrokes, the length of keystroke sequences, as well as actions performed when errors and corrections occur during free and natural typing.

The ability to accurately predict gender is not limited solely to the personalization of initial user activity. A statistical input dataset needs to be formulated. The ability to accurately predict gender is not limited solely to the personalization of user activity. According to the tested models, an accuracy level of up to 82% [1] [5] can be achieved, demonstrating the ability to determine the gender of unknown users based on statistical input data.

The present study examines a conceptual model that incorporates a statistical set of user input metrics.

2. Related work

Research on gender identification through keystroke dynamics indicates that temporal and behavioral characteristics of typing can contain statistically significant information for user classification. Indicators such as key press duration, intervals between consecutive keystrokes, typing speed, use of the Backspace key, and other characteristics of keyboard input have been analyzed [1]. Other approaches combine keystroke dynamics with stylometric and linguistic features, with results demonstrating the possibility of classification above chance level, although with varying effectiveness depending on the amount of available data [2]. Review studies in the field also consider gender and age as so-called soft biometric characteristics that can be extracted through the analysis of typing behavior.

However, results from individual experimental settings demonstrate considerable variability in classification accuracy. For example, accuracies above 95% have been reported for certain datasets [1], whereas other experiments have achieved approximately 70% accuracy or a maximum of 63% [4]. Additional studies have examined typing dynamics in real-world chats and the influence of features related to linguistic and social behavior [3], as well as natural keyboard input when using smartphones [5]. In summary, the existing research supports the feasibility of developing machine-learning classification models that use keystroke dynamics to predict gender categories; however, their effectiveness depends on the characteristics of the data, language, input device, sample size, and the selected methods for feature extraction and classification.

3. Problem description

The primary objective of this study is to develop a set of indicators whose numerical values could predict, with a high degree of reliability, the gender of an unknown user based on arbitrary user input. The second major objective is to develop a protocol for using this dataset, with the aim of making the protocol applicable across a broad range of software models designed to predict the gender of an unknown user based on their input behavior.

Based on the studies reviewed above, it can be concluded that developing a reliable model for predicting the gender of an unknown user from user-input data involves a number of interrelated statistical factors. First, there is no universal set of input features that provides consistent predictive performance under different conditions. The effectiveness of a model depends on the selected keystroke-dynamics features, as well as on the language, content, amount of available data, input device, and context of interaction. Significant differences are also observed depending on the classification algorithm and methodology employed, which makes direct comparisons of results across individual studies difficult.

Another major hurdle is generalization to previously unseen users, since high accuracy on a specific dataset does not necessarily guarantee comparable performance for new users or under different conditions. An additional issue concerns the distinction between behavioral characteristics and linguistic or stylometric features, as a model may exploit information that does not directly originate from the user's typing behavior. Therefore, the evaluation of such a model should rely on statistical metrics and an analysis of class distributions.

Ultimately, the key effort lies in constructing a statistical input set that contains sufficiently informative features while maintaining predictive robustness across variations in users and input conditions. As a result, a summary of the preliminary conclusions can be presented as follows:

- There is no universal set of predictors that performs equally well under all conditions;
- Results depend on the specific dataset and experimental setup;
- Generalization to previously unseen users represents a significant challenge;

- The language and content of the entered text can affect the results;
- Language independence is not guaranteed and requires additional validation;
- The input device and interaction environment can alter user-input characteristics;
- The choice of classification algorithm has a significant impact on the resulting accuracy;
- The amount of available input data affects the statistical reliability of the prediction;
- Relevant predictive signals must be distinguished from confounding factors that may influence typing behavior;
- Accuracy alone is insufficient for evaluating the model and should be complemented by additional performance metrics and analyses.

4. Conceptual model

The proposed model can be presented as a multi-stage statistical model for determining the gender of an unknown user based on characteristics extracted from user input. The formal logic of the conceptual model defines a sequential process in which user input is transformed into a statistical representation through feature extraction, followed by the identification and control of potentially confounding factors. The resulting statistical feature set is then provided to a classification model, which produces the predicted gender class of the unknown user. Formally, the conceptual model can be represented as:

$$\hat{G} = f_4(f_3(f_2(f_1(U), C), M)) \tag{1}$$

The model comprises distinct stages, formally represented by functions:

$$f1: U \to F \tag{2}$$
$$f2: F \to S \tag{3}$$
$$f3: S \times C \to X \tag{4}$$
$$f4: X \times M \to G \tag{5}$$

where:
U — the set of allowable user inputs;
F — the set of extracted features;
S — the set of statistical vectors;
C — the set of contextual and potentially confounding factors;
X — the set of controlled/reduced statistical features;
G — the set of gender classes;
M — the set of classification models;
P — the set of performance and reliability metrics.

Where the preliminary statistical input assessment could be presented as:

$$f_s: G \times G_{true} \to P \tag{6}$$

where G_{true} denotes the actual class of the user, used only during the evaluation stage. The conceptual model is built by two logical components:

Prediction process

$$U \to F \to S \to X \to \hat{G} \tag{7}$$

Evaluation process.

$$(\hat{G}, G_{true}) \to \mathrm{P} \tag{8}$$

Where:

$$P = \left\{ \begin{array}{c} Accuracy \\ Precision \\ Recall \\ F1 - score \\ Specificity \\ Sensitivity \\ Confusion\ Matrix \\ ROC - AUC \\ Confidence\ Estimates \\ Probability\ Estimates \\ Class\ Distribution\ Analysis \end{array} \right\} \quad (9)$$

Thus, P represents the set of all evaluation metrics, while each P_i represents an individual metric. This provides a cleaner formal representation than placing all metrics within a single array structure. The meaning of all elements of the metric set is:

Accuracy — the proportion of all correctly classified cases relative to the total number of cases.

Precision — the proportion of correctly predicted cases belonging to a given class relative to all cases predicted by the model as belonging to that class.

Recall — the proportion of correctly identified cases belonging to a given class relative to all cases that actually belong to that class.

F1-score — the harmonic mean of *Precision* and *Recall*, providing an overall measure of the balance between them.

Specificity — the proportion of correctly identified cases that do not belong to the class under consideration.

Sensitivity — the ability of the model to correctly identify cases belonging to the class under consideration; in binary classification, it generally corresponds to Recall.

Confusion Matrix — a matrix showing the number of correctly and incorrectly classified cases for each class.

ROC-AUC — the area under the ROC curve, which evaluates the model's ability to discriminate between classes at different classification thresholds [6].

Confidence Estimates — estimates of the model's confidence in its prediction.

Probability Estimates — probability estimates of an observation belonging to each of the possible classes.

Class Distribution Analysis — an analysis of the distribution of observations across classes, used to assess class balance or imbalance and to ensure the correct interpretation of the results.

5. Conclusion

The main characteristics of the proposed conceptual model and their relationships to previous studies can be identified (Table 1). This leads to the assertion that the proposed model is not constructed from entirely new components; rather, its essential distinction lies in the structuring and integration of these components into a unified framework that incorporates statistical representation, control of potential confounding factors, generalization to previously unseen users, and multidimensional assessment of performance and reliability.

Table 1. Relationship between the proposed conceptual model and previous approaches

Aspect	Previous Studies	Proposed Conceptual Model
Input data	Keystroke and user-input data	User input U
Feature extraction	Applied in individual studies	Explicit stage $U \rightarrow F$
Statistical feature set	Defined according to the specific study	Explicitly represented as $F \rightarrow S$
Feature selection/reduction	Applied in a number of studies	Explicit stage $S \rightarrow X$
Classification	Applied using selected algorithms	Classifier-independent stage $X \rightarrow \hat{G}$
Classification algorithm	Often based on specific algorithms or algorithm combinations	Represented as a variable component M
Language	Usually determined by the dataset or experimental setup	Explicitly considered as a contextual/confounding factor
Text content	Determined by the experimental task or dataset	Explicitly considered as a potential confounding factor
Input device	Usually defined by the experimental setup	Explicitly considered as a contextual factor
Interaction environment	Depends on the experimental conditions	Explicitly incorporated into the model
Confounding factors	Not consistently represented as a separate modelling stage	Dedicated control stage $S, C \rightarrow X$
Unseen-user generalization	Depends on dataset partitioning and experimental protocol	Explicitly incorporated as a validation requirement
Performance evaluation	Varies across individual studies	Defined as a separate evaluation stage
Accuracy	Frequently used as a principal performance measure	Included but not considered sufficient on its own
Class-specific evaluation	Varies between studies	Precision, Recall, Specificity and Sensitivity explicitly included
Error analysis	Depends on the selected evaluation measures	Confusion Matrix explicitly included
Discrimination assessment	Applied in studies using appropriate classifier outputs	ROC-AUC included where applicable
Confidence/probability estimates	Dependent on the classification algorithm	Explicitly considered as model-output information
Class distribution	Not consistently included as a separate evaluation component	Explicitly included in the evaluation stage
Overall framework	Frequently focused on a specific dataset, feature set or classifier	Unified, classifier-independent conceptual and statistical framework

The main contribution of the proposed conceptual model is the formulation of a unified conceptual and statistical framework for gender classification from user-input data, integrating feature extraction, statistical feature construction, confounding-factor control, classifier-independent prediction, unseen-user generalization, and multidimensional performance and reliability assessment.

The identified research gap lies in the fragmentation of the existing literature. Although previous studies demonstrate that gender classification from keystroke and user-input

dynamics is feasible, they typically investigate specific feature sets, devices, datasets, classifiers, or experimental protocols. The proposed work addresses this fragmentation by formulating a unified conceptual framework that explicitly represents feature construction, contextual and potentially confounding factors, user-independent validation, classifier-independent prediction, and multidimensional performance assessment within a common statistical pipeline.

References

[1] Tsimperidis, I., Arampatzis, A., Karakos, A., Keystroke dynamics features for gender recognition, Digital Investigation, Volume 24, 2018, pp: 4-10, ISSN 1742-2876, https://doi.org/10.1016/j.diin.2018.01.018.

[2] Brizan, D., G.,Goodkind, A., Koch, P., et al., Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics, International Journal of Human-Computer Studies, Volume 82, 2015, pp: 57-68, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2015.04.005.

[3] Bukeer, A., Roffo, G., Vinciarelli, Type Like a Man! Inferring Gender From Keystroke Dynamics in Live-Chats, IEEE Intelligent Systems, Volume 34, Issue 6, 2019, pp:53-59, ISSN: 1941-1294, doi:10.1109/MIS.2019.2948514.

[4] Asvesta, O., D., Tsimperidis, I., Yucel, C., Combining Keystroke Dynamics and Dempster–Shafer Theory to Recognize the Gender of an Unknown User, AI, Computer Science and Robotics Technology, Volume 4, 2025, pp: 1-19, https://doi.org/10.5772/acrt.20250038.

[5] Fernandez-Lanvin, D., Gonzalez-Rodriguez, M., Andres, J., Automatically Gender Identification from Smartphone with Undefined Keystroke Sequence, Journal of Organizational Computing and Electronic Commerce, 2026, pp: 1-35, ISSN: 1532-7744, https://doi.org/10.1080/10919392.2026.2699514.

[6] Hanley, J., A, McNeil, B., J., The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology, Volume 143, Issue 1, 1982, pp:29-36, https://doi: 10.1148/radiology.143.1.7063747. PMID: 7063747.