Comparison of the impact of input data diversity on user recognition in a behavioral-biometric system powered by machine learning method

Czuba Krzysztof, Matys Erwin, Skowroński Piotr, Stefański Oliwier

[1] Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44-100 Gliwice, Poland

Abstract. The article deals with the issue of comparing the impact of the diversity of input data on the recognition of users in a biometric system based on the machine learning method. Researchers conducted the analysis with two types of behavioral data in mind, such as the user's keyboard behavior when entering identical strings of characters and when entering different strings of characters. The aim of the research was to compare how the variation in this data affects the accuracy of recognizing users.

Experimental results showed that the diversity of input data plays a key role in the accuracy of the biometric system. Researchers have noticed that the use of different types of behavioral data can significantly increase the effectiveness of identifying users.

In addition, the article also presents a comparison of different machine learning methods for keystroke dynamics systems. Researchers tested different algorithms and models to determine which were more effective at recognizing users based on a variety of behavioral data.

The conclusion of the research is that the introduction of a greater variety of input data and the use of appropriate machine learning methods can significantly improve the effectiveness of biometric systems based on user behavior.

Keywords: Biometrics, Behavior, Key-stroke Dynamics, Authentication Algorithms; Security; KNN Classifier; Machine Learning.

1 Introduction

Along with the development of computer science and computing methods, systems have been created to automate the work of processing huge data sets. One of such systems are machine learning methods. This paper discusses the use of such a method to program a security system based on the concept of biometric security in the form of behavioral verification, where the element verifying the user is the way he uses the device's keyboard.

The work, however, is not a mere presentation of such a system, but deals with a less popular topic among keystroke dynamics research. During the research and study of the topic itself, it was noticed that most scientific articles describe experiments performed on a research sample that assumes testing while entering a constant, unchanging string of characters, which, however, may affect the quality of the results obtained or their narrowing down to the concept of assumptions. In view of this state of affairs, i.e. the deficit of research works devoted to the behavioral science of keystroke dynamics based on more real scenarios, i.e. on introducing variable strings of characters composed of words familiar to the examined person, the idea of this work was born. The experiment described here compares the number of user authentications that pass or fail when entering a fixed string to the number of authentications that pass and fail when entering a variable string. The exact research process, assumptions and data are described later in the article.

Another reason for this work is the need for the development of behavioral verification methods. The method based on keystroke dynamics is a fresh method, but it has an extremely large potential, because its implementation is simple, so it easily allows you to enrich the security system, which is extremely needed at the present time, while the role of cyber security is constantly increasing in the face of crises and increased digital threats, and in the face of a future that may be dominated by quantum computers as their high computing power poses a risk to the currently used security systems. However, the relationship between behavioral security and the potential of quantum computing is a topic that requires further research and scientific work.

2 The concept of using biometric and behavioral security

2.1 Biometric security systems

Biometrics is a science dealing with the measurement of living beings, mainly to control access to information systems. Biometrics is based on the use of biological or behavioral characteristics of a person, such as the anatomy of the iris, fingerprints, geometry of facial features, geometry of veins or even eye color. These features are completely unique to each person.[1]

The first attempts to use biometrics in IT security took place in the 1960s. However, only in recent years has biometrics become a popular tool in IT security. Thanks to it, it is possible to replace traditional authentication methods, such as passwords or magnetic cards.

In **the General Data Protection Regulation (GDPR)** we find the following definition:

- **Biometric data** - personal data that result from special technical processing, relate to the physical, physiological or behavioral characteristics of a natural person and enable or confirm the unambiguous identification of that person, such as facial image or dactyloscopic data. Art. 3 point 14 GDPR.

Systems using security based on the analysis of biometric features consist of:

- "Benchmark" - data representing the biometric measures of an enrollee, extracted from the enrollee's biometric sample, typically stored in the biometric system and used by the biometric system to verify compliance with subsequent submissions of matching benchmarks.
- "Raw Biometrics" - is raw digital biometric data collected from a measuring device (e.g. fingerprint image or audio stream), suitable for subsequent processing in order to create a biometric sample or pattern.
- "Match pattern" - data representing biometric measures of a person, extracted from a biometric sample for comparison with reference patterns.

Advantages:

- Speed and Convenience: Biometrics allows you to quickly and easily confirm a person's identity, without having to enter passwords or codes
- Security: Biometrics can provide a higher level of security than traditional authentication methods such as passwords or PINs.
- Uniqueness: Biometric features are unique to each individual and cannot be copied or counterfeited.

Disadvantages:

- Costs: Biometric systems are typically more expensive than traditional authentication methods.
- Errors: Biometrics are not perfect and can lead to identification errors.
- Privacy: Biometrics require the storage of personal data, which may pose a threat to users' privacy.

2.2 Behavioral security systems

Behavioral characteristics are a type of biometric characteristics that are directly related to human behavior.[2] This fact immediately indicates that these characteristics may change over time as a result of development or experience. Examples of characteristics are the voice, the way we sign, the way we type on the keyboard, the places we tend to visit, or even the way our brain reacts.

The advantages of using security systems based mainly on behavioral features are:

- Higher resistance to fraud attempts and data falsification.
- Ability to identify the user in a discreet and non-invasive way.
- No need to have specialized equipment for collecting biometric data.
- Ability to easily update biometric data.

The disadvantages of using security systems based mainly on behavioral characteristics are:

- High cost of implementation and maintenance.
- Poor performance for some users (e.g. elderly or disabled people).
- Sensitivity to changes in user behavior (e.g. illnesses, stress).

2.3 The concept of uniqueness of keystroke dynamics

Keystroke dynamics are not expected to be perfectly unique to each individual since there are likely to be similarities between individuals' typing style, particularly on mobile devices, but it is known to be sufficiently different between users to be useful enough as a method of verifying a user's identity, as is evidenced by the low error rates. **Keystroke dynamics has the potential to be used as an authenticator but it is not powerful enough to be used as an identifier.**[1]

In the experiment described in this paper, the similarity of typing styles of some users as well as differences when using different keyboards or devices by the same user were taken into account. However, due to the similarity of keystroke dynamics for some users, a number of incorrect authentications are a minority in the statistics obtained.

3 Machine learning method selection

3.1 Popular dataset selection

In order to create our own behavioral authentication system based on keystroke dynamics, it was necessary to choose a machine learning method that compares the given data in order to give a verdict.

It was decided that the selection would be based on the result of comparing the quality of the application of many different methods to a data set that would be identical or similar in content and data schema to the set on which the system used in the experiment would operate. Such a dataset turned out to be a publicly available dataset from *Kaagle*.

As we can read from **author’s description of the database**:

The data consist of keystroke-timing information from 51 subjects (typists), each typing a password (.tie5Roanl) 400 times.

The data are arranged as a table with 34 columns. Each row of data corresponds to the timing information for a single repetition of the password by a single subject. The first column, subject, is a unique identifier for each subject (e.g., s002 or s057). Even though the data set contains 51 subjects, the identifiers do not range from s001 to s051; subjects have been assigned unique IDs across a range of keystroke experiments, and not every subject participated in every experiment. For instance, Subject 1 did not perform the password typing task and so s001 does not appear in the data set. The second column, sessionIndex, is the session in which the password was typed (ranging from 1 to 8). The third column, rep, is the repetition of the password within the session (ranging from 1 to 50).

The remaining 31 columns present the timing information for the password. The name of the column encodes the type of timing information. Column names of the form H.key designate a hold time for the named key (i.e., the time from when key was pressed to when it was released). Column names of the form DD.key1.key2 designate a keydown-keydown time for the named digraph (i.e., the time from when key1 was pressed to when key2 was pressed). Column names of the form UD.key1.key2 designate a keyup-keydown time for the named digraph (i.e., the time from when key1 was released to when key2 was pressed). Note that UD times can be negative, and that H times and UD times add up to DD times.

In the set, we will find data related to the time of using keyboard keys. The times are based on the intervals from pressing a key to its release, from pressing to pressing the next one, from releasing to pressing and pressing the next one to releasing it. The relationship of these time values is called *dynamics*.

3.2 Datasets testing and results for particular machine learning methods

A test of each method was carried out, where the main result data that are taken into account is the quality of the method, i.e. the success of the authentication attempt, and the time, i.e. the duration of testing the entire set.

In order to make sure that the methods work, the study was also conducted in parallel on popular online databases used to learn machine learning methods. These databases were databases with species of irises, with wine species, with data on breast cancer or data on recognizing digits by their shape.

Fig. 1. Accuracy comparision of considered methods

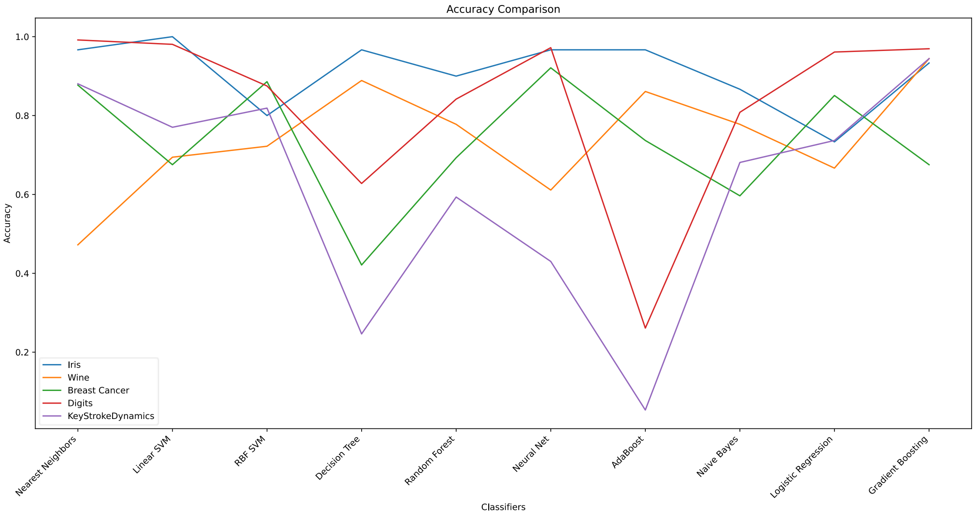
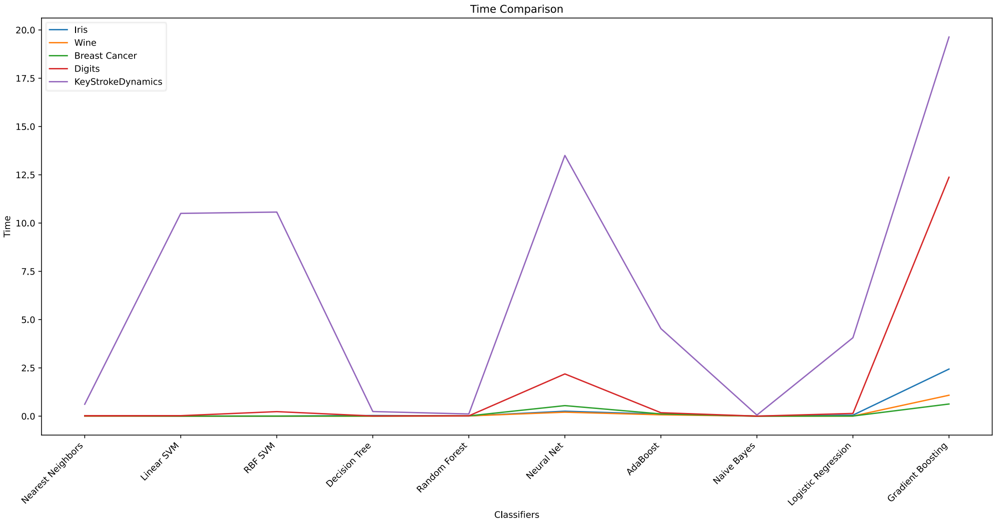


Fig. 2. Time comparison of considered methods



When compiling the obtained results, the KNN method was chosen, which is one of the simpler machine learning methods.

4 KNN characteristics

The k-nearest neighbors method is a machine learning algorithm that is a non-linear and unsupervised classifier. This algorithm is used to classify data against similar data. In both cases, the input is the k closest training examples in dataset.[3]

In the case of classification, the k-nearest neighbors algorithm assigns a new point to the class that is most often represented by the k nearest training points. In the case of regression, the k-nearest neighbors algorithm assigns a new point with the average of the values of the k nearest training points.

The k-nearest neighbor algorithm is one of the simplest machine learning algorithms and is often used in practice due to its simplicity and efficiency.

The k-nearest neighbors (KNN) algorithm uses “similarity of features” to predict the values of new data points. The new data point will be assigned a value based on how closely it matches the points in training set.

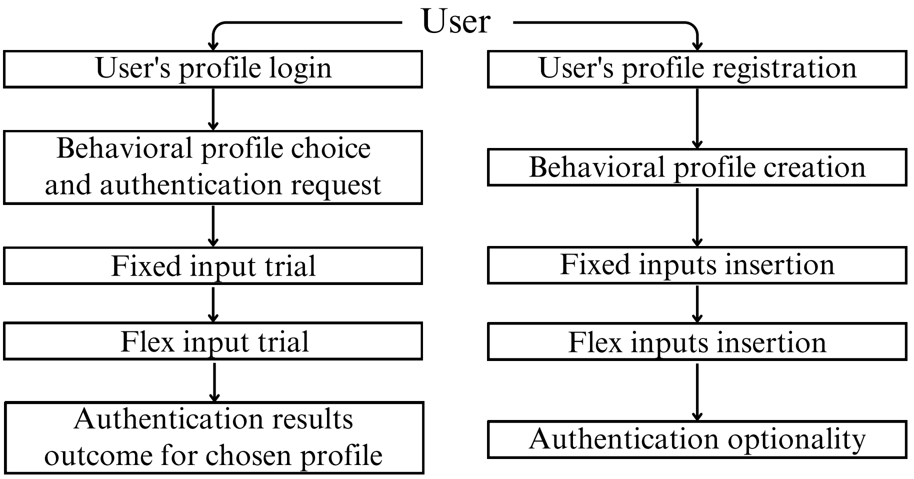
In classification, the K-nearest neighbor algorithm basically comes down to creating a majority of votes between the K most similar cases to a given "invisible" observation. Similarity is defined by the distance metric between two data points. A popular method is the Euclidean or Manhattan distance method. We have used the second one.

5 Original dataset for experimental use

5.1 Schema of the data acquisition application

In order to conduct the experiment, it was decided to prepare an original database consisting of data collected during the use of a special application that allows forging a behavioral profile operating on keystroke dynamics, and then trying to authenticate as one of the available, previously generated profiles. The application made it possible to attempt to authenticate the user in relation to his own profile and the profile of other testers.

Fig. 3. The schema of the application


User
User's profile login
User's profile registration
Behavioral profile choice and authentication request
Behavioral profile creation
Fixed input trial
Fixed inputs insertion
Flex input trial
Flex inputs insertion
Authentication results outcome for chosen profile
Authentication optionality

The registration phase consisted of the user entering 20 identical strings of characters in the form of a word in Polish with a length of 10. Then the user enters 20 Polish words with a length of 10 in the number of 20.

During this time, the application measures the time values of keystroke dynamics. The biometric profile consists of time values placed in geometric space.

After logging in, the user selects the profile as which he would like to authenticate. In this part of the program, using the keyboard, he types a single 10-letter word, identical to the fixed word that was typed in the first part of the behavioral profile creation. Then the user enters a single flex input, randomly selected from a set of inputs composed entirely of Polish-language words of identical length. Polish words were selected due to the fact that the study was conducted entirely on people who use Polish on a daily basis. This involved providing users with strings that are composed of linguistic fragments that these users use naturally. This naturalness in the use of the language translates into a naturalness in the use of the keyboard, which helps to reproduce the natural conditions necessary for the correct collection of keystroke dynamics data.

Originally, as part of the experiment, strings of characters composed of random characters were used, which in no way reflected the natural words of any language, which made users use the keyboard in a non-intuitive way, which could affect the quality of the downloaded data.

As part of standardizing data and giving the user a chance to get used to entering words to ensure behavioral authenticity, during registration and creating a biometric profile, the user enters 10 of the same words and then 10 random ones.

In the section related to the authentication request, using the k nearest neighbors machine learning method, the system compares the set time data with the data in the database and, starting from the set data, searches for the 9 geometrically closest other data. If most of the adjacent data belongs to the authenticated profile, the system returns information about the positive consideration of the request. If a major part of the data that does not belong to the selected profile is found, the system rejects the user's request.

5.2 Dataset characteristics

When acquiring data, the application places them in a database, which is then read as a geometric space using a machine learning method containing data obtained from each input. Each attempt to type a string creates a time relationship between two consecutive characters. The time of pressing the first key, the time from pressing the first key to pressing the second key, the time between releasing and pressing, the time from pressing the first key to releasing the second key, and the time from releasing the first key to releasing the second key are collected. [4,5]

The data is obtained for each pair of consecutive characters and the identical length of words allows for easy comparison of the work with the keys for each user. For each attempt is assigned the ID of the user logged into the system, which in the next section, using the KNN method, allows you to compare the data from the attempts.

6 Data comparison

On the basis of 202 proper records, i.e. samples obtained from 25 testers, four types of authentication attempts were selected for two different types of inputs (fixed and flex), i.e. a total of 8 types.

- **True positive type** - the user authenticates himself using a profile belonging to himself and receives a positively considered application.
- **False negative type** - the user authenticates himself using a profile belonging to himself, but receives a negatively considered application.
- **False positive type** - the user authenticates to a profile that does not belong to him and receives a positively considered application.
- **True negative type** - the user authenticates himself on a profile that does not belong to him and receives a negatively considered application.

Fig. 4. Results for fixed method

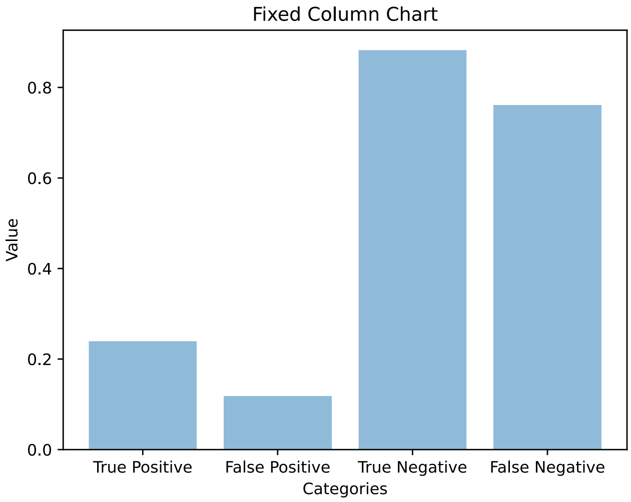
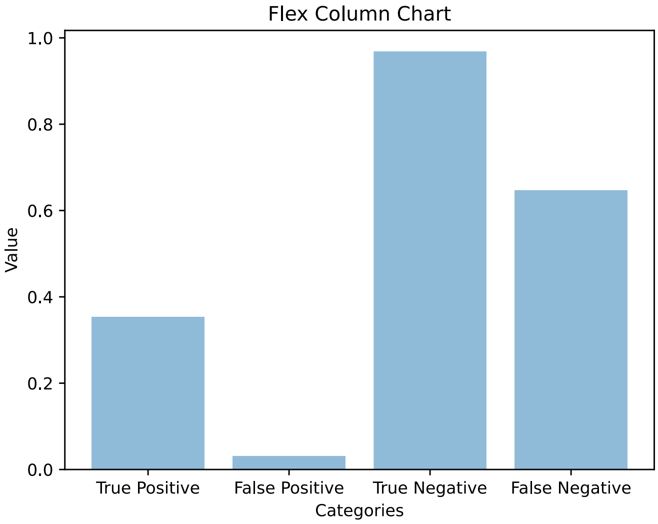


Fig. 5. Results for flex method



7 Conclusion

As we know, we cannot rely solely on a system based on behavioral security, nor is such a system used to identify you. Therefore, the most important results for us are those that belong to the false positive and true negative types. As we can see, the flex method, taking into account much more diverse data, performs better than the fixed method using the KNN classifier.

However, we must also take into account the high percentage of rejections, i.e. the type of false negative. Perhaps during using a different classifier or using a larger dataset this result would be lower.

In the context of the development of biometric and behavioral methods, it is also worth looking at issues related to cybernetics and the analysis of the human being. As part of further research, the database can be extended with tables related to the natural and everyday behavior of the surveyed people along with their psychological profiles (the context of personality, temperament, felt emotions) and their environmental profiles.

Collecting so much information would certainly be a long-term process, and their analysis would be a big-data issue, and the conducted experiment would have to be supported by expert knowledge in the field of behavioral psychology and even sociology. However, such a large database and taking into account all possible factors affecting the success of authentication could give many interesting conclusions directly affecting the development of security methods and better understanding of the human as the central unit in the cybernetic security system.

References

1. Crawford, Heather. "Keystroke dynamics: Characteristics and opportunities." *2010 Eighth International Conference on Privacy, Security and Trust*. IEEE, 2010.
2. Bergadano, Francesco, Daniele Gunetti, and Claudia Picardi. "User authentication through keystroke dynamics." *ACM Transactions on Information and System Security (TISSEC)* 5.4 (2002): 367-397.
3. Hidiyanto, Fitra, and Abdul Halim. "Knn methods with varied k, distance and training data to disaggregate nilm with similar load characteristic." *Proceedings of the 3rd Asia Pacific Conference on Research in Industrial and Systems Engineering*. 2020.
4. Ilonen, Jarmo. "Keystroke dynamics." *Advanced Topics in Information Processing–Lecture* (2003): 03-04.
5. Monrose, Fabian, and Aviel Rubin. "Authentication via keystroke dynamics." *Proceedings of the 4th ACM Conference on Computer and Communications Security*. 1997.
6. Roy Maxion – Keystroke Dynamics – Benchmark dataset