IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 August 2023, accepted 5 September 2023, date of publication 8 September 2023,
date of current version 13 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3313510

RESEARCH ARTICLE

The Way We Type Reveals Our Native Language

IOANNIS TSIMPERIDIS, DENITSA GRUNOVA, AND GEORGE A. PAPAKOSTAS
MLV Research Group, Department of Computer Science, International Hellenic University, 65404 Kavala, Greece
Corresponding author: George A. Papakostas (gpapak@cs.ihu.gr)

This work was supported by Hellenic Academic Libraries Link (HEAL-Link).

ABSTRACT Knowing some characteristics of an unknown user is useful information for security and commercial purposes. One of the acquired characteristics is the user's native language, and its recognition can be achieved with data derived from the text he/she types, since text is the most widespread means of communication between Internet users. Keystroke dynamics, which leverages data derived from how user types, ensures that no sensitive data are leaked. In this work, data from the daily typing of users of five different native languages are collected, keystroke dynamics features are extracted, the most suitable ones are selected using a feature selection algorithm, well-known machine learning models and a boosting algorithm are used, and a rate of correct prediction that exceeds 90% is achieved. Knowing a user's native language can help strengthen authentication systems, make better use of online services, and protect unsuspecting users from falling victim to fraud.

INDEX TERMS Biometrics, data mining, keystroke dynamics, machine learning, native language identification.

I. INTRODUCTION

Effective communication is pivotal for personal, social, and economic development. The Internet has expanded and accelerated the means of communication, and the average person uses it for work, education, information, and entertainment. Although phone calls, video conferencing, and voice messaging are available for communication between two or more users, text messaging remains the primary form of communication. That is, despite the increase in Internet speed and the consequent offer of more sophisticated communication applications, users prefer to communicate using text-based platforms such as email, Messenger, Viber, and Twitter. The reason is probably that users feel less exposed. Of course, communicating via text messaging comes with some problems. First, that the mood of the person writing the text message cannot be fully captured, with the result that many messages are misinterpreted, Second, that the author of the message is not identified, with the result that he/she has the possibility to hide or falsify his/her identity.

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

The latter can ensure to a certain extent the right of the user to remain anonymous, but at the same time, it is a factor of risks and difficulties. For example, it can be the tool of a malicious user to mislead the unsuspecting interlocutor, or it can prevent Internet companies from reaching the right people to inform them about their services. A solution to this problem has been proposed by various researchers, such as by using NLP [1], creating a user profile achieved through behavior in social networks, etc. Another suggestion is by using keystroke dynamics, which is the field of study that analyzes the patterns and rhythms of a user's typing behavior [2]. It is based on the idea that each individual types in a unique way and that this typing style can be used as a form of biometric identification, such as a fingerprint or facial recognition scan, as a means of classifying users, or as a means of identifying the mental and physical state of users. One of the main advantages of keystroke dynamics is that it does not require special equipment or hardware. Instead, it uses the simple data already generated by a user's keyboard as they type. This means that keystroke dynamics can be easily implemented in a variety of settings, including personal computers, mobile devices, and even public terminals. Another advantage is that it does not require the collection of sensitive

data. Unlike other forms of biometric recognition, such as facial recognition, keystroke dynamics does not utilize personal information or images, instead, it simply analyzes the way a user types, which is non-invasive and non-threatening. The fact that keystroke dynamics is based on the most widely used type of communication, text, is another advantage. Text-based communication, such as email, instant messaging, and social media, is ubiquitous in today's world. This means that keystroke dynamics algorithms can be applied to a wide range of online interactions, making it a powerful tool across different platforms and applications.

The most prominent and widely employed features in keystroke dynamics analysis are keystroke durations and digram latencies. Keystroke durations refer to the amount of time a key is held down before it is released, and digram latencies refer to the amount of time between the use of two consecutive keys.

The main goal of this paper is to develop a new approach to determine the native language of a user through keystroke dynamics, by using free-text data, in order to reveal part of the identity of an unknown person. This is achieved by extending the research presented in [3]. The proposed approach is based on the idea that different native languages have a unique effect on the way people type. The hypothesis is that individuals tend to transfer specific language characteristics to their typing behavior, such as typing speed and delays of digrams. For example, the different frequency of occurrence of characters and digrams in different languages is reflected in the way users type. By analyzing these characteristics, it is possible to develop a method for determining a user's native language. In this way, this work aims to extend the usage of keystroke dynamics from user authentication to user's native language identification, exploiting language-dependent and user-independent typing patterns.

The following sections of this paper provide a comprehensive overview of this topic. First, a thorough review of the relevant literature is presented. Second, the methodology used in this research is discussed in detail. Next, the findings of the study, which aimed to determine a user's native language, are presented and analyzed. Finally, the paper is concluded with a discussion of practical applications of this research and possible avenues for future research.

II. BACKGROUND

Keystroke dynamics is the computer science field of measuring and analyzing the timing and rhythm of keystrokes made on a keyboard or other input device. Keystroke dynamics can be used for a variety of purposes, including authentication, user classification, and the physical and mental estimation of a user.

Authentication using keystroke dynamics is based on the fact that every person has a unique typing rhythm or pattern. By analyzing keystroke dynamics, a system can determine whether the person typing is the authorized user or an imposter.

The study by Raul et al. [4] presented a comprehensive review of keystroke dynamics-based authentication mechanisms. The authors focused on the use of keystroke dynamics as a biometric authentication technique, which involves analyzing a user's typing rhythm and other behavioral patterns to verify their identity. The study provided an overview of the existing research on keystroke dynamics, including the datasets and methods used in previous studies as well as the limitations and challenges of this approach. The authors also proposed possible solutions to address these challenges and highlight the potential applications of keystroke dynamics-based authentication mechanisms in various domains, including e-commerce, online banking, and computer security. The study was based on a review of the literature and uses a qualitative approach to analyze and synthesize the findings from previous research studies.

In another work, Alsultan et al. [5] introduced an approach for user authentication using free-text data that incorporates the use of unconventional typing features. Semi-timing features and processing features were extracted from the user's keystroke stream. Decision trees were utilized to classify each of the user's data. In parallel for comparison, support vector machines were used for classification along with an ant colony optimization feature selection technique. The results obtained from this study are encouraging, as a low false acceptance rate (FAR) and false rejection rate (FRR) were achieved in the experiment phase.

El-Kenawy et al. [6] proposed a new authentication mechanism for smartphone users that combines meta-heuristic optimization techniques with keystroke dynamics. The authors aimed to improve the accuracy and efficiency of user authentication on smartphones, which is a critical issue given the increasing use of mobile devices for sensitive tasks such as mobile banking and e-commerce. The authors analyzed the data using meta-heuristic optimization algorithms and machine learning techniques to develop a dynamic keystroke-based authentication model that can adapt to individual users' keystroke patterns over time. The results showed that the proposed approach achieves a high level of accuracy and reduces the time required for user authentication compared to traditional methods.

Moreover, Ayotte et al. [7] presented a novel algorithm called "instance-based tail area density metric" to reduce the number of keystrokes required for user authentication using keystroke dynamics. The authors also investigated the effectiveness of keystroke dynamics features and found that all features contributed information about who was typing, but keystroke durations, down-down digram latencies, and up-up digram latencies were the most important. They used a random forest classifier to validate their approach across two publicly available datasets.

Similarly, Li et al. [8] focused on authentication based on features derived from keystroke dynamics in the free-text case. They found that dividing the sequence into a

number of fixed-length subsequences was an effective feature design strategy. Furthermore, they developed and analyzed image-like structures of constructed features that they reported as keystroke dynamics image (KDI) and keystroke dynamics sequence (KDS). KDI was used as input for the experiments with CNN, while KDS served as input data for the CNN-RNN experiments. In both cases, they applied cutoff normalization. The experimental results reported show that the pure CNN architecture outperforms the combination of CNN and RNN, while the cutoff significantly improves the performance of both models.

Moreover, Alsuhibany and Almuqbil [9] aimed to investigate the impact of using three different sizes of touch keyboard layouts in the user authentication process when typing Arabic free text. The experiment's results demonstrated that using the time feature for keystroke dynamics-based authentication offers a feasible technique for multi-device environments, since they achieved average FAR and FRR scores of 1.1% and 18.2%, respectively.

Most of the research in keystroke dynamics has focused on user authentication, resulting in the presentation of methods with increasingly better results in terms of FAR and FRR. The consequence of these was the development of authentication systems both for computers and mobile devices, as well as for both Internet applications and economic interest services.

Regarding user classification, Tsimperidis et al. [10] proposed a new machine learning model, the randomized radial basis function network, or R2BN, which combined characteristics of both radial basis function network and AdaBoost algorithm. The goal was to recognize and record the educational level of a person standing behind the keyboard. The performance of the proposed model is evaluated using the empirical data obtained from the volunteers' keystroke logging during daily computer use.

In another study, Nascimento et al. [11] investigated the use of feature selection techniques to identify the most relevant keystroke dynamics features for gender prediction. The authors used several feature selection techniques, including, recursive feature elimination (RFE) and correlation-based feature selection (CFS). The results of the experiments showed that the use of feature selection techniques improved the accuracy of gender prediction based on keystroke dynamics. In particular, the RFE technique was found to be the most effective.

The study by Cascone et al. [12] explored the use of touch keystroke dynamics for demographic classification, including age and gender. The authors aimed to investigate the potential of touch typing, which includes information about the amount of pressure applied to the keys, for determining user demographic information. The authors analyzed the data using machine learning algorithms, including k-nearest neighbours and support vector machines. One of the findings was that younger people tend to type faster and with more errors, while older people tend to type slower and with fewer errors.

In the area of user age search, Tsimperidis et al. [13] used a dataset of 387 logfiles and extracted 700 features from it, which were keystroke durations and digram latencies. They divided the users into four age groups and they tested their dataset with different classifiers. Finally, the experiments resulted in the creation of a system that was able to distinguish with about 90% accuracy the age group of an unknown user.

In their work, Sahu et al. [14] presented a localization-based algorithm to solve a multi-user classification problem in keystroke dynamics. The proposed algorithm performed dimensionality reduction using PCA or kernel-PCA, processed training data, nominated representative points as anchors, and treated testing data samples as targets at unknown locations. Using keystroke data from each of the known users, user-to-user and user-to-target cross-distances were computed using scaled Manhattan distances in the keystroke space.

There are already several studies in the literature on user classification using keystroke dynamics, based on some intrinsic (e.g., gender, handedness, etc.) or acquired (e.g., educational level, political beliefs, etc.) characteristic. The systems proposed, which can help, among other things, in digital forensics and targeted advertisement, show increasingly better accuracy.

Moreover, keystroke dynamics can be used for the physical and mental estimation of a user. Research has shown that typing rhythm can be used to detect physical and mental conditions such as Parkinson's disease, depression, and stress. By analyzing keystroke dynamics, a system can detect changes in typing patterns and provide an early warning for potential health issues.

The study by Arroyo-Gallego et al. [15] investigated the use of keystroke dynamics as a biomarker to detect mental fatigue. For this work, they used TypeNet, a state-of-the-art deep neuron network originally intended for user authentication at a large scale using keystroke dynamics. They adapted TypeNet for fatigue detection by exploiting the person identification information embedded in it. All experiments were conducted using three keystroke databases, including different environments and data collection protocols. Results showed performance ranging between 72.2% and 80.0% for classifying fatigue versus resting samples, which is aligned with previously published models on daily alertness and circadian cycles.

With different goals, López-Carral et al. [16] analyzed the typing patterns of a large sample of participants who were asked to describe a set of images selected from the OASIS normative database for affective research. Analyzing the keystroke dynamics data obtained by recording the participants, they found highly significant negative correlations of both digram latencies and keystroke durations with both arousal and valence, as well as between time to start and valence. Then, they checked for generalizations on

the content itself, finding significant negative correlations between keystroke durations and valence and between digram latencies and arousal.

The research conducted by Alfalahi et al. [17] aimed to investigate keystroke dynamics as a digital biomarker to detect fine motor decline in neuropsychiatric disorders. The researchers conducted a systematic review and meta-analysis of previous studies that used keystroke dynamics for the diagnosis of fine motor apoptosis in neuropsychiatric disorders, including Parkinson's disease, Huntington's disease, and schizophrenia. The research analyzed data from 25 different studies. The meta-analysis revealed that keystroke dynamics could accurately distinguish between healthy witnesses and patients with neuropsychiatric disorders.

Moreover, the study by Tripathi et al. [18] proposed a new method for detecting Parkinson's disease (PD) based on keystroke dynamics using data from an unsupervised population at home. They extracted several keystroke dynamics features, namely keystroke durations and digram latencies, created a new model that combines a new type of keystroke dynamics signal representation using these features, and used machine learning algorithms to classify participants as having PD or being healthy based on their keystroke data.

Recognizing a speaker's native language is a critical task in many applications, such as language learning, speech recognition, and forensic linguistics. Several methods have been proposed to identify a speaker's native language based on different linguistic features, including acoustic, lexical, and syntactic features. Keystroke dynamics have also been explored as a potential source of information for determining a typist's native language.

In their paper, Buckley et al. [19] build upon the notion of keystroke dynamics to infer an anonymous user's name and predict their native language. For this purpose, they asked volunteers to type a paragraph, generated according to their names, three times. The researchers found that there is a discernible difference between the ranking of digrams (based on their timing) contained within the name of a user and those that are not. As a result, they propose that individuals will reliably type information they are familiar with in a discernibly different way. They also found that it is possible to identify approximately a third of the digrams forming an anonymous user's name purely from how (not what) they type. Regarding language classification, they collected data from users from five different native languages and achieved an accuracy of 45%.

In another study, where native language is a crucial user characteristic, Bours and Brahmanpally [20] examined how well a keystroke dynamics system worked when users were asked to type specific words. They gathered data from participants who spoke six different native languages and had them type random words ranging from 8 to 12 characters in length in each of these six languages. The participants also typed random French and English words, with English assumed to be a language familiar to everyone, while French was not a native language for any of the participants, and many were likely not fluent in it. The researchers discovered that using language-specific words led to better performance of the challenge-based keystroke dynamics system compared to one based solely on English words. When using words in their native language, participants who spoke that language performed similarly to a system based on all English words, but non-native speakers had significantly lower false match rates than native speakers.

In a concluding paper, Shadman et al. [21] provided an overview of the concepts, techniques, and applications of keystroke dynamics. It covers the background and history of keystroke dynamics analysis, different approaches to collecting and processing keystroke data, different types of keystroke dynamics analysis, and their applications in different domains such as security, human-computer interaction, and health monitoring.

As made clear from the above review, keystroke dynamics is mostly used for authentication systems, and today it is a mature technology with many commercially available products. Beyond authentication, there is much research into user classification. Almost all of user classification works were aimed at finding the gender and/or age of a user. Very few studies focus on user characteristics other than gender and age, and only three, [3], [19], and [20], deal with users' native language. In the latter two of these studies, the fixed text method was followed, in which volunteers are asked to type specific text. In [19] a percentage of 45% is achieved in the correct prediction of the user's native language, among five different options, while in [20] some conclusions are drawn about the typing habits of users, related to their native language, so that they can be used in authentication systems.

Therefore, to the best of authors knowledge, in this work the user classification according to their native language is attempted for the first time, with data derived from their daily use of their computer. That is, using free text method.

A. CONTRIBUTION OF WORK

The problem that this research tries to solve is the recognition of the native language of an unknown Internet user, for purposes of profiling or authentication.

Since text is the simplest means of communication between users, but also the most widespread, its use is a very good solution for extracting the necessary information.

Two technologies applied to the text, or to the process of writing it, are natural language processing and keystroke dynamics. The second does not require the content of the text to be known and therefore can protect the privacy of users.

With this in mind, in this work, the solution of the problem using keystroke dynamics was chosen. However, in the literature it appears that this way of solving the problem has not been extensively attempted. Specifically, in addition to the present research, there are three other works that use

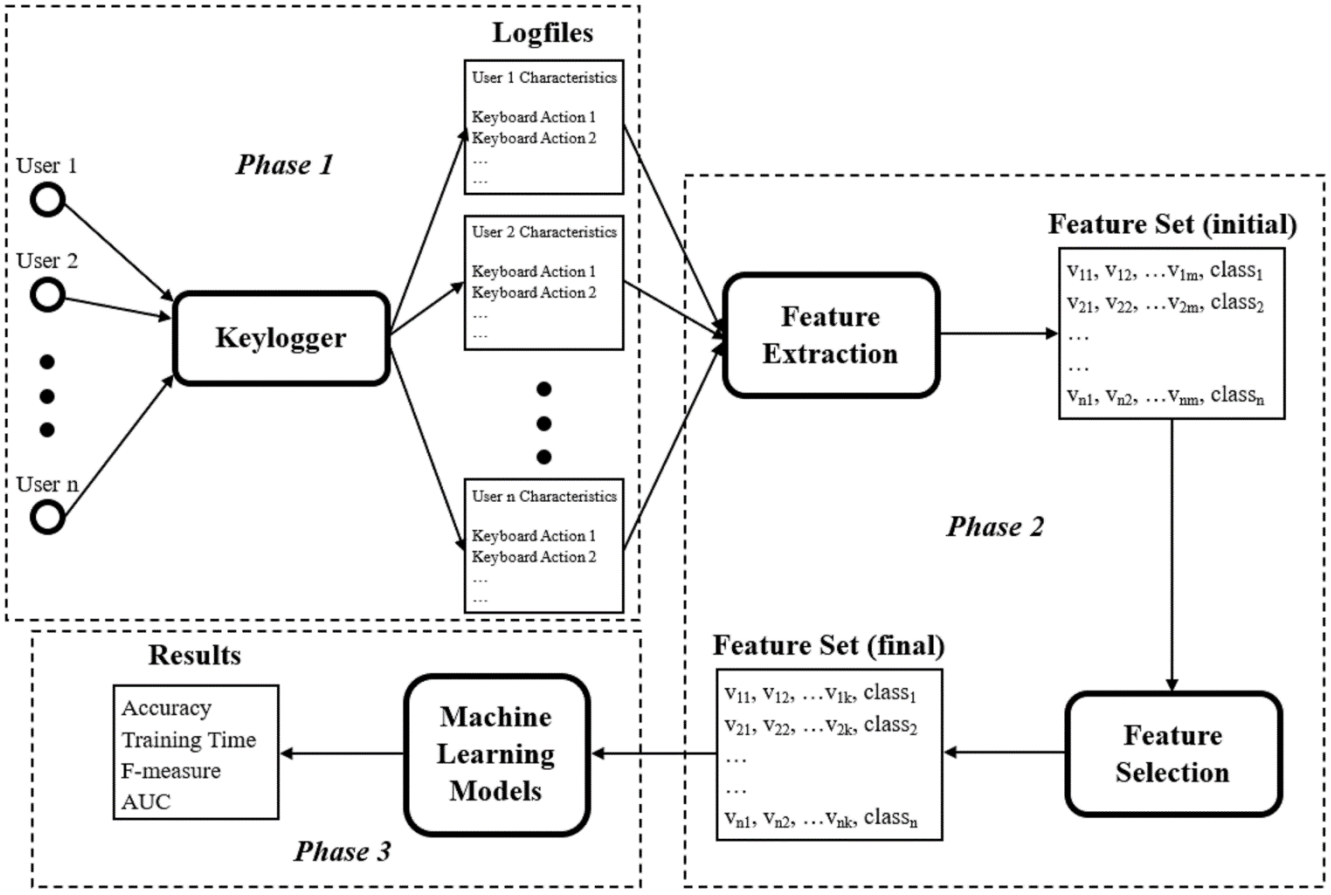

FIGURE 1. The block diagram of the methodology.

keystroke dynamics methodologies related to the native language.

In [19] the familiarity that a user acquires when typing the digrams contained in his/her name was exploited, while the data were collected with a fixed text method. In a similar way, the data were collected in [20], in which the researchers studied the way users type words in languages they are familiar with, as well as in those they are not.

An important difference of the present work, which as mentioned above is an extension of [3], is that the data were collected with a free text method, which better approximates the conditions of real typing. From the data collected, in [3] keystroke durations were used for native language classification. The differences of the methodology of this work are that firstly, many more keystroke dynamics features were extracted, secondly a feature selection algorithm was used to use only the most suitable features and reduce the time required to extract the result, and thirdly a boosting algorithm was used to improve system performance.

Accordingly, the present study appears to be the state-of-the-art in finding the native language of an unknown user through keystroke dynamics.

III. METHODOLOGY

The methodology employed in this study encompasses three distinct phases. The first phase involves data collection, where keystroke data are gathered from the participants. In the second phase, feature extraction and selection techniques are applied to the collected data. Finally, in the third phase, machine learning models are trained and evaluated using the selected features. In addition, boosting algorithms, such as bagging, are applied to enhance the performance of the chosen model.

The extensions of the methodology in relation to that of the paper [3] are found, firstly to the use of more keystroke dynamics features, secondly to the use of feature selection algorithm, and thirdly to the use of meta-algorithms to improve the performance of the system.

The methodology is illustrated in the block diagram of Fig. 1, in which the three phases are distinguished. In the first phase, the keylogger installed on the volunteers' computers appears, with the volunteers creating logfiles. A number of keyboard actions have recorded in each logfile. In the second phase, there is the creation of the file with the feature set and then the creation of the file with the features selected as appropriate. The value of the j-th feature of the i-th record

of the feature set, in Fig. 1, denoted by v_{ij}. Finally, in the third phase, the extraction of results after the use of machine learning models is presented.

A. DATA COLLECTION

During the data collection phase, hundreds of users were approached to participate in the project. To protect the personal data of the volunteers who eventually participated, some measures were taken. First, volunteers chose the times at which their typing was recorded, second, they had the right to review the recorded data, third, they had the right to refuse to hand over their data, and fourth, assurances were given by the researchers that the data will be encrypted, used only for the needs of this research, and will not be shared with any third party. All of the above was included in a consent form, which was signed by the participants, and which described in detail the obligations and commitments of the researchers, the possible risks, as well as the objectives of the research.

The volunteers who eventually took part in the study were given keylogger software with no further instructions, except that they would have to use their computer exclusively themselves when the keylogger was active. The reason why this particular way of recording typing was chosen was to approximate realistic keyboard usage as much as possible. Thus, it was not possible to check which applications the users were using when they were being recorded, what times of day they were being recorded, and if they were typing in languages other than their native language. All this was not necessary to be controlled, because the question was that the typists behave as naturally as possible. The only thing that needed to be cross-checked was the correct declaration of the native language of each volunteer, which was done diligently.

After a process that lasted from April to July 2022, 194 logfiles were collected, from users of five different native languages. This is a relatively small dataset, but it appears to be the only dataset in the literature derived from volunteer recording with free-text method, in terms of native language classification. Table 1 summarizes the number of logfiles per native language, as well as their percentage of the total.

TABLE 1. Number of logfiles per native language.

Native Language	Logfiles	%
Albanian	51	26.3
Bulgarian	46	23.7
English	17	8.8
Greek	55	28.3
Turkish	25	12.9
Total	194	100.0

Each logfile contains data from approximately 3,500 keystrokes, which were recorded in any application the volunteer wanted, typing whatever they wanted, at any time of the day (free-text). Also, each logfile contains the characteristics of the volunteer, among which is the native language. The data are stored in a standard CSV format, which includes the exact time when an action was performed on the keyboard and the virtual key code of the key used in that action.

The five languages of interest in this study, Albanian, Bulgarian, English, Greek, and Turkish, are all from different language families. Albanian is an Indo-European language spoken by about 7.6 million people, mostly in the west of the Balkans in countries such as Albania and Kosovo [22]. Bulgarian is a Slavic language spoken by about 7 million people, mostly in Bulgaria [23]. English is a West Germanic language spoken by over 1.5 billion people worldwide, mainly in the United Kingdom, the United States, Canada, Australia, and New Zealand [24]. Greek is the only surviving member of the Hellenic branch of the Indo-European language family. A vast majority of the Greek population, around 95%, speaks Greek as their native language, and about 840,000 Greek Cypriots also converse in the language [25]. Turkish is a Turkic language spoken by over 70 million people mostly in Turkey but also in other countries such as Cyprus and Azerbaijan [26]. Each of these languages has a unique grammar, vocabulary, and pronunciation system, which can affect the way people type. The fact that they come from different language families means that they have distinct linguistic characteristics that could potentially be used to determine a user's language based on keystroke dynamics.

B. FEATURE EXTRACTION AND SELECTION

In the second phase of the research, a procedure was followed to extract the most widely used keystroke dynamics features from the dataset created in the previous phase, namely keystroke durations and digram latencies [27]. This was done by calculating the difference between the release and press timestamps of each key, for keystroke durations, and by calculating the difference between the press timestamp of a key and the press timestamp of the next key, for digram latencies. But the use of a key or a digram, most of the time at least, is found several times in a logfile. For this reason, the value of each feature was derived from the average of the times calculated from all its appearances. However, when the use of a key appears less than 5 times in a logfile and when the use of a digram appears less than 3 times, then the value of the corresponding feature is not considered representative and is not taken into account. In those cases, the missing values were imputed with the mean of the numerical distribution.

Finally, the feature set was created, consisting of records, each of which is mapped to a logfile. In the records, separated by commas, are the values of the features calculated with the previous procedure, with the last one being the class to which the logfile belongs, which can take as a value one of the five native languages (''Albanian'', ''Bulgarian'', ''English'', ''Greek'', and ''Turkish''). The creation of the feature set is also shown in the block diagram of Fig. 1.

This process resulted in the extraction of 10,920 features, which is a very large number. Using all the features would entail very long training times for machine learning models.

TABLE 2. Top 20 Features with the highest information gain (IG).

#	Feature	Keys	IG	#	Feature	Keys	IG
1	73-78	I-N	0.3406	11	82	R	0.2087
2	83	S	0.2767	12	78-84	N-T	0.2076
3	66	B	0.2759	13	84-79	T-O	0.2061
4	70	F	0.2715	14	32	(space)	0.2024
5	68	D	0.2600	15	65-76	A-L	0.1961
6	82-79	R-O	0.2559	16	69-32	E-(space)	0.1960
7	84	T	0.2468	17	75	K	0.1911
8	76	L	0.2454	18	79-78	O-N	0.1869
9	65-78	A-N	0.2173	19	71	G	0.1835
10	86	V	0.2135	20	78	N	0.1829

For this reason, a feature selection algorithm was used to reduce the dimensionality. From the available algorithms, information gain is chosen because it seems to have better performance in a number of problems [28]. The information gain algorithm [29] works by evaluating each feature in the dataset according to its ability to provide information about the target variable. The algorithm calculates the information gain for each feature. The features with higher information gain are considered more informative and are therefore more likely to be selected for inclusion in subsequent analyses. It is worth noting that the selected features include both keystroke durations and digram latencies. After applying the feature selection algorithm to the dataset, 711 features with no-zero information gain were selected for use in the subsequent analysis. The top 20 features with the highest information gain are presented in Table 2.

As can be seen from Table 2, the features that contain the most information concern letters, which was expected, since numbers, punctuation, and other symbols are used in almost the same way by users of different native languages. A second conclusion is that in Table 2 features from several different letters are found, but the letter ''N'' seems to play an important role, both on its own and in the digrams it participates in.

C. EVALUATION

As part of investigating the performance of the system, a series of experiments were conducted using various machine learning models. To evaluate the accuracy of each model, the 10-fold cross-validation method was applied. The 10-fold cross-validation method is a method of repeatedly training and evaluating machine learning models that allows the use of all available data for training and evaluating the model. This is achieved by dividing the data into 10 equal non-overlapped parts (known as folds), training the model on 9 of them, and testing its performance on the part not used for training. This process is repeated 10 times, so that each part of the data is used once for testing and training on 9/10 of the data [30].

The criteria for selecting the most appropriate machine learning models, apart from accuracy, are F-measure and Area Under the ROC curve (AUC). The F-measure is a metric that combines precision and recall and is often used in binary classification tasks [31]. The ROC curve is a graphical plot that shows the tradeoff between the true positive rate and the false positive rate for different classification thresholds, and the Area Under the ROC curve is a metric that takes values between 0 and 1 [32].

TABLE 3. The performance of machine learning models, in terms of accuracy (Acc.), training time (Time to build model, TBM) in seconds, f-measure (F1), and area under the ROC curve (AUC).

Model	Acc.	TBM	F1	AUC
SVM	89.2%	0.12	0.891	0.945
SL	79.9%	0.69	0.799	0.932
NB	80.4%	0.02	0.804	0.925
BNC	85.6%	0.05	0.854	0.898
RBFN	86.1%	0.74	0.860	0.916

Using these metrics, the performance of many and different machine learning models, such as neural networks, decision trees, Bayesian classifiers, deep learning models, etc., was evaluated, and those that achieved the best results were selected. Five models performed well on the dataset, and these are Support Vector Machine (SVM), Simple Logistic (SL), Naïve Bayes (NB), Bayesian Network Classifier (BNC), and Radial Basis Function Network (RBFN).

IV. RESULTS

After the execution of a large number of experiments, the values of the classifiers' parameters from which their best performance is derived emerged. For SVM it is C-value equals to 1.25 and Polykernel as kernel type. For SL it is maximum number of boosting iterations equals to 500, number of iterations for early stopping of LogitBoost equals to 70, and weight trimming equals to 5%. For BNC it is initial count for estimating the probability tables equals to 0.06 and maximum number of parents of each node in Bayes network equals to 1. For RBFN it is number of clusters for K-Means equals to 20 and the minimum standard deviation for the cluster equals to 1.6.

Based on the results of the experiments, it can be seen that the SVM model performed the best overall. It achieved an accuracy of 89.2%, an F1 score of 0.891, and an AUC of 0.945. These values indicate that the SVM model was able to correctly identify the native language of the users with a high degree of accuracy. The second-best performing model was the RBFN model, which achieved an accuracy of 86.1%, an F1 score of 0.860, and an AUC of 0.916. However, the problem with this model is that it has a large Time to Build Model (TBM), compared to the rest. In particular, it is 6 times slower than SVM, which is the model with the best performance, and 37 times slower than NB, which is the model with the shortest training time. The BNC model performed moderately well, achieving an accuracy score of 85.6%, but having the worst AUC score of 0.898. Finally, the NB model has the lowest accuracy, but it is the fastest, requiring only 0.02 seconds for training, followed by BNC and SVM.

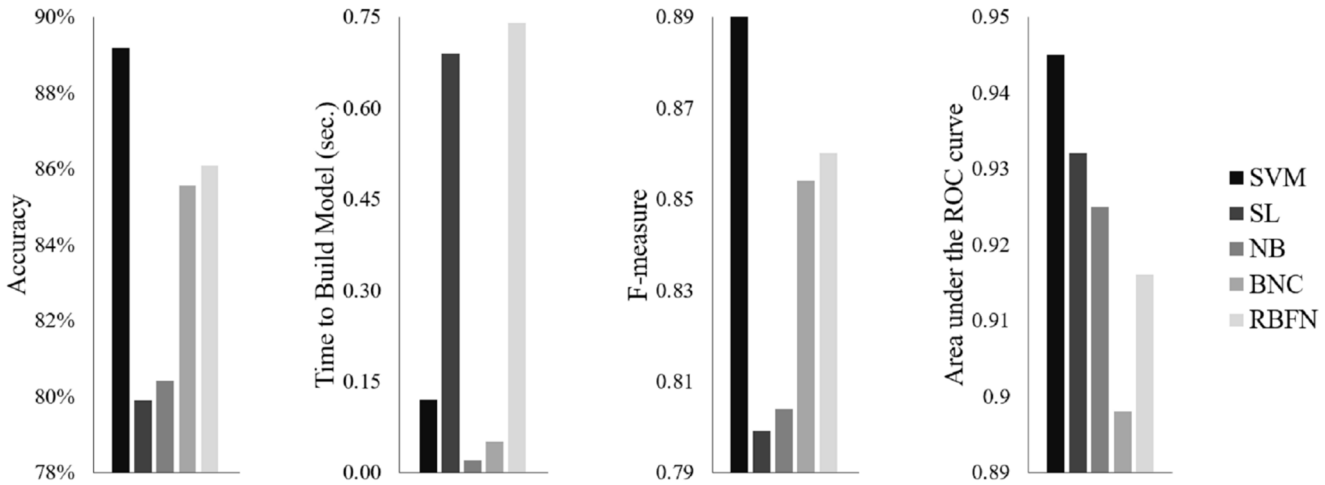

FIGURE 2. The performance of the five machine learning models.

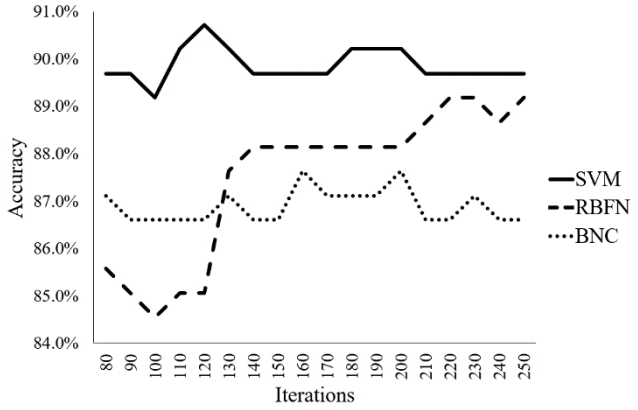

FIGURE 3. Accuracy using Bagging algorithm with different number of iterations.

The summary of results for all five machine learning models can be found in Table 3.

In conclusion, the results of the experiments suggest that the SVM model is the most effective at identifying the native language of users based on keystroke dynamics. However, because the dataset is relatively small and because of the small statistical differences between SVM, RBFN, and BNC, all three of these models are considered to be the most successful.

The graphical representation of the results is presented in Fig. 2.

Subsequently, to improve the accuracy of the method, boosting algorithms were used, using as base classifiers the SVM, RBFN, and BNC, which proved to be the most successful in this particular problem. The bagging algorithm turned out to have the best results.

Bootstrap Aggregating (bagging for short), as introduced by [33], is a technique used to improve the performance and stability of machine learning algorithms. It works by creating multiple bootstrap samples of the original training data, with each sample having the same size as the original data but allowing for repeatable instances. A base classifier is then trained on each bootstrap sample, and the predictions from all the base classifiers are combined to make the final prediction. By pooling the results from multiple classifiers, bagging can reduce variance and improve the accuracy of the final prediction.

TABLE 4. Accuracy, F1, and AUC, for each native language at the best run (SVM + bagging).

Language	Acc.	F1	AUC
Albanian	90.2%	0.902	0.969
Bulgarian	84.8%	0.848	0.971
English	94.1%	0.941	0.996
Greek	92.7%	0.927	0.965
Turkish	96.0%	0.960	0.980

After conducting a large number of experiments to find the parameters that lead to the highest accuracy, for different number of iterations in the algorithm, the results are shown in Fig. 3.

According to Fig. 3, the highest accuracy of the method is achieved at 120 iterations of Bagging algorithm that uses SVM as base classifier. In this best performance of the proposed method, accuracy, F1, and AUC, for each of the native languages of the dataset, are presented in Table 4.

The results presented in Table 4 show English and Turkish as the languages with the highest percentages of correct prediction. It is reminded, however, that these two languages have the least representation in the dataset, and thus any conclusion drawn may not be safe. On the contrary, the languages with the lowest accuracy are Albanian and Bulgarian. A possible explanation for this is that the countries where these languages are mainly spoken, Albania and Bulgaria, have experienced a high rate of immigration in recent years.

TABLE 5. Comparison between keystroke dynamics works for native language recognition.

Work	Number of Languages.	Outcome
Buckley *et al.* [19]	5	Accuracy 45.0%
Bours and Brahmanpally [20]	6	Important conclusions for authentication systems
Tsimperidis *et al.* [3]	5	Accuracy 82.5%
This work	5	Accuracy 90.7%

It is therefore possible that a larger percentage of users with Albanian and Bulgarian as their native languages are quite familiar with other languages as well.

In general, it can be said, based on Fig.2 and Fig.3, that a user's native language can be correctly identified, approximately 9 out of 10 times, using various machine learning models. This is also a strong indication that keystroke dynamics can be used to solve this problem.

V. DISCUSSION

The dataset created for the needs of this research comes from the recording of the typing of users from five different native languages. In the random prediction of what a user's native language is, the percentage of correct prediction would be 28.3%, i.e., the percentage of the logfiles of the largest class in the dataset.

The proposed method increases the correct prediction rate to 90.7%, which is a significant improvement compared to the results presented in [3], as well as compared to any other related research. It is recalled that the highest accuracy presented in [3] is 82.5%, and that only two other studies exist in the literature, at least as far as is known, in which the native language of users is searched, with data derived from the way they type, and the highest percentage of correct prediction achieved is 45%. Finally, it is worth mentioning that the important innovation of this method is that the data are obtained in free-text mode, which is much closer to the actual use of the computer by users. The use of free-text data in keystroke dynamics research, in addition to a better approximation of daily computer use, can help to develop continuous user authentication systems, in which the user is authenticated even after entering the password. As for example described in the works of Kim et al. [34] and Kim and Kang [35].

In Table 5 the works of keystroke dynamics on users' native language are presented. Their outcomes are compared with those obtained from the present research.

The high accuracy achieved in this study is an indication and not proof of the existence of a correlation between the native language and keystrokes. At the moment it can be said that the hypothesis that some characteristics of native language use are transferred to typing mode seems to be verified. In fact, this finding is used in the problem of language identification, which is encountered in studies of natural language processing, since some features identical to those of this study are used [36]. However, more data needs to be collected and further research conducted.

An explanation of the above is for example that the most frequently used digrams in the English language are ''TH'', ''HE'', and ''IN''. The most frequently used digrams in the Turkish language, as they are mapped to the QWERTY keyboard, are ''AR'', ''LA'', and ''AN'', and the corresponding ones in the Greek language are ''TO'', ''OY'', and ''TH''. These differences extend to other languages, as well as to other characteristics of the languages, such as the average size of words, the frequency of trigrams, etc. Therefore, a person with a particular native language who from a young age becomes familiar with a specific way of using the keyboard corresponding to that language, he/she acquires specific typing habits, which may betray his/her native language. The present research takes advantage of this fact and utilizes, among others, features related to the use of digrams, as can be seen in Table 2.

Once the correlation between native language and the way users type is proven, it will be possible to create a system that will recognize the native language of users, in addition to their other characteristics, using keystroke dynamics. However, when implementing such a system, considerable care must be taken in data collection so that user privacy is not violated. Keystroke dynamics leverages data from how users type, not what they type, meaning no access to sensitive information is required. But indirectly, the text typed by the user can be reconstructed from the keystroke logfiles, thus revealing passwords and private messages. One solution to this problem is to collect the keystroke data locally, on the user's device, then extract the keystroke dynamics features, which do not contain any sensitive information, also locally, and finally transfer only the keystroke dynamics features to dedicated servers.

VI. CONCLUSION

The native language is one of the acquired characteristics of a person. Knowing this characteristic for unknown Internet users is a valuable tool for safety, convenience, and commercial reasons. In this work, a method is proposed for revealing the native language of a completely unknown user, through keystroke dynamics. That is, using data derived from the way he/she types.

For the purposes of research, a new dataset was created, consisting of 194 logfiles, by recording volunteers from five native languages during daily use of their computer. From the dataset, the most widely used keystroke dynamics features, namely keystroke durations and digram latencies, were extracted, and then a procedure was followed to select the appropriate features to reduce the dimensionality of the problem. In the experiment stage, several machine learning models were tested and five of them performed well. Specifically, the SVM, SL, NB, BNC, and RBFN models, of which the first had the highest accuracy. Finally, using

SVM as a base classifier in meta-algorithms for boosting the performance, an accuracy of 90.7% was achieved.

This is quite a promising achievement for creating systems that will reveal a user's native language from data derived from the way they type. The reason for creating such systems is for example to facilitate a forensic investigation when trying to create the profile of the culprit, or for example to improve targeted advertising.

The future goals for extending the research are, firstly, the expansion of the dataset by recording more volunteers from more native languages, with the help of which questions such as for example what happens to languages that belong to the same language family (e.g., English and German, or Spanish and Italian) will be answered, secondly, the exploitation of other keystroke dynamics features, thirdly, the use of more up-to-date machine learning models, and fourthly, the investigation of the relationship between native languages and the use of digrams.

REFERENCES

[1] P. Vashisth and K. Meehan, "Gender classification using Twitter text data," in *Proc. 31st Irish Signals Syst. Conf. (ISSC)*, Letterkenny, Ireland, Jun. 2020, pp. 1–6, doi: 10.1109/ISSC49989.2020.9180161.

[2] K. Shekhawat and D. P. Bhatt, "Recent advances and applications of keystroke dynamics," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE)*, Dubai, United Arab Emirates, Dec. 2019, pp. 680–683, doi: 10.1109/ICCIKE47802.2019.9004312.

[3] I. Tsimperidis, D. Grunova, S. Roy, and L. Moussiades, "Keystroke dynamics as a language profiling tool: Identifying mother tongue of unknown internet users," *Telecom*, vol. 4, no. 3, pp. 369–377, Jul. 2023, doi: 10.3390/telecom4030021.

[4] N. Raul, R. Shankarmani, and P. Joshi, "A comprehensive review of keystroke dynamics-based authentication mechanism," in *Proc. Int. Conf. Innov. Comput. Commun.*, in Advances in Intelligent Systems and Computing, vol. 1059, A. Khanna, D. Gupta, S. Bhattacharyya, V. Snasel, J. Platos, and A. E. Hassanien, Eds. Singapore: Springer, 2020, pp. 149–162, doi: 10.1007/978-981-15-0324-5_13.

[5] A. Alsultan, K. Warwick, and H. Wei, "Non-conventional keystroke dynamics for user authentication," *Pattern Recognit. Lett.*, vol. 89, pp. 53–59, Apr. 2017, doi: 10.1016/j.patrec.2017.02.010.

[6] E.-S.-M. El-Kenawy, S. Mirjalili, A. A. Abdelhamid, A. Ibrahim, N. Khodadadi, and M. M. Eid, "Meta-heuristic optimization and keystroke dynamics for authentication of smartphone users," *Mathematics*, vol. 10, no. 16, p. 2912, Aug. 2022, doi: 10.3390/math10162912.

[7] B. Ayotte, M. Banavar, D. Hou, and S. Schuckers, "Fast free-text authentication via instance-based keystroke dynamics," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 377–387, Oct. 2020, doi: 10.1109/TBIOM.2020.3003988.

[8] J. Li, H.-C. Chang, and M. Stamp, "Free-text keystroke dynamics for user authentication," in *Artificial Intelligence for Cybersecurity* (Advances in Information Security), vol. 54, M. Stamp, C. A. Visaggio, F. Mercaldo, and F. Di Troia, Eds. Cham, Switzerland: Springer, 2022, pp. 357–380, doi: 10.1007/978-3-030-97087-1_15.

[9] S. A. Alsuhibany and A. S. Almuqbil, "Impact of using different-sized touch keyboards on free-text keystroke dynamics authentication in the Arabic language," *Sci. Rep.*, vol. 12, no. 1, p. 15866, Sep. 2022, doi: 10.1038/s41598-022-20099-6.

[10] I. Tsimperidis, P. D. Yoo, K. Taha, A. Mylonas, and V. Katos, "R2BN: An adaptive model for keystroke-dynamics-based educational level classification," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 525–535, Feb. 2020, doi: 10.1109/TCYB.2018.2869658.

[11] T. M. L. Nascimento, A. V. M. Oliveira, L. E. A. S. Santana, and M. Da Costa-Abreu, "Investigating the use of feature selection techniques for gender prediction systems based on keystroke dynamics," in *Proc. 11th Int. Conf. Pattern Recognit. Syst. (ICPRS)*, Mar. 2021, pp. 115–120, doi: 10.1049/icp.2021.1446.

[12] L. Cascone, M. Nappi, F. Narducci, and C. Pero, "Touch keystroke dynamics for demographic classification," *Pattern Recognit. Lett.*, vol. 158, pp. 63–70, Jun. 2022, doi: 10.1016/j.patrec.2022.04.023.

[13] I. Tsimperidis, S. Rostami, K. Wilson, and V. Katos, "User attribution through keystroke dynamics-based author age estimation," in *Proc. 12th Int. Netw. Conf.*, in Lecture Notes in Computer Science, vol. 180, B. Ghita and S. Shiaeles, Eds. Cham, Switzerland: Springer, 2021, pp. 47–61, doi: 10.1007/978-3-030-64758-2_4.

[14] C. Sahu, M. Banavar, and S. Schuckers, "A novel distance-based algorithm for multi-user classification in keystroke dynamics," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2020, pp. 63–67, doi: 10.1109/IEEECONF51394.2020.9443407.

[15] T. Arroyo-Gallego, A. A. Ayala, A. Morales, R. Vera-Rodriguez, J. Fierrez, and I. Mondesire-Crump, "Evaluating keystroke dynamics as a biomarker for mental fatigue detection," Tech. Rep., May 2022, doi: 10.21203/rs.3.rs-1580509/v1.

[16] H. López-Carral, D. Santos-Pata, R. Zucca, and P. F. M. J. Verschure, "How you type is what you type: Keystroke dynamics correlate with affective content," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 1–5, doi: 10.1109/ACII.2019.8925460.

[17] H. Alfalahi, A. H. Khandoker, N. Chowdhury, D. Iakovakis, S. B. Dias, K. R. Chaudhuri, and L. J. Hadjileontiadis, "Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: A systematic review and meta-analysis," *Sci. Rep.*, vol. 12, no. 1, p. 7690, May 2022, doi: 10.1038/s41598-022-11865-7.

[18] S. Tripathi, T. Arroyo-Gallego, and L. Giancardo, "Keystroke-dynamics for Parkinson's disease signs detection in an at-home uncontrolled population: A new benchmark and method," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 1, pp. 182–192, Jan. 2023, doi: 10.1109/TBME.2022.3187309.

[19] O. Buckley, D. Hodges, J. Windle, and S. Earl, "CLICKA: Collecting and leveraging identity cues with keystroke dynamics," *Comput. Secur.*, vol. 120, Sep. 2022, Art. no. 102780, doi: 10.1016/j.cose.2022.102780.

[20] P. Bours and S. Brahmanpally, "Language dependent challenge-based keystroke dynamics," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2017, pp. 1–6, doi: 10.1109/CCST.2017.8167838.

[21] R. Shadman, A. A. Wahab, M. Manno, M. Lukaszewski, D. Hou, and F. Hussain, "Keystroke dynamics: Concepts, techniques, and applications," 2023, *arXiv:2303.04605*.

[22] J. Klein, B. Joseph, and M. Fritz, *Handbook of Comparative and Historical Indo-European Linguistics*. Berlin, Germany: De Gruyter, 2017, doi: 10.1515/9783110261288.

[23] M. Videnov, "The present-day bulgarian language situation: Trends and prospects," *Int. J. Sociol. Lang.*, vol. 135, no. 1, pp. 1–10, 1999, doi: 10.1515/ijsl.1999.135.11.

[24] P. S. Rao, "The role of English as a global language," *Res. J. English*, vol. 4, no. 1, pp. 65–79, 2019.

[25] F. R. Adrados, *A History of the Greek Language: From Its Origins to the Present*. Leiden, The Netherlands: Brill, 2005.

[26] L. Johanson and É. Á. Csató, *The Turkic Languages* (Routledge Language Family Series), 2nd ed. London, U.K.: Routledge, 2022.

[27] R. Moskovitch, C. Feher, A. Messerman, N. Kirschnick, T. Mustafic, A. Camtepe, and B. Lohlein, "Identity theft, computers and behavioral biometrics," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, Richardson, TX, USA, 2009, pp. 155–160, doi: 10.1109/ISI.2009.5137288.

[28] S. Vora and H. Yang, "A comprehensive study of eleven feature selection algorithms and their impact on text classification," in *Proc. Comput. Conf.*, Jul. 2017, pp. 440–449, doi: 10.1109/SAI.2017.8252136.

[29] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, Mar. 2019, doi: 10.2478/cait-2019-0001.

[30] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.

[31] K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, and M. Sebban, "From cost-sensitive classification to tight F-measure bounds," in *Proc. 22nd Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 89, no. 1. Naha, Japan, Apr. 2019, pp. 1245–1253.

[32] A. C. J. W. Janssens and F. K. Martens, "Reflection on modern methods: Revisiting the area under the ROC curve," *Int. J. Epidemiol.*, vol. 49, no. 4, pp. 1397–1403, Aug. 2020, doi: 10.1093/ije/dyz274.

[33] T.-H. Lee, A. Ullah, and R. Wang, "Bootstrap aggregating and random forest," in *Macroeconomic Forecasting in the Era of Big Data*, vol. 52, P. Fuleky, Ed. Cham, Switzerland: Springer, 2020, pp. 389–429, doi: 10.1007/978-3-030-31150-6_13.

[34] J. Kim, H. Kim, and P. Kang, "Keystroke dynamics-based user authentication using freely typed text based on user-adaptive feature extraction and novelty detection," *Appl. Soft Comput.*, vol. 62, pp. 1077–1087, Jan. 2018, doi: 10.1016/j.asoc.2017.09.045.

[35] J. Kim and P. Kang, "Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107556, doi: 10.1016/j.patcog.2020.107556.

[36] C. Goutte, S. Léger, and M. Carpuat, "The NRC system for discriminating similar languages," in *Proc. 1st Workshop Applying NLP Tools Similar Lang., Varieties Dialects*, Dublin, Ireland, 2014, pp. 139–145, doi: 10.3115/v1/W14-5316.

IOANNIS TSIMPERIDIS received the bachelor's degree from the Department of Electrical and Computer Engineering, AUTH, in 1997, and the master's and Ph.D. degrees from the Department of Electrical and Computer Engineering, DUTH, in 2002 and 2017, respectively. Until September 2004, he was an executive in industry. From September 2006 to September 2021, he was a permanent secondary education teacher. Since October 2021, he has been the Laboratory Teaching Staff with the Computer Science Department, IHU. His research interests include keystroke dynamics, user classification, machine learning, and data mining.

DENITSA GRUNOVA received the degree in computer science from International Hellenic University. Her educational journey with university provided her with a strong foundation in computer science, equipping her with knowledge in programming languages, data structures, algorithms, and various other core concepts. During her time with International Hellenic University, she actively participated in projects and practical assignments that enhanced her problem-solving and critical thinking skills.

GEORGE A. PAPAKOSTAS received the Diploma, M.Sc., and Ph.D. degrees in electrical and computer engineering from the Democritus University of Thrace (DUTH), Greece, in 1999, 2002, and 2007, respectively.

He has 15 years of experience in large-scale systems design as a senior software engineer and the technical manager. He is the Head of the Machine Learning and Vision (MLV) Research Group. He is also a tenured Full Professor with the Department of Computer Science, International Hellenic University, Greece. He has (co)authored more than 220 publications in indexed journals, international conferences, book chapters, one book (in Greek), two edited books, and six journal special issues. His publications have over 3600 citations with an H-index of 34 (Google Scholar). His research interests include machine learning, computer/machine vision, pattern recognition, and computational intelligence.

Dr. Papakostas is a member of ACM, IAENG, MIR Laboratories, EUCogIII, and the Technical Chamber of Greece (TEE). He is a reviewer of numerous journals and conferences.