applied sciences

MDPI

*Article*

# Online Kanji Characters Based Writer Identification Using Sequential Forward Floating Selection and Support Vector Machine

**Md. Al Mehedi Hasan, Jungpil Shin * and Md. Maniruzzaman**

School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Fukushima, Japan
* Correspondence: jpshin@u-aizu.ac.jp

**Abstract:** Writer identification has become a hot research topic in the fields of pattern recognition, forensic document analysis, the criminal justice system, etc. The goal of this research is to propose an efficient approach for writer identification based on online handwritten Kanji characters. We collected 47,520 samples from 33 people who wrote 72 online handwritten-based Kanji characters 20 times. We extracted features from the handwriting data and proposed a support vector machine (SVM)-based classifier for writer identification. We also conducted experiments to see how the accuracy changes with feature selection and parameter tuning. Both text-dependent and text-independent writer identification were studied in this work. In the case of text-dependent writer identification, we obtained the accuracy of each Kanji character separately. We then studied the text-independent case by considering some of the top discriminative characters from the text-dependent case. Finally, another text-dependent experiment was performed by taking two, three, and four Kanji characters instead of using only one character. The experimental results illustrated that SVM provided the highest identification accuracy of 99.0% for the text-independent case and 99.6% for text-dependent writer identification. We hope that this study will be helpful for writer identification using online handwritten Kanji characters.

**Keywords:** writer identification; handwritten Kanji characters; feature extraction; feature selection; sequential forward floating selection; support vector machine



**Citation:** Hasan, M.A.M.; Shin, J.; Maniruzzaman, M. Online Kanji Characters Based Writer Identification Using Sequential Forward Floating Selection and Support Vector Machine. *Appl. Sci.* **2022**, *12*, 10249. https://doi.org/10.3390/app122010249

Academic Editor: António J. R. Neves

Received: 14 September 2022
Accepted: 9 October 2022
Published: 12 October 2022



## 1. Introduction

Writer identification has become a hot research topic in the fields of pattern recognition, forensic document analysis, the criminal justice system, etc. It is one kind of biometric recognition that can easily identify the writer based on their handwritten characters. The handwriting style of each individual, which appears identical at first glance and yet has its own originality, is an interesting research topic of writer identification. Handwritten characters play a significant role for forensic experts in identifying the writers. In the last twenty years, many attempts have been undertaken to use handwritten characters for writer verification and identification. Moreover, handwritten characters can be used for the identification of age groups [1,2]. Writer identifications are utilized for verification and authentication, for example, signature verification and identification in a court of law and bank. With the development of the information society, the importance of writer identification technology to verify the legitimacy of users is increasing. There are three types of authentication methods: knowledge authentication, property authentication, and biometric authentication [3]. The challenges in identifying writers based on handwriting include the following: (i) developing a suitable model to identify the handwriting of different individuals; (ii) extracting and identifying potential handwriting features; (iii) testing the performance of proposed methods. There are numerous studies conducted on writer

identification using different language-based characters such as Arabic [4], Bangla [5], Chinese [6,7], French [8], Japanese [8–10], and English [11–15].

In the field of handwriting analysis, the goal is to propose an efficient method that can be compared with existing approaches that use veins, fingerprints, and other biometrics. There are two ways of collecting handwritten samples for writer identification: offline and online character recognition methods. In offline-based identification or static method, writing-based samples are collected from documents or images using scanners. Generally, this method is built on several attributes, including lines, characters, words, and so on [16]. The dynamic features are not present in the offline mode. Moreover, there is a lack of sequential information in handwriting and huge intra-class variation. As a result, offline writer recognition is considered more complex. Research on writer identification based on offline methods is a very challenging issue for security and forensic purposes [7]. As a result, many studies have been conducted using offline methods [8,17,18].

Online handwritten characters are considered less challenging due to having various information such as pen pressure, pen altitude, pen azimuth, and pen position [7]. Moreover, handwritten characters or samples can be easily collected using tablets, smart phones, magnetic input, and so on. Writer-identification-based approaches can be categorized into two parts: text-independent and text-dependent. The text-independent methods analyze the entire image without knowing anything about the script's content, whereas text-dependent approaches analyze the entire image based on such knowledge. The summary of previous research on the identification of writers using handwriting is presented in Table 1. Only a few types of research on writer identification have been conducted on handwritten-based Kanji characters. In some sense, our research focuses on both online text-independent and text-dependent writer identification approaches for handwritten-based Kanji characters. In this study, we use x-coordinates, y-coordinates, pen pressure, pen altitude, pen azimuth, and the time taken to acquire the data. We collected online handwritten-based Kanji characters using a pen tablet. We extracted 47,520 samples from 33 people who wrote different 72 Kanji characters 20 times each and used efficient features to achieve a higher identification rate. Furthermore, we compared how the accuracy varies depending on whether the type is text-dependent or text-independent.

The organization of this paper is as follows: Section 2 presents related work; Section 3 presents the materials and methods, including dataset preparation, proposed methodology, feature extraction, feature normalization, feature selection, and classification using SVM. The experimental results and discussion are discussed in Section 4. Finally, the conclusion and future work direction are discussed in Section 5.

**Table 1.** Previous research on writer identification.

| Authors | DT | SS | Methods | Language | Writers | ACC(%) |
|---|---|---|---|---|---|---|
| Nakamura et al. [9] | Online | 1230 | ANOVA | Kanji | 41 | 90.4 |
| Soma et al. [11] | Offline | 5000 | Majority Voting | Kanji | 100 | 99.0 |
| Namboodiri and Gupta [13] | Online | 400 | NN | English | 6 | 88.0 |
| Li et al. [14] | Online | 1500 | k-NN | English | 242 | 93.6 |
| Wu et al. [15] | Online | 1700 | HMM | English | 200 | 95.5 |
| Nguyen et al. [17] | Offline | 2965 | CNN | Kanji | 480 | 93.8 |
| Rehman et al. [6] | Online | 52,800 | P2P | Chinese | 48 | 99.0 |
| Abdi et al. [4] | Offline | 4800 | k-NN | Arabic | 82 | 90.2 |
| Nasuno and Arai [10] | offline | 50,000 | CNN | Japanese | 100 | 90.0 |

DT—data types; SS—sample size.

## 2. Related Work

Writer identification based on online Kanji characters can achieve a greater identification rate than offline Kanji characters because the online characters provide more information such as the length of the stroke order and brushstrokes. We reviewed various existing papers on writer identification, which were mainly conducted on only online data

or offline data types based on various handwritten characters, such as English, Arabic, Japanese, Chinese, and so on. We summarized these existing papers and presented in Table 1. Our study mainly focuses on writer identification based on online handwritten Kanji characters. Efficient feature extraction and feature selection play a significant role in writer identification. Various local features were extracted from the characters to show the discriminative information of the writer. For example, Dargan et al. [19] used four features (transition, diagonal, zoning, and peak-extent-based) in order to develop a writer identification system using Devanagari characters. Bensefia and Djeddi [20] proposed a novel feature selection approach for writer identification and obtained an identification rate of 96.0%. Moreover, some studies used both local and global features for writer identification [11,21].

In existing studies, there have been various writer identification methods such as SVM [22,23], distance-based [9,24], deep learning [25], and so on. For example, Nakamura et al. [9] used online text-independent data of handwritten-based Kanji characters and collected 1230 samples from 41 respondents. They extracted different kinds of features using a one-way analysis of variance (ANOVA). Moreover, the distance-based method was adopted to separate the writers and obtained an identification accuracy of 90.4%. Soma et al. [11] utilized offline handwritten-based Kanji characters for writer identification and obtained 99.0% identification accuracy using a voting method for three Kanji characters. Namboodiri and Gupta [13] used a neural network (NN) to identify the writers using online handwritten-based English characters and achieved a classification accuracy of 88.0%. Li et al. [14] proposed k-nearest neighbors (k-NN) for writer identification based on online handwritten-based English characters and obtained an identification accuracy rate of 93.6%. Wu et al. [15] employed a hidden Markov model (HMM) for the identification of writers and achieved a 95.5% identification accuracy rate. Nguyen et al. [17] also adopted the convolution neural network (CNN)-based model for writer identification and reported 93.8% identification accuracy. Rehman et al. [6] also proposed deep learning for the identification of writers using handwritten-based Chinese characters. They collected 52,800 samples and obtained 99.0% identification accuracy. Abdi et al. [4] also proposed k-NN method for writer identification using offline handwritten-based Arabic characters and achieved 90.2% identification accuracy. Nasuno and Arai [10] conducted an experiment on Japanese handwritten characters that contained 100 Japanese characters written 50 times by 100 subjects. They trained AlexNet CNN for 90 Japanese characters, tested model performance on 10 characters, and obtained an identification accuracy of 90.0%.

## 3. Materials and Methods

### *3.1. Device for Data Collection*

Handwritten-based Kanji character data were collected using a pen tablet system (Cintiq Pro 16, Wacom Co., Ltd., Saitama, Japan). The pen tablet was connected to a laptop PC with Windows 10. The screen size of the pen tablet was 15.6 inches, and the resolution size was 2560 × 1440 pixels. The coordinates of generated parameters using a pen tablet are shown in Figure 1.

### *3.2. Dataset Preparation*

The dataset utilized in this study was collected using a pen tablet system. The dataset consisted of 33 people who wrote 72 Kanji characters 20 times each. A total of $33 \times 72 \times 20 = 47{,}520$ samples were used for this study. The handwriting data contained six pieces of information: the x-coordinate and y-coordinate, which are the positions where the characters were written; the writing pressure; the azimuth angle; the altitude; and the time taken to acquire the data.

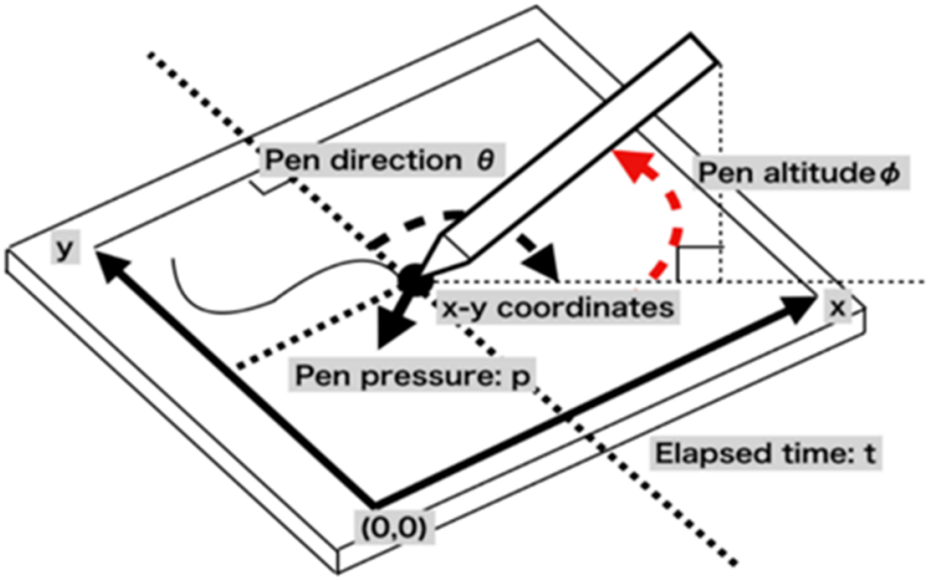


**Figure 1.** Pen tablet device.

### *3.3. Proposed Methodology*

The goal of this study is to propose an efficient writer identification system using handwritten-based Kanji characters. The proposed model of writer identification has the following tasks: feature extraction, feature normalization, feature selection, tune the hyperparameters and train the SVM-based model, and writer identification. The block diagram of the proposed model for writer identification is depicted in Figure 2. Every step or phase is more clearly explained in the following subsections.

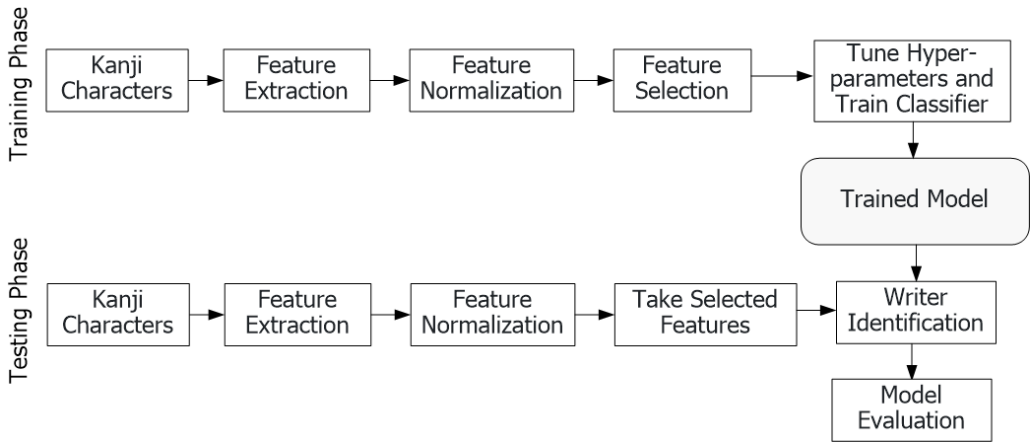


**Figure 2.** Block diagram of the proposed model for writer identification.

### *3.4. Feature Extraction*

We collected six raw features as the x-coordinate and y-coordinate, which are the positions where the characters were written, the writing pressure, the azimuth angle, the altitude, and the time taken to acquire the data. Based on these six raw features, we computed forty features. The calculation of speed is based on [26], and peak instantaneous speed, acceleration, peak instantaneous acceleration, positive pressure change, and negative pressure change are based on [22]. The mean and standard deviation (std) of the beginning and end states of writing pressure, altitude, azimuth, speed, and acceleration were taken from previous studies [27–29]. An explanation of each feature, along with their descriptions and calculation formula, is given in Table 2.

**Table 2.** List of extracted features names, their descriptions, and calculation formulas.

| SN | Features | Descriptions | Calculation Formula |
|---|---|---|---|
| 1 | Speed mean | Mean of speed. | $\bar{s} = \frac{1}{n}\sum_{i=0}^{n} s_i$ |
| 2 | Speed std | Std of speed. | $s\_\sigma = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(s_i - \bar{s})^2}$ |
| 3 | Max speed | Maximum speed. | $s\_max = Max(s_i)$ |
| 4 | First speed mean | Mean of the speed of the first 10% of the whole. | $\bar{s} = \frac{1}{k}\sum_{i=0}^{k} s_i; (k = \frac{n}{10})$ |
| 5 | First speed std | Std of the speed of the first 10% of the whole. | $s\_\sigma = \sqrt{\frac{1}{k}\sum_{i=0}^{k}(s_i - \bar{s})^2}; (k = \frac{n}{10})$ |
| 6 | Last speed mean | Mean speed of the last 10% of the whole. | $\overline{s\_last} = \frac{1}{n-k}\sum_{i=k}^{n} s_i;\ (k = \frac{9}{10}n)$ |
| 7 | Last speed std | Std of the speed of the last 10% of the whole. | $s\_last\_\sigma = \sqrt{\frac{1}{n-k}\sum_{i=k}^{n}(s_i - \bar{s})^2};\ (k = \frac{9}{10}n)$ |
| 8 | Piv | Maximum speed recorded at any time. | $V_{max}\ (V_i = \frac{L_i}{T_i})$ |
| 9 | Accel. mean | Mean of accel. | $\overline{ac} = \frac{1}{n}\sum_{i=0}^{n} ac_i$ |
| 10 | Accel. std | Std of accel. | $ac\_\sigma = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(ac_i - \overline{ac})^2}$ |
| 11 | Max accel. | Maximum accel. | $ac\_max = Max(ac_i)$ |
| 12 | First accel. mean | Mean of accel. of the first 10% of the whole. | $\overline{ac\_1st} = \frac{1}{k}\sum_{i=0}^{k} ac_i; k = \frac{n}{10}$ |
| 13 | First accel. std | Std of accel. of the first 10% of the whole. | $ac\_1st\_\sigma = \sqrt{\frac{1}{k}\sum_{i=0}^{k}(ac_i - \overline{ac})^2}; k = \frac{n}{10}$ |
| 14 | Last accel. mean | Mean of accel. of the last 10% of the whole. | $\overline{ac\_last} = \frac{1}{n-k}\sum_{i=k}^{n} ac_i;\ k = \frac{9}{10}n$ |
| 15 | Last accel. std | Std of accel. of the last 10% of the whole. | $ac\_last\_\sigma = \sqrt{\frac{1}{n-k}\sum_{i=0}^{n}(ac_i - \overline{ac})^2}; (k = \frac{9}{10}n)$ |
| 16 | Pia | Maximum accel. recorded at any point. | $A_{max}\ (A_i = \frac{V_i}{T_i})$ |
| 17 | Pressure mean | Mean of pen pressure. | $\bar{p} = \frac{1}{n}\sum_{i=0}^{n} p_i$ |
| 18 | Pressure std | Std of pen pressure. | $p\_\sigma = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(p_i - \bar{p})^2}$ |
| 19 | Max pressure | Maximum of pen pressure. | $p\_max = Max(p_i)$ |
| 20 | First pressure mean | Mean of pen pressure of the first 10% of the whole. | $\overline{p\_1st} = \frac{1}{n}\sum_{i=0}^{n} p_i;\ (k = \frac{n}{10})$ |
| 21 | First pressure std | Std of pen pressure of the first 10% of the whole. | $p\_1st\_\sigma = \sqrt{\frac{1}{k}\sum_{i=0}^{k}(p_i - \bar{p})^2};\ (k = \frac{n}{10})$ |
| 22 | Last pressure mean | Mean of pen pressure of the last 10% of the whole. | $\overline{p\_last} = \frac{1}{n-k}\sum_{i=k}^{n} p_i;\ (k = \frac{9}{10}n)$ |
| 23 | Last pressure std | The std of pen pressure of the last 10% of the whole. | $p\_last\_\sigma = \sqrt{\frac{1}{n-k}\sum_{i=k}^{n}(p_i - \bar{p})^2};\ (k = \frac{9}{10}n)$ |
| 24 | Azimuth mean | Mean of azimuth. | $\overline{az} = \frac{1}{n}\sum_{i=0}^{n} az_i$ |
| 25 | Azimuth std | Std of azimuth. | $az\_\sigma = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(az_i - \overline{az})^2}$ |
| 26 | First azimuth mean | Mean of the azimuth of the first 10% of the whole. | $\overline{az\_1st} = \frac{1}{k}\sum_{i=0}^{k} az_i;\ (k = \frac{n}{10})$ |
| 27 | First azimuth std | Std of the azimuth of the first 10% of the whole. | $az\_1st\_\sigma = \sqrt{\frac{1}{k}\sum_{i=0}^{k}(az_i - \overline{az})^2};\ (k = \frac{n}{10})$ |
| 28 | Last azimuth mean | Mean of the azimuth of the last 10% of the whole. | $\overline{az\_last} = \frac{1}{n-k}\sum_{i=k}^{n} az_i; (k = \frac{9}{10}n)$ |
| 29 | Last azimuth std | Std of the azimuth of the last 10% of the whole. | $az\_last\_\sigma = \sqrt{\frac{1}{n-k}\sum_{i=k}^{n}(az_i - \overline{az})^2};\ (k = \frac{9}{10}n)$ |
| 30 | Altitude mean | Mean of altitude. | $\overline{alt} = \frac{1}{n}\sum_{i=0}^{n} alt_i$ |
| 31 | Altitude std | Std of altitude. | $alt\_\sigma = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(alt_i - \overline{alt})^2}$ |
| 32 | First altitude mean | Mean of the altitude of the first 10% of the whole. | $\overline{alt\_1st} = \frac{1}{n}\sum_{i=0}^{k} alt_i;\ (k = \frac{n}{10})$ |
| 33 | First altitude std | Std of the altitude of the first 10% of the whole. | $alt\_1st\_\sigma = \sqrt{\frac{1}{k}\sum_{i=0}^{k}(alt_i - \overline{alt})^2}; (k = \frac{n}{10})$ |
| 34 | Last altitude mean | Mean of the altitude of the last 10% of the whole. | $\overline{alt\_last} = \frac{1}{n-k}\sum_{i=k}^{n} alt_i; (k = \frac{9}{10}n)$ |
| 35 | Last altitude std | Std of the altitude of the last 10% of the whole. | $alt\_last\_\sigma = \sqrt{\frac{1}{n-k}\sum_{i=k}^{n}(alt_i - \overline{alt})^2};\ (k = \frac{9}{10}n)$ |
| 36 | Positive pressure change mean | Mean of increase in pen pressure between two time points. | $\text{Mean}(p_{i+1} - p_i / t_{i+1} - t_i, \text{where } p_{i+1} > pi)$ |
| 37 | Positive pressure changes std | Std of increase in pen pressure between two time points. | $\text{Std}(p_{i+1} - p_i / t_{i+1} - t_i, \text{where } p_{i+1} > pi)$ |
| 38 | Max positive pressure change | Maximum increase in pen pressure between two time points. | $\text{Max}(p_{i+1} - p_i / t_{i+1} - t_i \text{ for } p_{i+1} > pi)$ |
| 39 | Negative pressure change mean | Mean decrease in pen pressure between two time points. | $\text{Mean}(p_{i+1} - p_i / t_{i+1} - t_i, \text{where } p_{i+1} < pi)$ |
| 40 | Negative pressure changes std | Std of decrease in pen pressure between two time points. | $\text{Std}(p_{i+1} - p_i / t_{i+1} - t_i, \text{where } p_{i+1} < pi)$ |

### *3.5. Feature Normalization*

Data normalization is a technique that minimizes redundancy and improves the efficiency of the data. Mathematically, it is defined as follows:

$$z = \frac{X - \mu}{\sigma} \tag{1}$$

where $X$ is the original feature vector, $\mu$ is the mean of that feature vector, and $\sigma$ is the standard deviation. The value of $z$ lies between 0 and 1.

### *3.6. Feature Selection*

Feature selection is the process of removing irrelevant features and improving the model's performance. We applied sequential forward floating selection (SFFS) to select the best combination of features. It is one kind of greedy algorithm to reduce the initial $d$-dimensional feature space into a k-dimensional feature subspace $(k < d)$ [30]. The pseudo-code is shown below:

**Input:** $Y = \{y_1, y_2, \ldots, y_d\}$
**Output:** $X_k = \{x_j | j = 1, 2, \ldots, k; x_j \in Y\}$, where $k \in (0, 1, 2, \ldots, d)$
**Initialize:** $X_o = \emptyset, k = 0$

**Step 1:** $x^+$ = argmax $J(X_k + x)$, where $x \in Y - X_k$, $J$ is an evaluation index and $x^+$ is the feature with the highest evaluation when it is chosen.

**Step 2:** $x^-$ = argmax $J(X_k - x)$, where $x \in X_k$ and $x^-$
is the feature with the best performance when the feature is deleted.
If $J(X_k - x) > J(X_k)$ :
$X_{k-1} = X_k - x^-$
$k = k - 1$
Go to **Step 1.**

In **Step 1**, include the feature of the feature space that leads to the best performance increase from the feature subset. Then, go to **Step 2**. In **Step 2**, remove only a feature in the case of increasing the performance of the result subset. If k = 2 or the improvement cannot be made, go back to **Step 1**; else, repeat **Step 2**. When k = k, terminate this algorithm. In this study, we implemented SFFS using the SequentialFeatureSelctor of the "mlxtend" library [31]. We selected the combination of the best features with the highest identification accuracy.

### *3.7. Classification Using SVM*

SVM is a supervised learning method that is used to solve problems in classification and regression. It is an effective learning method in high-dimensional spaces and when the number of dimensions is larger than the number of samples. Hyperplanes can be constructed in high-dimensional or infinite-dimensional spaces [32,33]. Intuitively, separation is achieved by the hyperplane with the largest distance to the nearest training data point in the class—the so-called margin-maximizing hyperplane. In this study, we implemented it in scikit-learn support vector classification (SVC) [34]. The main objective of SVM is to find the hyperplane in the feature space that can easily separate the classes [19], which needs to solve the following constraint problem:

$$\max \alpha \sum_{i=1}^{n} \alpha_i \ - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, \ x_j) \tag{2}$$

subject to

$$\sum_{i=1}^{n} {y_i}^T \alpha_i = 1, \ 0 \le \alpha_i \le C, i = 1, \ \ldots, \ n \ \forall \ i = 1, 2, 3, ..., n \tag{3}$$

The final discriminate function takes the following form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \quad (4)$$

where $b$ is the bias term.

SVM can be used for both binary classification problems and multiclass classification problems [35]. For multiclass problems, there are two different approaches: (i) one-vs-one approach and (ii) one-vs-all approach. In the case of this study, we have used a one-vs-all approach.

## 4. Results and Discussion

In this study, we used SVM with two kernels, either linear or radial basis function (RBF), for classification. These kernels had some additional parameters called hyperparameters. We tuned these hyperparameters using the grid search method. In the current study, we performed two experiments: (i) text-dependent and (ii) text-independent. The details are explained in the following sections.

### *4.1. Selected Features Using SFFS*

In the current study, we adopted SFFS to identify the potential features for writer identification. We selected 30 out of 40 features using SFFS. The list of selected features is presented in Table 3.

**Table 3.** List of selected features using SFFS.

| SN | Features | SN | Features | SN | Features |
|---|---|---|---|---|---|
| 1 | Accel. mean | 11 | Altitude mean | 21 | Altitude std |
| 2 | Azimuth mean | 12 | Azimuth std | 22 | Altitude mean first accel. mean |
| 3 | First azimuth mean | 13 | First pressure mean | 23 | First pressure std |
| 4 | First speed mean | 14 | First speed std | 24 | Last accel. mean |
| 5 | Last accel. std | 15 | Last altitude mean | 25 | Last azimuth mean |
| 6 | Last speed mean | 16 | Last pressure mean | 26 | Last pressure std |
| 7 | Last azimuth mean | 17 | Last speed std | 27 | Max accel. |
| 8 | Max pressure | 18 | Negative pressure change mean | 28 | Negative pressure changes std |
| 9 | Pressure mean | 19 | Positive pressure change mean | 29 | Positive pressure changes std |
| 10 | Pressure std | 20 | Speed mean | 30 | Speed std |

### *4.2. Text-Dependent*

In the text-dependent type, we applied SVM for all 72 Kanji characters. We performed five-fold cross-validation for each character and computed the identification accuracy. The identification accuracy of SVM for each character is presented in Table 4. The identification accuracy of SVM per character ranged from 84.2% to 99.2%. SVM gave the highest accuracy of 99.2% for "避" and the lowest accuracy of 84.2% for "人". Compared with the characters with high and low accuracy, the characters with many strokes had high accuracy and the characters with few strokes had low accuracy.

We also completed an experiment by taking more than one character (called connected characters) at a time and trying to show the performance of SVM for writer identification. Table 5 shows the identification accuracy of SVM for two, three, and four connected characters. In order to make two, three, and four connected characters, we selected the top two (避and 担), three (避, 担, and 甫), and four characters (避, 担, 甫, and 還) from Table 4. SVM provided 99.4% and 99.6% identification accuracy for connecting two and three characters, respectively.

**Table 4.** Identification accuracy (in %) of SVM for each character.

| SN | Kanji | ACC | SN | Kanji | ACC | SN | Kanji | ACC | SN | Kanji | ACC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 避 | 99.2 | 19 | 臓 | 97.4 | 37 | 衛 | 96.6 | 55 | 甘 | 95.9 |
| 2 | 担 | 98.6 | 20 | 岡 | 97.2 | 38 | 右 | 96.5 | 56 | 鯛 | 95.7 |
| 3 | 甫 | 98.6 | 21 | 湖 | 97.1 | 39 | 肇 | 96.5 | 57 | 憲 | 95.7 |
| 4 | 還 | 98.3 | 22 | 旬 | 97.1 | 40 | 委 | 96.3 | 58 | 響 | 95.4 |
| 5 | 虚 | 98.1 | 23 | 完 | 97.1 | 41 | 込 | 96.3 | 59 | 麗 | 95.4 |
| 6 | 鐵 | 98.1 | 24 | 崩 | 96.9 | 42 | 末 | 96.3 | 60 | 蹟 | 95.3 |
| 7 | 旭 | 98.0 | 25 | 車 | 96.9 | 43 | 凰 | 96.3 | 61 | 圏 | 95.1 |
| 8 | 魅 | 98.0 | 26 | 鞠 | 96.9 | 44 | 雪 | 96.2 | 62 | 藁 | 95.1 |
| 9 | 鳥 | 98.0 | 27 | 性 | 96.9 | 45 | 方 | 96.2 | 63 | 名 | 95.1 |
| 10 | 柿 | 97.8 | 28 | 釉 | 96.9 | 46 | 石 | 96.2 | 64 | 台 | 94.7 |
| 11 | 惹 | 97.7 | 29 | 亜 | 96.9 | 47 | 羊 | 96.2 | 65 | 四 | 94.5 |
| 12 | 讃 | 97.5 | 30 | 巣 | 96.9 | 48 | 曲 | 96.2 | 66 | 士 | 94.3 |
| 13 | 君 | 97.5 | 31 | 函 | 96.8 | 49 | 区 | 96.0 | 67 | 今 | 93.6 |
| 14 | 鋒 | 97.5 | 32 | 闌 | 96.8 | 50 | 寵 | 96.0 | 68 | 升 | 93.1 |
| 15 | 策 | 97.4 | 33 | 忘 | 96.8 | 51 | 機 | 96.0 | 69 | 川 | 92.5 |
| 16 | 歯 | 97.4 | 34 | 渡 | 96.6 | 52 | 禾 | 96.0 | 70 | 上 | 92.5 |
| 17 | 武 | 97.4 | 35 | 憾 | 96.6 | 53 | 刊 | 95.9 | 71 | 入 | 91.2 |
| 18 | 費 | 97.4 | 36 | 滝 | 96.6 | 54 | 彌 | 95.9 | 72 | 人 | 84.2 |

**Table 5.** Identification accuracy (in %) of connected characters.

| Kanji | ACC | Cost | Gama | Kernel |
|---|---|---|---|---|
| 避担 | 99.4 | 0.1 | 0.00001 | Linear |
| 避担甫 | 99.6 | 10 | 0.01 | RBF |
| 避担甫還 | 99.4 | 0.1 | 0.00001 | Linear |

### *4.3. Text-Independent*

For text-independent, we split the dataset into two sets as training and test sets. We took 70% of the dataset for training and another 30% for the test set. In this section, we used the top 5, top 10, top 15, top 20, top 25, top 30, top 35, top 40, top 45, top 50, top 55, top 60, top 65, top 70, and all 72 characters to show how the identification accuracy of SVM changes depending on whether features are selected or not over characters. Here, the person can write any combination of the top 5, top 10, etc. characters in order to make it text-independent writer identification. Table 6 shows the identification accuracy of SVM without SFFS (SVM-WO-SFFS) and SVM with SFFS (SVM-W-SFFS). It was observed that SVM-WO-SFFS provided 96.2% identification accuracy for the top 5 characters, whereas SVM-W-SFFS also provided 99.0% identification accuracy. It was also observed that the identification accuracy of SVM-W-SFFS decreased (99.0% to 94.3%) when increasing the number of characters.

### *4.4. Comparison of Our Proposed Identification System and Similar Existing Studies*

In this section, we made a comparison between our proposed writer identification system against similar existing studies. This comparative study is shown in Table 7, which presents various parameters such as author names, data type (DT), study type, sample size, classification methods, language, number of characters and writers, number of selected characters and writers, and identification accuracy rate. Some existing studies were conducted for writer identification using Kanji characters as follows:

**Table 6.** Identification accuracy (in %) of Kanji characters used for each of the five letters.

| SN | Characters | SVM-WO-SFFS | SVM-W-SFFS |
|---|---|---|---|
| 1 | Top 5 | 96.2 | 99.0 |
| 2 | Top 10 | 96.5 | 97.7 |
| 3 | Top 15 | 96.0 | 97.0 |
| 4 | Top 20 | 96.2 | 97.4 |
| 5 | Top 25 | 95.6 | 97.1 |
| 6 | Top 30 | 96.1 | 96.7 |
| 7 | Top 35 | 95.6 | 96.9 |
| 8 | Top 40 | 95.5 | 96.9 |
| 9 | Top 45 | 95.5 | 96.3 |
| 10 | Top 50 | 95.3 | 96.1 |
| 11 | Top 55 | 94.7 | 95.9 |
| 12 | Top 60 | 94.8 | 95.6 |
| 13 | Top 65 | 94.2 | 95.3 |
| 14 | Top 70 | 94.1 | 95.0 |
| 15 | All 72 | 93.9 | 94.3 |

**Table 7.** Comparison of our proposed system against similar existing studies.

| Authors | Data Types | Study Types | Samples | Methods | Language | Characters | Writers | ACC (%) |
|---|---|---|---|---|---|---|---|---|
| Nakamura et al. [9] | Online | Text-independent | 1230 | Based on Distance | Kanji | 4 | 41 | 96.5 |
| Soma and Arai [24] | Offline | Text-dependent | 50,000 | Based on Distance | Kanji | 100 | 100 | 95.2 |
| Soma et al. [11] | Offline | Text-dependent | 50,000 | Majority Voting | Kanji | 100 | 100 | 99.0 |
| Nguyen et al. [17] | Offline | Text-independent | 2965 | CNN | Kanji | 100 | 400 | 93.8 |
| Our proposed | Online | Text-independent | 47,520 | SVM | Kanji | 72 | 33 | 99.0 |
| | | Text-dependent | 47,520 | SVM | Kanji | 72 | 33 | 99.6 |

Nakamura et al. [9] collected a total of 1230 samples using pen tablets from 41 subjects (males: thirty-five; females: six). They asked each of the subjects to write only four Kanji characters five times, and repeated this six times $(5 \times 6 \times 41)$. They extracted 563 features and evaluated their discriminative power using one-way ANOVA and Kruskal–Wallis test. They separated the writers in feature space based on distance and achieved an identification accuracy of 90.4% for all 563 features; an identification accuracy of 96.5% was also obtained for 270 selected features.

Soma and Arai [24] also proposed a recognition system for text-dependent writer identification except using character recognition features. They obtained five features, such as start and end points of both x and y coordinates, and the angle of the stroke from each stroke. They conducted their study on 50,000 samples $(100 \times 100 \times 50)$, which had 100 Kanji characters, 100 subjects, and 50 samples of each character's class for a writer. They used 5000 samples (100 writers $\times$ 50 samples) for the experiment of each character class. They used 4999 samples for the dictionary and one sample for validation. Then, Euclidian distance was used to determine the writers and achieved an identification accuracy of 99.6%, 97.0%, and 95.2% for 10 writers, 50 writers, and 100 writers, respectively. Here, the authors obtained good identification accuracy when they considered only 10 writers.

Soma et al. [11] conducted another study on the same database [24] and proposed efficient character features and a recognition system for only text-dependent writer identification. They extracted various local and global features from each character. These extracted features were fed into a majority voting approach with a leave-one cross-validation protocol for writer identification and obtained an identification accuracy rate of 99.0% for three-character classes. Nguyen et al. [17] proposed a CNN-based approach for offline text-independent writer identification. They illustrated that their proposed method

obtained an identification accuracy of 99.9% for 200 characters and 100 writers; 92.8% identification accuracy for 50 characters and 100 writers; and 93.8% identification accuracy for 100 characters and 400 writers. Although the authors achieved good identification accuracy, it required writing 200, 50, or 100 characters to identify a person, which is very time-consuming and laborious.

As mentioned, the above studies were conducted for writer identification using online or offline handwritten Kanji characters based on only text-independent [9,17] or text-dependent methods [11,24] (Soma et al., 2013; Soma et al., 2014). Moreover, Nakamura et al. [9] conducted their study on only four online Kanji characters. As shown in Table 7, we used both text-dependent and text-independent writer identification methods in this work. Our experimental results demonstrated that SVM achieved the highest identification accuracy of 99.6% for only the three connected characters text-independent and 99.0% identification accuracy for only the top five characters text-dependent, which are shown in Table 7.

## 5. Conclusions and Future Work Direction

In this study, we proposed an efficient method for writer identification using online handwritten-based Kanji characters. The text-dependent and text-independent writer identification methods were used. First, we extracted different features from handwriting data; then, we applied SVM for writer identification. Parameter tuning was performed to increase the identification accuracy of SVM for writer identification. SVM provided the highest identification accuracy of 99.0% for the text-independent and 99.6% for the text-dependent methods. Furthermore, the characters with the highest accuracy were not simple characters with few strokes, such as "人" and "入", but characters with distinctive strokes, such as "避" and "担". As a result of connecting the top characters with high accuracy, we were able to efficiently obtain a high-performing system. The results of the experiment showed that some of these features, such as the relative position of the start of writing, the length of writing, and the tilt of the pen and its movement, are significant enough to identify and verify the writer. We hope that our proposed method will be helpful in various applications, such as providing evidence to a forensic expert in identifying the writer or verifying and authenticating the writer in a court of law or bank. In the future, we will extend this work by adding more subjects or users and use a deep-learning-based approach to identify writers.

**Author Contributions:** Conceptualization, J.S., M.A.M.H., M.M. and M.A.M.H.; software, M.M. and M.A.M.H.; validation, J.S. and M.A.M.H.; formal analysis, M.M. and M.A.M.H.; investigation, J.S. and M.A.M.H.; resources, J.S.; data curation and collection, J.S., M.A.M.H. and M.M.; writing—original draft preparation, J.S., M.M. and M.A.M.H.; writing—review and editing, M.M. and M.A.M.H.; visualization, M.M. and M.A.M.H.; supervision, J.S. and M.A.M.H.; project administration, J.S.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (KAKENHI), Japan (Grant Numbers JP20K11892).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest for this research.

## References


1. Shin, J.; Hasan, M.A.M.; Maniruzzaman, M.; Megumi, A.; Suzuki, A.; Yasumura, A. Online Handwriting Based Adult and Child Classification using Machine Learning Techniques. In Proceedings of the 2022 IEEE 5th Eurasian Conference on Educational Innovation (ECEI), Taipei, Taiwan, 10–12 February 2022; pp. 201–204.
2. Shin, J.; Maniruzzaman, M.; Uchida, Y.; Hasan, M.; Mehedi, A.; Megumi, A.; Suzuki, A.; Yasumura, A. Important features selection and classification of adult and child from handwriting using machine learning methods. *Appl. Sci.* **2022**, *12*, 5256. [CrossRef]
3. Huang, X.; Xiang, Y.; Chonka, A.; Zhou, J.; Deng, R.H. A generic framework for three-factor authentication: Preserving security and privacy in distributed systems. *IEEE Trans. Parallel Distrib. Syst.* **2010**, *22*, 1390–1397. [CrossRef]
4. Abdi, M.N.; Khemakhem, M. Off-Line Text-Independent Arabic Writer Identification using Contour-Based Features. *Int. J. Signal Image Process.* **2010**, *1*, 4–11.
5. Adak, C.; Chaudhuri, B.B. Writer identification from offline isolated Bangla characters and numerals. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 486–490.
6. Rehman, A.; Naz, S.; Razzak, M.I.; Hameed, I.A. Automatic visual features for writer identification: A deep learning approach. *IEEE Access* **2019**, *7*, 17149–17157. [CrossRef]
7. Shin, J.; Liu, Z.; Kim, C.M.; Mun, H.J. Writer identification using intra-stroke and inter-stroke information for security enhancements in P2P systems. *Peer-Netw. Appl.* **2018**, *11*, 1166–1175. [CrossRef]
8. Tan, G.X.; Viard-Gaudin, C.; Kot, A.C. Automatic writer identification framework for online handwritten documents using character prototypes. *Pattern Recognit.* **2009**, *42*, 3313–3323. [CrossRef]
9. Nakamura, Y.; Kidode, M. Individuality analysis of online kanji handwriting. In Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, Korea, 29 August–1 September 2005; pp. 620–624.
10. Nasuno, R.; Arai, S. Writer identification for offline japanese handwritten character using convolutional neural network. In Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing, Hawaii, HI, USA, 7–12 September 2017; pp. 94–97.
11. Soma, A.; Mizutani, K.; Arai, M. Writer identification for offline handwritten Kanji characters using multiple features. *Int. J. Inf. Electron. Eng.* **2014**, *4*, 331–336. [CrossRef]
12. Grębowiec, M.; Protasiewicz, J. A neural framework for online recognition of handwritten kanji characters. In Proceedings of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Poznan, Poland, 9–12 September 2018; pp. 479–483.
13. Namboodiri, A.; Gupta, S. Text independent writer identification from online handwriting. In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, France, 23–26 October 2006; pp. 287–292.
14. Li, B.; Sun, Z.; Tan, T. Hierarchical shape primitive features for online text-independent writer identification. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 986–990.
15. Wu, Y.; Lu, H.; Zhang, Z. Text-independent online writer identification using hidden markov models. *IEICE Trans. Inf. Syst.* **2017**, *100*, 332–339. [CrossRef]
16. Khalid, S.; Naqvi, U.; Siddiqi, I. Framework for human identification through offline handwritten documents. In Proceedings of the 2015 International Conference on Computer, Communications, and Control Technology (I4CT), Kuching, Sarawak, Malaysia, 21–23 April 2015; pp. 54–58.
17. Nguyen, H.T.; Nguyen, C.T.; Ino, T.; Indurkhya, B.; Nakagawa, M. Text-independent writer identification using convolutional neural network. *Pattern Recognit. Lett.* **2019**, *121*, 104–112. [CrossRef]
18. 京相雅樹. その他の生体特徴による個人認証. 生体医工学 **2006**, *44*, 47–53.
19. Dargan, S.; Kumar, M.; Garg, A.; Thakur, K. Writer identification system for pre-segmented offline handwritten Devanagari characters using k-NN and SVM. *Soft Comput.* **2020**, *24*, 10111–10122. [CrossRef]
20. Bensefia, A.; Djeddi, C. Feature's Selection-Based Shape Complexity for Writer Identification Task. In Proceedings of the 2020 International Conference on Pattern Recognition and Intelligent Systems, Athens, Greece, 30 July–2 August 2020; pp. 1–6.
21. Bulacu, M.; Schomaker, L. Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 701–717. [CrossRef] [PubMed]
22. Saranya, K.; Vijaya, M. Text dependent writer identification using support vector machine. *Int. J. Comput. Appl.* **2013**, *65*, 1–6.
23. Thendral, T.; Vijaya, M.; Karpagavalli, S. Analysis of Tamil character writings and identification of writer using Support Vector Machine. In Proceedings of the 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 8–10 May 2014; pp. 1407–1411.
24. Soma, A.; Arai, M. Writer identification for offline handwritten Kanji without using character recognition features. In Proceedings of the 2013 International Conference on Information Science and Technology Applications (ICISTA-2013), Macau, China, 17–19 June 2013; pp. 96–98.
25. Semma, A.; Hannad, Y.; Siddiqi, I.; Djeddi, C.; El Kettani, M.E.Y. Writer identification using deep learning with fast keypoints and harris corner detector. *Expert Syst. Appl.* **2021**, *184*, 115473. [CrossRef]
26. 青木隆浩. バイオメトリクス. 映像情報メディア学会誌 **2016**, *70*, 307–312. [CrossRef]
27. Drotár, P.; Mekyska, J.; Rektorová, I.; Masarová, L.; Smékal, Z.; Faundez-Zanuy, M. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artif. Intell. Med.* **2016**, *67*, 39–46. [CrossRef]

28. Diaz, M.; Moetesum, M.; Siddiqi, I.; Vessio, G. Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and BiGRUs. *Expert Syst. Appl.* **2021**, *168*, 114405. [CrossRef]
29. Muramatsu, D.; Matsumoto, T. Effectiveness of pen pressure, azimuth, and altitude features for online signature verification. In *International Conference on Biometrics*; Springer: Berlin, Germany, 2007; pp. 503–512.
30. Pudil, P.; Novovičová, J.; Kittler, J. Floating search methods in feature selection. *Pattern Recognit. Lett.* **1994**, *15*, 1119–1125. [CrossRef]
31. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw.* **2018**, *3*, 638–640. [CrossRef]
32. Jan, S.U.; Lee, Y.D.; Shin, J.; Koo, I. Sensor fault classification based on support vector machine and statistical time-domain features. *IEEE Access* **2017**, *5*, 8682–8690. [CrossRef]
33. Hasan, M.A.M.; Nasser, M.; Pal, B.; Ahmad, S. Support vector machine and random forest modeling for intrusion detection system (IDS). *J. Intell. Learn. Syst. Appl.* **2014**, *2014*, 45–52. [CrossRef]
34. Nelli, F. Machine Learning with scikit-learn. In *Python Data Analytics*; Springer: Berlin, Germany, 2018; pp. 313–347.
35. Hsu, C.W.; Lin, C.J. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425. [PubMed]