



## Article

# User Authentication Based on Handwriting Analysis of Pen-Tablet Sensor Data Using Optimal Feature Selection Model <sup>†</sup>

Nasima Begum <sup>1</sup>, Md Azim Hossain Akash <sup>1</sup>, Sayma Rahman <sup>1</sup>, Jungpil Shin <sup>2</sup>, Md Rashedul Islam <sup>1,\*</sup> and Md Ezharul Islam <sup>3,\*</sup>

<sup>1</sup> Department of Computer Science and Engineering, University of Asia Pacific, Dhaka 1216, Bangladesh; nasima.cse@uap-bd.edu (N.B.); md.azimhossainakash@gmail.com (M.A.H.A.); saymamrahman@gmail.com (S.R.)

<sup>2</sup> School of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu, Fukushima 965-8580, Japan; jpshin@u-aizu.ac.jp

<sup>3</sup> Department of Computer Science and Engineering, Jahangirnagar University, Dhaka 1342, Bangladesh  
\* Correspondence: rashed.cse@gmail.com (M.R.I.); ezharul.islam@juniv.edu (M.E.I.)

<sup>†</sup> This paper is an extended version of “User Authentication Through Pen Tablet Data Using Imputation and Flatten Function” published in the Proceedings of 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), Kaohsiung, Taiwan, 21–23 August 2020.



**Citation:** Begum, N.; Akash, M.A.H.; Rahman, S.; Shin, J.; Islam, M.R.; Islam, M.E. User Authentication Based on Handwriting Analysis of Pen-Tablet Sensor Data Using Optimal Feature Selection Model. *Future Internet* **2021**, *13*, 231. <https://doi.org/10.3390/fi13090231>

Academic Editor:  
Efsthios Stamatatos

Received: 1 August 2021  
Accepted: 29 August 2021  
Published: 6 September 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Handwriting analysis is playing an important role in user authentication or online writer identification for more than a decade. It has a significant role in different applications such as e-security, signature biometrics, e-health, gesture analysis, diagnosis system of Parkinson’s disease, Attention-deficit/hyperactivity disorders, analysis of vulnerable people (stressed, elderly, or drugged), prediction of gender, handedness and so on. Classical authentication systems are image-based, text-dependent, and password or fingerprint-based where the former one has the risk of information leakage. Alternatively, image processing and pattern-analysis-based systems are vulnerable to camera attributes, camera frames, light effect, and the quality of the image or pattern. Thus, in this paper, we concentrate on real-time and context-free handwriting data analysis for robust user authentication systems using digital pen-tablet sensor data. Most of the state-of-the-art authentication models show suboptimal performance for improper features. This research proposed a robust and efficient user identification system using an optimal feature selection technique based on the features from the sensor’s signal of pen and tablet devices. The proposed system includes more genuine and accurate numerical data which are used for features extraction model based on both the kinematic and statistical features of individual handwritings. Sensor data of digital pen-tablet devices generate high dimensional feature vectors for user identification. However, all the features do not play equal contribution to identify a user. Hence, to find out the optimal features, we utilized a hybrid feature selection model. Extracted features are then fed to the popular machine learning (ML) algorithms to generate a nonlinear classifier through training and testing phases. The experimental result analysis shows that the proposed model achieves more accurate and satisfactory results which ensure the practicality of our system for user identification with low computational cost.

**Keywords:** user authentication; handwriting analysis; optimal feature; feature selection; machine learning; SVM; F-1 score; sensor data; SFFS

## 1. Introduction

In the modern age of information technology, user authentication is an important process for information security and IoT-based systems. User authentication is a useful technique to keep one’s networks secured by permitting only authenticated users to access one’s protected resources which may include databases, computer systems, websites, and other network-based services or applications. It plays an important role in the user’s

own authorized account which is only accessible by his or her in a protective manner. The authentication methods can be structured into three prime classes: knowledge-based, token-based, and biometric-based. Password-based user authentication is very common but not robust enough. Several researchers investigated user authentication processes based on Keystroke, Hand Geometry, Finger Print, Iris, Face, DNA, Voice, and others. However, every individual's handwriting has distinguishable attributes, which can be used to uniquely authenticate a person both offline or online.

In recent years, handwriting recognition and person identification has become a potential field of research. In handwriting recognition system, it takes the input from a touch screen, electronic pen, scanner, images, and paper documents to read, elaborate, and create features of the handwritten characters. Every individual has his or her own individuality in handwriting style. This reality is exercised in various applications like signature verification [1,2] or person identification [3,4]. Personal identification based on handwriting is a unique approach that has a variety of potential applications like security purposes, financial activities, forensic activities, demography-based writer authentication, and archaeological authentication (e.g., to identify ancient document writers). Handwriting can also be used to identify the author's identity relating to age, gender, handedness, and ethnicity. This process requires feature extraction from the parameters such as writing speed, pressure, direction, duration, height, width, slant angle, etc. Hence, designing a handwriting recognition system considering all sorts of handwriting styles with a high range of accuracy is challenging. Several approaches have been used for this purpose like Structural Methods, Convolutional Neural Networks, Syntactic Methods, Back Propagation, K-Nearest Neighbor Algorithm, Support Vector Machine (SVM), etc.

Most of the existing works in this area are image-based which focused on image pre-processing techniques and so on. In an image-based system, the quality of images depends on lights, effects, reflection, image quantization, brightness, camera attributes, camera frames, and image quality. Hence, the system accuracy is affected by the image quality, contrast, lighting, shading, noise suppression, filtering, etc. Depending on image pre-processing, the time complexity and computational processing cost are also high for the existing systems. Moreover, an image-based model does not work with the genuine and exact numerical values of an individual's handwriting data. Direct numerical values of a person's handwriting data provide more genuine and accurate features of an individual user which is crucial for the authentication of a writer.

Therefore, in this research, we concentrate on collecting the more genuine numerical values of the user's handwriting directly from the sensors signal of digital pen-tablet devices which motivates this work. The process of getting direct data values helps the proposed method to get more improve accuracy in the context of pen-tablet sensor data. Hence, in this research, we emphasized more robust and efficient data acquisition and analysis which is invariant of the camera frame, camera positioning, image quality, and image pre-processing techniques. The proposed model is constructed based on the user's pen tablet handwriting data analysis which automatically stores the numeric values of the user's handwriting attributes from the pen-tablet sensor signals. In the proposed system, since the data values are directly extracted from the sensor signal of digital pen and tablet devices, it ensures the robustness of the identification of handwriting data.

In this research, we emphasize real-time handwritten sensor data analysis of individuals by using pen tablet devices. In the proposed user authentication model, firstly, six completely separate parameters of handwritten data are collected from the sensor signals of digital pen-tablet devices. Before feature extraction, data pre-processing has been done. In the feature extraction phase, several discriminant kinematic and statistical features are extracted from the parameters of handwriting data. However, all the features extracted from high dimensional sensor data are not equally important to identify a user. Hence, to keep our trained model focused only on the vital features, we introduce the optimal feature selection model which minimizes the overfitting problem.

Then, the most informative and vital features are selected using the optimal feature selection model. Finally, machine learning techniques are applied for training and testing the selected optimal features for the user authentication process. The proposed system is implemented using different classification algorithms and the experimental analysis shows that we got satisfactory results which ensure the practicality of our system with low implementation cost. Our main contributions in this research work are:

- a. User authentication using motion sensor data of pen-tablet devices.
- b. A quantitative analysis of pen-tablet sensor data using kinematic and statistical features extraction model.
- c. Introduce optimal feature selection model combination of filter-based approach and wrapper approach.
- d. An efficient and robust writer identification model using support vector machine (SVM), logistic regression (LR), and random forest (RF) classifier.

The full paper is organized as follows: the next section provides the literature review. Section 3 explains the proposed system model, data acquisition, data pre-processing, feature extraction, and optimal feature selection process. Section 4 discusses the experimental result analysis. And Section 5 concludes this research works with future scopes. Each section includes the necessary diagrams, tables, and graphical representations for an easy and clear understanding of this research work.

## 2. Literature Review

In this section, the theoretical and methodological approaches related to our work are explained. Multiple works on image and pattern analysis-based handwriting recognition and person identification have been reported. Saini et al. in [5], proposed a three-step biometric authentication model based on sitting, walking, and relaxing patterns using mobile devices. Here, Random Forest and KNN are used for classification. However, this model achieves less than 1% False Acceptance Rate (FAR) and 2.2% EER Equal Error Rate (EER). Several papers based on touchscreen gestures while using mobile devices in scenarios such as document reading, keystroke dynamics, web surfing, or free tasks have been achieved popular impact nowadays [6–8]. In [9], a multistage cascading system to serve the field of offline Arabic handwriting recognition is proposed where deep learning techniques are used. The similarity between characters and inconstancy of the writing styles makes Arabic character recognition more challenging. They utilize Hierarchical Agglomerative Clustering (HAC) technique to form the dataset into partially inter-related clusters. The IFN/ENIT Arabic dataset has been used for their experiment. Their process is divided into three consecutive stages. However, the inter-related clusters represent the database as a big search tree model. A behavioral authentication method for mobile devices based on browsing behaviors is proposed in [10]. They develop a news APP using Web View Library and news API of Extreme Speed Data. Their result depends on the external environment model (EEM), the screen-sliding behavior model (SSBM), and the browsing behavior model (BBM). The average DR of the method in this paper is 86% which needs to be improved. In [11], a novel forensic hand radiograph-based human authentication is proposed using a deep neural network. Three-layered convolutional deep neural network architecture is used for the feature extraction and recognition of hand radiographs using KNN and SVM classifiers. The performance of the cross-validation accuracy is 97.60% for KNN and 99.20% for SVM. However, the system required a large amount of data and high computational cost.

Image Processing Features are widely used at the document level or the paragraph level. Here features are divided into three groups: pen pressure, writing movement, and stroke information. In [12,13], the author used image acquisition, image pre-processing, image segmentation, feature extraction, and classification techniques in their proposed system. But the problem of these systems is lighting, shading effect, noise suppression, etc. which are dependent on image pre-processing techniques. In [14], a method concerning to develop the performance of a palm print-based verification system by integrating hand

geometry features has been proposed. The hand shape and palm print images are used to extract features. The result is then examined for their individual and combined verification performances. This feature extraction process has been done by image pre-processing techniques which are costly. M. Kosugi and T. Suzuki proposed an image-based user authentication method for touch screen devices by using the latest image shot by the user as the pass-image [15]. Their authentication could resist smudge attacks, which is one of the major threats for touch screen devices. However, the security strength of the method is low. Some systems on handwriting recognition are done depending on some image pre-processing techniques like image contrast, lighting, shading, illumination, noise suppression, filtering, etc. [16–18]. Hence, the quality of the image affects the robustness of the data. Image quality also varies depending on pre-processing techniques [19,20]. From [21–26], the research works have been done on handwriting recognition and person identification. These systems are also based on image and pattern analysis which suffers from high computational cost.

The researcher in [27] presents a hybrid feature selection model with a discriminant feature distribution analysis-based feature subset evaluation. This method is focused on online bearing fault diagnosis in induction motors. The proposed model better performance than the state-of-the-art average distance-based approaches by a performance margin of about 5%.

Paper [28] has presented an overview of automatic writer identification modules of an embedded biometric system based on small-scale handwritten data. They evaluated both static and dynamic features and proposed a new feature selection algorithm based on likeness coefficients. The identification result of 95% validates the use of handwriting in embedded personal identification devices. However, this system focused only on the small-scale handwriting samples of single handwritten words. The author considered the minimum distance classifier which is sensitive to the differences in variance among the categories.

In [29], the authors have pointed out future trends and challenges in biometric research on signature and handwriting. Special emphasized the use of handwriting signals not only on biometric traits but also on e-security and e-health. Some challenges are identified which should attract the interest of the research community towards a more secure society.

Chahi et al. in [30,31] proposed classic approaches emphasizing on extracting the desirable features. Authors in [30] proposed a Block Wise Local Binary Count (BW-LBC) operator stimulated by traditional LBP that characterizes multiple histograms. Based on the distribution of the pixels in small blocks, the histograms are generated. They utilized the nearest-neighbor classification using the Hamming distance and presented that their approach is better than the modern approaches. For writer identification, the authors in [31], proposed another classical feature extraction method. Their descriptor illustrates a salient feature for local writing structure and is applied to small connected regions of the sample. These feature maps have been used as inputs for the nearest neighbor classifier to classify the query writer. These two works show that traditional features are still useful for writer identification and superior in some tasks. However, as the authors of this research are resizing the handwriting images, there is a possibility of losing the writers' personalities due to stretching the images.

He and Schomaker have recently proposed two deep architectures [32,33]. The former one represents the handwriting which contains two implicit and explicit information where the explicit information refers to the length, number of characters, and lexical contents of single words. In contrast, the implicit information is the writer's behavioral information that can be used for author identity. The author proposed using both implicit and explicit information so that explicit information can be used along with the implicit features to make extra information. The proposed CNN architecture is based on the AlexNet, including two parallel pathways. However, a challenging issue of this research relies on resizing handwriting patches into  $120 \times 140 \times 1$ , where such an affine transformation can lose the writer's intrinsic information.

In [34], the authors represent a deep learning framework for offline text-independent writer identification based on the conjugation of the deep and traditional features. Proposed deep architecture is an extended version of ResNet, which is done using the auxiliary information of handwriting thickness descriptor (HTD). The HTD computed the thickness of handwriting as an essential and preliminary feature for human handwriting analysis. However, the authors do not resize the handwriting images, rather they suggest to crop patches of the handwriting and propose a descriptor to explain each handwriting image regarding the script thickness. Therefore, their model is a conjugating approach to fill the gap of using classical and modern features and achieves appropriate results. However, for future scopes, more reasonable features and multimodal descriptions can be utilized.

### 3. Proposed Model

In this section, the proposed system model of writer identification is explained in a step-by-step procedure. The proposed system consists of six phases: data acquisition, data analysis and preprocessing, feature extraction, optimal feature selection, and classification. In the data collection phase, handwriting data is collected from the sensor signal of a digital pen tablet device which generates six attributes (Time, Pressure, X-axis, Y-axis, Horizontal angle, and Vertical angle) which represents different unique numerical values for different types of handwriting. To make a balanced format of dataset some functions such as Imputation Function and Flatten Function are utilized. Our handwriting dataset consists of real-time handwritten data and every data is important. After data analysis, it is observed that some data values are missing and some became very high which can be a problem for our machine learning model. Using a direct mean of the data may cause the loss of some attributes of the data such as pressure, angle, etc. Since, missing values of the handwritten data are not ignorable; data imputation has the highest potential to preserves the mean of the non-missing data. In our research, we have utilized mean imputation to replace the missing data values. On the other hand, Flatten function is utilized to convert the 2D data into a one-dimensional array for passing it to the next layer. We flatten the dataset to keep the level same for all the data. It makes our computational process simple and easier. Hence, to enhance our dataset and to make it normalize, we have utilized some pre-processing techniques such as imputation function and flatten function. These pre-processing makes our dataset balanced (ready) for the implementation and experiment. That is why we called it as a balanced format of dataset.

After data pre-processing, different statistical and kinematic features are extracted. The feature selection model is introduced to find the optimal features which are vital for the identification of a user. Three different and most widely used machine learning algorithms are used for classification which are: Support Vector Machine (SVM), Linear Regression (LR), and Random Forest (RF) classifiers.

A general block diagram of the user identification model is shown in Figure 1. The model is divided into two major parts. In this research, the first part represents our main contribution to the experimental analysis. The first part consists of extracting different statistical and kinematic features from the dataset and the optimal feature selection process. The dataset is split into two subsets. The first subset is used to generate different optimal features using a hybrid feature selection process. Several kinematic and statistical features are calculated from the six parameters (time, pressure,  $x$ -axis,  $y$ -axis, horizontal angle, and vertical angle) of sensor signals from the first data subset. In this research, for better accuracy, we have considered both statistical and kinematic features. The proposed model introduces a hybrid feature selection algorithm to find out the important feature sets. In the second phase, training and testing are completed using the second data subset. Two third of the dataset is used for feature selection and training process and one-third of the dataset is used for the testing process. A machine learning technique is deployed to identify the person's handwriting based on the classified optimal features set. We have used SVM, LR, and RF classifiers to compare the accuracy of the classified features in this research and to show the stability of our proposed model.

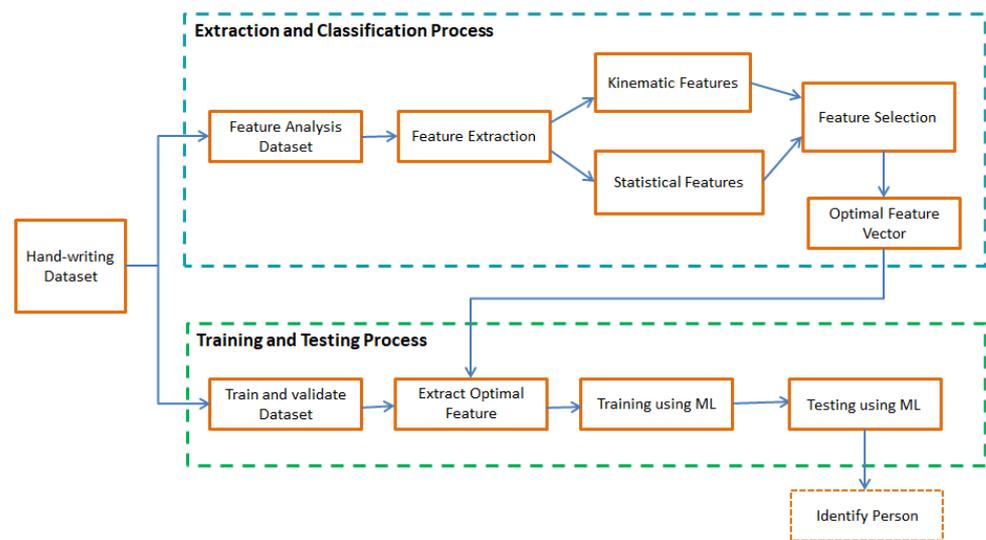


Figure 1. General Architecture of Proposed Model.

### 3.1. Pen Tablet Handwriting Data Collection

In this research, we have collected handwriting samples from different individuals of age 19~40 using digital pen-tablet devices. Initially, we have collected hand-writing samples from 25 different persons both male and female. The sample keywords are written on a plane platform of the tablet within a margin. When the persons are writing on the tablet surface using a digital pen, then an automatic dataset is generated with the corresponding numeric values in an excel sheet. The dataset consists of six parameters (time, pressure,  $x$ -axis,  $y$ -axis, horizontal angle, and vertical angle) which are automatically generated from the pen and tablet sensor signals. Therefore, our dataset is more genuine and robust which contains unique attributes of different person’s handwriting. For collecting the handwriting dataset, we have used Wacom Tablet which is shown in Figure 2.

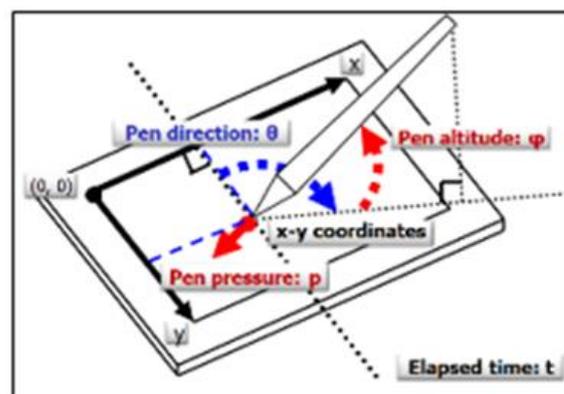


Figure 2. Wacom Tablet Sample Model.

For ensuring the data values of our system, every keyword has been collected 5 times from each person. Initially, 25 persons writing samples are collected. Every person was asked to write 10 defined keywords (Basic Research, Computer Vision, Pattern Processing, Machine Learning, Japan, Tokyo, Fukushima, Aizu University, Hello World, and Thank You) and each of them was repeated 5 times. Thus, 1250 sample data are collected to conduct this research. Since the numeric data values are directly storing in an excel sheet without any filtering or pre-processing, therefore after data analysis, we claim that we have got more genuine and robust data than the existing image-based system. Table 1 shows an example of our dataset which is automatically generated from the sensor signals of pen and tablet devices. The automatic numeric data collection process of handwriting through

sensor signals makes our proposed model significant to get better accuracy and open up a new direction for further research in this field. Additionally, the initial data pre-processing is limited in the proposed system, which does not affect the processing time and hardware cost.

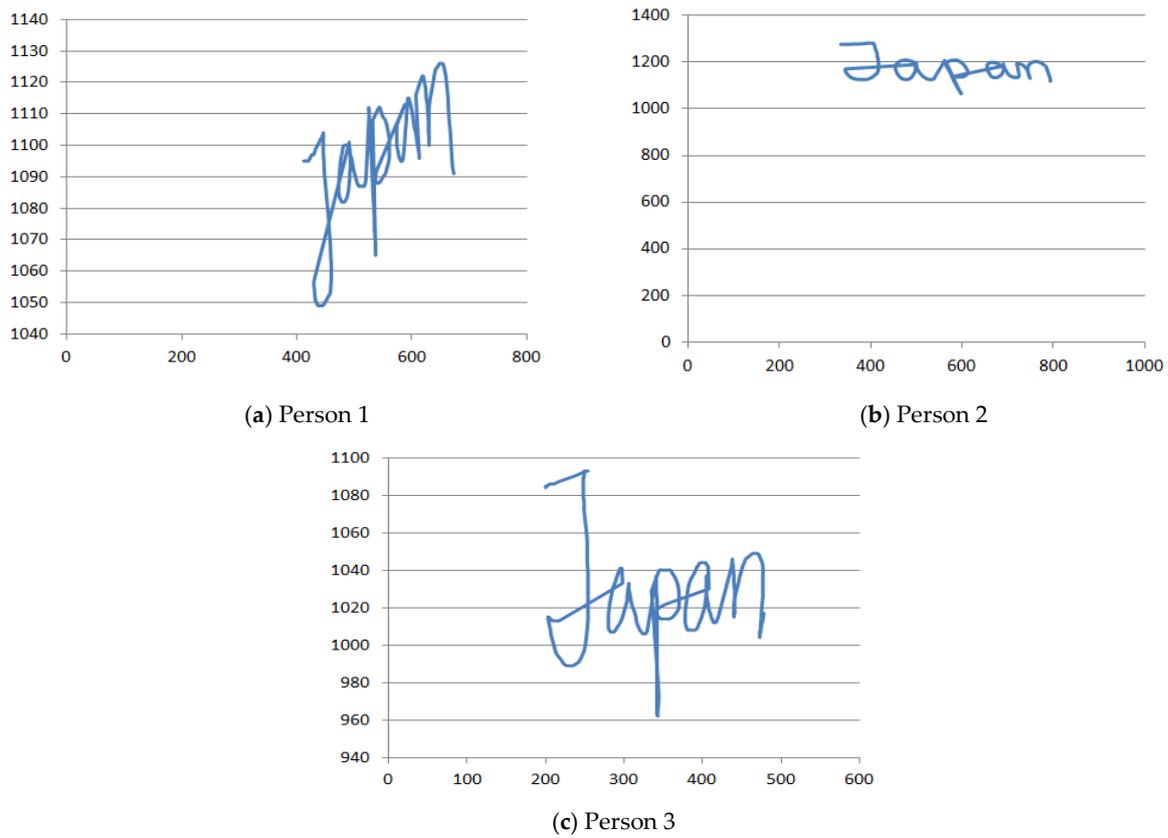
**Table 1.** Sample Dataset.

Time	Pressure	X-Axis	Y-Axis	Horizontal Angle	Vertical Angle
6143.52	6787	346	1214	160	410
6151.46	8707	346	1214	160	420
6159.44	15,671	346	1214	161	420
6166.47	15,847	346	1214	160	430
6174.52	16,391	346	1214	161	430

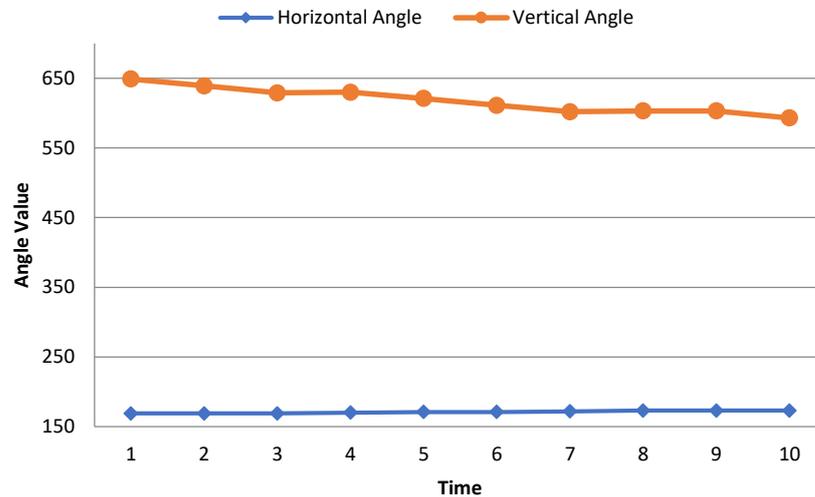
### 3.2. Parameters of Pen Tablet Handwriting Data

In this research work, the dataset has six important parameters or attributes from which different kinematic and statistical features are calculated. The detailed explanations of these attributes are given below:

1. **Writing Time:** Every person consumes a different amount of time for his/her writing depending on the writing speed. Someone writes very slowly, someone writes moderately and someone writes very fast. After analysis, it is found that for those who write very fast, their handwriting style changes like their baseline changed over time such as rising, straight, falling, erratic, etc.
2. **Pen Pressure:** Pen pressure is the most vital attribute of individual handwriting. Pen pressure of every individual is different from each other in terms of heavy or light. In our collected, dataset, the pressure of a person is not the same for each iteration. The pressure range of some person is (2345–6595) and some other person is (1978–22905).
3. **X-axis:** The X-axis represents the writing position of a person from X-axis. We have taken a single keyword 5 times by the same person. It is noticed that the X-axis position value for each of the 5 cases is very close to each other which indicates that the X-axis value remains almost the same for each case for the same person. It ensures that the X-axis value is a promising attribute that can uniquely identify one person from another.
4. **Y-axis:** The Y-axis indicates the writing position of a person from Y-axis. This position is different from each other. Like the X-axis position value, the Y-axis position value remains very similar in each iteration of the same person. The range of Y-axis value for some person is (1110–1201) and that of some other person has (1200–1350). Figure 3 shows  $x$ -axis and  $y$ -axis values for different persons for the same keyword.
5. **Horizontal Angle:** The horizontal angle is the dimension of an angle within two lines, rising from the same spot. This angle is automatically measured by the pen tablet. To uniquely identify a user, the horizontal angle is one of the major attributes of our work. The maximum range of horizontal angles in the data set is below 500. Figure 4 shows that the horizontal angle remains constant over time.
6. **Vertical Angle:** The vertical angles are the opposite angles to each other after passing two lines. This is also a major attribute of this research work to identify a user uniquely. In the data set, the maximum vertical angle is below 800. Figure 4 shows the horizontal and vertical angle of the same person which is kind of constant over time.



**Figure 3.** Writing graph with the x-axis, y-axis values for different persons for the same keyword: (a) Person 1 (b) Person 2 (c) Person 3.



**Figure 4.** Horizontal angle and vertical angle of the same person.

Figure 4 represents how the writing angle (horizontal angle and vertical angle) of a person is changing over time. Writing angle is an attribute of the writer’s handwritten data, this is not a model feature. From Figure 4, we can see that the horizontal and vertical angles of a person remain constant over time which can be considered as a vital attribute for a person. From here, we extracted some kinematic and statistical features such as horizontal velocity, vertical velocity, average writing start velocity, average writing end velocity, etc. Then those extracted features are used in our model. Therefore, there is no direct relation of this attribute (writing angle) with the model.

### 3.3. Handwriting Data Preprocessing

In this section, the dataset preprocessing techniques are described which have been utilized in the proposed method. To get a more balanced dataset and improve efficiency, some preprocessing techniques are applied in the proposed model. In our dataset, data in a row has multiple column attributes and each of the columns is supposed to carry a valid numeric value. However, there are some missing data values in our dataset. Due to the speed difference between the data capture device and data acquisition device the writing signal can be lost which may cause some null values. Because whenever the pen is detached from the tablet there can be a signal lost which may lead to some null values. Therefore, we have utilized the Imputation Function to fill up the missing values in a column attribute. We have filled the null values with the computed mean values of the same column attributes shown in Figure 5.

225	Time	Pressure	X-axis	Y-axis	Horizontal	Vertical
226	5043.37	18632	481	1170	165	460
227	5050.63	18440	481	1169	165	460
228	5058.4	18264	482	1169	NaN	460
229	5066.44	18568	482	1168	165	460
230	5074.57	18680	484	1167	165	460
231	5081.47	18824	NaN	1165	NaN	460
232	5089.49	19000	487	1163	165	460
233	5096.36	NaN	490	1162	165	460
234	5104.55	19801	492	1160	166	460
235	5112.63	20057	495	1158	166	460
236	5119.45	20441	497	1156	166	470
237	5127.57	20601	500	1154	166	470
238	5135.55	20745	502	1153	166	470

(a)

225	Time	Pressure	X-axis	Y-axis	Horizontal	Vertical
226	5043.37	18632	481	1170	165	460
227	5050.63	18440	481	1169	165	460
228	5058.4	18264	482	1169	165	460
229	5066.44	18568	482	1168	165	460
230	5074.57	18680	484	1167	165	460
231	5081.47	18824	485	1165	165	460
232	5089.49	19000	487	1163	165	460
233	5096.36	19384	490	1162	165	460
234	5104.55	19801	492	1160	166	460
235	5112.63	20057	495	1158	166	460
236	5119.45	20441	497	1156	166	470
237	5127.57	20601	500	1154	166	470
238	5135.55	20745	502	1153	166	470

(b)

Figure 5. (a) Before Imputation; (b) After Imputation.

There are various techniques for imputing the missing data values. However, in our research, we have utilized mean imputation to replace the missing data values. This is because; mean imputation is a very simple and popular technique to understand and to apply with the basic knowledge of statistics. It helps us to make our general ML audience to understand our process. It calculates a statistical value for each column (such as a mean) and replaces all missing values for that column with the immediate statistics by calculating the data close to the missing values. Besides, mean imputation keeps the full dataset while not reducing the sample size. And multiple imputations for missing data allow us to obtain good estimates of the standard errors. In addition, our dataset consists of numerical data rather than categorical data, and mean imputation often shows good performance for numerical data values. On the other hand, the ML-based imputation process requires heavy learning process which increases the computational overhead. Thus, to tradeoff the performance as well as computational cost, we used mean imputation in our proposed model.

During the data preprocessing phase, we have also utilized Flatten Function for converting the two-dimensional data into one dimension. Flatten is a useful technique for converting 2D data into 1D. A Flatten function reshapes the dataset to have a shape that is equal to the number of elements contained in the file. It makes our computational process simple and easier to extract the required feature.

### 3.4. Feature Extraction

The detailed feature extraction process is described in this section. The mathematics behind statistics plays a vital role in data analysis. The feature extraction process can

extract some specific parameters from a person’s handwriting that can discriminate it from others. The features are extracted to uniquely identify each individual from his/her handwritten dataset. Handwriting is quite an unstable process, and writing styles also vary. Similarly, there is a great variety of parameters involving environmental, psychological circumstances within the handwriting of one person. For feature extraction, different statistical and kinematic features are extracted from the parameters of pen tablet sensor signals.

### 3.4.1. Statistical Features

Statistics is an effective mathematical technique to perform technical data analysis. Statistics help us to get deeper and meaningful information along with the graphical visualization of data. Every person has unique feature values. Sometimes some feature values are very close but not exactly the same to each other. These features are genuine and unique to identify a writer. Here, we consider both time domain and frequency domain statistical parameters. For this, some statistical feature has been extracted from the sensor signal of handwriting dataset. Table 2 shows total of 20 features along with their corresponding statistical formulas. A brief explanation of these statistical features is given below:

**Table 2.** Statistical Features from Pen-Tablet Sensor Data.

Features	Equation	Features	Equation
Mean	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$	Cross-Correlation	$\rho_{ij} = \frac{X_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}$
Median	$M = \left( \frac{n-cf}{f} \right) (w) + L_m$	Absolute Mean value	$A_m = \mu(abs_x)$
Maximum	MAX(M)	Shape Factor	$SF = \frac{X_{rms}}{A_m}$
Minimum	Minimum MIN(M)	Energy	$E = \sum_{i=1}^N  x_i ^2$
Standard Deviation of Pressure	$P_{sd} = \sqrt{\frac{\sum_{n=1}^N (P_n - \bar{P})^2}{N}}$	RMS Frequency	$RMS_{fr} = \mu \left( \sqrt{(frx_1^2) + frx_2^2 + \dots + frx_n^2} \right)$
Skewness	$P_{skw} = \frac{\sum_{n=1}^N (P_n - \bar{P})^3 / N}{P_{sd}^3}$	Peak to Peak	$PPV = MAX(M) - MIN(M)$
Kurtosis	$P_k = \frac{\sum_{n=1}^N (P_n - \bar{P})^4 / N}{(P_{sd})^4}$	SRA	$SF = \mu \left( \sqrt[2]{abs\bar{X}} \right)$
Variance	$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$	Impulse Factor	$i = \frac{x_{peak}}{x_{mean}}$
Correlation Coefficient	$\frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$	Margin Factor	$MF = x_{peak} / x_{sra}$
Root mean Square	$\sqrt{\frac{X_{rms}}{N} (x_1^2 + x_2^2 + \dots + x_n^2)}$	Energy Center	$FC = \sqrt{f_1 f_2}$

The mean is commonly known as the average which is the total of all the data values divided by the number of data points. We can understand the standard level of attributes from the mean value. Here, mean represents average value for writing attributes such as angle, pressure, axis value, etc. When the data is sorted in ascending or descending order then the middle value is known as median. The median is sometimes used as opposite to the mean when there are outliers in the sequence that might skew the average value.

Standard deviation (SD) is a measure of how a set of data is spread out. A low SD states that the data is closely clustered around the mean or average while a high SD indicates that the data is dispersed over a wider range of values. Standard deviation is used to realize whether a specific data point is standard or unusual.

Skewness is a measure of the degree of asymmetry of a frequency distribution. In our research, its value of skewness became as negative, positive or zero. A positive value indicates right-skewed, a negative value indicates left-skewed and zero-skewed indicates symmetric distribution. It helps to understand on which side of the data we need to work. Skewness is measured by using the average of attributes, and their standard deviation. Kurtosis is a measure of the peakedness of a distribution. Kurtosis generally represents a positive value. It is used as a measure of error or risk. A large kurtosis is represented with a high error and low indicates low error. Variance is used to determine how well the mean represents an entire set of data. In our research, the variances of several attributes are measured such as for pressure, angle, time, etc. Standard deviation and variance are related to each other.

The peak value represents the maximum value of an attribute. Impulse Factor is measured by using the average of attributes and their peak value by using a statistical formula. Margin Factor depends on the peak value and Statistical Reasoning Assessment (SRA). Maximum value is measured for the maximum data boundary level for every attribute. Minimum value is measured for minimum data boundary level for every attribute. These levels aid us to understand the attribute levels. Peak to peak value is measured by subtracting the Maximum attribute value to the Minimum attribute value.

### 3.4.2. Kinematic Features

In our experiment, for better classification, 10 different kinematic features are extracted as shown in Table 3. The X-axis value has been measured by the difference of the X-axis's value by different variations of the period. We measured the x-axis velocity (horizontal velocity) by subtracting the x-axis's value of  $X_n$  from that of  $X_{n-1}$ . The mathematical formula of X-axis velocity is:

$$\text{Horizontal velocity} = \sum_{n=1}^N \frac{x_{n+1} - x_n}{t_{n+1} - t_n} \tag{1}$$

where,  $N$  represents the length of the input data of each person,  $t_n$  is the elapsed time after the start of the test,  $P_n$  is the writing pressure,  $X_n$  is the positional coordinate in the horizontal direction, and  $Y_n$  is the positional coordinate in the vertical direction. The Y-axis velocity (vertical velocity) has been measured from its Y-axis value difference with its time variation at each iteration. The average velocity has been located using the length and square root of the velocity of the X-axis and Y-axis. The mean pressure has been calculated through the mean of writing pressure. The peak pressure has been calculated from the highest pressure which is greater than the mean pressure. The peak pressure is different from person to person.

**Table 3.** Kinematic Features from Sensor Data.

Features	Equation	Features	Equation
Average Pressure	$P_{\text{mean}} = \frac{1}{N} * \sum_{n=1}^N (P_n)$	Peak Pressure	$P_{\text{peak}} = 1 \leq n \leq N_{\text{max}} (P_n)$
Average Velocity	$V_{\text{mean}} = \frac{\sqrt{\{x_n - x_{n+1}\}^2 + \{y_n - y_{n+1}\}^2}}{t_{n+1} - t_n}$	Peak Velocity	$V_{\text{max}} = \max_{1 \leq n \leq N} \frac{\sqrt{\{x_n - x_{n+1}\}^2 + \{y_n - y_{n+1}\}^2}}{t_{n+1} - t_n}$
Average Start Position of Pen	$P_{\text{start}} = \frac{1}{N * 0.05} \sum_{n=1}^{N * 0.05} P_n$	Average End Position of Pen	$P_{\text{End}} = \frac{1}{N * 0.05} \sum_{n=1}^{N * 0.95} P_n$
Average Writing Start Velocity	$V_{\text{start}} = \frac{1}{N * 0.05} \sum_{n=1}^{N * 0.05} \frac{\sqrt{\{x_n - x_{n+1}\}^2 + \{y_n - y_{n+1}\}^2}}{t_{n+1} - t_n}$	Average Writing End Velocity	$V_{\text{End}} = \frac{1}{N * 0.05} \sum_{n=1}^{N * 0.95} \frac{\sqrt{\{x_n - x_{n+1}\}^2 + \{y_n - y_{n+1}\}^2}}{t_n - t_{n+1}}$
Horizontal Velocity	$V_h = \sum_{n=1}^N \frac{x_{n+1} - x_n}{t_{n+1} - t_n}$	Vertical Velocity	$V_v = \sum_{n=1}^N \frac{y_{n+1} - y_n}{t_{n+1} - t_n}$

Writing Velocity is the rate of change of the position of the pen tip over time. The writing velocity has been measured from horizontal direction and vertical direction. The average start position is the rate of change of the start position of the pen tip over time. The average end position is the rate of change of the end position of the pen tip over time. The horizontal velocity and vertical velocity of each person are not changing much i.e., it also remains constant for each individual.

### 3.5. Optimal Feature Selection

In our experiment, 20 statistical and 10 kinematic features are extracted from the handwriting sensor signals for each person. However, all the features are not equally important to identify a user or writer. Thus the training algorithm should be trained focusing only on the important features while ignoring the non-vital features. Hence, in our experiment, we introduced a hybrid feature selection algorithm to reduce the high dimension problem of the original feature set and to achieve higher classification results.

#### 3.5.1. Hybrid Feature Selection Model

The feature selection algorithm consists of two parts. The first part is a filter-based approach. It uses sequential forward floating search (SFFS) where the feature subset evaluation process does not depend on the classification algorithm. The second part is a wrapper-based approach. This part takes the feature sets derived from the filter-based approach as input and evaluates the best feature sets among them.

The whole process of optimal feature selection is shown in Figure 6. The filter-based feature analysis is an iterative approach that is conducted by SFFS. Initially, 25 users' first and second datasets are applied for the training and validation respectively. The dataset is again randomly splitted into two equal parts for the training and testing process. The first equal part is then again divided into two equal parts to use in the filter-based approach. For testing in the wrapper approach, the second equal part is used. SFFS is applied in the first equal part where the process iterates five times and brings out the 10 sub-optimal feature sets. This process is repeated for the rest of the sensor data. Then the selected optimal feature sets are forwarded to the wrapper method. The wrapper method analyzes different optimal feature sets by using different machine learning algorithms to evaluate the best features.

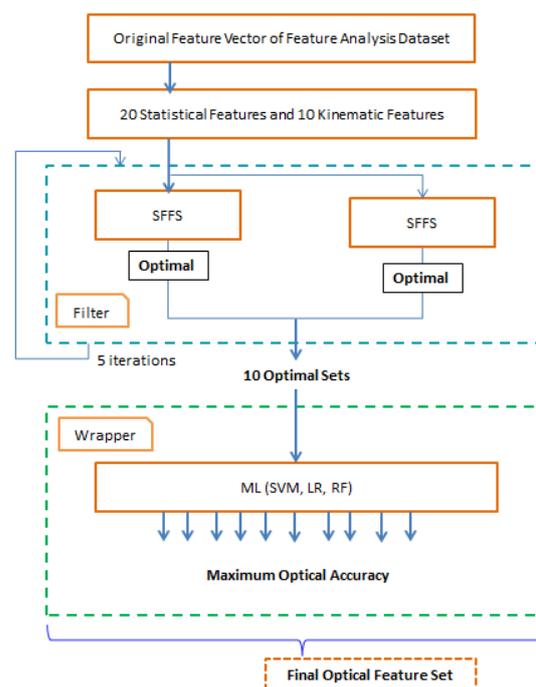


Figure 6. Block diagram of the optimal feature selection process.

The SFFS algorithm (Algorithm 1) that we utilized in our algorithm is given below:

---

**Algorithm 1: SFFS Algorithm**

---

Input: The set of all features  $X = \{x_1, x_2, \dots, x_n\}$   
 Output: A subset of features  $Y = \{y_i \mid i = 1, 2, 3, \dots, n; y_i \in X\}$   
 Where,  $n = (0, 1, 2, \dots, m)$   
 Steps:  
 1.  $X_0 = \{\emptyset\}$   
 2. Select the best feature  $Y^+$   
 Update:  $X_{N+1} = X_N + Y^+ = +1$   
 3. Select the best feature  $Y^-$   
 4. If  $I(X_N - Y^-) > I(X_N)$ ; [ $I(Y) = \text{criterion function}$ ]  
 Then,  $X_{N+1} = X_N - Y^-$ ;  $N = N + 1$   
 Go to step 3.

---

For training and testing of the second equal part of the dataset, the SVM algorithm is applied in the wrapper-based approach. Then the 10 sub-optimal features which are obtained from the filter-based approach are forwarded to the SVM machine learning algorithm to find the best feature sets in the wrapper approach. To show the stability of our proposed model of writer identification, we have done the feature selection procedure and the classification process using two additional machine learning algorithms as logistic regression (LR) and random forest (RF) classifier. For that, we repeated the above procedure for feature selection with LR and RF, and also the evaluation is done using the same classifier algorithm. From the experimental evaluation, it is observed that the performance of our proposed model with LR and RF is also satisfactory which makes our claim stronger. However, among the 3 machine learning algorithms (SVM, LR, and RF), SVM gives us the best feature sets from the 10 optimal feature sets in terms of classification accuracy.

### 3.5.2. Objective Function Based on Discriminant Feature

In this research, the proposed hybrid feature selection approach is carried out in two parts, i.e., filter approach and wrapper approach. In the wrapper approach, the evaluation process is based on a classification model, which is very much efficient but computationally expensive. Thus, the filter approach is used before the wrapper approach to minimize the cost and optimized the performance. The filter approach selects some optimal subsets of features using SFFS with the help of an efficient feature subset evaluation model. In this research, we adopted Discriminant Feature Distribution Analysis Based Objective Function proposed in [27] as a feature subset evaluation model. In this evaluation model, the power of discrimination of features are measured by calculating the within-class compactness and between-class distance. The distance measurement has a significant role in feature analysis and machine learning techniques. There are several criteria to measure distribution of features, i.e., distance, correlation, entropy, information gain, and others. However, distance measurement is one of the most used techniques in data science and machine learning. Several machine learning and classification models are depending on distance measurement, i.e., k-NN, k-means, self-organizing map (SOM). Many kernel-based algorithms also use distance measurement, such as, SVM uses the distance measurement for classification. Also, the distance measurement expresses the geographical distribution in high dimension feature space efficiently. Among the several distance measurements, Euclidean distance is mostly used technique. Thus, in this measurement process, the Euclidean distance is considered to evaluate the distance of samples in a class and/or other classes. The Euclidean distance is formulated as Equation (2), where  $x$  and  $y$  are two points,  $n$  is the number of features, and  $x_i$  and  $y_i$  are features values of points.

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{t=1}^n (x_t - y_t)^2} \quad (2)$$

To calculate the within-class compactness, firstly, the class median is determined based on the distances of all points of a class. After that, the maximum distance is calculated from the class medium to the farthest point of a class. Finally, within-class compactness is calculated by averaging all class compactness of all classes. To calculate the between-class

distance, the minimum distance between two classes is calculated from all distances between each point from one class to each point to the other class. The minimum distance of two classes is considered as class separability. The overall between-class distance is calculated by averaging all class separability of different classes. Finally, the evaluation/objective value is calculated as Equation (3). Figure 7 presents the process of calculation of objective value.

$$objectiv\_valu = \frac{between\_class\_distance}{within\_class\_compactness} \tag{3}$$

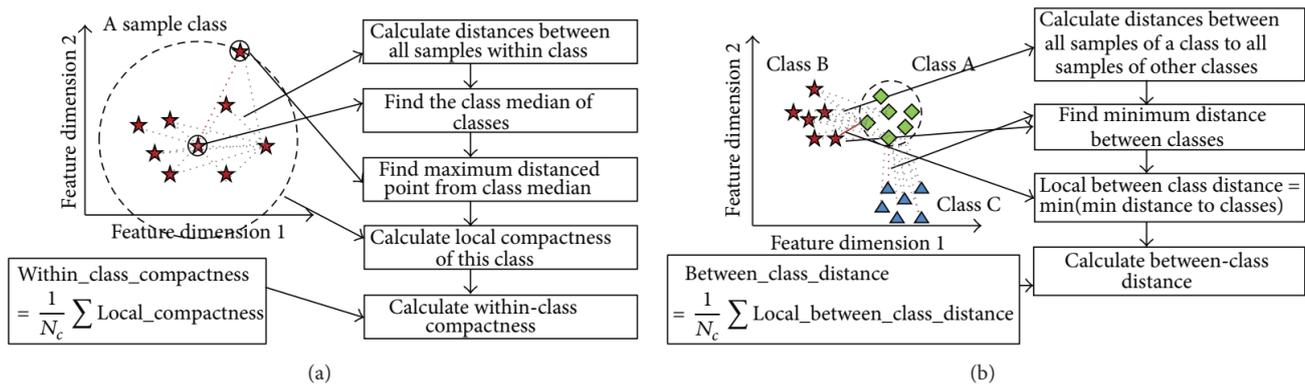


Figure 7. Process of objective value calculation: (a) Within-class compactness (b) between-class distance.

### 3.6. User Authentication Using Classification Algorithm

In this section, the user authentication process using different classifier models (SVM, LR and RF) is described. Our main goal is to make our system faster and real time system for the future. As we know Deep learning is very efficient and provides better accuracy. But to make a proper deep learning model, we need huge amount of data. However, in this type of single person identification or authentication system, huge amount of person data collection is complex. In this research, we are working with one dimensional single data and our dataset is limited. Hence, if we can extract the good or optimal features which can classify the classes very well, then we can use less computation overhead classifier to get a good result. Considering this point or concept in mind, we chooses the less computational classifiers (SVM, LR, RF) which can classify the classes very well with less amount of data in the training process. If we can provide good features from the data, then these classifiers can train the model very well and can provide good and stable accuracy result with less computational overhead. To check the robustness of the proposed model, in the authentication phase, the selected optimal features are validated using different classifiers to identify the writes. Figure 8 represents the user authentication process.

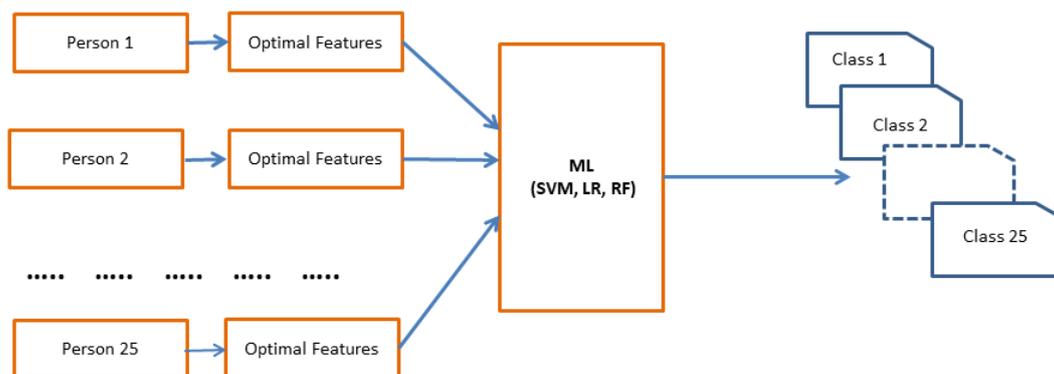


Figure 8. Block Diagram of User Authentication Process.

A short description of the 3 ML classifiers is given below:

Logistic Regression (LR): The logistic regression is one of the widely used linear and statistical models for discriminant analysis. Logistic regression can occasionally lead to outperforming other sophisticated nonlinear models such as ensemble learners or support vector machines due to simplicity and interoperability [35].

$$\pi_i = \beta_0 + \beta_1 + \dots + \beta_n X_n \quad (4)$$

For a binary classifier, the logistic regression model is expressed by summing over the linear combinations of input features and a corresponding weight ( $w$ ) plus a bias term ( $b$ ) for each instance as shown in Equations (5) and (6).

$$P(y^{(i)} = 1 | x^{(i)}, w) = 1 - \frac{1}{1 + \exp(w^T x^{(i)} + b)} \quad (5)$$

$$P(y^{(i)} = 0 | x^{(i)}, w) = 1 - \frac{1}{1 + \exp(w^T x^{(i)} + b)} \quad (6)$$

Random Forest (RF): Random forests is an ensemble learning algorithm for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time. RF is comprised of  $n$  collections of de-correlated decision trees. It uses multiple trees to average (regression) or compute majority votes (classification) in the terminal leaf nodes when making a prediction [35]. The following equation is needed for the calculation of RF model:

$$Entropy = -p \log_2(p) - q \log_2(q) \quad (7)$$

Support Vector Machine (SVM): SVM is an efficient and popular and machine learning technique which is widely used for classification due to its simplicity and computational efficiency. Different kernel techniques are there for SVM. The RBF non-linear kernel is formulated as Equation (8).

$$k(sv_i, sv_j) = \exp\left(-\frac{\|sv_i - sv_j\|^2}{2\sigma^2}\right) \quad (8)$$

where  $k(sv_i, sv_j)$  is the kernel function, and  $sv_i$  and  $sv_j$  are the input data, and parameter  $\sigma$  is a set by the user. The  $\sigma$  used here to determine the width of the kernel function  $k$ . Note that, if  $\sigma$  values are small, then overtraining may occur. Again, if  $\sigma$  values are large, then the basis function puts an oval around the points without describing their shapes or patterns [36].

#### 4. Experimental Result Analysis

For the experiment of this research, initially, we have collected pen-tablet handwriting samples from 25 different persons of age 19~40. Users in the experiment were given 10 different keywords (Japan, Machine Learning, Tokyo, Hello World, Fukushima, Aizu University, Basic Research, Computer Vision, Pattern Processing and Thank You) to write using the pen-tablet devices. After data collection, for labeling the training and testing dataset, we have separated our dataset into two portions. A total 60% of the dataset is separated for training the system and 40% of the dataset for testing purpose. In our research, during data collection, each key word is written 5 times by each person. Thus from (10 key words  $\times$  5 times) = 50 data samples of one person, total (10  $\times$  3) = 30 data samples are randomly selected for feature selection and training process and rest of the (10  $\times$  2) = 20 data samples are selected for testing process. The same procedure is repeated for the 25 classes. Thus a total 1250 writing samples and 25 classes are used in the experiment. After data collection, the test dataset and train dataset kept separated. The test data is not mixed up with the training data. Table 4 presents the summary of dataset labeling.

**Table 4.** Summary of Training and Testing Data Labeling.

	No. of Keywords	No. of Writing for Each Keywords	Data Samples for One Person	Training and Feature Selection Data	Test Data
Class 1	10	5	50	30	20
Class 2	10	5	50	30	20
Class 3	10	5	50	30	20
...	...	...	...	...	...
Class 25	10	5	50	30	20
			Total Data: 1250	Total Train Data: 750	Total Test Data: 500

All the experiments are implemented on a laptop computer Intel Core i5 (2.20 GHz) and 8GB RAM with operating system windows ( $\times 64$ ) version 10, using Python programming language, Anaconda software, and Wacom Tablet Device. In our research, we utilized 3 different machine learning algorithms (SVM, LR, and RF) to show the stability of our proposed model by measuring the classification results. There are two phases of our experiment: (1) Classification without feature selection (all features) and (2) Classification with feature selection (selected features). To evaluate the performance of the proposed model, four evaluation metrics are used for which we need to compute the parameters of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). A confusion matrix is generated by these four evaluation metrics of classification results. These four evaluation metrics are computed as follows:

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \quad (9)$$

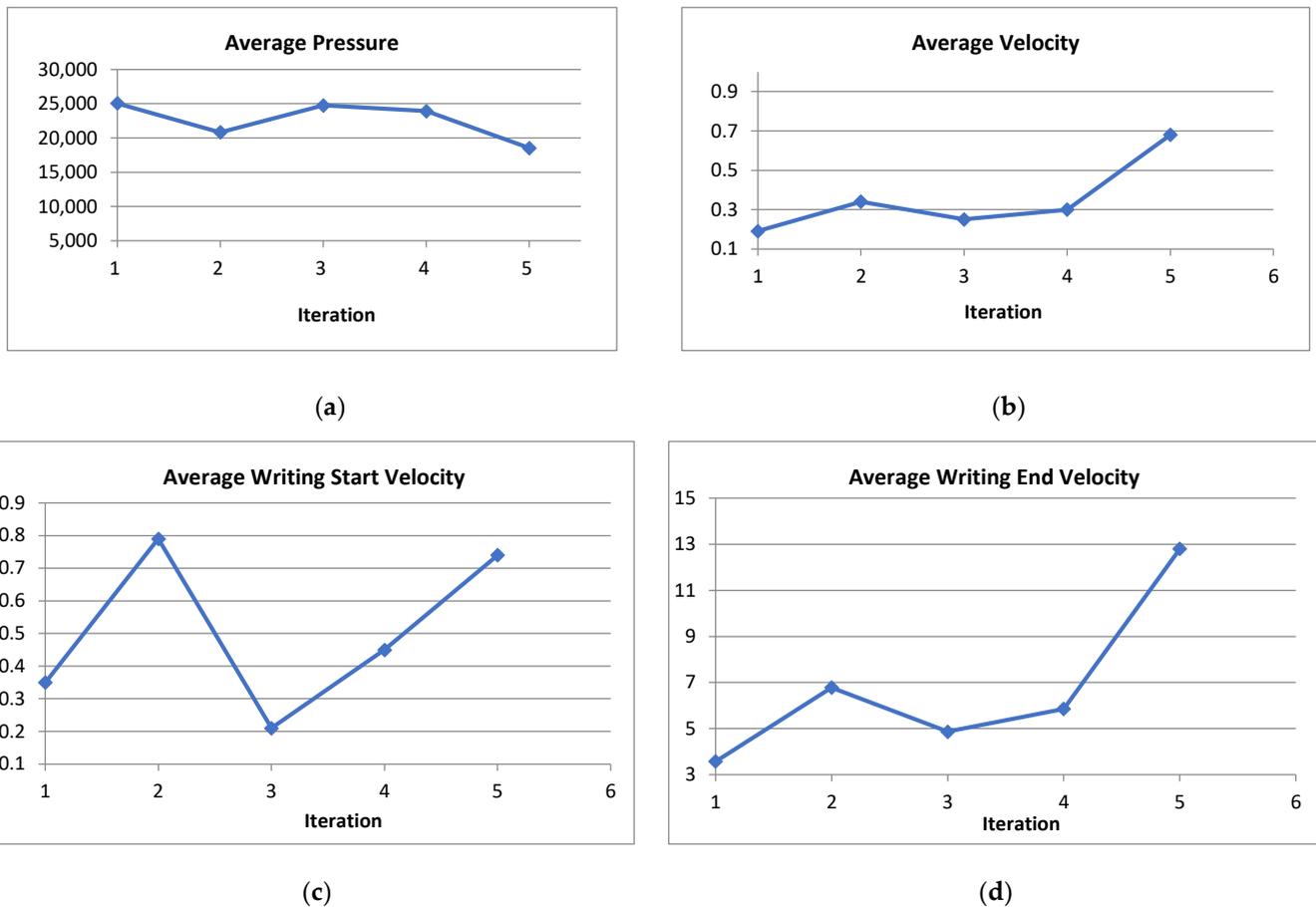
$$Precision (P) = \frac{T_p}{T_p + F_p} \quad (10)$$

$$Recall (R) = \frac{T_p}{T_p + F_n} \quad (11)$$

$$F_1 Score = \frac{2 \times (P \times R)}{P + R} \quad (12)$$

The writing samples of different users are not the same. Hence, there is visible discrimination between the same feature values of different persons. Figure 9 shows different features of writing samples for different persons.

There are 30 main features in the experiment ( $f_1 \sim f_{30}$ ). These feature vectors are directly applied to the filter approach for optimal feature selection. The filter technique is executed by two different SFFS algorithms on the two randomly splitted datasets containing equal numbers of data. Individual SFFS algorithm generates separate optimal feature sets from both datasets. This whole process is repeated for 4 iterations. Finally, the filter technique generates 10 sub-optimal feature sets from the 30 datasets. Then the 10 sub-optimal feature sets are forwarded to the wrapper approach, where the classification accuracies are measured using SVM, LR, and RF for each of the sub-optimal feature vectors. Table 5 shows the 10 sub-optimal feature sets after the filter approach. In our experiment, firstly, we have done the best optimal feature selection process with SVM and the evaluation is also done with the same classifier that is SVM. The classification accuracy using SVM is about 98% which shows the practicality of our system. However, to prove the stability of our system, we implemented our model with two additional machine learning algorithms such as LR and RF classifier. For that, the feature selection and evaluation processes are done using LR and RF accordingly following the same procedure of SVM. From the experimental analysis, it is found that the LR and RF provide satisfactory results and SVM perform outstanding which makes our claim stronger for the proposed model.



**Figure 9.** Sample Handwriting Features for Different Persons: (a) average pressure (b) average velocity (c) average writing start velocity (d) average writing end velocity.

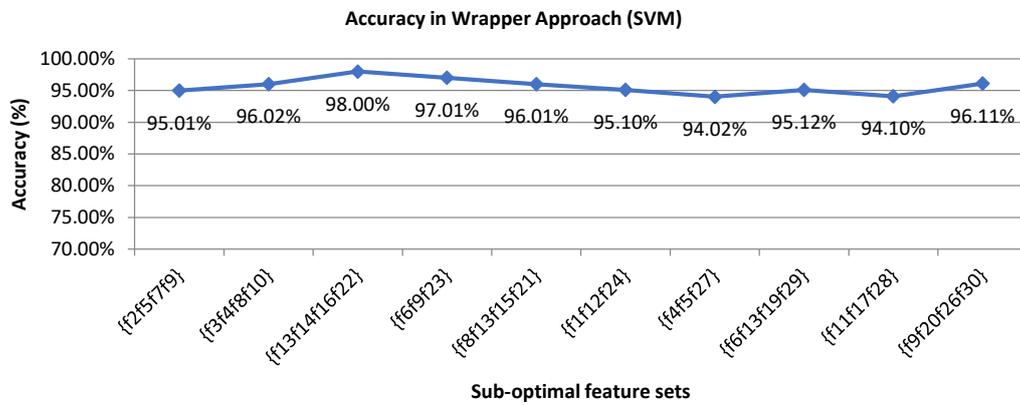
**Table 5.** Sub-optimal Feature Sets after Filter Approach.

Sub-Optimal Feature Sets in Filter Approach									
1	2	3	4	5	6	7	8	9	10
{f <sub>2</sub> f <sub>5</sub> f <sub>7</sub> f <sub>9</sub> }	{f <sub>3</sub> f <sub>4</sub> f <sub>8</sub> f <sub>10</sub> }	{f <sub>13</sub> f <sub>14</sub> f <sub>16</sub> f <sub>22</sub> }	{f <sub>6</sub> f <sub>9</sub> f <sub>23</sub> }	{f <sub>8</sub> f <sub>13</sub> f <sub>15</sub> f <sub>21</sub> }	{f <sub>1</sub> f <sub>12</sub> f <sub>24</sub> }	{f <sub>4</sub> f <sub>5</sub> f <sub>27</sub> }	{f <sub>6</sub> f <sub>13</sub> f <sub>19</sub> f <sub>29</sub> }	{f <sub>11</sub> f <sub>17</sub> f <sub>28</sub> }	{f <sub>9</sub> f <sub>20</sub> f <sub>26</sub> f <sub>30</sub> }

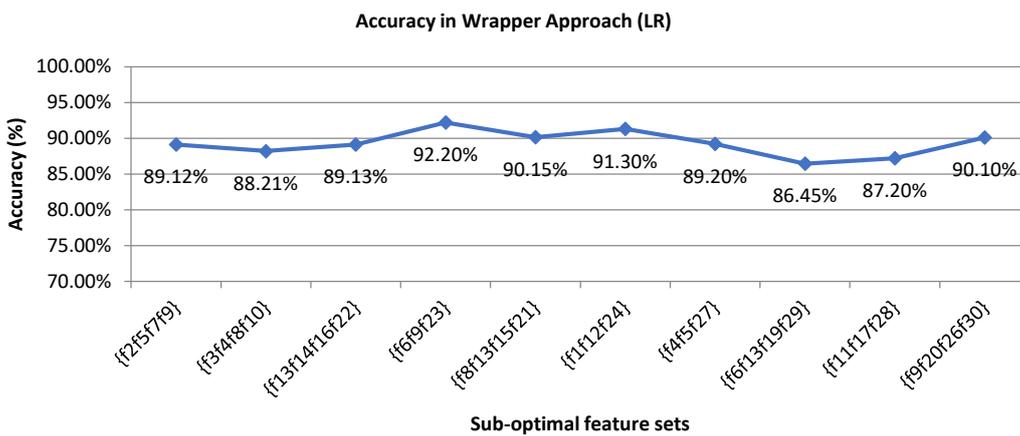
Figure 10 presents the classification accuracies of all sub-optimal feature vectors for 3 machine learning algorithms in the wrapper approach. From the 10 sub-optimal features, the wrapper approach identifies the best feature for higher classification results. Table 6 presents the best optimal feature sets and the classification accuracy for the selected best optimal feature set.

Table 6 presents the best accuracy of finally selected best optimal features in wrapper approach using support vector machine, logistic regression, and random forest. From Table 6, we can see that SVM provides the best accuracy result compare to LR and RF. It is found that our proposed system provides a higher accuracy outcome with the combination of best optimal features {f<sub>13</sub>, f<sub>14</sub>, f<sub>16</sub>, f<sub>22</sub>} with the best accuracy of 98.0% whereas the accuracy without feature selection is 93% using SVM. Using LR with the combination of best optimal features {f<sub>6</sub>, f<sub>9</sub>, f<sub>23</sub>} provides the best accuracy of 92.2% whereas the accuracy without feature selection is 86% only. Again, using RF with the combination of best optimal features {f<sub>4</sub>, f<sub>5</sub>, f<sub>27</sub>} provides the best accuracy of 94.6% whereas the accuracy without feature selection is 90%. These selected best features are then used in the validation process to check the overall accuracy of our proposed model of user identification. Our proposed

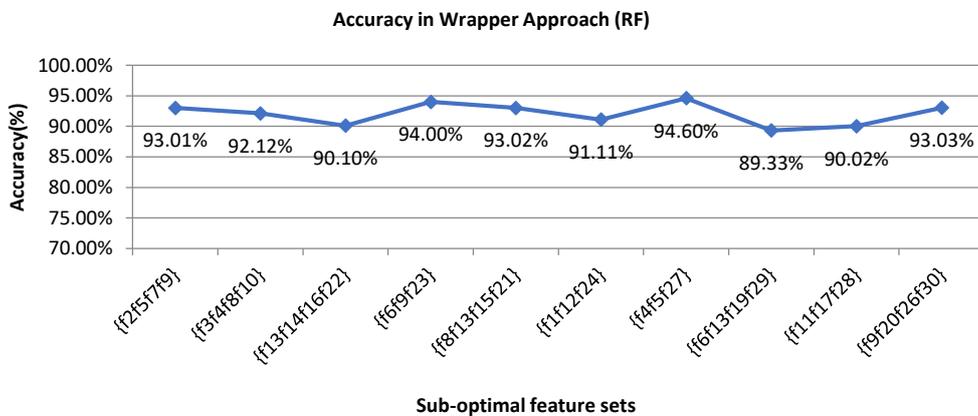
model provides efficient and satisfactory writer identification with limited computational resources and hardware cost which prove the practicality of our system.



(a)



(b)



(c)

**Figure 10.** Classification accuracy of different sub-optimal feature sets in wrapper approach: (a) estimated by support vector machine (SVM) (b) estimated by logistic regression (LR) (c) estimated by random forest (RF).

**Table 6.** Classification Accuracy of Best Optimal Features after Feature Selection with Wrapper Approach.

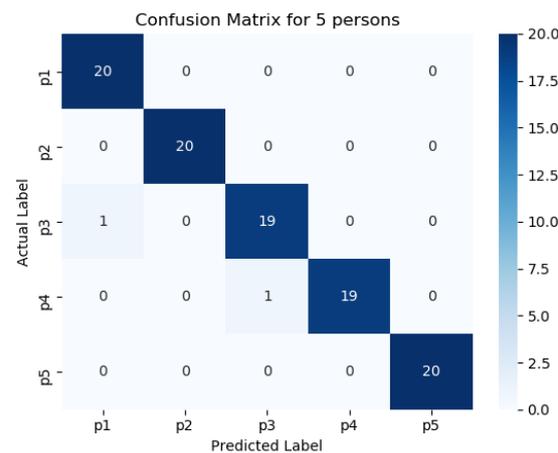
	Machine Learning Algorithms		
	Using SVM	Using LR	Using RF
Best Optimal Feature	{f <sub>13</sub> , f <sub>14</sub> , f <sub>16</sub> , f <sub>22</sub> }	{f <sub>6</sub> , f <sub>9</sub> , f <sub>23</sub> }	{f <sub>4</sub> , f <sub>5</sub> , f <sub>27</sub> }
Best Accuracy	98.0%	92.2%	94.1%

To evaluate the performance of the proposed model, we used four evaluation metrics namely accuracy, precision, recall, and F1 score. We measured the values of performance matrices for 5 different classes utilizing the SVM classifier. Table 7 shows the performance of the SVM classifier for a different number of persons with optimal feature sets.

**Table 7.** Confusion Matrix of Classification using Optimal Feature Sets.

No. of Persons	SVM Classifier (with Feature Selection)			
	Accuracy	Precision	Recall	F1 Score
5	98.00	98.00	98.00	98.00
10	97.00	97.13	97.00	97.06
15	97.30	97.25	97.35	97.30
20	96.75	97.00	96.20	96.60
25	96.20	96.00	96.30	96.15

Figures 11–15 represents the confusion matrix to visualize the performance of the implemented model in a contingency table for different number of persons with optimal feature sets. Figure 16 presents the ROC curve of classification of 5 persons. From the experiment, it is observed that in case of optimal feature selection, when the no. of person’s increases then the values of the performance metrics are not changing much which is satisfactory and stable. However, in case of no feature selection, these values are changing significantly. This is because, the distribution of large number of data in high dimensional feature space is more complicated and inefficient for classification.



**Figure 11.** Confusion matrix of the proposed model for the five persons.

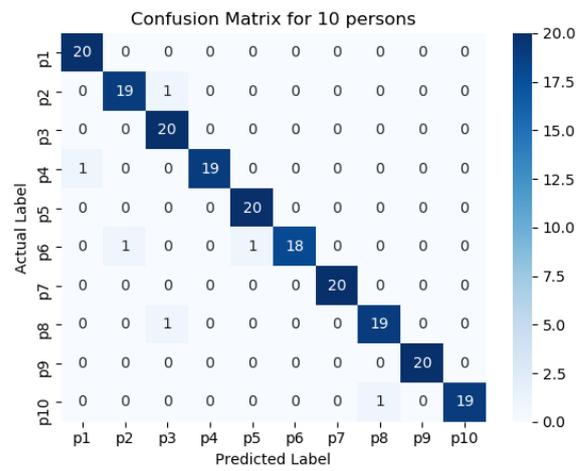


Figure 12. Confusion matrix of the proposed model for the ten persons.

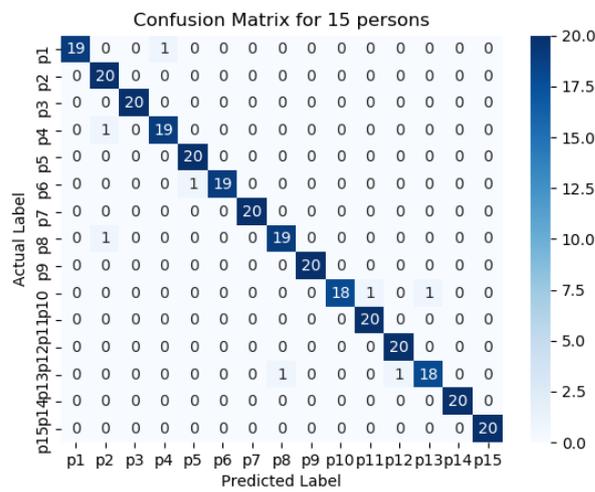


Figure 13. Confusion matrix of the proposed model for the fifteen persons.

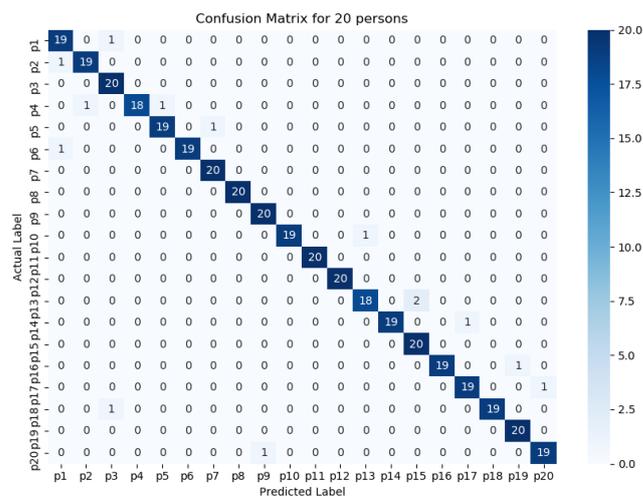


Figure 14. Confusion matrix of the proposed model for the twenty persons.

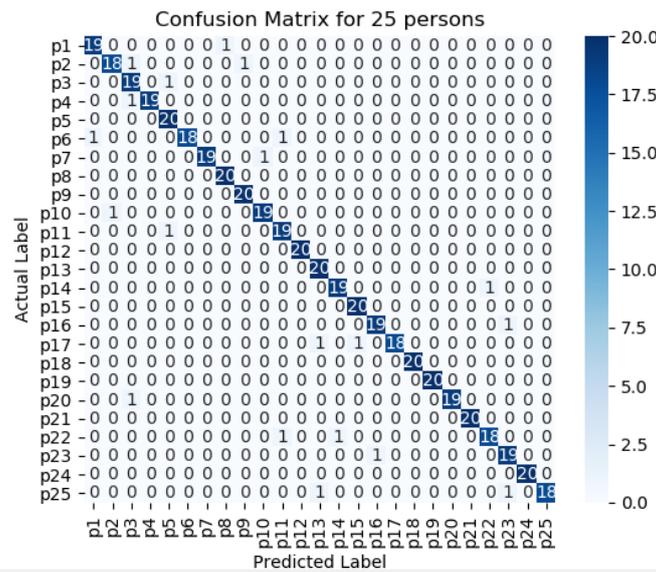


Figure 15. Confusion matrix of the proposed model for the twenty-five persons.

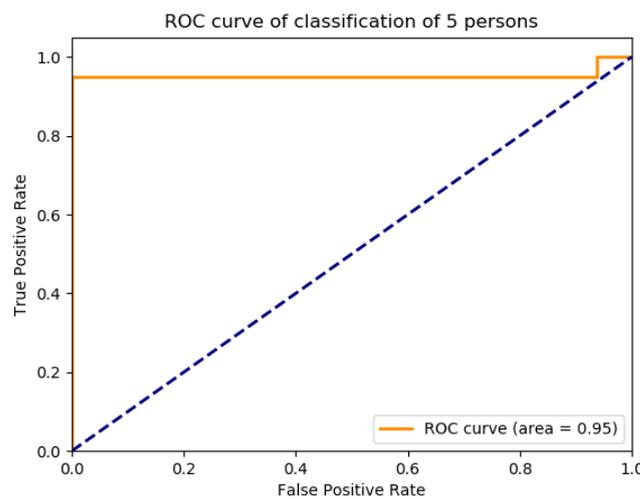


Figure 16. ROC curve of classification of 5 persons.

The performance matrices, accuracy, precision, recall, and F1 score achieve higher values such as 98%, 98%, 98%, and 98% respectively when the number of persons is smaller. However, when the number of persons is 15, then the values of accuracy, precision, recall, and F1 score are 97.3%, 97.25%, 97.35%, and 87.3% respectively where the performance is changing insignificantly which is practical.

Figure 17 presents the overall accuracy of different groups of persons with feature selection and without feature selection using the SVM linear kernel technique. From the experimental analysis, it is found that our proposed system achieves higher results with the feature selection method than no feature selection process. For SVM, the highest accuracy with feature selection is 98.0% where that of which is 92.20% with no feature selection technique.

Three different machine learning algorithms are used to measure the classification performance of the proposed system. Figure 18 shows the comparison of classification results of different algorithms that are used in the proposed system for different groups of persons. We have compared the classification performance without feature selection and with the feature selection (FS) technique. SVM provides the best accuracy outcome than logistic regression and random forest. The accuracy values are 98%, 92%, and 94% in case of SVM, in case of LR, and in case of RF respectively with feature selection process.



Figure 17. Classification accuracy for different persons with and without feature selection using SVM.

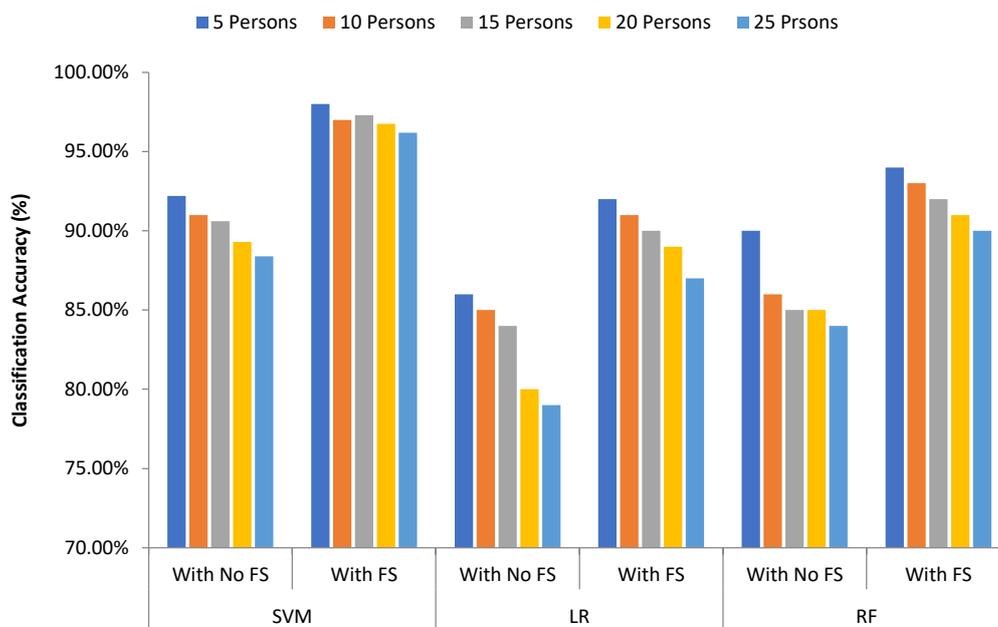


Figure 18. Comparison of accuracy for different classification algorithms with and without FS.

In [20], the researcher proposed a writer identification system using image-based handwriting feature analysis using K-NN and neural networks. Here, the overall recognition rate is 74.2%, where all feature vectors in the training set are used as prototypes. They varied the number of neighbor’s k from 1 up to 25. In our case (pen-tablet sensor data-based), initially, we considered 25 persons, and the proposed system has 98% accuracy with the feature selection approach. Figure 19 depicts the intuitive comparison between the state-of-art (image-based) system [20] and our proposed system.

Table 8 presents a comparison summary of our proposed system with some existing studies. From Table 8, we can see that our self-collected pen-tablet data are more robust in terms of noise effect compare to image-based. And our model outperforms with low computational costs.

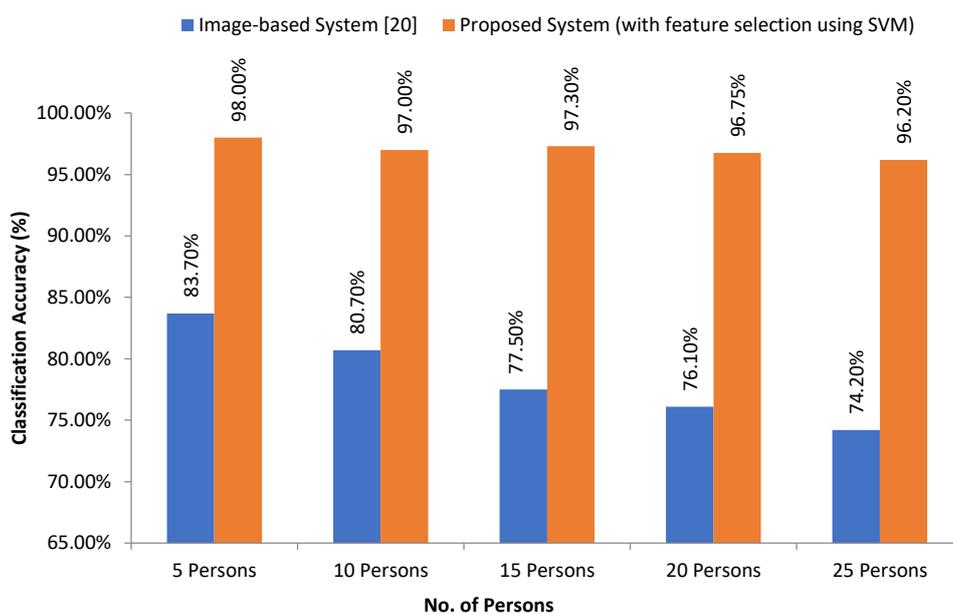


Figure 19. Class-wise accuracy (%) comparison of the proposed model with state-of-the-art image-based system.

Table 8. Comparison of Proposed Scheme with State of Art.

Ref #	Classifier	Feature	Result/Accuracy	Input Data Type	Comments
[5]	RF, KNN	Keystrokes Dynamic	Achieved an EER of 2.9%	Image-based	Considered only 3 typing positions
[11]	KNN, SVM	Hand Radiographs	97%	Image-based	Suffers from high computational Cost
[28]	Minimum Distance, Bayes	Dynamic Features	95%	Pen-tablet data	Focused only on the small scale writing samples
[30]	Hamming Distance, Nearest neighbor	Histogram	91.17%	Image-based	Writer’s information lost due to stretching the images
[32]	AlexNet	Implicit and explicit information	92%	Image-based	Resizing handwriting patches loss write’s intrinsic information
[34]	ResNet	Handwriting thickness descriptor (HTD)	97%	Image-based	High Computational cost, Multimodal descriptions needed
Proposed Model	SVM, LR, RF	Optimal Features (kinematic and statistics)	98%	Pen-tablet sensor data (numerical value)	Pen-tablet data are more robust in terms of noise effect compare to image-based, low computational cost

### 5. Conclusions

This paper proposed a user authentication system by analyzing digital pen-tablet sensor data by optimal feature selection model which can successfully identify the writer. In this research, several statistical and kinematic features are extracted for classifying the users. However, not all the features are vital for the authentication of a user. Moreover, sometimes, extra features may degrade the performance of classifiers. Therefore, a hybrid optimal feature selection algorithm based on the filter and wrapper approach has been introduced to get the best optimal features from original high dimension feature vector. The selected features are then used for training and authenticating the writes using SVM, LR, and RF to identify based on person’s handwriting. To validate the proposed model, different handwritings from different users are collected using digital pen tablet. Ten specific keywords were written by the users iteratively. The proposed system shows 90% and 92% average classification results with feature selection techniques which are around 7% and 6% improved and higher performance than without feature selection technique using LR and RF respectively. The proposed system also shows 97% average classification results with feature selection technique which is around 7% improved and

higher performance than the no feature selection phase using SVM. Among the 3 machine learning algorithms, the SVM performs outstanding compared to the rest.

The proposed system can successfully classify the given handwriting dataset. Thus, the system can be applied in some real applications like signature verification in the banking sector, proving the writer in forensic analysis, analysis of mental states of the writer, and so on. As the proposed system uses more genuine and robust sensor data for analyzing the handwriting, it can accurately predict the writer with limited computational resources and hardware costs which can be a good application in the case of small mobile devices.

**Author Contributions:** Conceptualization, N.B. and M.R.I.; Data curation, N.B., M.A.H.A. and S.R.; Formal analysis, N.B., M.A.H.A. and S.R.; Funding acquisition, N.B., M.E.I.; Investigation, M.R.I. and M.E.I.; Methodology, N.B., M.A.H.A., J.S. and M.R.I.; Software, N.B., M.A.H.A. and S.R.; Supervision, M.R.I. and M.E.I.; Validation, M.E.I., J.S.; Visualization, M.A.H.A., S.R. and M.R.I.; Writing—original draft, N.B., M.A.H.A. and S.R.; Writing—review & editing, M.R.I., J.S., and M.E.I. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is partially supported by the Institute of Energy, Environment, Research, and Development (IEERD), University of Asia Pacific (UAP), Bangladesh.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** Special thanks to University of Asia Pacific for supporting to do research. Thanks to IEERD, University of Asia Pacific for supporting for research and publication fund.

**Conflicts of Interest:** There is no conflict of interest.

## References

1. Al Emran, M.; Naief, S.; Hossain, M. Handwritten Character Recognition and Prediction of Age, Gender and Handedness Using Machine Learning. Ph.D. Thesis, BRAC University, Dhaka, Bangladesh, 2018.
2. Plamondon, R.; Lorette, G. Automatic signature verification and writer identification—The state of the art. *Pattern Recognit.* **1989**, *22*, 107–131. [\[CrossRef\]](#)
3. Gupta, S. Automatic person identification and verification using online handwriting. *Int. Inst. Inf. Technol.* **2008**, *408*, 356–361.
4. Azim, H.A.; Nasima, B.; Sayma, R.; Jungpil, S.; Amiruzzaman, M.; Rashedul, I. User Authentication Through Pen Tablet Data Using Imputation and Flatten Function. In Proceedings of the ICKII 2020, Kaohsiung, Taiwan, 21–23 August 2020; pp. 208–211. [\[CrossRef\]](#)
5. Saini, B.S.; Singh, P.; Nayyar, A.; Kaur, N.; Bhatia, K.S.; El-Sappagh, S.; Hu, J.-W. A Three-Step Authentication Model for Mobile Phone User Using Keystroke Dynamics. *IEEE Access* **2020**, *8*, 125909–125922. [\[CrossRef\]](#)
6. Li, L.; Zhao, X.; Xue, G. Unobservable Re-Authentication for Smartphones. Available online: <https://optimization.asu.edu/papers/XUE-CNF-2013-NDSS.pdf> (accessed on 30 August 2021).
7. Sae-Bae, N.; Memon, N.; Isbister, K.; Ahmed, K. Multitouch gesture-based authentication. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 568–582. [\[CrossRef\]](#)
8. Fierrez, J.; Pozo, A.; Martinez-Diaz, M.; Galbally, J.; Morales, A. Benchmarking touchscreen biometrics for mobile authentication. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2720–2733. [\[CrossRef\]](#)
9. Ghanim, T.M.; Khalil, M.I.; Abbas, H.M. Comparative Study on Deep Convolution Neural Networks DCNN-Based Offline Arabic Handwriting Recognition. *IEEE Access* **2020**, *8*, 95465–95482. [\[CrossRef\]](#)
10. Chen, D.; Ding, Z.; Yan, C.; Wang, M. A behavioral authentication method for mobile based on browsing behaviors. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 1528–1541. [\[CrossRef\]](#)
11. Joshi, S.V.; Rajendra, D.K. Deep Learning Based Person Authentication Using Hand Radiographs: A Forensic Approach. *IEEE Access* **2020**, *8*, 95424–95434. [\[CrossRef\]](#)
12. Wang, X.; Ding, X.; Liu, H. Writer identification using directional element features and linear transform. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 3–6 August 2003; pp. 942–945.
13. Hamid, N.A.; Sjarif, N.N.A. Handwritten recognition using SVM, KNN and neural network. *arXiv* **2017**, arXiv:1702.00723.
14. Kumar, A.; Wong, D.C.; Shen, H.C.; Jain, A.K. Personal authentication using hand images. *Pattern Recognit. Lett.* **2006**, *27*, 1478–1486. [\[CrossRef\]](#)
15. Kosugi, M.; Suzuki, T.; Uchida, O.; Kikuchi, H. SWIPASS: Image-Based User Authentication for Touch Screen Devices. *J. Inf. Process.* **2016**, *24*, 227–236. [\[CrossRef\]](#)
16. Russ, J.C. Image processing. In *Computer-Assisted Microscopy*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 33–69.
17. Masui, T. Image Processing Apparatus, User Authentication Method and Storage Medium Storing Program for User Authentication. U.S. Patent No. 2007/0101.415A1, 3 May 2007.

18. Savvides, M.; Kumar, B.V. Illumination normalization using logarithm transforms for face authentication. In *International Conference on Audio-and Video Based Biometric Person Authentication*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 549–556.
19. Yicong, W.; Xu, H. Image Analysis for User Authentication. U.S. Patent US009202105B1, 1 December 2015.
20. Adobe, H. Communication System, Image Processing Apparatus, Image Processing Method, Authentication Server, Image Managing Method, Image Managing Program, and Image Processing System. U.S. Patent 20060015734A1, 19 January 2006.
21. Marti, U.-V.; Messerli, R.; Horst, B. Writer identification using text line based features. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, Seattle, WA, USA, 10–13 September 2001; pp. 101–105.
22. Plamondon, R.; Srihari, S. Online and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 63–84. [[CrossRef](#)]
23. Madhvanath, S.; Govindaraju, V. The role of holistic paradigms in handwritten word recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 149–164. [[CrossRef](#)]
24. Park, J.; Govindaraju, V.; Srihari, S.N. Ocr in a hierarchical feature space. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 400–407. [[CrossRef](#)]
25. Patvarczki, J.; Kornafeld, A.; Tamas, E. Method for Image-Based Authentication. U.S. Patent 20120005483A1, 5 January 2012.
26. Ritter, D.; Schaub, F.; Walch, M.; Weber, M. Miba: Multitouch image-based authentication on smartphones. In *Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems*, Paris, France, 27 April–2 May 2013; pp. 787–792.
27. Islam, R.; Khan, S.A.; Kim, J.-M. Discriminant Feature Distribution Analysis-Based Hybrid Feature Selection for Online Bearing Fault Diagnosis in Induction Motors. *J. Sens.* **2015**, *2016*, 7145715. [[CrossRef](#)]
28. Chapran, J. Biometric writer identification: Feature analysis and classification. *Int. J. Pattern Recognit. Artif. Intell.* **2006**, *20*, 483–503. [[CrossRef](#)]
29. Faundez-Zanuy, M.; Fierrez, J.; Ferrer, M.A.; Diaz, M.; Tolosana, R.; Plamondon, R. Handwriting Biometrics: Applications and Future Trends in e-Security and e-Health. *Cogn. Comput.* **2020**, *12*, 940–953. [[CrossRef](#)]
30. Chahi, A.; El Khadiri, I.; El Merabet, Y.; Ruichek, Y.; Touahni, R. Block wise local binary count for off-Line text-independent writer identification. *Expert Syst. Appl.* **2018**, *93*, 1–14. [[CrossRef](#)]
31. Chahi, A.; El Merabet, Y.; Ruichek, Y.; Touahni, R. Cross multi-scale locally encoded gradient patterns for off-line text-independent writer identification. *Eng. Appl. Artif. Intell.* **2020**, *89*, 103459. [[CrossRef](#)]
32. He, S.; Schomaker, L. Deep adaptive learning for writer identification based on single handwritten word images. *Pattern Recognit.* **2018**, *88*, 64–74. [[CrossRef](#)]
33. He, S.; Schomaker, L. FragNet: Writer Identification Using Deep Fragment Networks. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3013–3022. [[CrossRef](#)]
34. Javidi, M.; Mahdi, J. A deep learning framework for text-independent writer identification. *Eng. Appl. Artif. Intell.* **2020**, *95*, 103912. [[CrossRef](#)]
35. Kirasich, K.; Trace, S.; Bivin, S. Random forest vs. logistic regression: Binary classification for heterogeneous datasets. *SMU Data Sci. Rev.* **2018**, *1*, 9.
36. Islam, R.; Amiruzzaman, M.; Nasim, S.; Shin, J. Smoke Object Segmentation and the Dynamic Growth Feature Model for Video-Based Smoke Detection Systems. *Symmetry* **2020**, *12*, 1075. [[CrossRef](#)]