Keystroke Dynamics Recognition Using an Eight-Variate Prediction Ellipsoid for Normalized Data with a Reduced Feature Set

Sergiy Prykhodko[1,2,†] and Artem Trukhov[1,†]

[1] *Admiral Makarov National University of Shipbuilding, Heroes of Ukraine Ave., 9, Mykolaiv, 54007, Ukraine*
[2] *Odesa Polytechnic National University, Shevchenko Ave., 1, Odesa, 65044, Ukraine*

Abstract

Keystroke dynamics recognition is a prominent technique in behavioral biometrics, offering continuous and non-intrusive user authentication based on individual typing patterns. It leverages temporal features such as key press durations and inter-key latencies, which are distinctive for each user. Despite advantages like hardware independence and applicability in security-critical contexts, many existing methods rely on the assumption of multivariate normality, which real-world keystroke data often violates, leading to reduced recognition accuracy, recall, and specificity.

To address this limitation, this study explores the impact of data-normalizing transformations designed to transform keystroke features toward a multivariate normal distribution. Specifically, it examines the decimal logarithm, univariate Box-Cox, and multivariate Box-Cox transformations. The univariate transformations normalize each feature independently, without considering relationships between them. In contrast, the multivariate Box-Cox transformation processes all features jointly, taking into account their correlations, though it requires more complex computations. Parameters for Box-Cox transformations are estimated using the maximum likelihood method.

In contrast to previous work, which utilized a nine-dimensional feature vector composed exclusively of key hold durations, the current approach employs a reduced eight-dimensional feature set that combines key hold times and inter-key latencies. Despite the lower dimensionality, this set offers a more informative and representative characterization of user typing behavior.

Results demonstrate that applying normalizing transformations improves recognition accuracy, recall, and specificity, highlighting the importance of data normalization. Among the constructed models, the eight-variate prediction ellipsoid based on normalized data using the multivariate Box-Cox transformation shows the best recognition accuracy, recall, and specificity, significantly outperforming models built on non-normalized data. In addition, the use of a reduced eight-dimensional feature vector, combining hold durations and inter-key intervals, resulted in a more informative and compact representation of typing behavior, which further improved recognition accuracy, recall, and specificity. These findings emphasize the key role of data normalization and correct feature selection in enhancing the accuracy, recall, and specificity of keystroke dynamics recognition systems.

Future work should explore the application of alternative normalizing transformations and evaluate the proposed approach on larger, more diverse datasets that better capture the variability of user behavior.

Keywords

keystroke dynamics, multivariate normal distribution, normalizing transformation, mathematical modeling

1. Introduction

Keystroke dynamics recognition is a behavioral biometric technique that identifies or verifies individuals based on their unique typing patterns [1]. It analyzes temporal characteristics such as key press durations and latencies between successive keystrokes, which are difficult to replicate and remain relatively stable over time. This approach offers a non-intrusive, continuous method of user authentication and is commonly applied in areas such as secure system access, fraud detection, and

---

ICST-2025: Information Control Systems & Technologies, September 24-26, 2025, Odesa, Ukraine

† These authors contributed equally.

sergiy.prykhodko@nuos.edu.ua (S. Prykhodko); artem.trukhov@gmail.com (A. Trukhov)
0000-0002-2325-018X (S. Prykhodko); 0000-0002-7160-8609 (A. Trukhov)

user behavior monitoring, especially in environments where traditional authentication methods may be insufficient or inconvenient [2].

Since keystroke dynamics systems are typically used for user authentication, they are often modeled as one-class classification problems, where models are trained only on data from the target user and aim to detect deviations that may indicate unauthorized access [3].

Many existing methods used for recognition rely on statistical models that assume the underlying data follow a multivariate normal distribution [4]. However, in practice, this assumption for keystroke dynamics data is often violated [5] due to natural variability in human typing behavior, noise, and external factors. As a result, the accuracy, recall, and specificity of recognition can decrease significantly when these models are applied to real-world data.

Given the widespread use of keystroke dynamics recognition, including in security-sensitive applications, there is a clear need to improve recognition methods to ensure higher recognition accuracy, recall, and specificity. In particular, enhancing mathematical models to better accommodate deviations from the multivariate normal distribution can lead to more accurate systems. This study addresses this challenge by applying normalizing transformations that aim to adjust keystroke data toward a multivariate Gaussian distribution, thereby aligning more closely with the assumptions of statistical recognition models.

2. Literature review

Keystroke dynamics recognition has become a well-established approach in behavioral biometrics, used for verifying individuals based on how they type. It typically extracts temporal features such as key press durations and inter-key latencies, which reflect a person's unique typing pattern. These features are widely used due to their stability and ability to differentiate users.

A variety of methods have been employed to recognize typing patterns, including tree-based models [6, 7], support vector machines [8, 9], neural networks [10, 11], and others.

In practical applications such as authentication, keystroke dynamics is frequently modeled as a one-class classification problem. In this setting, the system is trained solely on data from the target class and attempts to detect whether new samples match this profile. This approach is particularly suited to real-world conditions, where data from unauthorized users may not be available during training. One-class classification is conceptually linked to outlier detection, as it identifies deviations from the normal behavior of the target user [12].

Several techniques are commonly used in one-class classification, including one-class SVM [13, 14], autoencoders [15, 16], GANs [17], and prediction ellipsoids [18]. Prediction ellipsoids construct a statistical boundary in the form of a multivariate ellipsoid that encompasses the known data. If a new point lies outside the ellipsoid, it is considered an outlier. This method is attractive due to its simplicity and well-defined mathematical formulation, particularly when the input data is assumed to follow a multivariate normal distribution.

However, the accuracy of prediction ellipsoids depends on how closely the data distribution matches the Gaussian assumption [19]. In practice, keystroke dynamics data often deviates from normality due to factors such as individual variability, typing inconsistencies, device differences, and external noise. These deviations can lead to incorrect estimation of the ellipsoid boundary and result in decreased recognition accuracy, recall, and specificity.

To address this issue, data normalization transformations are used to adjust feature distributions closer to a multivariate Gaussian [20, 21]. Common transformations include the decimal logarithm, univariate Box-Cox, and multivariate Box-Cox transformations. The univariate approaches adjust each feature independently, while the multivariate Box-Cox transformation jointly transforms all features while preserving their correlation structure.

In [22], a nine-variate prediction ellipsoid for normalized data was constructed using the Box-Cox transformation. However, the feature set in that study included only key hold times, which in some cases led to insufficient recognition accuracy. This limitation is addressed in the present study by using a broader set of temporal features that better represent user behavior.

This study investigates the impact of normalizing transformations on recognition accuracy, recall, and specificity in keystroke dynamics. The aim is to improve these metrics by applying transformations that adjust the data distribution to better align with the Gaussian assumption.

3. Materials and methods

Keystroke dynamics recognition relies heavily on the quality, structure, and temporal precision of the input data. To evaluate the proposed approach, this study uses the CMU Keystroke Dynamics Benchmark Dataset [23], a widely used [24] and publicly available dataset designed for authentication research. It contains keystroke data from 51 users, each of whom typed the same fixed password string, ".tie5Roanl", over eight sessions, with at least one day between each session. During every session, the password was typed 50 times, resulting in a total of 400 samples per subject and 20,400 samples in total.

Each entry includes timestamps for key press and key release events, recorded with sub-millisecond precision. From these events, 31 timing features are derived, including key hold durations (the time from pressing to releasing a key), keydown–keydown (DD) intervals, and keyup–keydown (UD) intervals. Feature names follow a standard notation: H.k for hold time of key k, DD.k1.k2 for the interval between pressing keys k1 and k2, and UD.k1.k2 for the delay between releasing k1 and pressing k2. UD values may be negative, and the sum of a key's hold time and UD equals the corresponding DD value.

In this study, a reduced subset of eight features was selected to create vector X: H.o, UD.o.a, H.a, UD.a.n, H.n, UD.n.l, H.l, and UD.l.Return. These features represent a continuous segment of the password near its end and include both hold times and inter-key latencies [25]. This selection captures individual key behavior and transitions between keys, providing a compact but informative representation of typing patterns suitable for multivariate modeling.

After extracting the feature vectors, an essential preprocessing step involves the identification and removal of outliers. Such anomalous observations may arise due to involuntary user behavior, external distractions, hardware inconsistencies, or inaccuracies in event logging. If not addressed, these outliers can introduce bias into parameter estimation, distort statistical decision boundaries, and negatively impact the accuracy of recognition models. The removal of outliers ensures that the training dataset more accurately reflects typical user behavior, thereby enhancing the stability and interpretability of the resulting model.

For this purpose, the squared Mahalanobis distance (SMD) is employed, which measures the distance between each sample and the center of the distribution, accounting for the covariance between features. Under the assumption that the data follows a multivariate normal distribution, the SMD of inliers is expected to follow a chi-square distribution with degrees of freedom equal to the number of features [26]. This statistical property enables the selection of a significance level to define a threshold beyond which samples are considered outliers.

However, this approach relies on the assumption that the input data follows an approximately multivariate Gaussian distribution, which is often not true for raw keystroke dynamics features due to inherent variability in human typing behavior. To evaluate whether the distribution significantly deviates from the Gaussian, the Mardia test is applied. This test evaluates the joint distribution of the features based on two measures: multivariate skewness β_1 and multivariate kurtosis β_2. If the test indicates that the data significantly deviates from multivariate normality, a normalization transformation must be applied to convert a non-Gaussian data $X = X_1, X_2, \ldots, X_8{}^T$ to a Gaussian vector $Z = Z_1, Z_2, \ldots, Z_8{}^T$.

Normalizing transformations can be categorized into univariate and multivariate approaches. Univariate transformations, such as the logarithmic and the Box-Cox, operate on individual features independently. The logarithm is typically used to compress large values and reduce positive skewness.

The Box-Cox transformation (BCT) generalizes this adjustment through a power function parameterized by λ, allowing for greater flexibility based on the distribution's shape. The transformation is defined as:

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

where X_j is a *j*-th non-Gaussian variable; Z_j is a normalized variable; λ_j is *j*-th transformation parameter.

The univariate BCT is applied element-wise to each feature using a single parameter λ. When dealing with multivariate data, this transformation is typically applied independently to each of the k features, resulting in k separate transformations, one per feature, with individual parameters drawn from the *k*-dimensional vector $\Theta = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$. Determining the optimal values of λ is a crucial step aimed at transforming each feature's distribution to approximate normality as closely as possible. Maximum likelihood estimation (MLE) is commonly employed to identify these optimal parameters:

$$l(\lambda) = (\lambda - 1)\sum_{i=1}^{N} \ln(x_i) - \frac{N}{2} ln \sum_{i=1}^{N} \frac{(x(\lambda)_i - \overline{x(\lambda)})^2}{N}, \tag{2}$$

where x_i is the *i*-th data point; λ is the transformation parameter; *N* is the number of observations.

While univariate methods can improve the separate distributions of individual features, they do not account for the correlations between variables. This limitation is addressed by multivariate normalization, particularly the multivariate BCT, which applies a transformation to the entire feature vector. This method preserves the dependency structure among features. The components of the transformed vector Z are defined as in equation (1). However, it requires the estimation of multiple transformation parameters by maximizing a multivariate log-likelihood function, which may introduce additional computational complexity:

$$l(X, \Theta) = \sum_{j=1}^{k} (\lambda_j - 1) \sum_{i=1}^{N} \ln(x_{ji}) - \frac{N}{2} \ln[\det(S_Z)], \tag{3}$$

where *k* is the number of variables; x_{ji} is the value of variable *j* for observation *i*; λ_j is *j*-th transformation parameter; S_Z is the covariance matrix of the transformed data.

To evaluate whether the data distribution has been successfully approximated to multivariate normality, the Mardia test is applied to assess multivariate skewness and kurtosis before and after normalization. If the results confirm sufficient normality, the data can be reliably used in subsequent modeling stages. Otherwise, further preprocessing or the application of more robust modeling techniques may be necessary.

The next step involves constructing the prediction ellipsoid, which defines a decision boundary for one-class classification. The left-hand side of the comparison is the squared Mahalanobis distance, which measures how far a data point deviates from the distribution's center. The right-hand side is a critical value derived from the chi-square distribution, based on the dimensionality of the feature space and the desired significance level.

$$(X - \bar{X})^T S_X^{-1} (X - \bar{X}) = \chi^2_{8,\ 0.005}, \tag{4}$$

where X is a non-Gaussian random vector; S_X is a sample covariance matrix for initial data; $\bar{X}$ is a vector of sample means of the variables; $\chi^2_{8,\ 0.005}$ is the chi-square distribution quantile with 8 degrees of freedom and significance level 0.005.

Assuming multivariate normality, the SMD follows a chi-square distribution with degrees of freedom equal to the number of features, in this study, 8. This allows the threshold to be determined according to a selected significance level; for one-class classification, a commonly used value is

0.005 [19]. If a point's squared Mahalanobis distance exceeds the critical value, it is classified as an anomaly, likely representing a different class. If the distance is below the threshold, the point is accepted as part of the target class.

In cases where the raw data does not follow a Gaussian distribution, normalization is applied first to bring the data closer to multivariate normality. After this step, the eight-variate prediction ellipsoid is constructed according to (4):

$$(Z - \bar{Z})^T S_Z^{-1} (Z - \bar{Z}) = \chi^2_{8,\ 0.005}, \tag{5}$$

where Z is a Gaussian random vector; $\bar{Z}$ is a vector of sample means of the variables; S_Z is a sample covariance matrix for normalized data; $\chi^2_{8,\ 0.005}$ is the chi-square distribution quantile with 8 degrees of freedom and significance level 0.005.

At a significance level of 0.005 and with 8 degrees of freedom, the corresponding critical value from the chi-square distribution is 21.96. Any sample with a Mahalanobis distance below this value is considered to lie within the ellipsoid and thus to belong to the target class.

To evaluate the constructed models, several widely used metrics are applied: accuracy, specificity, precision, recall, and the F1 score [27, 28]. These metrics are calculated based on four possible recognition outcomes. A true positive occurs when a sample from the target class is correctly recognized as a target. A false positive refers to an outlier that is incorrectly identified as belonging to the target class. A true negative is an outlier sample that is correctly rejected, and a false negative is a sample that is mistakenly rejected as an outlier.

Accuracy measures the overall proportion of correctly recognized samples, encompassing both target and outlier classes. While it provides a general view of system performance, it may be less informative when class distributions are imbalanced. Recall, also known as sensitivity, indicates the proportion of target samples that are correctly identified, reflecting the system's ability to minimize false rejections. Precision expresses the proportion of samples predicted as target that are indeed from the target class, and is particularly relevant when the cost of false acceptances is high. Specificity measures the system's ability to correctly reject outlier inputs, ensuring robust protection against unauthorized access. Lastly, the F1 score is the harmonic mean of precision and recall. It provides a single balanced metric that accounts for both false positives and false negatives, which is particularly useful in scenarios where both types of errors are costly [29].

Each of the evaluation metrics captures a different aspect of the model's recognition performance. Accuracy provides a general measure of overall correctness by considering all samples, both target and outlier. The probability of recognition is reflected by the combination of recall and specificity.

4. Experiments

In this study, the data associated with subject identifier s036 was randomly selected to represent the target class, while the data from s022 was used to represent an outlier class. Each subject contributed 400 keystroke samples. The first step in the analysis involved identifying and removing outliers from the target class data. Data preprocessing, normalization, and statistical calculations were performed using built-in Microsoft Excel formulas and matrix operations.

To evaluate whether the target data followed a multivariate normal distribution, the Mardia test was applied. The results indicated significant deviation from normality. Specifically, the test statistic for multivariate skewness $N\beta_1/6$ was 5207.45, exceeding the quantile of the chi-square distribution, which is 163.65 for 120 degrees of freedom at a 0.005 significance level. Similarly, the multivariate kurtosis statistic β_2 was 211, which is higher than the corresponding normal quantile value of 83.25, based on a mean of 80 and a variance of 1.6 at the same significance level. These results suggest that the raw data significantly deviates from multivariate normality, indicating the need for normalization before further analysis.

At this stage, where the primary goal is to detect and remove outliers from the target class data, the multivariate BCT is applied. This approach is chosen because it adjusts all features simultaneously while preserving their correlations. In contrast to univariate transformations that process each feature separately, the multivariate form ensures consistent scaling across the feature space, which is especially important for computing Mahalanobis distances. Since this distance metric assumes that the data follows an approximately multivariate normal distribution, the applied transformation helps bring the feature distribution closer to Gaussian, making it suitable for outlier detection.

As a result of applying the maximum likelihood method to the log-likelihood function (3), the following parameter estimates were obtained: $\widehat{\lambda_1}$ = 0.0523, $\widehat{\lambda_2}$ = 0.1285, $\widehat{\lambda_3}$ = 2.0907, $\widehat{\lambda_4}$ = 0.1095, $\widehat{\lambda_5}$ = 0.6797, $\widehat{\lambda_6}$ = -2.5205, $\widehat{\lambda_7}$ = -0.8415, $\widehat{\lambda_8}$ = -0,8439.

After applying the eight-variate BCT, the resulting feature set with component (1) was evaluated using the Mardia test. The test statistic for multivariate skewness, $N\beta_1/6$ = 396.16, exceeds the critical value of 163.64 from the chi-square distribution with 120 degrees of freedom at a 0.005 significance level. Similarly, the test statistic for multivariate kurtosis, β_2 = 94.04, is higher than the corresponding quantile value of 83.25, given a mean of 80 and a variance of 1.6.

These results indicate that, despite the transformation, the data still deviates from multivariate normality, mainly due to the presence of outliers, which distort the distribution. Nevertheless, working with the transformed dataset remains advantageous, as it shifts the distribution closer to a Gaussian form, thereby improving the reliability of Mahalanobis distance calculations in the outlier detection process.

Next, the SMD is computed for each feature vector to detect potential outliers. These distances are compared against the critical value of 21.96 from the chi-square distribution with 8 degrees of freedom at a significance level of 0.005. In each iteration, only the feature vector with the highest SMD, if it exceeds the critical value, is removed from the dataset.

In the first iteration, the 24th vector, which had the maximum SMD of 55.43, was identified as an outlier and excluded. This process is repeated iteratively: after each removal, the normalization is applied again using the multivariate BCT, and the SMDs are recalculated for the updated dataset. The procedure continues until all remaining vectors have SMD values below the threshold. Table 1 presents the indices and corresponding SMD values of all removed outliers. In total, 13 vectors were excluded.

Table 1
Removed outliers

№	SMD	Vector index	№	SMD	Vector index
1	55.43	24	8	26.89	124
2	33.22	3	9	26.77	36
3	27.83	37	10	24.88	2
4	27.69	146	11	24.21	110
5	27.67	103	12	24.29	281
6	27.47	35	13	23.82	36
7	28.18	131			

The final dataset, obtained after the removal of outliers, consists of 387 feature vectors. The vector of means: $\bar{X}$ = {0.0435; 0.2103; 0.0568; 0.2973; 0.0392; 0.5510; 0.0412; 0.8041}. The corresponding covariance matrix is presented in Table 2.

The sample was randomly shuffled to avoid any bias from the original order of the data. This step helps make sure that both the training and test sets are representative of the whole dataset. After shuffling, the data was divided into two equal parts: 193 samples were used for the training set, and the remaining 194 samples were used for testing. The training set was used to build the prediction ellipsoid by estimating the average values and relationships between features. The test set, which was not used during training, helped check how well the model works on new, unseen data. This approach provides a fair way to measure recognition accuracy, recall, and specificity, and helps prevent overfitting to the training data.

Table 2
The covariance matrix of the set s036 after the removal of outliers

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	$0.0^{4}29$	$0.0^{4}57$	$0.\ 0^{5}65$	$-0.0^{4}33$	$0.0^{5}59$	$-0.0^{4}47$	$0.0^{5}65$	$-0.0^{4}66$
X_2	$0.0^{4}57$	0.012	$-0.0^{3}24$	$-0.0^{2}29$	$0.0^{4}37$	$0.0^{2}69$	$0.0^{4}83$	0.02
X_3	$0.0^{5}65$	$-0.0^{3}24$	$0.0^{4}41$	$0.0^{4}28$	$0.0^{5}51$	$-0.0^{3}21$	$-0.0^{6}62$	$-0.0^{3}6$
X_4	$-0.0^{4}33$	$-0.0^{2}29$	$0.0^{4}28$	$0.0^{2}47$	$-0.0^{4}12$	$0.0^{3}55$	$-0.0^{4}27$	$0.0^{3}22$
X_5	$0.0^{5}59$	$0.0^{4}37$	$0.0^{5}51$	$-0.0^{4}12$	$0.0^{4}28$	$0.0^{4}91$	$0.0^{4}1$	$-0.0^{4}4$
X_6	$-0.0^{4}47$	$0.0^{2}69$	$-0.0^{3}21$	$0.0^{3}55$	$0.0^{4}91$	0.036	$0.0^{3}11$	0.041
X_7	$0.0^{5}65$	$0.0^{4}83$	$-0.0^{6}62$	$-0.0^{4}27$	$0.0^{4}1$	$0.0^{3}11$	$0.0^{4}36$	$0.0^{3}47$
X_8	$-0.0^{4}66$	0.02	$-0.0^{3}6$	$0.0^{3}22$	$-0.0^{4}4$	0.041	$0.0^{3}47$	0.3

The training set has a mean vector given by $\bar{X}$ = {0.0434; 0.2063; 0.0570; 0.3009; 0.0392; 0.5437; 0.0414; 0.7779}. The corresponding covariance matrix is presented in Table 3.

Table 3
The covariance matrix of the training set

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	$0.0^{4}33$	$0.0^{4}61$	$0.0^{5}92$	$-0.0^{4}27$	$0.0^{5}71$	$-0.0^{4}16$	$0.0^{5}81$	$-0.0^{5}98$
X_2	$0.0^{4}61$	0.011	$-0.0^{3}24$	$-0.0^{2}24$	$0.0^{4}42$	$0.0^{2}54$	$0.0^{4}91$	0.02
X_3	$0.0^{5}92$	$-0.0^{3}24$	$0.0^{4}46$	$0.0^{4}27$	$0.0^{5}46$	$-0.0^{3}28$	$-0.0^{5}21$	$-0.0^{3}8$
X_4	$-0.0^{4}27$	$-0.0^{2}24$	$0.0^{4}27$	$0.0^{2}45$	$0.0^{4}13$	$0.0^{3}55$	$0.0^{5}55$	$-0.0^{3}86$
X_5	$0.0^{5}71$	$0.0^{4}42$	$0.0^{5}46$	$0.0^{4}13$	$0.0^{4}25$	$0.0^{3}1$	$0.0^{4}1$	$-0.0^{3}12$
X_6	$-0.0^{4}16$	$0.0^{2}54$	$-0.0^{3}28$	$0.0^{3}55$	$0.0^{3}1$	0.023	$0.0^{3}29$	0.041
X_7	$0.0^{5}81$	$0.0^{4}91$	$-0.0^{5}21$	$0.0^{5}55$	$0.0^{4}1$	$0.0^{3}29$	$0.0^{4}43$	$0.0^{3}64$
X_8	$-0.0^{5}98$	0.02	$-0.0^{3}8$	$-0.0^{3}86$	$-0.0^{3}12$	0.041	$0.0^{3}64$	0.295

The Mardia test indicates that the training set does not follow a multivariate normal distribution. The test statistic for multivariate skewness $N\beta_1/6$ was calculated as 1849.77, which exceeds the quantile of the chi-square distribution of 163.65 at 120 degrees of freedom and a 0.005 significance level. Additionally, the test statistic for multivariate kurtosis, β_2, was found to be 149.04, surpassing the corresponding normal distribution threshold of 84.69, based on a mean of 80, a variance of 3.32, and the same significance level. These results indicate a significant deviation from multivariate normality, supporting the need to apply a normalizing transformation.

As a result of applying the univariate decimal logarithm transformation, the normalized training set has a mean vector of $\bar{Z}$ = {-1.3663; -0.7339; -1.2478; -0.5333; -1.4102; -0.2767; -1.3874; -0.1627}. The corresponding covariance matrix is presented in Table 4.

Table 4
The covariance matrix of the normalized training set by decimal logarithm

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Z_1	$0.0^{2}32$	$0.0^{3}8$	$0.0^{3}73$	$-0.0^{3}49$	$0.0^{3}79$	$0.0^{4}14$	$0.0^{3}75$	$0.0^{2}12$
Z_2	$0.0^{3}8$	0.043	$-0.0^{2}38$	$-0.0^{2}81$	$0.0^{3}97$	$0.0^{2}70$	$0.0^{2}22$	0.016
Z_3	$0.0^{3}73$	$-0.0^{2}38$	$0.0^{2}31$	$0.0^{3}56$	$0.0^{3}36$	$-0.0^{2}14$	$-0.0^{3}2$	$-0.0^{2}27$
Z_4	$-0.0^{3}49$	$-0.0^{2}81$	$0.0^{3}56$	0.011	$0.0^{3}12$	$0.0^{3}56$	$0.0^{4}77$	$-0.0^{3}47$
Z_5	$0.0^{3}79$	$0.0^{3}97$	$0.0^{3}36$	$0.0^{3}12$	$0.0^{2}32$	$0.0^{3}6$	$0.0^{2}12$	$-0.0^{3}23$
Z_6	$0.0^{4}14$	$0.0^{2}70$	$-0.0^{2}14$	$0.0^{3}56$	$0.0^{3}6$	$0.0^{2}92$	$0.0^{2}17$	0.011
Z_7	$0.0^{3}75$	$0.0^{2}22$	$-0.0^{3}2$	$0.0^{4}77$	$0.0^{2}12$	$0.0^{2}17$	$0.0^{2}4$	$0.0^{2}32$
Z_8	$0.0^{2}12$	0.016	$-0.0^{2}27$	$-0.0^{3}47$	$-0.0^{3}23$	0.011	$0.0^{2}32$	0.038

The normalized training set by decimal logarithm does not satisfy the assumption of multivariate normality, as determined by Mardia's test. The test statistic for multivariate skewness, $N\beta_1/6$ =

546.59, exceeds the quantile of the chi-square distribution, which is 163.65 at 120 degrees of freedom and a significance level of 0.005. In addition, the test statistic for multivariate kurtosis, β_2 = 101.39, is greater than the value of 84.69, based on a normal distribution with a mean of 80, a variance of 3.32, and a 0.005 significance level. These results confirm significant deviation from multivariate normality. However, the transformation improved the distribution, bringing it closer to the Gaussian.

The univariate BCT was applied to the training set. Using the maximum likelihood estimation method for the logarithmic function (2), the following parameter estimates were obtained: $\widehat{\lambda_1}$ = 0.0501, $\widehat{\lambda_2}$ = 0.1127, $\widehat{\lambda_3}$ = 2.0209, $\widehat{\lambda_4}$ = 0.9140, $\widehat{\lambda_5}$ = 0.9446, $\widehat{\lambda_6}$ = -2.6803, $\widehat{\lambda_7}$ = -1.1806, $\widehat{\lambda_8}$ = -0.9382. As a result of applying the univariate BCT with components (1), where each variable is transformed independently of the others, the normalized training set has the following mean vector: $\bar{Z}$ = {-2.9101; -1.5283; -0.4933; -0.7298; -1.0090; -1.9666; -36.4872; -0.5659}, the corresponding covariance matrix is presented in Table 5.

Table 5
The covariance matrix of the normalized training set by univariate BCT

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Z_1	0.0125	$0.0^{2}33$	$0.0^{5}97$	$-0.0^{3}6$	$0.0^{3}2$	$0.0^{2}31$	0.1373	0.0112
Z_2	$0.0^{2}33$	0.1550	$-0.0^{4}45$	$-0.0^{2}94$	$0.0^{3}2$	0.1492	0.3963	0.0887
Z_3	$0.0^{5}97$	$-0.0^{4}45$	$0.0^{6}12$	$0.0^{5}12$	$0.0^{6}32$	$-0.0^{4}87$	$-0.0^{4}87$	$-0.0^{4}46$
Z_4	$-0.0^{3}6$	$-0.0^{2}94$	$0.0^{5}12$	$0.0^{2}55$	$0.0^{5}16$	$0.0^{2}73$	$0.0^{2}91$	$0.0^{3}5$
Z_5	$0.0^{3}2$	$0.0^{3}2$	$0.0^{6}32$	$0.0^{5}16$	$0.0^{4}36$	$0.0^{3}4$	0.0127	$0.0^{4}36$
Z_6	$0.0^{2}31$	0.1492	$-0.0^{4}87$	$0.0^{2}73$	$0.0^{3}4$	0.9331	1.3638	0.3221
Z_7	0.1373	0.3963	$-0.0^{4}87$	$0.0^{2}91$	0.0127	1.3638	38.0950	0.9601
Z_8	0.0112	0.0887	$-0.0^{4}46$	$0.0^{3}5$	$0.0^{4}35$	0.3221	0.9601	0.3178

The training set normalized using the univariate BCT does not deviate from multivariate normality. According to Mardia's test, the test statistic for multivariate skewness, $N\beta_1/6$ = 163.03, does not exceed the quantile of the chi-square distribution, which is 163.65 for 120 degrees of freedom at a 0.005 significance level. Similarly, the test statistic for multivariate kurtosis, β_2 = 80.98, remains below the critical threshold of 84.69, calculated for a normal distribution with a mean of 80 and a variance of 3.32. These results indicate that the distribution of the training set normalized using the univariate BCT follows an approximately multivariate normal distribution.

The eight-variate BCT was applied to the training dataset. The parameter estimates were obtained by maximizing the log-likelihood function (3), resulting in the following values: $\widehat{\lambda_1}$ = 0.0668, $\widehat{\lambda_2}$ = 0.2385, $\widehat{\lambda_3}$ = 1.9483, $\widehat{\lambda_4}$ = 0.7088, $\widehat{\lambda_5}$ = 0.8871, $\widehat{\lambda_6}$ = -2.5191, $\widehat{\lambda_7}$ = -1.1207, $\widehat{\lambda_8}$ = -0.9131. Unlike the univariate approach, where each variable is transformed independently, the eight-variate BCT accounts for the joint structure of the data, preserving the correlations between features.

Table 6
The covariance matrix of the normalized training set by the eight-variate BCT

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Z_1	0.0113	$0.0^{2}28$	$0.0^{4}11$	$-0.0^{3}73$	$0.0^{3}19$	$0.0^{2}25$	0.108	0.01
Z_2	$0.0^{2}28$	0.102	$-0.0^{4}46$	-0.011	$0.0^{3}2$	0.109	0.261	0.072
Z_3	$0.0^{4}11$	$-0.0^{4}46$	$0.0^{6}19$	$0.0^{5}21$	$0.0^{6}47$	$-0.0^{4}98$	$-0.0^{4}91$	$-0.0^{4}56$
Z_4	$-0.0^{3}73$	-0.011	$0.0^{5}21$	$0.0^{2}9$	$0.0^{4}24$	$0.0^{2}8$	$0.0^{2}89$	$0.0^{3}27$
Z_5	$0.0^{3}19$	$0.0^{3}2$	$0.0^{6}47$	$0.0^{4}24$	$0.0^{4}52$	$0.0^{3}45$	0.0126	$0.0^{4}37$
Z_6	$0.0^{2}25$	0.109	$-0.0^{4}98$	$0.0^{2}8$	$0.0^{3}45$	0.76	1.029	0.289
Z_7	0.108	0.261	$-0.0^{4}91$	$0.0^{2}89$	0.0126	1.029	25.985	0.785
Z_8	0.01	0.072	$-0.0^{4}56$	$0.0^{3}27$	$0.0^{4}37$	0.289	0.785	0.312

After transformation with components (1), the normalized training set was characterized by the following mean vector: $\bar{Z}$ = {-2.8368; -1.3729; -0.5113; -0.8118; -1.0637; -1.8200; -31.5424; -0.5597}, the corresponding covariance matrix is provided in Table 6.

To assess the deviation of the distribution of the normalized training set from multivariate normality, Mardia's test was employed. The test statistic for multivariate skewness, $N\beta_1/6$ = 162.07, does not exceed the quantile of the chi-square distribution, which is 163.65 with 120 degrees of freedom at a significance level of 0.005. The test statistic for multivariate kurtosis, β_2 = 80.95, is lower than the corresponding quantile of the normal distribution, 84.69, calculated for a mean of 80 and a variance of 3.32. Based on these results, the distribution of the normalized training set is considered approximately multivariate normal.

5. Results

The next step is the construction of mathematical models in the form of prediction ellipsoids and the comparison of recognition metrics. The first model corresponds to the prediction ellipsoid constructed for non-Gaussian data (4) (PENGD), the original training set without any normalization. The second model is a prediction ellipsoid for normalized data (5) using the univariate decimal logarithm transformation (PE-Log). The third model is a prediction ellipsoid for normalized data (5) transformed by the univariate Box-Cox transformation (PE-UBCT). The fourth model is a prediction ellipsoid for normalized data (5) by the eight-variate Box-Cox transformation (PE-EBCT). Recognition metrics for all models are presented in Table 7.

Table 7
Comparison of prediction ellipsoids

Model	Specificity	Recall	Precision	F1 score	Accuracy
PENGD	0.9625	0.9588	0.9254	0.9418	0.9613
PE-Log	0.9750	0.9639	0.9492	0.9565	0.9714
PE-UBCT	0.9900	0.9845	0.9795	0.9820	0.9882
PE-EBCT	0.9900	0.9897	0.9796	0.9846	0.9899

PENGD uses non-Gaussian data without any normalization. While this model demonstrates relatively high performance, with an accuracy of 96.13%, its precision (92.54%) is notably lower than its recall (95.88%), indicating a higher rate of false positives relative to false negatives.

Applying the univariate decimal logarithm transformation (PE-Log) leads to a noticeable improvement across all metrics, particularly in precision (94.92%) and F1 score (95.65%), suggesting better balance between false positives and false negatives. This transformation brought the data closer to a Gaussian distribution, but the distribution still showed significant deviation from normality. As a result, recognition metrics improved, but not as much as with a more powerful normalizing transformation.

The univariate Box-Cox transformation (PE-UBCT) further enhances performance, achieving an F1 score of 98.20% and accuracy of 98.82%. The improvements in both precision (97.95%) and recall (98.45%) reflect a strong ability to correctly recognize both target and negative instances.

The highest performance is achieved by the prediction ellipsoid for normalized data using the eight-variate Box-Cox transformation (PE-EBCT). It shows the highest recall (98.97%) and F1 score (98.46%), as well as a specificity of 99.00% and overall accuracy of 98.99%. These findings indicate that multivariate transformations provide better recognition metrics, as they preserve the correlations between features.

The results show that applying prediction ellipsoids for normalized data greatly improves recognition accuracy, recall, and specificity. While univariate transformations lead to noticeable improvements, the eight-variate Box-Cox transformation provides the most accurate and balanced results, as it takes into account the relationships between features.

6. Discussion

The results demonstrate that applying a prediction ellipsoid for normalized data significantly improves the recognition accuracy, recall, and specificity when the training set deviates from multivariate normality. Among the tested transformations, the eight-variate BCT achieved the most accurate and balanced results, as it preserves the correlations between features.

The univariate decimal logarithm transformation brought the data closer to multivariate normality, but Mardia's test indicated that the distribution still exhibited a statistically significant deviation. As a result, the corresponding model PE-Log showed improvement in recognition metrics compared to the non-normalized PE-NGD, but its performance was lower than that of the prediction ellipsoid for normalized data using other transformations. The PE-UBCT model showed further improvement, as it more effectively corrected skewness and kurtosis in individual variables. However, since it transforms features independently, it does not preserve inter-feature dependencies, which results in lower recognition accuracy, recall, and specificity compared to the multivariate transformation.

In all models, prediction ellipsoids were constructed using a significance level of 0.005. This threshold is commonly used in one-class classification and outlier detection tasks, where high specificity is desirable. A stricter significance level results in a smaller ellipsoid, reducing the false positive rate, but may also exclude borderline true positives. The balance between sensitivity and specificity must therefore be considered when selecting this parameter.

Another important aspect is the size and representativeness of the training data. Since the prediction ellipsoid relies on estimating the mean vector and covariance matrix, it requires a sufficient number of samples to ensure stability. In this study, 13 data points were removed as outliers before model construction. While this removal improved the approximation to multivariate normality, it may also have excluded meaningful variation in the data, especially if the outliers reflected rare but valid behavior.

Although the eight-variate transformation provided the best results in this experiment, selecting an appropriate normalization method remains a nontrivial task, particularly when the data have complex, multimodal, or heavy-tailed distributions.

The dataset used in this study is limited to input sequences of fixed structure and moderate length, which may not fully capture the diversity observed in real-world typing behavior. In practical scenarios, longer input sequences, such as passwords or phrases containing 20–22 characters, are generally more suitable, as they allow for more comprehensive feature extraction. Moreover, in this work, only eight features were selected from the available data, which may restrict the model's ability to fully represent the user's typing dynamics. In addition to the length and content of the typed password, other important factors were not included in this study. Contextual influences, such as the user's physical condition, emotional state, time of day, or even environmental conditions like temperature or humidity, can affect how a person types. These factors may lead to changes in typing rhythm or key timing and could have a noticeable impact on recognition accuracy in real-world applications [30]. Although this study focused only on timing features from controlled input, future work could benefit from considering these real-life conditions, as they may help improve the reliability and robustness of keystroke-based biometric systems.

7. Conclusions

This study explored the impact of non-Gaussian data and examined how the application of prediction ellipsoids for normalized data affects keystroke dynamics recognition accuracy, recall, and specificity. A new, reduced feature set was used, consisting of hold time and inter-key time features. This combination reflects different aspects of user behavior, key press duration, and transition time between keys, allowing the model to capture more diverse typing characteristics and improve recognition outcomes.

The results demonstrated that the application of prediction ellipsoids for normalized data significantly enhances the recognition accuracy, recall, and specificity. Among the evaluated transformations, the eight-variate Box-Cox transformation achieved the most accurate and balanced results. This confirms the advantage of multivariate normalization, which preserves correlations between features and brings the data closer to a multivariate normal distribution. In contrast, univariate transformations: decimal logarithm and univariate Box–Cox, showed only moderate improvements, as they treat each feature independently and do not account for inter-feature dependencies.

Despite the improvements achieved in this study, there are still some limitations and disadvantages.

One disadvantage is the need for a large and representative dataset. Building a reliable prediction ellipsoid requires enough data to accurately estimate both the average values and the relationships between features, typically at least 100 samples in the training set. Another disadvantage is the difficulty of selecting an appropriate normalizing transformation. While the multivariate Box-Cox transformation gave the best results, choosing the most suitable normalization method is not always an easy task, especially when the data contains outliers or has a complex structure.

A key limitation of this work is that 13 outliers were removed before model construction. Although this improved the fit to a multivariate normal distribution, it may have excluded rare but valid typing patterns.

Future research will focus on several directions to further improve keystroke dynamics recognition. First, more advanced normalization techniques could be explored. While the multivariate Box-Cox transformation showed the best results in this study, it may not be optimal for all datasets. The Johnson transformation, for example, may provide better adaptability to complex or multimodal data distributions, especially when the deviation from normality is strong. Second, it would be beneficial to use a more comprehensive dataset that includes a larger number of features. In this study, only eight features were used, which may limit the model's ability to fully represent the user's typing behavior. Additionally, the password in the current dataset contains only 10 characters, while longer sequences, such as those with 20 to 22 characters, are generally preferred, as they allow for more detailed feature extraction and better capture of individual typing patterns. It would also be useful to consider contextual factors that can influence keystroke dynamics, such as the user's physical or emotional state, time of day, or environmental conditions.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Shadman, A. Wahab, M. Manno, M. Lukaszewski, D. Hou, F. Hussain, Keystroke Dynamics: Concepts, Techniques, and Applications, ACM Computer Survey 57 11 (2025) 283. https://doi.org/10.1145/3733103

[2] P. Kasprowski, Z. Borowska, K. Harezlak, Biometric Identification Based on Keystroke Dynamics, Sensors 22 9 (2022) 3158. https://doi.org/10.3390/s22093158

[3] S. Roy, J. Pradhan, A. Kumar, D. Adhikary, U. Roy, D. Sinha, R. Pal, A systematic literature review on latest keystroke dynamics based models. IEEE Access, 10 (2022) 92192-92236. http://doi.org/10.1109/ACCESS.2022.3197756

[4] O. Rippel, P. Mertens, D. Merhof, 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, pp. 6726-6733. http://doi.org/ 10.1109/ICPR48806.2021.9412109.

[5] O. Oyebola, 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, pp. 613-620. DOI: https://doi.org/10.1007/978-981-99-0835-6_43

[6] B. Saini, P. Singh, A. Nayyar, N. Kaur, K. Bhatia, S. El-Sappagh, J. Hu, A three-step authentication model for mobile phone user using keystroke dynamics. IEEE Access, 8 (2020) 125909-125922. https://doi.org/10.1109/ACCESS.2020.3008019
[7] A. Tanapat, Strengthening Password Authentication using Keystroke Dynamics and Smartphone Sensors, in: Proceedings of the 9th International Conference on Information Communication and Management (ICICM'19), Association for Computing Machinery, New York, NY, USA, 2019, pp. 70-74. https://doi.org/10.1145/3357419.3357425
[8] Q. Li, H. Chen, 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, pp. 447-452. https://doi.org/10.1145/3316615.3316691
[9] L. Chen, Y. Zhong, W. Ai, D. Zhang, Continuous Authentication Based on User Interaction Behavior, in: Proceedings of the 2019 7th International Symposium on Digital Forensics and Security (ISDFS), Barcelos, Portugal, 2019, pp. 1-6, https://doi.org/10.1109/ISDFS.2019.8757539
[10] A. Alvin, M. Jayabalan, V. Thiruchelvam, Keystroke Dynamics Based User Authentication using Deep Multilayer Perceptron, International Journal of Machine Learning and Computing 10 2020 134-139. https://doi.org/10.18178/ijmlc.2020.10.1.910
[11] H. AbdelRaouf, SA. Chelloug, A. Muthanna, N. Semary, K. Amin, M. Ibrahim, Efficient Convolutional Neural Network-Based Keystroke Dynamics for Boosting User Authentication. Sensors 23 10 (2023) 4898. https://doi.org/10.3390/s23104898
[12] H. Marques, L. Swersky, J. Sander, On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles, Data Min Knowl Disc 37 (2023) 1473-1517. https://doi.org/10.1007/s10618-023-00931-x
[13] N. Seliya, A. Abdollah Zadeh, T. Khoshgoftaar, A literature review on one-class classification and its potential applications in big data, J Big Data 8 122 (2021). https://doi.org/10.1186/s40537-021-00514-x
[14] S. Yerima, A. Bashar, Semi-supervised novelty detection with one class SVM for SMS spam detection, in: 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP), Sofia, Bulgaria, 2022, pp. 1-4. https://doi.org/10.1109/IWSSIP55020.2022.9854496
[15] A. He, X. Jin, Deep Variational Autoencoder Classifier for Intelligent Fault Diagnosis Adaptive to Unseen Fault Categories, in: IEEE Transactions on Reliability, vol. 70, no. 4, pp. 1581-1595, 2021. https://doi.org/10.1109/TR.2021.3090310
[16] Y. Patel, K. Ouazzane, V. Vassilev, I. Faruqi, G. L. Walker, Keystroke Dynamics using Auto Encoders, in: Proceedings of the 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 2019, pp. 1-8, https://doi.org/10.1109/CyberSecPODS.2019.8885203
[17] I. Eizagirre, L. Segurola-Gil, F. Zola, R. Orduna, 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, pp. 119-126. https://doi.org/10.1145/3571697.3571714
[18] S. Kim, D. Park, J. Jung, Evaluation of One-Class Classifiers for Fault Detection: Mahalanobis Classifiers and the Mahalanobis–Taguchi System, Processes, 2021, 9, 1450. https://doi.org/10.3390/pr9081450
[19] S. Prykhodko, L. Makarova, K. Prykhodko, A. Pukhalevych, Application of Transformed Prediction Ellipsoids for Outlier Detection in Multivariate Non-Gaussian Data, in: Proceedings of the 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 2020, pp. 359-362. https://doi.org/10.1109/TCSET49122.2020.235454
[20] K. Lam, K. Meijer, F. Loonstra, E. Coerver, J. Twose, E. Redeman, B. Moraal, F. Barkhof, B. Uitdehaag, J. Killestein, Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis, Multiple sclerosis (Houndmills, Basingstoke, England) 27 9 (2021) 1421-1431. https://doi.org/10.1177/1352458520968797

[21] S. Prykhodko, A. Trukhov, Application of a ten-variate prediction ellipsoid for normalized data and machine learning algorithms for face recognition, in: Selected Papers of the Seventh International Workshop on Computer Modeling and Intelligent Systems (CMIS-2024). Workshop Proceedings (CMIS-2024), Zaporizhzhia, Ukraine, May 3, 2024. CEUR Workshop Proceedings, vol. 3702, 2024, pp. 362-375. https://ceur-ws.org/Vol-3702/paper30.pdf
[22] S. Prykhodko, A. Trukhov, Application of a Nine-Variate Prediction Ellipsoid for Normalized Data and Machine Learning Algorithms for Keystroke Dynamics Recognition, in: Proceedings of the Information Technology and Implementation (IT&I) Workshop: Intelligent Systems and Security (IT&I-WS 2024: ISS), CEUR Workshop Proceedings, vol. 3933, 2024, pp. 51-64. https://ceur-ws.org/Vol-3933/Paper_5.pdf
[23] K. Killourhy, R. Maxion, CMU Keystroke Dynamics Benchmark Dataset, Carnegie Mellon University. URL: https://www.cs.cmu.edu/~keystroke/.
[24] HC. Chang, J. Li, CS. Wu, M. Stamp, Machine Learning and Deep Learning for Fixed-Text Keystroke Dynamics, in: Stamp, M., Aaron Visaggio, C., Mercaldo, F., Di Troia, F. (eds) Artificial Intelligence for Cybersecurity. Advances in Information Security 54 (2022). Springer, Cham. https://doi.org/10.1007/978-3-030-97087-1_13
[25] I. Tsimperidis, O. Asvesta, E. Vrochidou, G. Papakostas, IKDD: A Keystroke Dynamics Dataset for User Classification. Information 15 9 (2024) 511. https://doi.org/10.3390/info15090511
[26] T. Etherington, 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 9:e11436, (2021). https://doi.org/10.7717/peerj.11436
[27] S. Nurmaini, A. Darmawahyuni, A. Sakti Mukti, M. Rachmatullah, F. Firdaus, B. Tutuko, Deep Learning-Based Stacked Denoising and Autoencoder for ECG Heartbeat Classification, Electronics, 9 135 (2020). https://doi.org/10.3390/electronics9010135
[28] H. Waleed, S. Gadsden, J. Yawney, Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances, Expert Systems with Applications 193 (2021) 116429. https://doi.org/10.1016/j.eswa.2021.116429
[29] X. Wang, D. Hou, Enhancing Keystroke Dynamics Authentication with Ensemble Learning and Data Resampling Techniques, Electronics 13 22 (2024) 4559. https://doi.org/10.3390/electronics13224559
[30] S. Bilan, M. Bilan, A. Bilan, Interactive biometric identification system based on the keystroke dynamic, in: Proceedings of the Biometric Identification Technologies Based on Modern Data Mining Methods. Springer, Cham, 2021, pp. 39-58. https://doi.org/10.1007/978-3-030-48378-4_3