DOI: https://doi.org/10.15276/aait.09.2026.30
UDC 004.93

Keystroke dynamics recognition using cluster-based prediction ellipses

Sergiy B. Prykhodko[1)],
ORCID: https://orcid.org/0000-0002-2325-018X; sergiy.prykhodko@nuos.edu.ua. Scopus Author ID: 55225622100
Artem S. Trukhov[1)]
ORCID: https://orcid.org/0000-0002-7160-8609; artem.trukhov@gmail.com
[1)] Admiral Makarov National University of Shipbuilding, 9, Heroes of Ukraine Ave. Mykolaiv, 54007, Ukraine

ABSTRACT

Keystroke dynamics recognition is a biometric approach that enables user identification based on individual typing patterns. The **relevance** of the study is determined by the need to improve the accuracy of keystroke dynamics recognition under conditions of variability, heterogeneity, and deviation of data from a multivariate normal distribution. The object of this research is the keystroke dynamics recognition process, while the subject of the research is a mathematical model based on cluster-based prediction ellipses constructed in a two-variate feature space. **The purpose of the article** is to improve the accuracy of keystroke dynamics recognition by applying cluster-based prediction ellipses that take local data structures into account. The **objectives** of the study are to form cluster-based prediction ellipses constructed on local training samples and to compare the proposed approach with other one-class recognition methods. The **methods** of the study include Mardia's test for assessing multivariate normality, the Box-Cox transformation for data normalization, k-means clustering, the elbow method and cluster size analysis, the construction of prediction ellipses based on the squared Mahalanobis distance, and robust covariance estimation based on the minimum covariance determinant. The **scientific novelty** consists in developing a cluster-based approach to constructing local prediction ellipses for keystroke dynamics recognition, which makes it possible to account for data heterogeneity and form more precise decision boundaries. The **practical significance** of the study lies in the possibility of improving the reliability of biometric user authentication systems without switching to more complex multiclass models. The **results** of the study demonstrate that the proposed approach significantly outperforms the one-class support vector machine and the prediction ellipse constructed for non-Gaussian data, while also providing better recognition of samples belonging to another class and recognition accuracy comparable to the prediction ellipse for normalized data. This improvement is achieved through adaptive modeling of local data distributions and more precise decision boundaries in the feature space. The **conclusions** confirm the effectiveness of clustering as a preprocessing stage for keystroke dynamics recognition based on cluster-based prediction ellipses and demonstrate the suitability of adaptive local modeling for heterogeneous keystroke dynamics data. Further research may focus on the application of alternative clustering methods, as well as on the use of advanced data normalization methods, such as the Johnson transformation. In addition, extending the feature space to a higher dimensionality by incorporating a larger number of keystroke dynamics characteristics may further improve recognition accuracy.

Keywords: Keystroke dynamics recognition; two-variate data; normalizing transformation; clustering; prediction ellipse; Box-Cox transformation

***For citation*: Prykhodko S. B., Trukhov A. S. "Keystroke dynamics recognition based on cluster-oriented prediction ellipses". *Applied Aspects of Information Technology*. 2026; Vol.9 No.3: 451–464. DOI: https://doi.org/10.15276/aait.09.2026.30**

INTRODUCTION

Keystroke dynamics recognition is a behavioral biometric method used to identify users based on their typing behavior. Because it is non-intrusive and does not require additional hardware, this approach is widely studied in information security, continuous authentication, and user verification systems [1]. However, the development of accurate and reliable recognition models remains challenging due to the variability of typing patterns and the complexity of keystroke data. Keystroke dynamics data are highly diverse and can change under the influence of many factors, such as the user's physical condition, mood, environmental conditions, and others [2], [3]. Therefore, typing behavior cannot be treated as statistically uniform.

At the same time, many statistical recognition models are based on the assumption of a multivariate normal distribution [4]. In practice, keystroke dynamics data often deviate from this assumption [5], which reduces the effectiveness of global statistical models and leads to lower recognition accuracy.

A common approach in biometric pattern recognition is the use of decision rules based on prediction ellipsoids constructed [6], [7]. These methods provide clear and interpretable decision boundaries in a feature space. However, their performance strongly depends on how well the data satisfy the assumed distribution. When normality is violated, the probability of recognition is significantly reduced.

To address these limitations, this paper explores normalizing transformations [7], [8] and a cluster-based approach [9] that adapts the construction of

prediction ellipses to local data structures. By dividing the keystroke feature space into clusters, it becomes possible to describe local statistical properties more accurately and to construct separate prediction ellipses for each cluster. This approach allows the recognition model to better capture intra-class variability and reduces the negative impact of data heterogeneity.

The object of this research is the keystroke dynamics recognition process. The subject of the research is a mathematical model based on cluster-based prediction ellipses constructed in a two-variate feature space. The purpose of this work is to improve recognition accuracy by adapting decision boundaries to local distributions of keystroke features using cluster-based prediction ellipses.

LITERATURE REVIEW AND PROBLEM STATEMENT

Keystroke dynamics recognition has been actively studied as a behavioral biometric method for user identification and authentication. Existing approaches can be broadly divided into statistical methods and machine learning techniques. Statistical approaches often rely on probabilistic models and distance-based decision rules, while machine learning methods apply classifiers such as support vector machines [10], [11], neural networks [12], [13], autoencoders [14], GANs [15], and others [16]. Although many of these methods demonstrate good performance, their effectiveness strongly depends on data quality and feature selection.

A large group of statistical recognition methods is based on the construction of prediction regions under the assumption of normal data distribution [6], [7]. In the case of two-variate feature spaces, prediction ellipses are commonly used as decision boundaries. The construction of a prediction ellipse is based on the squared Mahalanobis distance (SMD), which measures the distance between a data point and the mean of the distribution while taking feature correlation into account. Under the assumption of bivariate normality, the probability that a data point lies inside the prediction ellipse is equal to the probability that the squared Mahalanobis distance does not exceed a certain threshold [17], [18].

This threshold is determined by the quantile of the Chi-square distribution for a given confidence level. The squared Mahalanobis distance follows a Chi-square distribution with k degrees of freedom, where k is the number of features [18], [19]. For the two-dimensional case considered in this study, k=2.

Therefore, the prediction ellipse can be defined as the set of points satisfying the following inequality:

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})^T \boldsymbol{S}_X^{-1} (\boldsymbol{X} - \bar{\boldsymbol{X}}) = \chi^2_{p,\alpha}, \quad (1)$$

where

$$\mathrm{S_X} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{X}_i - \overline{\mathrm{X}}\right)\left(\mathrm{X}_i - \overline{\mathrm{X}}\right)^T .$$

These methods are attractive due to their mathematical interpretability and relatively low computational complexity. However, real keystroke dynamics data often deviate from normality, which limits the applicability of such models when they are built on non-normalized data.

To address this issue, various normalization transformations have been proposed to bring keystroke dynamics data closer to a normal distribution [6], [7], [20]. In particular, previous studies have shown that prediction ellipsoids for normalized data significantly outperform those based on raw features. In the conference paper (https://ceur-ws.org/Vol-3933/Paper_5.pdf), it was demonstrated that a nine-variate prediction ellipsoid built for normalized keystroke data achieved higher recognition accuracy than machine learning models. This result highlights the potential of statistically grounded approaches when appropriate data normalization is applied.

Data normalization techniques used in keystroke dynamics recognition can generally be divided into univariate and multivariate transformations. Univariate transformations are easier to apply but do not take into account the correlations between features, which may limit their effectiveness. Common examples include the univariate logarithmic transformation and the univariate Box-Cox transformation. In contrast, multivariate normalization methods explicitly consider inter-feature dependencies and therefore tend to provide better results, although they are more complex to implement. The use of multivariate normalization allows the statistical properties of the data to be improved more consistently, which is especially important for the construction of prediction regions based on normality assumptions.

Another important characteristic of keystroke dynamics data is their heterogeneity and multimodal structure. Global models that describe the entire feature space with a single set of parameters often fail to capture local variations in typing behavior. To overcome this limitation, clustering techniques have been explored as a way to partition the data into

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

more homogeneous subsets. By grouping similar samples, clustering allows local statistical properties to be modeled more accurately.

Among clustering methods, the k-means algorithm is widely used due to its simplicity and efficiency. However, it is well known that k-means is sensitive to feature scaling, distance measures, and outliers [21], and therefore performs better when applied to normalized data [22]. Normalization not only improves clustering quality but also leads to more compact and well-separated clusters, which is especially important for subsequent statistical modeling [23].

One limitation of cluster-based approaches is the possibility of obtaining clusters with an insufficient number of samples for reliable covariance matrix estimation. In multivariate statistical analysis, the stability of the covariance matrix is strongly dependent on the ratio between the number of observations and the dimensionality of the feature space. Small local sample sizes may lead to unstable covariance estimates, which negatively affect the construction of prediction ellipses and reduce the reliability of classification boundaries. To address this problem, robust statistical techniques such as the Minimum Covariance Determinant (MCD) estimator are commonly used in the literature to obtain more stable covariance estimates under conditions of limited or heterogeneous data [24].

Despite the reported benefits of normalization and clustering, their combined use with prediction ellipses in keystroke dynamics recognition has not been sufficiently investigated. In particular, there is a lack of studies focusing on the construction of cluster-based prediction ellipses in a two-variate feature space using normalized data. Most existing works either apply normalization without adapting decision boundaries locally or use clustering without integrating it into a statistical recognition framework.

The main problem addressed in this study is the limited effectiveness of global prediction ellipse models for keystroke dynamics recognition in the presence of heterogeneous and non-normally distributed data. There is a need for a recognition approach that adapts decision boundaries to local data characteristics while preserving the interpretability and simplicity of statistical models. This work aims to fill this gap by developing and evaluating a keystroke dynamics recognition method based on cluster-based prediction ellipses constructed from cluster-specific training data.

MATERIALS AND METHODS

The experimental evaluation of the proposed approach was conducted using the CMU Keystroke Dynamics Benchmark Dataset (https://www.cs.cmu.edu/~keystroke/), which is a publicly available and widely used dataset in keystroke dynamics research [25], [26]. The dataset contains keystroke timing data collected from 51 users during repeated typing of a fixed password string, ".tie5Roanl". Each participant typed the password multiple times across several sessions, resulting in a large number of samples that capture both short-term and long-term variability in typing behavior.

For each keystroke event, precise timestamps of key press and key release actions are recorded. Based on these timestamps, several temporal features are derived, including key hold times, keydown–keydown intervals, and keyup–keydown intervals. These features reflect both individual key behavior and transitions between consecutive keys, making them suitable for modeling typing dynamics.

To simplify the analysis and enable visual interpretation of the decision boundaries, only two features were selected in this study: X = {H.o, UD.o.a}, where H.o denotes the hold time of the key O, and UD.o.a represents the interval between releasing the key O and pressing the key A.

After constructing the feature vectors, a crucial preprocessing step involves the identification and removal of outliers. In keystroke dynamics data, anomalous samples may appear due to accidental keystrokes, fluctuations in user concentration, environmental factors, or technical inaccuracies during data acquisition. If such observations are retained, they can distort statistical estimates and negatively affect the reliability of subsequent recognition models. Therefore, removing outliers allows the dataset to better represent stable and typical typing behavior.

Outlier detection is performed using the SMD, which evaluates how far each observation deviates from the center of the multivariate distribution while accounting for dependencies between features. It is important to note that this method relies on the assumption that the data approximately follow a multivariate normal distribution. However, raw keystroke dynamics features frequently violate this assumption. To verify whether the normality assumption is satisfied, Mardia's test is applied [27]. This test evaluates multivariate skewness β_1 and kurtosis β_2.

$$\beta_1 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^T \boldsymbol{S}_X^{-1}\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}\right)\right\}^3,$$

$$\beta_2 = \frac{1}{N}\sum_{j=1}^{N}\left\{\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}\right)^T \boldsymbol{S}_X^{-1}\left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}\right)\right\}^2.$$

When Mardia's test indicates significant deviation from multivariate normality, the squared Mahalanobis distance may become unreliable. In such cases, a normalization transformation is required before proceeding with outlier removal and further modeling. The normalization step transforms the original feature vectors into a representation that more closely follows a multivariate normal distribution.

In this study, a multivariate Box-Cox transformation was applied (BCT) [28]:

$$Z_j = x(\lambda_j) = \begin{cases} \left(X_j^{\lambda_j} - 1\right)/\lambda_j, & \lambda_j \neq 0; \\ ln(X_j), & \lambda_j = 0. \end{cases} \quad (2)$$

Unlike univariate transformations, which process each feature independently, the multivariate Box–Cox transformation considers the joint distribution of the features and preserves their correlation structure. After normalization, Mardia's test was applied again to evaluate the effectiveness of the transformation.

After normalization and outlier removal, the keystroke data were partitioned into clusters to capture local statistical structures in the feature space. Clustering was performed using the k-means algorithm, which aims to group the data into clusters such that points within the same cluster are similar to each other, while points in different clusters are as distinct as possible. K-means iteratively assigns points to the nearest cluster centroid and updates centroids based on the mean of points in each cluster, repeating this process until cluster assignments stabilize.

Since k-means relies on distance measures, its performance improves when applied to normalized data, which ensures that each feature contributes comparably to the clustering process and prevents features with larger scales from dominating the results.

The optimal number of clusters was selected using the Elbow Method [29], [30], which examines how the total within-cluster variation changes as the number of clusters increases. Initially, adding more clusters significantly reduces the within-cluster variation, but after a certain point, the improvement slows down. This point of diminishing returns, called the "elbow," is chosen as the optimal number of clusters. This approach provides an objective and interpretable way to determine the number of clusters that adequately represent the local structure of the data without overcomplicating the model.

To ensure statistical reliability of the obtained clusters, an additional cluster size constraint is applied. If at least one cluster contains fewer than $5p$ samples, where p is the dimensionality of the feature space, the number of clusters is iteratively reduced until all clusters satisfy this condition. This requirement is motivated by the fact that robust covariance estimation using the MCD method requires a minimum sample size proportional to the feature dimensionality to ensure reliable estimation of the covariance matrix [26]. This guarantees a sufficient sample size for robust covariance estimation within each cluster.

After clustering the normalized dataset, a separate statistical model is constructed for each cluster in the form of a prediction ellipse based on (1). For a given cluster k, the prediction ellipse is defined based on the mean vector and covariance matrix estimated from the corresponding training samples.

For clusters containing a sufficient number of samples $N_k \geq p + 30$, where p is the dimensionality of the feature space, the standard estimation of the mean vector and covariance matrix is used.

The squared Mahalanobis distance for a feature vector Z with respect to cluster k is given by:

$$E_k = (\boldsymbol{Z} - \bar{\boldsymbol{Z}}_k)^T \boldsymbol{S}_{Zk}^{-1}(\boldsymbol{Z} - \bar{\boldsymbol{Z}}_k) = \chi^2_{p,\alpha}, \quad (3)$$

where $\bar{\boldsymbol{Z}}_k$ is the mean vector of the k-th cluster, $\boldsymbol{S}_{Zk}^{-1}$ is the corresponding covariance matrix, p is the number of features, and α is the selected significance level.

For clusters with a limited number of samples $N_k \leq p + 30$, the standard estimates become unreliable due to insufficient statistical support. In this case, a robust estimation approach based on the MCD is applied. The corresponding cluster representation is then defined using robust estimates $\widehat{\boldsymbol{Z}}_k$ and $\widehat{\boldsymbol{S}}_{Zk}$, the squared Mahalanobis distance is computed as:

$$E_k = \left(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}_k\right)^T \widehat{\boldsymbol{S}}_{Zk}^{-1}\left(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}_k\right) = \chi^2_{p,\alpha}, \quad (4)$$

where $\widehat{\boldsymbol{Z}}_k$ is the mean vector of the k-th cluster estimated using the MCD, $\widehat{\boldsymbol{S}}_{Zk}$ is the corresponding covariance matrix estimated using the MCD.

In addition to cluster-specific models, a global prediction ellipse is constructed using all training samples of the target class. This global model

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

provides an additional constraint that reflects the overall distribution of the data.

The corresponding expression is based on (1) and defined as:

$$E_{global} = (\boldsymbol{Z} - \bar{\boldsymbol{Z}})^T \boldsymbol{S}_Z^{-1} (\boldsymbol{Z} - \bar{\boldsymbol{Z}}) = \chi^2_{p,\alpha}. \quad (5)$$

These ellipses represent the expected range of typical observations for each cluster and serve as local decision regions in the feature space. A test sample is classified as belonging to the target class only if it satisfies both conditions: it lies inside at least one of the cluster-specific prediction ellipses and simultaneously belongs to the global prediction ellipse.

Based on (3), (4), and (5), the decision rule can be expressed as:

$$\exists k \in \{1, \dots, K\}: E_k{}^T \le \chi^2_{p,\alpha} \wedge E_{global} \le \chi^2_{p,\alpha}. \quad (6)$$

In addition to statistical models based on prediction ellipses, a machine learning approach was also considered for comparison. One of the most commonly used methods for one-class recognition is the one-class support vector machine (OCSVM). This method is widely applied in anomaly detection, novelty detection, one-class recognition, and biometric verification tasks, where only samples from a single target class are available during training.

The main idea of OCSVM is to learn a decision boundary that encloses the majority of the target-class samples while separating them from the rest of the feature space.

The decision boundary that OCSVM learns is defined by the following equation [31]:

$$g(x) = \omega^T \varphi(x) - \rho,$$

where ω represents the normal vector of the hyperplane, and ρ is the bias term.

During training, the algorithm constructs a compact region that contains most of the training data and treats samples lying outside this region as anomalies or representatives of another class.

EXPERIMENTS

The data corresponding to subject identifier s036 was randomly selected to represent the target class; it contributed 400 keystroke samples [27]. The mean vector of the target set is $\bar{X}$ = {0.0434, 0.2137}, where the first feature corresponds to key hold time and the second feature corresponds to inter-key latency, and the corresponding covariance matrix is presented in Table 1.

Table 1. **Covariance matrix of the initial sample**

2.833E-05	5.532E-05
5.532E-05	0.0137

Source: **compiled by the authors**

At this stage of the experiment, the main objective was to prepare the target class data for further analysis by identifying and removing outliers. Keystroke dynamics data are known to contain anomalous samples caused by unstable typing behavior, momentary loss of concentration, or external factors. The presence of such outliers can significantly distort statistical estimates and negatively affect the performance of recognition models.

Outlier detection was performed using a method based on the squared Mahalanobis distance. This approach is widely used in multivariate statistical analysis, as it accounts for both feature variances and correlations. However, its correct application relies on the assumption that the analyzed data follow a multivariate normal distribution. Therefore, before applying the SMD-based outlier detection, this assumption was verified for the target class data.

To evaluate the degree of deviation from multivariate normality, the Mardia test was applied to the initial feature sample. The test results indicated a significant deviation from normality. In particular, the test statistic for multivariate skewness $N\beta_1/6$ was 218.4, which exceeds the chi-square distribution quantile of 10.6 for 2 degrees of freedom at a 0.005 significance level. Additionally, the multivariate kurtosis statistic β_2 was 12.91, exceeding the corresponding normal quantile value of 9.03, based on a mean of 8 and a variance of 0.16 at the same significance level. These results confirm that the initial data significantly deviates from multivariate normality, indicating the need for normalization before further analysis.

Thus, the direct application of methods based on the Mahalanobis distance to the initial data is not appropriate. Since the SMD-based method requires approximately normal data, a normalization step was introduced. The target set was normalized using the multivariate Box–Cox transformation (2), which preserves correlations. By maximizing the log-likelihood function, the following parameter estimates were obtained: $\widehat{\lambda_1}$ = 0.0686, $\widehat{\lambda_2}$ = 0.0456.

After normalization, the Mardia test was applied again to the transformed feature vectors of the s036 dataset to reassess normality. In this case, the multivariate skewness statistic $N\beta_1/6$ was 5.22,

which does not exceed the chi-square threshold of 10.6 for 2 degrees of freedom at a 0.005 significance level. The multivariate kurtosis statistic β_2 was 8.81, which is below the corresponding normal quantile value of 9.03, based on a mean of 8 and a variance of 0.16, at a 0.005 significance level.

After applying the multivariate Box–Cox transformation, the data distribution does not deviate from multivariate normality. This means that the method based on the square of the Mahalanobis distance can be used for the normalized feature set. The SMD value was calculated for each normalized feature vector.

The outlier removal process was carried out iteratively. At each iteration, the feature vector with the maximum SMD value was identified. If this value exceeded the critical threshold of 10.6, derived from the chi-square distribution with 2 degrees of freedom at a 0.005 significance level, the corresponding vector was removed from the dataset. After removing one outlier, the normalization procedure was applied again, and the Mahalanobis distances were recalculated for the updated dataset. This iterative process was repeated until all remaining feature vectors had SMD values below the threshold.

The use of an iterative strategy is important, as the presence of outliers affects both the mean vector and the covariance matrix. Removing all anomalous samples at once may lead to incorrect parameter estimates, while gradual removal allows the statistical characteristics of the dataset to stabilize. Table 2 shows the removed outliers, their IDs in the set, and SMD.

Table 2. **Removed anomalies**

No.	SMD	Vector number
1	17.042	82
2	19.527	307
3	15.731	79
4	12.207	178
5	10.698	87
6	11.177	65
7	11.151	179
8	10.736	328

Source: **compiled by the authors**

The set without anomalies was obtained by removing 8 outliers. The mean vector is $\bar{X}$ = {0.0433, 0.2168}, and the corresponding covariance matrix is presented in Table 3.

After the iterative removal of outliers, the resulting dataset of the target class was further examined. Although the exclusion of anomalous samples reduced their influence on the statistical characteristics of the data, it did not fully ensure compliance with the assumption of multivariate normality. Since this assumption remains important for the next stages of the proposed approach, an additional verification step was required.

Table 3. **Covariance matrix of the set without anomalies**

2.728E-05	6.984E-05
6.984E-05	1.348E-02

Source: **compiled by the authors**

To reassess the distributional properties of the cleaned dataset, the Mardia test was applied again to the feature vectors after outlier removal. The obtained results indicate that the data still deviates from a multivariate normal distribution. In particular, the test statistic for multivariate skewness $N\beta_1/6$ = 229.69, which exceeded the corresponding chi-square quantile 10.6 at a significance level of 0.005 with 2 degrees of freedom. Similarly, the multivariate kurtosis statistic β_2 = 12.8, to the quantile of the normal distribution 9.04, defined by a mean of 8 and a variance of 0.16 at the same significance level.

These results indicate that, even after removing outliers, the dataset does not fully satisfy the assumption of multivariate normality. This behavior can be attributed to the internal structure of keystroke dynamics data, which may consist of several stable typing patterns within the same subject.

Due to the observed deviation from multivariate normality, a normalization step was introduced. The dataset was normalized using the multivariate BCT. The parameters of the transformation were estimated using the maximum likelihood approach. The following parameter estimates were obtained: $\widehat{\lambda_1}$ = -0,0122, $\widehat{\lambda_2}$ = -0.3019.

After applying the BCT with components (2), the normalized set has a mean vector $\bar{Z}$ = {-3.2075, -2.185}, and the covariance matrix is presented in Table 4.

Table 4. **Covariance matrix of the normalized set**

0.0155	0.0112
0.0112	0.582

Source: **compiled by the authors**

After applying the multivariate Box-Cox transformation, the distribution of normalized feature vectors became approximately Gaussian, according to the Mardia test. The test statistic for multivariate skewness $N\beta_1/6$ = 2.16, which does not exceed the corresponding chi-square quantile 10.6 at a significance level of 0.005 with 2 degrees of

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

freedom. The multivariate kurtosis statistic β_2 = 7.39, does not exceed the quantile of the normal distribution 9.04, defined by a mean of 8 and a variance of 0.16 at the 0.005 significance level.

Following normalization, the dataset was partitioned into clusters using the k-means algorithm. This method groups feature vectors based on their similarity in the multidimensional feature space and allows the identification of homogeneous subgroups within the data. To determine the optimal number of clusters, the elbow method was applied (Fig. 1). This method evaluates how the intracluster dispersion changes as the number of clusters increases.

The results of this analysis indicated an inflection point at k=7, which was selected as the initial number of clusters for the normalized dataset. However, due to the cluster size constraint, the final number of clusters was reduced to 5, since the smallest cluster contained fewer than 5p samples.

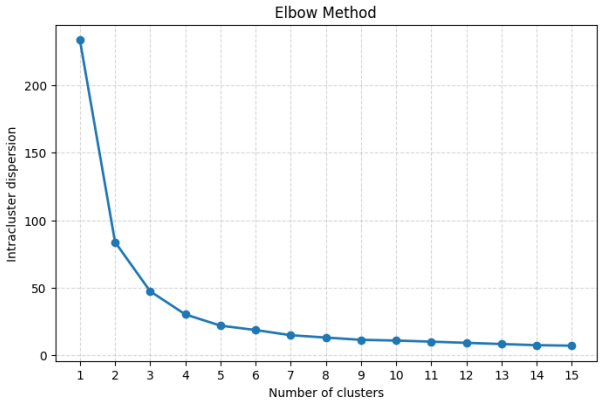

Fig. 1. **Elbow method**
Source: **compiled by the authors**

A representation of the distribution of points by clusters and the prediction ellipse constructed based on the entire set of normalized data is shown in Fig. 2.

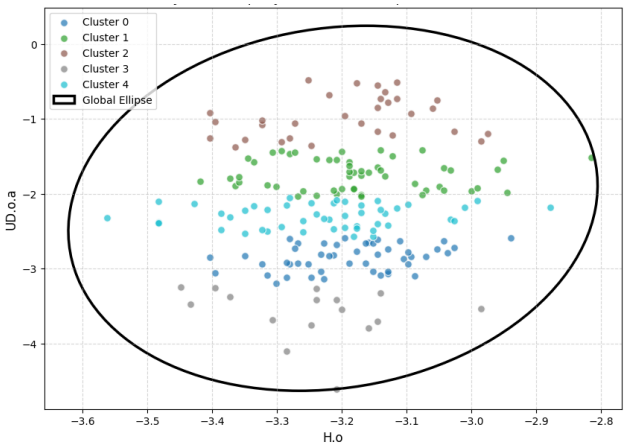

Fig. 2. **Distribution of points by clusters**
Source: **compiled by the authors**

After clustering, the dataset was prepared for training and testing. Instead of performing a global split, the division was carried out independently within each cluster. For every cluster, the feature vectors were randomly shuffled and then split into two equal parts: 50 % of the samples were assigned to the training set, and the remaining 50% were assigned to the test set.

This cluster-based splitting strategy ensures that both the training and test sets contain representative samples from all clusters, preserving the internal structure of the data and accounting for intra-class variability. The training subsets were used to estimate the parameters required for building prediction models, while the test subsets were used exclusively to build the prediction ellipses.

For each cluster-specific training set, a prediction ellipse was constructed using the estimated mean vector and covariance matrix of the corresponding cluster (Fig. 3). Each ellipse defines a confidence region that characterizes the typical typing behavior within that cluster and serves as a local decision boundary in the feature space.

For clusters 2 and 3, the number of samples was below the threshold of p+30, specifically 30 and 15 samples, respectively. Therefore, robust covariance estimation based on the MCD method was applied for these clusters to construct the corresponding ellipses, ensuring stability of the covariance estimates under limited sample conditions.

In addition to the cluster-specific models, a global prediction ellipse was constructed using the entire training dataset of the target class. This global ellipse represents the overall distribution of the data and provides an additional constraint for the recognition process.

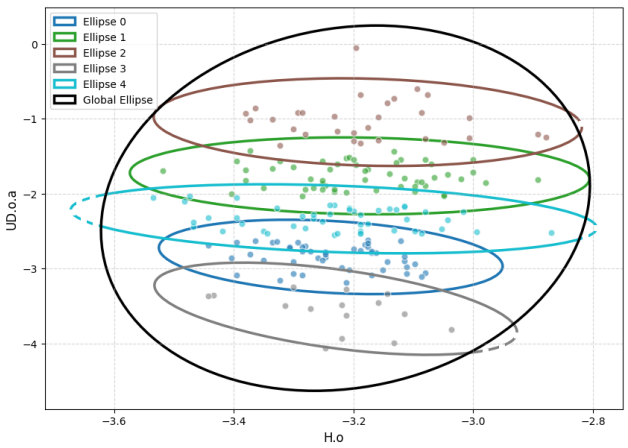

Fig. 3. **Constructed cluster-oriented prediction ellipses**
Source: **compiled by the authors**

As a result, instead of relying on a single global model, a set of cluster-based prediction ellipses was obtained and combined with the global prediction ellipse. A test vector is classified as a representative of the target class if it falls inside at least one of the cluster-based prediction ellipses and simultaneously lies within the global prediction ellipse. Otherwise, if the vector does not satisfy these conditions, it is classified as a representative of another class. This rule reflects the idea that the typing behavior of the target subject may exhibit multiple stable patterns, each captured by a separate cluster and its corresponding prediction ellipse, while still conforming to the overall distribution of the target data.

The experiments were implemented in Python using NumPy, Pandas, SciPy, and scikit-learn libraries. Initial data inspection and organization were performed using Microsoft Excel.

RESEARCH RESULTS

At the evaluation stage, the data from subject s018 was used to represent another class. This dataset was not involved in model construction and served only for testing purposes. The mean vector of the s018 is $\bar{X}$ = {0.0872, 0.0761}, and its covariance matrix is shown in Table 5.

Table 5. **Covariance matrix of the s018 set**

2.98E-04	-4.01E-04
-4.01E-04	7.73E-03

Source: **compiled by the authors**

The use of an independent subject as another class allows an assessment of the ability of each method to reject non-target typing patterns and to separate the target subject from other users. For comparison, several recognition models were constructed using the same training data and evaluated on identical test sets. The first model is cluster-based prediction ellipses (CBPE) (6), which represents the approach proposed in this study. The second model is the prediction ellipse for normalized data (PEND). This model consists of a single prediction ellipse for the normalized training data of the target class. The third model is the prediction ellipse for non-Gaussian data (PENGD) (1), where the prediction ellipse is built for non-normalized training data. This model is included to demonstrate the influence of distributional assumptions on recognition performance and to highlight the limitations of using prediction ellipses when the condition about data distribution is violated. In addition to statistical models, a one-class support vector machine was used as a representative machine learning approach for one-class recognition. The OCSVM was trained on target class data and learned a decision boundary that separates target samples from outliers.

The performance of all models was evaluated using the following standard metrics used in one-class recognition: recall measures the ability of a model to correctly identify samples belonging to the target class; specificity reflects how well the model rejects samples from another class; precision indicates the proportion of correctly accepted samples among all samples classified as target; *F*1-score provides a balanced measure that combines precision and recall; accuracy represents the overall proportion of correctly classified samples [32].

These metrics allow a comprehensive evaluation of both acceptance and rejection performance, which is especially important for biometric recognition systems. The results for all compared models are summarized in Table 6.

The obtained results demonstrate that both the CBPE and PEND models provide the highest overall recognition accuracy of 0.9749. However, despite identical accuracy, the two approaches exhibit different classification behavior. The PEND model achieves the highest specificity of 0.9949, indicating slightly better recognition of target user samples. At the same time, the CBPE model demonstrates a higher recall value of 0.97 compared to 0.965 for PEND, reflecting more accurate recognition of samples belonging to another class. As a result, the CBPE model also achieves the highest F1 score of 0.9557 among all compared methods.

The PENGD model demonstrates lower performance across all evaluated metrics compared to both CBPE and PEND. In particular, the accuracy decreases to 0.9631, while the F1 score reaches 0.9377. The reduction in performance confirms that constructing prediction ellipses directly on non-Gaussian data negatively affects covariance estimation and decreases the reliability of the resulting classification boundaries. These results further emphasize the importance of data normalization for statistical modeling in keystroke dynamics recognition.

The OCSVM model shows the lowest overall performance among the evaluated approaches, achieving an accuracy of 0.943 and an F1 score of 0.9155. In addition, both specificity and recall remain lower than those obtained for the ellipse-based methods. The obtained results suggest that the proposed cluster-based statistical modeling

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

Table 6. **Comparison of the results on the s018 test dataset**

Model	Specificity	Recall	Precision	*F*1 score	Accuracy
CBPE	0.9848	0.9700	0.9417	0.9557	0.9749
PEND	0.9949	0.9650	0.9333	0.9489	0.9749
PENGD	0.9746	0.9575	0.9187	0.9377	0.9631
OCSVM	0.9442	0.9425	0.8900	0.9155	0.9430

Source: **compiled by the authors**

provides more reliable decision boundaries for keystroke dynamics recognition in the considered feature space.

After evaluating the models on the s018 test dataset, an additional experiment was conducted to assess the robustness of the considered recognition approaches with respect to representatives of other classes. For this purpose, all four recognition models were sequentially tested against samples from 30 classes. Each class was evaluated independently to analyze the variability of recognition performance across different external typing patterns.

Since the target user samples remained unchanged across all experiments, the specificity metric remained constant for each model and was excluded from the comparative analysis. The analysis focused on the distribution of accuracy and recall values obtained for each non-target class. For this purpose, the maximum, minimum, mean, and standard deviation values of Accuracy and Recall were analyzed. The mean values characterize the average recognition performance across different external classes, while the minimum and maximum values reflect the best and worst-case recognition scenarios. The standard deviation is additionally used to evaluate the stability of the considered models with respect to variations in external typing patterns. The results are summarized in Table 7.

The empirical results presented in Table 7 show a strong overall performance of all evaluated models on the extended test set. All methods achieve a mean accuracy above 0.95, which confirms their general suitability for the task. The CBPE model provides the best overall balance, with a mean accuracy of 0.9861 and a mean recall of 0.9867. The PEND model shows almost the same average accuracy of 0.9856 and a slightly lower mean recall of 0.9810. This means that both approaches perform at a very similar level in terms of average recognition quality. At the same time, differences between the models become clearer when stability is taken into account. The CBPE model shows the most stable behavior, with the lowest standard deviation for accuracy of 0.0163 and for recall of 0.0244. The PEND model is slightly less stable, with higher standard deviations of 0.0236 for accuracy and 0.0353 for recall.

The PENGD model shows the largest variation, with 0.0605 for accuracy and 0.0902 for recall, which means its performance strongly depends on the specific class. The OCSVM model is more stable than PENGD but still less consistent than the ellipse-based approaches.

The analysis of minimum values highlights the worst-case behavior of the models. The CBPE model maintains a relatively high minimum accuracy of 0.9213 and a minimum recall of 0.8900, which shows that it remains reliable even in difficult cases. The PEND model reaches a slightly lower minimum accuracy of 0.8894 and a minimum recall of 0.8375, which indicates weaker performance in some classes despite strong average results.

The PENGD model shows a minimum accuracy of 0.6851 and a minimum recall of 0.5425, which confirms its instability on non-Gaussian data. The OCSVM model achieves a minimum accuracy of 0.8090 and a minimum recall of 0.7425.

When considering average values, stability, and worst-case results together, the CBPE model provides the most balanced performance. It combines high accuracy with low variation and strong results even in difficult cases. The PEND model is a close competitor, but shows slightly less stable behavior. The PENGD model performs the weakest overall, while OCSVM remains in the middle range with moderate but less stable results.

DISCUSSION AND FUTURE RESEARCH

The results confirm that modeling keystroke dynamics using cluster-based prediction ellipses can improve recognition accuracy and probability. By dividing the normalized feature space into local regions, the proposed approach adapts better to multimodal structures. This makes the method advantageous in situations where user behavior is not well represented by one Gaussian distribution. An important factor contributing to the improved recognition accuracy is the reduction of the decision region achieved through cluster-based local

Table 7. **Comparison of the results across an extended test set**

Model	Accuracy Mean	Accuracy Std	Accuracy Min	Accuracy Max	Recall Mean	Recall Std	Recall Min	Recall Max
CBPE	0.9861	0.0163	0.9213	0.9950	0.9867	0.0244	0.8900	1
PEND	0.9856	0.0236	0.8894	0.9983	0.9810	0.0353	0.8375	1
PENGD	0.9597	0.0605	0.6851	0.9916	0.9524	0.0902	0.5425	1
OCSVM	0.9559	0.0359	0.8090	0.9816	0.9617	0.0536	0.7425	1

Source: **compiled by the authors**

modeling. As a result, the overall acceptance region becomes more compact and better aligned with the actual distribution of target user data.

At the same time, the study reveals several challenges. First, the use of k-means clustering may not be optimal for this task. The algorithm assumes spherical clusters and relies on Euclidean distance, which may not reflect the true structure of keystroke dynamics data even after normalization. Therefore, alternative clustering strategies may provide better partitioning of the feature space and lead to further improvements in recognition accuracy.

Another important aspect of the proposed framework is the integration of robust covariance estimation for clusters containing limited numbers of samples. The application of the MCD estimator improves the stability of covariance matrices and reduces the influence of insufficient local sample sizes on prediction ellipse construction.

An additional limitation of the proposed approach is the complexity of the processing pipeline. The method requires multiple preprocessing stages, including testing for multivariate normality, applying normalizing transformations, performing clustering, and constructing several prediction ellipses. Such a multi-stage procedure may complicate practical implementation, since each stage introduces additional parameters that may influence the final recognition results.

Despite these facts, the results demonstrate that the cluster-based prediction ellipses approach provides an effective method for keystroke dynamics recognition. It combines the interpretability of classical statistical models with the flexibility of local data modeling, achieving a balance between recognition accuracy and robustness. Future research may focus on alternative clustering techniques, improved normalization strategies, and the evaluation of the method on larger and more diverse datasets.

CONCLUSIONS

This study proposed a cluster-based statistical approach for keystroke dynamics recognition that combines data normalization, clustering, and construction of cluster-based prediction ellipses. The method was designed to address the common problem that raw keystroke features rarely follow a multivariate normal distribution and often show diverse structures. By applying normalization, removing outliers, and partitioning the data into clusters, the proposed approach allows prediction ellipses to be constructed under conditions where their statistical assumptions are better satisfied. In particular, partitioning the normalized feature space into local clusters enables the model to represent multimodal behavioral patterns more effectively than a single global prediction ellipse.

To improve the stability of covariance estimation in clusters with limited numbers of samples, the MCD estimator was included. In addition, the proposed decision rule combines local cluster-based ellipses with a global prediction ellipse, which reduces the probability of accepting samples outside the overall behavioral profile of the target user.

The experimental results demonstrate that the proposed CBPE model achieves high recognition accuracy and stable performance in comparison with conventional prediction ellipses and the one-class support vector machine method. In particular, PEND shows slightly higher specificity due to a larger admissible decision region; however, CBPE achieves higher mean accuracy, as well as lower standard deviation and higher minimum accuracy, indicating more stable performance across different evaluation cases. In comparison with PENGD and OCSVM, both methods demonstrate lower performance across all evaluated metrics, including accuracy and stability, confirming the overall advantage of the proposed cluster-based approach.

At the same time, the study highlights several directions for future work. The current

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

implementation relies on k-means clustering, which may not optimally reflect the structure of keystroke data. Exploring alternative clustering strategies and investigating adaptive normalization procedures may further improve performance. In addition, future research should evaluate the proposed approach on datasets with a higher number of features in order to assess its scalability and effectiveness in higher-dimensional feature spaces.

Overall, the findings suggest that combining normalization, clustering, and statistical decision regions provides a promising approach for behavioral biometric recognition. The developed model preserves the interpretability of classical statistical models while improving flexibility in modeling heterogeneous biometric data distributions.

REFERENCES

1. Shadman, R., Wahab, A., Manno, M., Lukaszewski, M., Hou, D. & Hussain, F. “Keystroke dynamics: Concepts, techniques, and applications”. *ACM Computer Survey.* 2025; 57: 11, www.scopus.com/pages/publications/105011274996. DOI: https://doi.org/10.1145/3733103.

2. Williams, E. A., Warburton, M., Krzywinski, M. & Mushtaq, F. “The typability index: A tool for measuring and controlling for typing difficulty in text stimuli”. *Behavior Research Methods*. 2026; 58 (2): 61, www.scopus.com/pages/publications/105029972746. DOI: https://doi.org/10.3758/s13428-025-02877-y.

3. Yang, L., & Qin, S. “Identify user features impacting keystroke, mouse and touchscreen dynamics”. *Multimedia Tools and Applications*. 2025. 1–29, www.scopus.com/pages/publications/105013346175. DOI: https://doi.org/10.1007/s11042-025-21043-2.

4. Rippel, O., Mertens, P. & Merhof, D. “Modeling the distribution of normal data in pre-trained deep features for anomaly detection”. *In 25th IEEE International Conference on Pattern Recognition (ICPR)*, Milan, Italy. 2021. p. 6726-6733, www.scopus.com/pages/publications/85110412250. DOI: https://doi.org/10.1109/ICPR48806.2021.9412109.

5. Oyebola, O. “Examining the distribution of keystroke dynamics features on computer, tablet and mobile phone platforms”. In *Proceedings of the Mobile Computing and Sustainable Informatics ICMCSI*, 2023. p. 613-620. DOI: https://doi.org/10.1007/978-981-99-0835-6_43.

6. Kim, S., Park, D. & Jung, J. “Evaluation of One-Class classifiers for fault detection: Mahalanobis classifiers and the Mahalanobis–Taguchi”. *Processes*. 2021; 9: 1450, www.scopus.com/pages/publications/85113462897. DOI: https://doi.org/10.3390/pr9081450.

7. Prykhodko, S. & Trukhov, A. “Face recognition using Ten-Variate prediction ellipsoids for normalized data with different quantiles”. *Applied Aspects of Information Technology.* 2024; 7 (2): 151–161. https://aait.od.ua/index.php/journal/article/view/149. DOI: https://doi.org/ 10.15276/aait.07.2024.11.

8. Marimuthu, S., Mani, T., Sudarsanam, T. D., George, S. & Jeyaseelan, L. “Preferring Box-Cox transformation, instead of log transformation to convert skewed distribution of outcomes to normal”. In *Medical research, Clinical Epidemiology and Global Health*. 2022; 15: 101043, www.scopus.com/pages/publications/85128284476. DOI: https://doi.org/10.1016/j.cegh.2022.101043.

9. Oyewole, G. J. & Thopil, G. A. “Data clustering: application and trends”. *Artif Intell Rev.* 2023; 56: 6439–6475. DOI: https://doi.org/10.1007/s10462-022-10325-y.

10.Li, Q. & Chen, H. “CDAS: A Continuous Dynamic Authentication System”. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications (ICSCA'19), Association for Computing Machinery.* New York, NY, USA, 2019. p. 447–452, www.scopus.com/pages/publications/85066054180. DOI: https://doi.org/ 10.1145/3316615.3316691.

11.Chen, L., Zhong, Y., Ai, W. & Zhang, D. “Continuous authentication based on user interaction behavior”. In *Proceedings of the 2019 7th International Symposium on Digital Forensics and Security (ISDFS)*. Barcelos, Portugal. 2019. p. 1–6. DOI: https://doi.org/10.1109/ISDFS.2019.8757539.

12.Alvin, A., Jayabalan, M. & Thiruchelvam, V. “Keystroke dynamics based user authentication using deep multilayer perceptron”. *International Journal of Machine Learning and Computing*. 2020; 10: 134–139. DOI: https://doi.org/10.18178/ijmlc.2020.10.1.910.

13.AbdelRaouf, H., Chelloug, SA., Muthanna, A., Semary, N., Amin, K. & Ibrahim, M. “Efficient convolutional neural network-based keystroke dynamics for boosting user authentication”. *Sensors*. 2023; 23 (10): 4898. DOI: https://doi.org/10.3390/s23104898.

14.He, A. & Jin, X. “Deep variational autoencoder classifier for intelligent fault diagnosis adaptive to unseen fault categories”. In *IEEE Transactions on Reliability*. 2021; 70 (4): 1581–1595. DOI: https://doi.org/10.1109/TR.2021.3090310.

15.Eizagirre, I., Segurola-Gil, L., Zola, F. & Orduna, R. “Keystroke presentation attack: Generative adversarial networks for replacing user behaviour”. In *Proceedings of the 2022 European Symposium on Software Engineering (ESSE'22). Association for Computing Machinery*. New York, NY, USA. 2023. p. 119–126. DOI: https://doi.org/10.1145/3571697.3571714.

16.Chang, H., Li, J., Wu, C. & Stamp, M. “Machine learning and deep learning for fixed-text keystroke dynamics”. In *Proceedings of the Cybersecurity for Artificial Intelligence*. *Springer*. 2022; 54: 309–329, www.scopus.com/pages/publications/85134626360. DOI: https://doi.org/10.1007/978-3-030-97087-1_13.

17.Etherington, T. R. “Mahalanobis distances for ecological niche modelling and outlier detection: implications of sample size, error, and bias for selecting and parameterising a multivariate location and scatter method”. *PeerJ*. 2021; 9, www.scopus.com/pages/publications/85105860173. DOI: https://doi.org/10.7717/peerj.11436.

18.Prykhodko, S., N. Prykhodko, N. & Smykodub, T. “A joint statistical estimation of the RFC and CBO metrics for open-source applications developed in Java”. In *Proceedings of the 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*. *IEEE*. 2022. p. 442–445, www.scopus.com/pages/publications/85146307410.
DOI: https://doi.org/10.1109/CSIT56902.2022.10000457.

19.Sun, Y., Jiang, L., Pan, J., Sheng, S., & Hao, L. “A satellite imagery smoke detection framework based on the Mahalanobis distance for early fire identification and positioning”. *International Journal of Applied Earth Observation and Geoinformation*. 2023; 118: 103257. DOI: https://doi.org/10.1016/j.jag.2023.103257.

20.Lam, K., Meijer, K., Loonstra, F., Coerver, E., Twose, J., Redeman, E., Moraal, B., Barkhof, F., Uitdehaag, B. & Killestein, J. “Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis”. *Multiple sclerosis*. Houndmills, Basingstoke, England. 2021; 27 (9): 1421–1431. DOI: https://doi.org/10.1177/1352458520968797.

21.Yarema, O. R. & Babichev, S. A. “Current state of methods and algorithms for gene expression data clustering and biclustering: A survey”. *Herald of Advanced Information Technology*. 2024; 7 (4): 347–360.: https://doi.org/10.15276/hait.07.2024.24.

22.Vo, B. G., Van, H. D. S., Van, N. D. & Huynh, H. D. “Customer segmentation: Automatic K-optimization and RFM-Based K-Means Cclustering”. In *Proceedings of the 2025 10th International Conference on Intelligent Information Technology (ICIIT '25). Association for Computing Machinery,* New York, NY, USA. 2025. p. 173–178, www.scopus.com/pages/publications/105036823329. DOI: https://doi.org/ 10.1145/3731763.3731805.

23.Hassan, M., Eesa, A., Mohammed, A. & Arabo, W. “Oversampling method based on Gaussian distribution and K-Means clustering”. *Computers, Materials & Continua,* 2021; 69: 451–469, www.scopus.com/pages/publications/85107897344. DOI: https://doi.org/10.32604/cmc.2021.018280.

24.Hubert, M., Debruyne, M. & Rousseeuw, P. J. “Minimum covariance determinant and extensions”. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2018; 10 (3): e1421, www.scopus.com/pages/publications/85045418575. DOI: https://doi.org/10.1002/wics.1421.

25.Muralidharan, A., Eaman, A., & Hassan, E. “A comparative analysis of machine learning models for behavioral biometric authentication using keystroke dynamics”. *Procedia Computer Science*. 2025; 265: 116–123, www.scopus.com/pages/publications/105015150683.
DOI: https://doi.org/10.1016/j.procs.2025.07.163.

26.Raouf, H. A., Fouda, M. M. & Ibrahem, M. I. “Revolutionizing user authentication exploiting explainable AI and CTGAN-based keystroke dynamics”. In *IEEE Open Journal of the Computer Society*, 2025; 6: 97–108, www.scopus.com/pages/publications/85211972361.
DOI: https://doi.org/10.1109/OJCS.2024.3513895.

27.Khatun, N. “Applications of normality test in statistical analysis”. *Open Journal of Statistics*. 2021; 11 (01): 113. DOI: https://doi.org/10.4236/ojs.2021.111006.

28.Blum, L., Elgendi, M. & Menon, C. “Impact of Box-Cox transformation on machine-learning algorithms”. *Frontiers in Artificial Intelligence*. 2022; 5: 877569. DOI: https://doi.org/10.3389/frai.2022.877569.

29.Racolte, G., Marques, A., Scalco, L., Tonietto, L., Zanotta, D., Cazarin, C., Gonzaga, L. & Veronez, M. “Spherical K-Means and Elbow method optimizations with fisher statistics for 3D stochastic DFN from virtual outcrop models”. In *IEEE Access*. 2022; 10: 63723–63735, www.scopus.com/pages/publications/85132743666. DOI: https://doi.org/10.1109/ACCESS.2022.3182332.

30.Kobets, V. M., Gulin, O. V. & Nosov, P. S. “A cluster approach to matching the competences of data specialists with skills in demand on the labour market”. *Applied Aspects of Information Technology.* 2024; 7 (3): 231–241, https://aait.od.ua/index.php/journal/article/view/154.
DOI: https://doi.org/10.15276/aait.07.2024.16.

31.Chen, S., Ouyang, X. & Luo, F. “Ensemble One-Class support vector machine for Sea Surface target detection based on k-Means clustering”. *Remote Sensing*. 2024; 16 (13): 2401, www.scopus.com/pages/publications/85198396711. DOI: https://doi.org/ 10.3390/rs16132401.

32.Wang, X & Hou, D. “Enhancing keystroke dynamics authentication with ensemble learning and data resampling techniques”. *Electronics*. 2024; 13 (22): 4559, www.scopus.com/pages/publications/85210279575. DOI: https://doi.org/10.3390/electronics13224559.

Received 20.03.2026
Received after revision 10.06.2026
Accepted 17.06.2026

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

DOI: https://doi.org/10.15276/aait.09.2026.30
УДК 004.93

Розпізнавання клавіатурного почерку з використанням кластерно-орієнтованих еліпсів прогнозування

Приходько Сергій Борисович[1)]
ORCID: https://orcid.org/0000-0002-2325-018X; sergiy.prykhodko@nuos.edu.ua. Scopus Author ID: 55225622100
Трухов Артем Сергійович[1)]
ORCID: https://orcid.org/0000-0002-7160-8609; artem.trukhov@gmail.com
[1)] Національний університет кораблебудування імені адмірала Макарова, пр. Героїв України, 9. Миколаїв, 54007, Україна

АНОТАЦІЯ

Розпізнавання клавіатурного почерку є біометричним підходом, що забезпечує ідентифікацію користувача на основі індивідуальних особливостей набору тексту. **Актуальність** дослідження визначається потребою підвищення точності розпізнавання клавіатурного почерку в умовах варіативності, неоднорідності та відхилення даних від багатовимірного нормального розподілу. Об’єктом цього дослідження є процес розпізнавання клавіатурного почерку, тоді як предметом дослідження є математична модель, що базується на кластерно-орієнтованих еліпсах прогнозування, побудованих у двовимірному просторі ознак. **Метою** статті є підвищення точності розпізнавання клавіатурного почерку шляхом застосування кластерно-орієнтованих еліпсів прогнозування, які враховують локальні структури даних. **Завдання** дослідження полягають у формуванні кластерно-орієнтованих еліпсів прогнозування, побудованих на локальних навчальних вибірках, а також у порівнянні запропонованого підходу з іншими методами однокласового розпізнавання. **Методи** дослідження включають тест Мардіа для перевірки багатовимірної нормальності, перетворення Бокса-Кокса для нормалізації даних, кластеризацію k-means, метод ліктя та аналіз розміру кластерів, побудову еліпсів прогнозування за квадратом відстані Махаланобіса, робастне оцінювання коваріації на основі детермінанта мінімальної коваріації. **Наукова новизна** полягає у розробленні кластерного підходу до побудови локальних еліпсів прогнозування для розпізнавання клавіатурного почерку, що дає змогу враховувати неоднорідність даних та формувати точніші вирішальні межі. **Практична значимість** дослідження полягає у можливості підвищення надійності біометричних систем автентифікації користувачів без переходу до складніших багатокласових моделей. **Результати** дослідження демонструють, що запропонований підхід значно перевершує однокласову машину опорних векторів та еліпс прогнозування, побудований для негауссових даних, водночас забезпечуючи краще розпізнавання точок, що належать до іншого класу, та точність розпізнавання, порівнянну з еліпсом прогнозування для нормалізованих даних. Це покращення досягається за рахунок адаптивного моделювання локальних розподілів даних та більш точних вирішальних меж у просторі ознак. **Висновки** підтверджують ефективність кластеризації як етапу попередньої обробки для розпізнавання клавіатурного почерку на основі кластерно-орієнтованих еліпсів

прогнозування та демонструють доцільність адаптивного локального моделювання для неоднорідних даних клавіатурного почерку. Подальші дослідження можуть бути спрямовані на застосування альтернативних методів кластеризації, а також на використання розширених методів нормалізації даних, таких як перетворення Джонсона. Крім того, розширення простору ознак більшої розмірності шляхом включення більшої кількості характеристик клавіатурного почерку може додатково підвищити точність розпізнавання.

Ключові слова: Розпізнавання клавіатурного почерку; двовимірні дані; нормалізуюче перетворення; кластеризація; еліпс прогнозування; перетворення Бокса–Кокса

ABOUT THE AUTHORS

Sergiy B. Prykhodko - Doctor of Engineering Sciences, Professor, Head of Department of Software for Automated Systems. Admiral Makarov National University of Shipbuilding. 9, Heroes of Ukraine Ave. Mykolaiv, 54007, Ukraine
ORCID: https://orcid.org/0000-0002-2325-018X; sergiy.prykhodko@nuos.edu.ua. Scopus Author ID: 55225622100
Research field: Mathematical modeling; computer simulation; stochastic processes; multivariate statistical analysis; system identification; parameter estimation; nonlinear stochastic differential equations; software metrics estimation; pattern recognition

Приходько Сергій Борисович - доктор технічних наук, професор, завідувач кафедри Програмного забезпечення автоматизованих систем. Національний університет кораблебудування імені адмірала Макарова, пр. Героїв України, 9. Миколаїв, 54007, Україна

Artem S. Trukhov – PhD, Assistant, Department of Software for Automated Systems. Admiral Makarov National University of Shipbuilding. 9, Heroes of Ukraine Ave. Mykolaiv, 54007, Ukraine
ORCID: https://orcid.org/0000-0002-7160-8609; artem.trukhov@gmail.com
Research field: Mathematical modeling; multivariate statistical analysis; parameter estimation; pattern recognition
Трухов Артем Сергійович - доктор філософії, асистент кафедри Програмного забезпечення автоматизованих систем. Національний університет кораблебудування імені адмірала Макарова, пр. Героїв України, 9. Миколаїв, 54007, Україна