PERSON AND GENDER IDENTIFICATION FROM HANDWRITING USING MACHINE LEARNING

BY

Md. Imran Hossain

ID: 162-15-7741

Department of Computer Science and Engineering

Daffodil International University

Supervised By

Gazi Zahirul Islam

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

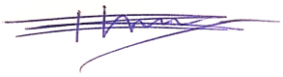
April 2021

APPROVAL

This Project/internship titled Person and Gender Identification from Handwriting Using Machine Learning, submitted by Md. Imran Hossain, ID No:162-15-7741 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 31-05-2021 .

BOARD OF EXAMINERS

Chairman



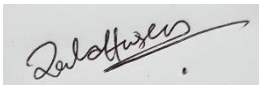
Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

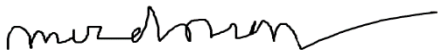
Zahid Hasan

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

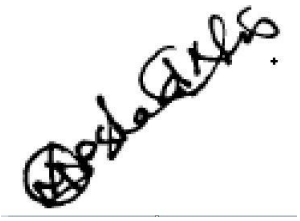
Md. Riazur Rahman

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

Dr. Md Arshad Ali

Associate Professor

Department of Computer Science and Engineering

Hajee Mohammad Danesh Science and Technology

University

DECLARATION

I hereby declare that, the work presented in this thesis paper based on research project is done by us under the supervision of Mr. Gazi Zahirul Islam, Assistant Professor of Department of Computer Science and Engineering, Daffodil International University, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering. I am declaring this report is my original work. I ensure that neither this report nor any part has been submitted elsewhere for the award of any degree.

Supervised by:

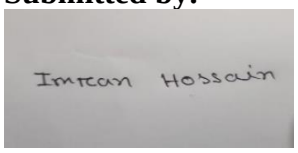
Gazi Zahirul Islam
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Aniruddha Rakshit

Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



Md. Imran Hossain
ID: 162-15-7741
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

I would like to express our gratitude to almighty Allah, the most merciful, beneficent, and generous, for always guiding us in the right direction and assisting us in completing our thesis. There are many people I would like to thank for their assistance in the process of planning for our thesis. I would like to express our gratitude to Gazi Zahirul Islam my thesis supervisor, for his unwavering support during the completion of my thesis. His direction and support for work have aided in our learning new skills and provided us with extraordinary experiences in the work. I would like to express my gratitude to all faculty and staff members of the Department of Computer Science and Engineering, Daffodil International University for directing me during my time at the university. I wish to express my heartfelt appreciation to my adored parents for the outpouring of love and care they show me. my sisters deserve my undying love and gratitude. Finally, I would like to express my gratitude to my friends who assisted me in completing my study, whether directly or indirectly.

ABSTRACT

Identification of an individual and gender from handwritten documents presents an intriguing research problem for researchers, as there has been relatively little research in this area. This research aims to examine a machine learning classification algorithm for recognizing the attributes and handwritten characters of various authors. The study proposes a scheme for user authentication that is based on data from pen tablets and handwriting. This research employed two techniques to ascertain the author's identity and to recognize handwritten characters. According to the study's analysis of experimental findings, the accuracy levels for SVM, LR, LDA, and RF are 87% 85%, 73%, and 77%, respectively. Both SVM and LR have an accuracy level greater than 80%.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 OVERVIEW OF THE CHAPTER

This paper endeavours to develop a system the will be able to identify person and gender from handwriting using machine learning. This introductory chapter provides discussions on the background of the study, research statement and research problem, objectives of the study and motivation of the study. The introductory chapter further contributions of the present study and outlines of the thesis.

1.2 BACKGROUND OF THE STUDY

In recent years, handwriting recognition and individual identification have emerged as a promising research area. Handwriting recognition and individual identification are two techniques used by the device to interpret and elaborate handwritten characters using input from paper documents, scanners, electronic pen and a touch screen. In terms of handwriting style, each individual has his or her own distinct personality. This fact is put to the test in a number of applications, including signature authentication [1], and individual recognition [2]. In recent years, this has emerged as a promising research field. Personal identification from handwriting is a novel method with a wide range of probable use, including defence, forensics, financial transactions, archaeological authentication and writer authentication based on demography. Handwriting may also reveal information about the author's ethnicity, handedness, gender and age. This procedure necessitates the extraction of features for parameters like writing speed, individual pressure, direction, length, height, width, and slant angle, among others. Both of these characteristics can be used to express a person's identity and predict personality characteristics [3]. It is difficult to create a handwriting recognition

system that can recognize a wide variety of handwriting styles with high accuracy. The main goal of an individual identification method, on the other hand, is to interpret the person's personal information from her or his writings and adjust those particularities to increase the identification system's validity. Since everyone writes at different pressure, at a different angle (horizontal or vertical), at a different speed, and takes a different amount of time for writing words, there are boundaries in everyone's writing style.

Authentication is now an essential part of any area of security-related work. Authentication is a human identification method based typically on handwriting, keystroke, hand geometry, user name, password, fingerprint, iris, face, DNA, Voice, and signature. User authentication is a technique that adapts a system to verify that someone who is associated with a network resource is recognized. Authentication plays a critical role in every area of safety-based work today. It has become an essential part of the safety check. Authentication is important because it allows an entity to rely on its protected networks and only identified users to access them. Authentication ensures that a user is securely entered into his/her safe tools, including databases, operating systems, websites and other network-based services or applications. It is crucial in the user's own registered account, which can only be accessed by his or her password-protected methods. This paper is based on this protocol, which takes into account these limitations and boundaries. Structural Methods, Convolutional Neural Networks, Syntactic Methods, Back Propagation, K-Nearest Neighbor Algorithm, Support Vector Machine (SVM), and other techniques have all been used for this function. Some recognition systems are capable of recognizing single characters or whole words. The machine functions as a human-computer interface. The system as a whole serves as a link between people and machines.

1.3 RESEARCH STATEMENT AND RESEARCH PROBLEM

Image and pattern analysis has been used in previous works on handwriting recognition and

individual identification. At the document or paragraph level, Image Processing Features are commonly used [4]. Handwriting recognition systems rely on image pre-processing techniques such as image contrast, lighting, shedding, illumination, noise reduction, and filtering, among others. As a result, it is difficult to get data that are more reliable if the image quality is not up to par. The consistency of the image is often affected by the pre-processing technique used. Text-based methods seem to be close to the text that would be used for authentication. Text-dependent writer recognition is used to identify signatures. Feature extraction methods are used to identify the writer in this scenario. On text-dependent offline handwriting, structural and directional features are calculated over single characters or multiple words in [5]. Text-dependent writer recognition based on text-line feature extraction is unable to provide us with the exact and proper value of features to be used for classification. As a result, our thoughts are focused on improving the authenticity and robustness of user authentication through handwriting recognition.

1.4 OBJECTIVES OF THE STUDY

The main objective of this research is to determine the identity of a consumer or an individual by analyzing real-time handwritten data. Our objective is to perform user authentication using data from a person's handwriting using a pen tablet. Our program is interested in the numerical value of a user's handwriting data obtained through pen tablet devices.

This work can be applied in a variety of sectors, including defence, the automotive industry, business field service personnel, education, writer recognition based on demography, signature authentication, postal address interpretation, and bank check processing. Handwriting identification is critical for resolving a variety of handwriting-related issues. I can speed up and improve the accuracy of handwriting recognition by using machine-learning algorithms. Our proposed model, I claim, will produce a more robust and accurate result when defining an

individual or a consumer.

1.5 MOTIVATION OF THE STUDY

The main motivation of the present study is to contribute something more authentic to the field of user authentication research. Historically, the majority of work in this field has been done using image pre-processing techniques and so on. These procedures cannot guarantee the precise numerical value of a person's writing samples. A Pen Tablet device is the device that allows us to extract the user's exact numerical value from their writing samples, which motivates us to do this work because existing authentication systems that rely on numerical values have limitations. The proposed method was tested on the handwriting data of the users using per tablet, which mechanically collects the numeric value of handwriting sample attributes. This data collection procedure aids in the robustness of our proposed system.

1.6 CONTRIBUTION OF THE STUDY

The majority of our contribution to user authentication using real-time hand-written data from each person via Pen Tablet is associated with various pre-processing and data analysis activities.

The main features are discovered following data pre-processing and analysis. These characteristics are primarily related to the essential characteristics of the handwriting of a user. To create a stable format for the dataset, many features Imputation functions are used. In addition, the focus of our contribution is on the use of Impute functions to create a balanced dataset format.

1.7 OUTLINE OF THE STUDY

The total thesis book contains five chapters. Contents of the chapters are briefly discussed

below.

Chapter one provides discussions on the background of the study, research statement and research problem, objectives of the study and motivation of the study. The introductory chapter further contributions of the present study and outlines of the thesis.

Chapter two presents a comprehensive analysis of the current literature on the subject and discusses the study's theoretical and methodological approaches. The chapter points the research gap based on the literature review and discusses how the current analysis can lead to closing it.

Chapter three provides a methodological overview of the study. This chapter, therefore, provided the system overview, discusses the proposed methods for the study, data collection methods, data analysis and processing methods. This chapter further provides the data classification process adopted for the study.

Chapter four presents the experimental result and discussion of the study, where findings of data analysis and discusses the findings with necessary tables, graphs and charts.

Chapter five presents the experimental result and discussion of the study, where findings of data analysis and discusses the findings with necessary tables, graphs and charts.

1.8 SUMMARY OF THE CHAPTER

This introductory chapter provides discussions on the background of the study, research statement and research problem, objectives of the study and motivation of the study. The introductory chapter further contributions of the present study and outline of the thesis. The next chapter will provide an extensive review of existing literature related to the present study.

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW OF THE CHAPTER

Numerous studies on personal identification and handwriting recognition have been conducted. This chapter conducts a comprehensive analysis of the current literature on the subject and discusses the study's theoretical and methodological approaches. The chapter points the research gap based on the literature review and discusses how the current analysis can lead to closing it.

2.2 RELATED STUDIES ON IDENTIFICATION FROM HANDWRITING USING MACHINE LEARNING

Text-dependent methods seemed to be analogous to the text that would be used for authentication. The identification of signatures is based on text-dependent author identification. Author recognition is accomplished by the use of feature extraction techniques. [5] examines the directional and structural characteristics of offline handwriting over single or multiple characters. For this generally, three types of algorithms are used to recognize objects [6], Multi-layer Perception Neural Network (MLP), Support Vector Machine (SVM) and]. K-Nearest Neighbor (KNN). MLP Neural Network was unable to make a clear forecast about features, while the other two achieved a satisfactory result. SVM and KNN are excellent at feature extraction-based prediction. MLP is an excellent function for studying non-linear models since it is a non-linear function. As a result, the percentage of accuracy can vary depending on the algorithm used to generate the desired outcome.

CNN is another form of neural network used in deep learning (Convolutional Neural Network). K. Fukushima proposed a hierarchical model called Neocognitron [7] based on a visual pattern

recognition mechanism. The proposed neural network model self-organizes by "unsupervised learning" and acquires the ability to identify stimulus patterns based on their geometrical similarity regardless of their location. After self-organization is complete, the network resembles the hierarchy model of the visual nervous system proposed by Hubel and Wiesel based on their ground-breaking findings in the cat [8]. The network is composed of an input layer (photoreceptor array) and a cascade of modular structures, each of which is composed of two layers of cells linked in a cascade. [7] Addressed stacked pairs of simple unit layer and complex unit layer in this proposed scheme.

Biometric verification systems based on touchscreen gestures performed when using mobile devices in scenarios such as web browsing, document reading or free tasks are gaining traction nowadays [9]. These techniques allow continuous or active authentication schemes that authenticate the user invisibly [10]. Numerous algorithms and features have been implemented in this area, all of which have demonstrated excellent performance from random attacks. In [11], the authors adopted a method that utilized several multitouch gestures based on the characteristics of finger and hand movements in conjunction with an algorithm that ensured a robust result. The proposed systems in [9] are based on SVM and Gaussian Mixture Models (GMM), which are considered for their ability to achieve high efficiency. In comparison to previous results, improved results have been obtained when verification algorithms such as Neural Networks, Random Forest, KNN and SVM were used.

The authors of [12] introduced three problems classified according to gender, handedness, and a combination of the two. In [12], the problem is solved using a Convolutional Neural Network (CNN). [13] has achieved the task of detecting repudiated documents from handwritten statements. Processing of Images at the document or paragraph level, features are frequently used [4]. Three categories of features are presented here: pen pressure, writing movement, and

stroke detail. The author of [14] proposed a method that utilized image acquisition, image pre-processing, image segmentation, feature extraction, and classification techniques. However, the previous method is reliant on image pre-processing techniques such as lighting and noise suppression. By scanning documents or taking a snapshot, a data collection for image acquisition has been compiled. Image pre-processing is used to remove noise from an image by erasing redundant points. It is often used to develop the image, which eliminates unwanted tortuosity, distortions, and so on.

In [15], handwriting uniqueness was developed through the use of pattern recognition techniques. Here, I see that their features were extracted using handwriting samples from 1,500 individuals representing every sector of the US population. The variations in handwriting samples from scanned images of handwriting were used to extract features in this work. Handwriting samples are analysed for attributes such as line separation, slant, and character form. By using machine-learning techniques, attributes were used to quantitatively determine individuality.

Gender classification from an Inertial Sensor-Based Gait Dataset is a proposed framework that is concerned with a variety of applications like protection, safety, entertainment, and billing. All of these applications are based on gait awareness and involve the recognition of people's gender and events. Numerous techniques may also be used to determine a person's gender and activities. The inertial sensor-based gait dataset is a novel addition to the gait analysis field. The gender of individuals from an inertial sensor-based gait dataset is explored in this proposed framework, which employs SVM, KNN, Bagging, and Boosting algorithms. In addition, the accuracy of gender identity obtained using the Bagging method is approximately 87 per cent; while the accuracy obtained using; SVM is approximately 86.09 per cent.

To ensure the system's protection, a password-based or biometric-based authentication system

can be used. Although using a password-based authentication method is simpler than using biometrics-based authentication, it can become tedious at times due to the user's need to memorize the password. Again, installing or modifying a password in a biometrics-based authentication system is not easy, as the process is hardware-dependent. Kin-Write is a framework that combines both password-based and biometric-based authentication to ensure security. To begin, the user must select a short, easily memorized password from the system's pool. The user is then required to write the password in three-dimensional space using a low-cost motion-sensing system called a Kinect. Kinect records the user's handwriting movement. The data collected by Kinect inherits several fundamental issues, such as low resolution and noise. Additionally, the inconsistency of in-space handwriting complicates properly verifying the consumer. When used to measure similarities between handwritten passwords, the Dynamic Time Wrapping (DTW) algorithm offers 100% precision and 70% recall for affirming honest users.

Existing research on handwriting recognition [16],[17],[18] and individual identification [19],[20],[21],[22],[23],[24],[25],[26],[27],[28],[29],[30] has been conducted using picture [19],[20],[21],[22],[23],[24],[25],[26],[27],[28],[29],[30] and pattern analysis.

In [19], this article introduces a novel approach to personal authentication based on hand image. By adding hand geometry features, the proposed method aims to improve the performance of a palm print-based verification framework. Extracting features from photographs of the hand shape and palm print is followed by an examination of their individual and combined verification performances. This method of feature extraction was accomplished using an image pre-processing technique. This experiment was conducted using a 100-user image database.

Article [20] describes a method for registering and authenticating users using image processing. They used hashing for authentication, which means that no images are stored in the system;

only hashed values are.

They present a content-based signature technique for ensuring the authenticity and credibility of images and videos in [21]. They present an interactive video authentication mechanism and suggest content-fragile watermarking, a concept that incorporates watermarking and content-based digital signatures to ensure copyright security and reputation infringement detection.

In [29], a novel approach for mobile user authentication is defined that makes use of statistical touch dynamics images. They suggest a statistical five-touch kinetics picture (statistic property model) trained on six graphical features of touch posture.

J. L. Traenkenschuh and D. W. Gilles [22] introduced implementation. A variety of icons is illustrated using their tool. A pattern is created by arranging the plurality of forms. M. Kosugi and T. Suzuki [23] proposed an image-based user authentication system for touch screen devices that makes use of the user's most recent image shot as the pass-image. Smudge attacks, one of the most serious threats to touch screen devices, are immune to the proposed authentication process. However, the method's protection is restricted.

Yicong WANG and Xu [24] suggested a model for a user authentication framework based on images. They used visual analysis, three-dimensional imaging, and thermal imaging to determine whether the known human user is interacting with the system. O'donoghue and Niall [25] suggested an identification-based approach for approving the use of an electronic device. They obtain a large number of images as a data set from a picture passkey library in order to implement their process. The method is checked by providing different images to users; if at least one of the selected images matches the predetermined identification data, the images in the picture passkey library are user-generated.

Samangouei Pouya and Patel [26] suggested a system for continuous authentication of

smartphone users based on facial attributes. They trained binary attribute classifiers on the Public Fig dataset and produced concise visual descriptions of faces. Shamik and Bandyopadhyay [27] suggested an image-based locking and unlocking paradigm for computing devices. The image they used in a model with multiple components is described and then identified and displayed to the user via a touch screen. S. V. Raghavan [28] proposed methods for facilitating user authentication through pattern recognition. This method works by comparing many matrix values; if the two sequences match, the consumer is marked.

L. L. Mizrah's [29] approach to user authentication is novel. They used an algorithm for random partial pattern recognition (RPPR). B. Fallah and H. Khotanlou [30] suggested using a neural network to identify human personality based on handwriting. They research the psychological concepts underlying disposition, character, attitude, and behaviour, as well as the nervous and social functions of the human brain. They suggest a method for determining an individual's personality based on their handwriting. They used the Minnesota Multiphasic Personality Inventory (MMPI) to determine the character parameters, and then used a neural network (MLP) and a hidden Markov model to perform classification on the handwriting to determine personality.

These handwriting recognition systems rely on image pre-processing techniques such as image contrast, lighting, shading, illumination [32], noise reduction, and filtering, among others. As a result, if the image quality is inadequate, it is difficult to obtain more vigorous data. Additionally, image quality changes according to the technique of pre-processing [33]. As a result, the proposed method was tested on handwriting data of user's using a pen tablet, that collects stores the numeric value of handwriting sample attributes automatically. This data collection procedure aids in the robustness of our proposed system.

2.3 RESEARCH GAPS AND CONTRIBUTION OF THE STUDY

This paper proposes a method for authenticating users based on their real-time handwritten data using a pen tablet computer. As part of the dataset, 24 individuals' writing samples are obtained. Each person has written ten keywords, five of which are repeated. This process collects fifty writing data values for each user. For creating a balanced dataset format, Imputation functions are used. The study conducted data analysis and implemented the proposed method using different algorithms. In this regard, the study focuses on six different attributes of handwritten data, namely Time, Pressure, X-axis, Y-axis, Horizontal angle, and Vertical angle. To address this research problem, the study used a variety of classification algorithms, including the Random Forest (RF) Classifier, Support Vector Machine (SVM), Logistic Regression (LR) and Linear Discriminant Analysis (LDA). Different algorithms exhibit varying degrees of accuracy based on their implementation. SVM, LR, LDA, and RF all have an accuracy rate of 87 per cent, 85 per cent, 73 per cent, and 77 per cent, respectively. The result analysis indicates that I obtained more stable and adequate performance, ensuring the system's practicality. The F-1 score is used to determine the accuracy of our test, and it is calculated using precision and recall. By combining precision and recall, the F-1 score will provide a more realistic representation of a test's results.

2.4 SUMMARY OF THE CHAPTER

This chapter has provided an extensive literature review based on the existing studies related to the present study and discusses the study's theoretical and methodological approaches. The next chapter will provide methodological overviews of the study.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 OVERVIEW OF THE CHAPTER

This chapter provides a methodological overview of the study. This chapter, therefore, provided the system overview, discusses the proposed methods for the study, data collection methods, data analysis and processing methods. This chapter further provides the data classification process adopted for the study.

3.2 SYSTEM OVERVIEW

Two methods are defined in this paper as part of the proposed framework. The first method is divided into four stages: data collection, data analysis, feature extraction, and classification. The proposed second method is divided into five stages: data collection, analysis, feature extraction, pre-processing, and classification. The data collection process utilized a pen tablet system that produces six attributes (Time, Pressure, X-axis, Y-axis, Horizontal angle, and Vertical angle), each of which represents a unique numerical value for each type of handwriting. Following the data collection steps, the data is analyzed using Mean Normalization, where X represents a single attribute's data, represents the average value of X1 in the training set, and represents the attribute's largest norm. This procedure is replicated for the remaining six attributes. Following these measures, characteristics are extracted from the analyzed data. To improve accuracy, I manually added a new function called velocity, which is denoted by the unit ms1. To create a balanced format for the dataset, functions such as Imputation are used.

In addition, for user authentication, four algorithms are used. The algorithms are used are the SVM, LR, LDA algorithm and RF classifier. As previously mentioned, SVMs have supervised learning models that use associated learning algorithms to analyse data for classification and

regression analysis. Our system employed the SVM, LDA, RF, and LR algorithms during the training and testing phases.

3.3 PROPOSED METHODS

Two approaches are used to incorporate the proposed scheme Method one and Method two. The first method (illustrates in Figure 3.2) consists of four stages: data collection, data analysis, feature extraction, and classification. Method 1. Method 2 consists of five steps: data collection, data analysis, feature extraction, data pre-processing, and classification. The second method is depicted in Figure 3.3. A graphologist analyses the handwriting on a piece of paper written by the individual writer, which is a lengthy process [15]. Numerous studies in this proposed area [16] and [16] used a four-step approach: data collection, feature extraction, pre-processing, and classification.

The proposed method 1 consisted of four major steps in this system: data collection, data analysis, feature extraction, and classification. The second method consisted of five steps: data collection, analysis, feature extraction, pre-processing, and classification. and Imputation functions are used to pre-process the datasets. Throughout the dataset selection process, the time difference and pressure applied to each keyword vary with each iteration. As a result, additional rows containing data values are generated. There is a discrepancy between the maximum and the minimum number of rows for these excessive rows of data. To accomplish this, I used the largest number of rows of individuals in Method-2 to ensure that all rows of values were equal in form. This is accomplished by adding all additional rows that are significantly less in number to the highest row and by satisfying the extended row values through the imputation function.

	A	B	C	D	E	F
1	Time	Pressure	X axis	Y axis	Horizontal Angle	Vertical angle
2	6143.52	6787	346	1214	160	410
3	6151.46	8707	346	1214	160	420
4	6159.44	15671	346	1214	161	420
5	6166.47	15847	346	1214	160	430
6	6174.5	16391	346	1214	161	430
7	6181.42	16839	346	1214	160	440
8	6189.64	16775	345	1217	161	440
9	6197.64	16183	344	1215	161	450
10	6205.05	16343	344	1210	162	460
11	6212.64	16935	343	1205	162	460
12	6220.8	17383	343	1198	161	460
13	6228.76	18024	342	1189	161	460
14	6235.74	18456	342	1180	161	460
15	6243.67	18504	341	1171	161	470
16	6251.88	18568	341	1162	161	470
17	6258.74	18648	341	1153	160	480
18	6267.33	18696	341	1146	159	480
19	6273.64	18872	341	1139	158	490
20	6281.65	18984	341	1133	158	490
21	6289.63	18936	341	1128	157	500
22	6297.74	18808	342	1124	157	500
23	6304.75	18856	342	1120	157	510
24	6312.73	18824	342	1117	157	510
25	6319.61	18808	342	1114	156	510

FIGURE 3.1: SAMPLE DATASET

3.3.1 OVERVIEW OF METHOD ONE

The first method entails four stages. Those are;

- Data collection
- Data analysing
- Feature extraction and
- Classification

To begin, the data set is gathered through a Pen tablet computer. The attribute is available in six distinct flavours. The data collection contains 24 individuals' data in random order. Equation 1 is then used to normalize the data set depicted in Table I. The Label Encoder Function is then used to convert the string data to a numeric value. Then, using equation

1, the numerical data. Only the SVM algorithm is used for training and testing in Method 1. Following that, I calculate the F1 score to determine the accuracy level. The proposed Method-1 is depicted in Figure 3.2.

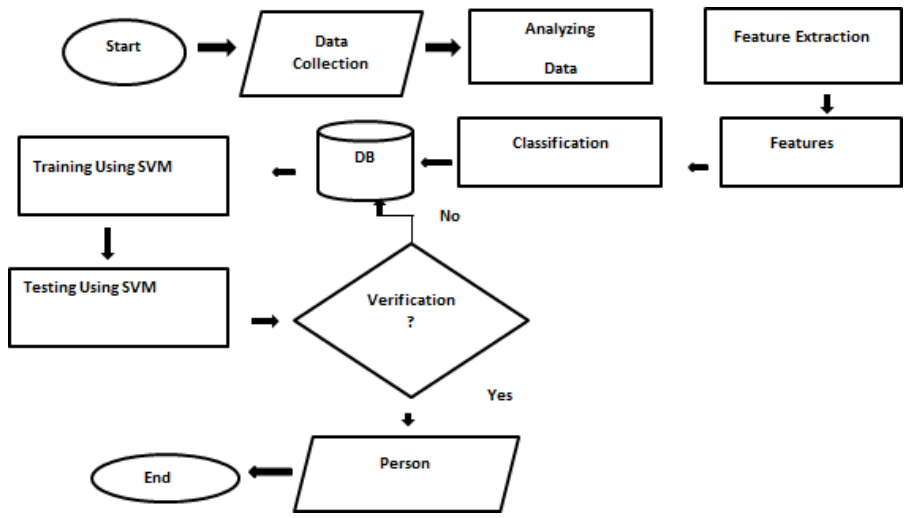


FIGURE 3.2: WORKING PROCEDURE OF METHOD 1

3.3.2 OVERVIEW OF METHOD TWO

The second method entails five stages. Those are;

- Data collection
- Data analysing
- Feature extraction
- Data pre-processing and
- Classification

I have prepared the data set first, and then converted the string data to a float value. Then, I have used the Function to convert our two-dimensional data to one-dimensional data.

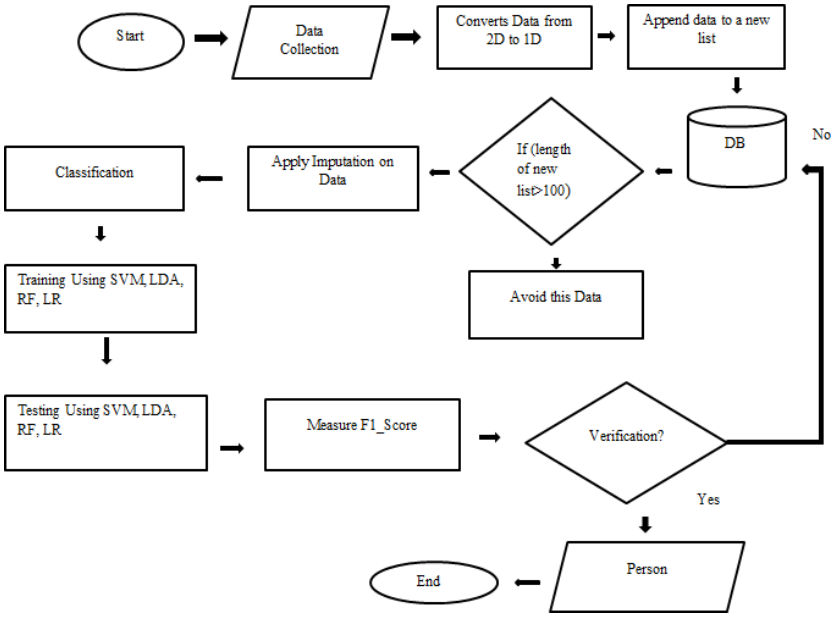


Figure 3.3: Working Procedure of Method 2

I have prepared the data set first and then converted the string data to a float value. Then, I used the Function to convert our two-dimensional data to one-dimensional data.

Following that, I assigned an index number to our newly created data set. There is a length check alternative. If the new length of our 1D data is less than 100, it will be deleted; otherwise, the data will be stored in a separate register. After that, I have use imputation to fill in the missing data value. The average value of the same column is used to replace the missing data value in this step. Then, I denote 23 classes for 24 data sets whose indexing begins at zero. I trained and tested the data set using SVM, LDA, RF, and Logistic Regression. Then, I determined the accuracy using the F1 20. This procedure was replicated for groups of 5, 10, 15, 20, and 24 people. The proposed Method-2 is depicted in Figure 3.3.

3.4 PEN TABLET DATASET COLLECTION

I have used handwriting samples from a variety of individuals. I have used handwriting samples from 24 separate individuals that were digitally obtained using pen tablet devices to create the proposed scheme. Ten distinct keywords are used in this section (Fukushima, Daffodil

University, Dhaka, Thank You, Pattern Processing, Machine learning, Computer Vision, Basic Research, Bangladesh, Hello world) with each keyword being written five times by each person. The dataset was compiled in numerical format. The samples are written inside a margin on the tablet's flat platform.

TABLE 3.1: NORMALIZED DATA SET

Time	Pressure	X-axis	Y-axis	Horizontal Angle	Vertical Angle
4578	15640	629	1147	183	476
3831	15855	569	1166	176	502
3358	15622	454	1196	181	397
3588	15935	506	1201	179	400
3023	15452	514	1173	177	407

3.5 DATA ANALYSIS PROCEDURES

The dataset is analysed by the following Mean Normalization equation:

$$X = \frac{X_1 - \mu}{S}$$

Where X denotes a single instance of an attribute's data represents the average value of X1 in the training set, and S denotes the attribute's maximum value. Table 3.1 contains the normalized dataset. By performing this operation, Table 3.1 is created; Table 3.1 contains a sample of the average value for a single keyword five times.

3.6 FEATURE EXTRACTION

This technique is used for reducing the dimensions of input data that contains a large number of dimensions. This reduced output data is a transformation of the feature vector-based high-dimensional input data. Those are six critical factors in our system on which a computer can

predict the user. These characteristics are discussed in detail below:

3.6.1 WRITING TIME

I am aware that each person's writing needs a unique amount of time. A user writes very slowly, someone else writes moderately, and someone else writes very quickly. After analysis, it was discovered that those who write quickly have a handwriting style that varies over time, such as straight, rising, dropping, irregular, etc.

3.6.2 PEN PRESSURE

The most critical characteristic in this study is pen strain. Each individual's pen pressure is unique. The pen tablet computer detects and records writing pressures automatically. Here, I found various forms of strain, including high, light, and so forth. Excessive pressure demonstrates dedication and a willingness to take things seriously [15]. Light pressure demonstrates attention to the environment as well as empathy for others [14]. After conducting a study, I have discovered that those who write under duress spend more time than others do. The pressure in our data set is not constant. For instance, some individuals have a pressure range of (2345-6595) Pascale, while others have a pressure range of (1978-22905) Pascale.

3.6.3 X-axis

The X-axis represents a person's writing angle relative to the X-axis. Each angle is unique. When I take a single keyword and repeat it five times by the each person, I find that the X-axis angle for each of the five cases is very similar, indicating that the same person's written X-axis value remains nearly identical in each case. Our collected data set's X-axis value is always the same. For instance, some individuals have an X-axis value range of (505-630), while others have (400-550).

3.6.4 Y-axis

The Y-axis represents a person's writing angle relative to the Y-axis. Each angle is unique. When the same individual uses a single keyword five times, it is clear that this angle is very

narrow. For instance, some individuals have a Y-axis value range of (1110–1201), while others have (1200-1350). There is no trigonometric formula used to generate numeric values. These numerical values can be automatically stored in a database by a pen tablet computer.

3.6.5 Horizontal Angle

The horizontal angle is the measurement of the angle formed by two parallel lines rising from the same point. The pen tablet computer automatically calculates this angle. This horizontal angle is a critical aspect of our work when it comes to user authentication. The horizontal angle range is limited to less than 500 degrees, as shown in Table I and Figure 3.1.

3.6.6 Vertical Angle

Vertical angles are those formed by the intersection of two sides. The pen tablet computer automatically measures this angle. This vertical angle is a critical aspect of our work when it comes to user authentication. As shown in Table I and Figure 3.1, the maximum vertical angle in our data set is less than 800.

3.7 DATA PRE-PROCESSING

Pre-processing methods are used in Method-2 to create a more balanced data collection. The data collection is reduced to a single dimension. Because the data set is not balanced and contains a large number of extra rows. This need to be done in a balanced manner. I the Function to convert the two-dimensional data to a one-dimensional format.

I performed data balancing in the Imputation Function. As our data in a row contains multiple column attributes, each column should be true and contain no null values, but in our scheme, I discover that some column sections contain null values. As a result, I have devised a method for populating the empty values in a column attribute. I filled in the null values by using the mean value of the same column attribute.

3.8 CLASSIFICATION

Classification is a term that refers to the process of categorizing data so that it can be used more effectively. It is capable of being applied to both structured and unstructured data. Classification facilitates the application and recovery of data. This classification procedure starts by predicting the group of data points that have been given. Classifications are sometimes abbreviated as goal, sticker, or divisions. In our proposed method, I use four distinct algorithms for classification: SVM, LR, LDA, and RF, as each algorithm is effective in a different way. In this method, I have used two distinct SVM kernels (linear and RBF). Our primary objective is to improve accuracy. This is accomplished by the use of four algorithms, as well as the F1 ranking, which indicates the model's accuracy. As a result, following the classification, preparation, and testing processes, I calculated the F1 score for each algorithm.

3.9 SUMMARY OF THE CHAPTER

This chapter discussed the study's methodology, including the framework description, proposed methods for the study, data collection methods, data analysis and processing methods, and classification methods. The following chapter will present and analyse the results of the experimental data analysis.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 OVERVIEW OF THE CHAPTER

This chapter presents experimental result and discussion of the study, where findings of data analysis and discusses the findings with necessary tables, graphs and charts.

4.2 EXPERIMENT

As mentioned earlier the paper has prepared two different methods as shown in Figure 3.2 and 3.3 respectively. I have implemented both methods. For this study Python is used as the programming language to develop this system. The programming language Python is used for our implementation. Four different types of algorithms are used for implementation. Those are SVM, LR, LDA & RF. For method one, only SVM algorithm is applied, while in method two all four algorithms are used. Though SVM has several kernels, this study used only two kernels Linear and Radial Basis Function (RBF). In Method 1, I have used a linear kernel in SVM. In Method, 2; I have used Linear and RBF kernel in SVM. The accuracy of both kernel shown in Table 5.1 and Table 5.5. Table 4.1 provides the list of the environments that I used for this system.

CPU	**Intel Core i5 (2.20GHz)**
Main memory	8.00 GBytes
OS	Windows 10 (Education)
System type	64-bit Operating system

TABLE 4.1: ENVIRONMENTS FOR IMPLEMENTATION OF BOTH METHOD.

4.3 ALGORITHMS

As mentioned earlier, in our implementation, I have used four classification algorithms.

LR is a classification algorithm used in machine learning that makes use of one or more independent variables to decide a result. The outcome is quantified using a dichotomous variable, which means it has only two potential outcomes: true or false.

The SVM used as a classifier to represent the training data as points in space divided into as many categories as possible by a gap. To determine the categories data points belong to and space they will be added to, new points are predicted and added to space. SVMs contain several kernels. The kernel act is used to enter data and then convert it to the required format. Different forms of kernel functions are used in the SVM algorithm: linear, nonlinear, polynomial, Radial Basis Function (RBF), and sigmoid. I used two kernels in our implementation: one in Radial Basis Function (RBF) and another in Linear Basis Function (LBF).

Random Forest or Random Decision Tree is a classification technique that uses ensemble learning. During preparation, it constructs outputs classes and substitutes decision trees. For supervised subspace learning, LDA is the most commonly used and operable technique. The objective of optimization is to build a complete skeleton for a novel method for clustering and unsupervised subspace learning in both ratio trace and trace ratio groups.

```
from sklearn import datasets
from sklearn import svm
import numpy as np
from sklearn.metrics import accuracy_score
import numpy as np
import pandas as pd
path = 'F:/Data'
dataset = pd.read_csv( path )
X = dataset.iloc[:, :-1 ].values
y = dataset.iloc[:, 2 ].values
```

FIGURE 4.1: SAMPLE CODE OF LIBRARY FUNCTION FOR METHOD 1

4.4 EVALUATION

The steps for Method 1 are as follows:

- Users insert information into the Pen Tablet computer using handwriting.
- All data input was stored in a.csv format.
- Data is read from .csv file.
- At this stage, I will refer to a.py file (created by python).
- The.py file will perform the normalization calculation and write the result to the.csv file.
- For preparation and testing, the Support Vector Machine (SVM) algorithm is used.
- After applying the SVM algorithm, display the accurate result and the F1 value.
- Now, I will explain how Python creates a.py file.
- When the.csv file is used to store the input data. Python is used to read the.csv file.

```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
#Enocdding the data set
labelencoder_X = LabelEncoder()
X[:, 0 ] = labelencoder_X.fit_transform(X[:, 0 ] )

labelencoder_y = LabelEncoder();
y = labelencoder_y.fit_transform( y )

onehotencoder = OneHotEncoder( categorical_features = [0]  )
X = onehotencoder.fit_transform(X).toarray()
```

FIGURE 4.2: SAMPLE CODE OF LABEL ENCODER FUNCTION FOR METHOD 1

- The input data is managed based on the user's input.
- Since I am using the Jupyter Notebook, our software is in the.ipynb format.
- I convert .ipynb files to.py files.

TABLE 4.2: TOOLS AND LIBRARIES FOR IMPLEMENTATION OF METHOD 1.

GUI	Anaconda Navigator
Platforms	Jupyter Notebook
Programming Language	Python
Libraries of python	Pandas
Algorithm	SVM
Device name	Wacom Tablet 6.3.32

I also included a list of all tools and libraries that I used in Method 1 in Table 4.2. I use the SVM algorithm exclusively for testing and training in the first process. The Linear kernel is used in this implementation phase, followed by the F1 score for accuracy. Figure 4.1, 4.2, and 4.3 illustrate sample codes for Method 1. In the first approach, two-thirds of the data samples are used as training data and one-third are used as testing data. I have extracted five times the amount of data for a single keyword. The first three datasets are used for instruction, while the final two datasets are used for research. Each of the 24 writers is asked to create a test keyword, which is automatically stored in the database. Following that, each individual analysis is conducted. The first two-thirds of the handwriting data sample is used to train the system, and the remaining one-third of the handwriting data sample is used to verify the system's accuracy.

```
In [7]: from sklearn.svm import SVC
        classifier = SVC( kernel = 'linear' , random_state = 0)
        classifier.fit( X_train, y_train )
        y_pred = classifier.predict( X_test )

        from sklearn.metrics import confusion_matrix
        import matplotlib.pyplot as plt
        import numpy as np
        cm = confusion_matrix( y_test, y_pred )
```

FIGURE 4.3: SAMPLE CODE OF SVM KERNEL FOR CREATING CONFUSION MATRIX IN METHOD 1

```
from sklearn import datasets
from sklearn import svm
import numpy as np
from sklearn.metrics import accuracy_score
import numpy as np
data=[]
Name =['Akash','Dipta','fahim','Jannatul_Ferdoush','Linza','Lubna','mahi','Mohiuddin'
PATH = "F:\Data"
EXT = "*.csv"
all_csv_files = [file
                 for path, subdir, files in os.walk(PATH)
                 for file in glob(os.path.join(path, EXT))]
```

FIGURE 4.4: SAMPLE CODE FOR READING. CSV FILE FROM F DRIVE-IN METHOD 2.

In the second step, I have used four distinct algorithms to improve the precision of the measurement. LR, SVM, RF, and LDA are the algorithms used. The steps for Method 2 are as follows:

- Users insert information into the Pen Tablet computer using handwriting.
- All data input was stored in a.csv format.
- Data is read from .csv file.
- At this stage, I will refer to a.py file (created by python).
- For training and research, SVM, LDA, RF, and LR algorithms are used.
- show the accuracy result obtained by the use of algorithms.

Now I will explain how Python creates a.py file.

- When the.csv file is used to store the input data. Python is used to read the.csv file. As shown in Figure 4.4.
- Input data is handled based on the user's input.
- To create a balanced shape, I have used the Function to convert the two-dimensional data to one-dimensional data and then saved it as a list. As illustrated in Figures 4.5 and 4.6.

```
for file in all_csv_files:
    print(file)
    with open(file, 'r') as f:
        reader = csv.reader(f)
        your_list = list(reader)
        #print(len(your_list))
        flatten_list = list(chain.from_iterable(your_list))
        #num = [ a for a in flatten_list if a.isnumeric() ]
        #print(len(flatten_list))
        #print(flatten_list)
```

FIGURE 4.5: SAMPLE CODE FOR FLATTEN FUNCTION IN METHOD 2.

FIGURE 4.6: SAMPLE DATA FOR A SINGLE KEYWORD AFTER CONVERT DATA SET FROM 2D TO 1D

- After applying the Flatten function, provided an index number.
- Following the conversion of the Indexing List to a data frame.
- The imputation Function was used to balance the results. As shown in Figure 4.7.
- StandardScaler () and fit transform () functions are used to scale the data.
- np.unique () is used to denote the class. As shown in Figure 4.8.
- Since I am using the Jupyter Note book, our software is in the .ipynb format.
- I converted the .ipynb files to.py files.

In Table 4.3, I have described all of the resources and libraries that I have used to construct this framework. I have use the LR, SVM, RF, and LDA algorithms in Method 2 for both training and research. The algorithm's sample code is shown in Figure 4.10, 4.11, 4.12, and 4.13. The linear and RBF kernels are used in this implementation phase, followed by the F1 score for accuracy is used.

```
In [7]: from sklearn.preprocessing import Imputer
        imp_mean = Imputer(missing_values=np.NaN, strategy='mean')
        imp_mean.fit(X)
        X=imp_mean.transform(X)
```

FIGURE 4.7: SAMPLE CODE FOR IMPUTATION FUNCTION

```
In [9]: np.unique(y)
Out[9]: array([  0.,   1.,   2.,   3.,   4.,   5.,   6.,   7.,   8.,   9.,  10.,
                11.,  12.,  13.,  14.,  15.,  16.,  17.,  18.,  19.,  20.,  21.,
                22.,  23.])

In [10]: X = StandardScaler().fit_transform(X.data)

In [ ]:

In [11]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1, stratify=y)
         X_train.shape, y_train.shape, X_test.shape, y_test.shape
Out[11]: ((1009, 23214), (1009,), (253, 23214), (253,))
```

FIGURE 4.8: SAMPLE CODE FOR DENOTE CLASS

```
In [5]: import pandas as pd
        df = pd.DataFrame(data, dtype = float)

In [6]: y = df.iloc[:,0].values

        X = df.iloc[:,1:].values
        X.shape, y.shape
Out[6]: ((1262, 23214), (1262,))

In [7]: from sklearn.preprocessing import Imputer
        imp_mean = Imputer(missing_values=np.NaN, strategy='mean')
        imp_mean.fit(X)
        X=imp_mean.transform(X)

In [8]: np.asarray(y)
        np.asarray(X)
```

FIGURE 4.9: SAMPLE CODE FOR SEPARATE TRAINING AND TESTING CSV FILE.

In this process, 80% of data samples are used as training data and 20% as testing data. I have collected data on a single keyword five times. Additionally, in a few instances, I took six times. Each of the 24 writers was asked to create a test keyword, which is automatically saved to the database. I currently have 1262 csv files. Eighty per cent of the handwriting data (1009 csv files) is used to train the system, while the remaining twenty per cent (253 csv files) is used to verify the system's accuracy. To begin, training accuracy, testing accuracy, and F1 score for

five individuals are determined. Then the same procedure is repeated for ten individuals, fifteen persons, twenty persons, and eventually, twenty-four persons.

TABLE 4.3: TOOLS AND LIBRARIES FOR IMPLEMENTATION OF METHOD 2.

GUI	Anaconda Navigator
Platforms	Jupyter Notebook
Programming Language	Python
Libraries of python	panda, numpy
Algorithm	SVM, LDA,RF & LR
Device name	Wacom Tablet 6.3.32

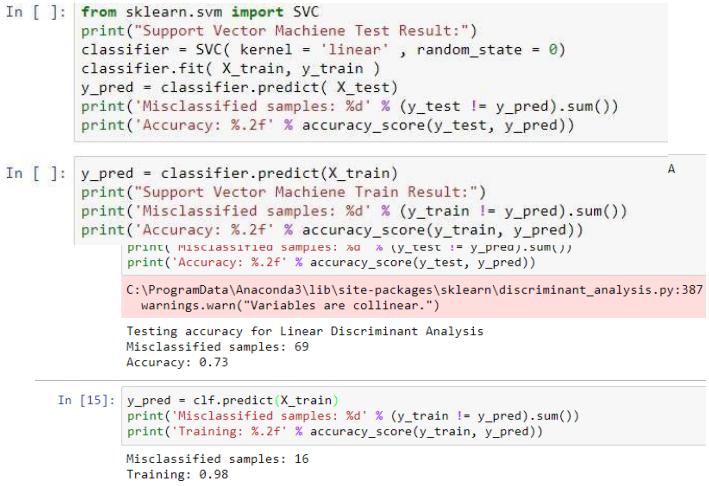
```
In [14]: from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
         clf = LDA()
         clf.fit(X_train, y_train)
         y_pred = clf.predict(X_test)
         print("Testing accuracy for Linear Discriminant Analysis  " )
         print('Misclassified samples: %d' % (y_test != y_pred).sum())
         print('Accuracy: %.2f' % accuracy_score(y_test, y_pred))

         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\discriminant_analysis.py:387
           warnings.warn("Variables are collinear.")
         Testing accuracy for Linear Discriminant Analysis
         Misclassified samples: 69
         Accuracy: 0.73

In [15]: y_pred = clf.predict(X_train)
         print('Misclassified samples: %d' % (y_train != y_pred).sum())
         print('Training: %.2f' % accuracy_score(y_train, y_pred))

         Misclassified samples: 16
         Training: 0.98
```

FIGURE 4.10: SAMPLE CODE FOR LR ALGORITHM.

```
In [12]: from sklearn.ensemble import RandomForestClassifier
         clf = RandomForestClassifier(n_estimators=10)
         clf = clf.fit(X_train, y_train)
         y_pred = clf.predict(X_test)
         print("Testing accuracy for Random Forest Classifier")
         print('Misclassified samples: %d' % (y_test != y_pred).sum())
         print('Accuracy: %.2f' % accuracy_score(y_test, y_pred))

         Testing accuracy for Random Forest Classifier
         Misclassified samples: 58
         Accuracy: 0.77

In [13]: y_pred = clf.predict(X_train)
         print('Misclassified samples: %d' % (y_train != y_pred).sum())
         print('Accuracy : %.2f' % accuracy_score(y_train, y_pred))

         Misclassified samples: 0
         Accuracy : 1.00
```

FIGURE 4.11: SAMPLE CODE FOR RF CLASSIFIER ALGORITHM.

FIGURE 4.12: SAMPLE CODE FOR LDA ALGORITHM.



```
In [ ]: from sklearn.svm import SVC
print("Support Vector Machiene Test Result:")
classifier = SVC( kernel = 'linear' , random_state = 0)
classifier.fit( X_train, y_train )
y_pred = classifier.predict( X_test)
print('Misclassified samples: %d' % (y_test != y_pred).sum())
print('Accuracy: %.2f' % accuracy_score(y_test, y_pred))

In [ ]: y_pred = classifier.predict(X_train)
print("Support Vector Machiene Train Result:")
print('Misclassified samples: %d' % (y_train != y_pred).sum())
print('Accuracy: %.2f' % accuracy_score(y_train, y_pred))
print('Accuracy: %.2f' % accuracy_score(y_test, y_pred))

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\discriminant_analysis.py:387
  warnings.warn("Variables are collinear.")

Testing accuracy for Linear Discriminant Analysis
Misclassified samples: 69
Accuracy: 0.73

In [15]: y_pred = clf.predict(X_train)
print('Misclassified samples: %d' % (y_train != y_pred).sum())
print('Training: %.2f' % accuracy_score(y_train, y_pred))

Misclassified samples: 16
Training: 0.98
```

FIGURE 4.13: SAMPLE CODE FOR SVM ALGORITHM.

4.5 RESULT ANALYSIS

To demonstrate the feasibility of our proposed scheme, I have quantified the accuracy of each algorithm. Our Result in Method 1 is very bad. SVM has very low accuracy, and the F1 score has a very low result of 54 per cent, as shown in Figure 5.1. The precision of Method 1 is shown in Figure 1, which was conducted on 24 individuals (1262 CSV files).

However, in Method two, our testing and training accuracy for all four algorithms is very high. The accuracy level of SVM is shown in Table 6. 1. The accuracy level for Logistic Regression (LR) is mentioned in Table 6. 2. The accuracy level for Linear Discriminant Analysis (LDA) is mentioned in Table 6. 3. The accuracy level of the Random Forest Classifier Algorithm (RF) is mentioned in Table 6. 4.

As shown in Table 5.6, the training accuracy is nearly 100% for all algorithms. In Method two, our testing accuracy is 73% for LDA, 85% for LR, and 82% for SVM (RBF kernel), and 77% for the RF algorithm.

```
1  from sklearn.metrics import f1_score
2
3  a =  f1_score(y_test, y_pred, average='macro')
4
5  b =  f1_score(y_test, y_pred, average='micro')
6  c =  f1_score(y_test, y_pred, average='weighted')
7
8  print( a , b , c  )
```

```
54.056789 52.453782 54.154237
```

FIGURE 5.1: F1 SCORE FOR METHOD 1

TABLE 5.1: ACCURACY FOR SVM ALGORITHM (KERNEL: RBF).

No of Persons	Training	Testing	F1 score
5	97%	82%	84%
10	98%	83%	83%
15	99%	80%	82%
20	98%	80%	81%
24	98%	81%	80%

TABLE 5.2: ACCURACY FOR LR ALGORITHM.

No of Persons	Training	Testing	F1 score
5	100%	94%	84%
10	100%	86%	80%
15	100%	86%	70%
20	100%	87%	73%
24	100%	85%	76%

TABLE 5.3: ACCURACY FOR LDA ALGORITHM.

No of Persons	Training	Testing	F1 score
5	97%	92%	89%
10	98%	89%	85%
15	99%	85%	82%
20	98%	79%	77%
24	98%	76%	69%

TABLE 5.4: ACCURACY FOR RF ALGORITHM.

No of Persons	Training	Testing	F1 score
5	100%	92%	78%
10	100%	75%	87%

15	100%	86%	81%
20	100%	82%	80%
24	100%	77%	78%

TABLE 5.5: ACCURACY FOR SVM ALGORITHM (KERNEL: LINEAR).

No of Persons	Training	Testing	F1 score
5	100%	96%	89%
10	100%	90%	85%
15	100%	92%	86%
20	100%	92%	87%
24	98%	87%	88%

```
In [11]: from sklearn.metrics import f1_score

         a =  f1_score(y_test, y_pred, average='macro')

         b =  f1_score(y_test, y_pred, average='micro')
         c =  f1_score(y_test, y_pred, average='weighted')

         print( a , b , c  )

0.869766925042 0.886153846154 0.8834271305
```

FIGURE 6.2: F1 SCORE FOR METHOD 2 BY USING SVM ALGORITHM

While I am aware that SVM has multiple kernels, in our method, I used the linear and radial basis functions. The linear kernel's accuracy in SVM is 87 per cent. The accuracy level obtained after applying the linear kernel to the SVM is shown in Table 5.5. The accuracy level of LR and SVM (Linear kernel) is significantly higher than the other two algorithms. The optimal F1 score is obtained using the SVM (linear kernel) algorithm depicted in Figure 5.2. Table 5.6 compares the four algorithms in aggregate. TR denotes training accuracy, Te denotes testing accuracy, and FS denotes F1 ranking. It is worth noting that after pre-processing the data set in Method-2, the outcome is superior to Method one without pre-processing of data.

TABLE 5.6: ACCURACY COMPARISON OF DIFFERENT ALGORITHM

No of Persons	Accuracy comparison											
	SVM (linear)			LR			LDA			RF		
	Tr	Te	FS	Tr	Te	FS	Tr	Te	FS	Tr	Te	FS
5	100%	96%	89%	100 %	94%	84%	97%	92%	89%	100 %	92%	78%
10	100%	90%	85%	100 %	86%	80%	98%	89%	85%	100 %	75%	87%
15	100%	92%	86%	100 %	86%	70%	99%	85%	82%	100 %	86%	81%
20	100%	92%	87%	100 %	87%	73%	98%	79%	77%	100 %	82%	80%
24	98%	87%	88%	100 %	85%	76%	98%	76%	69%	100 %	77%	78%

4.6 SUMMARY OF THE CHAPTER

This chapter provided experimental result and a discussion of the study; the next chapter will conclude the study with concluding remarks.

CHAPTER FIVR

CONCLUSION AND FUTURE WORK

5.1 OVERVIEW OF THE CHAPTER

This is the concluding chapter of the thesis, this chapter provides discussions on how this future studies in this area can be done and conclude the paper with concluding remarks.

5.2 FUTURE WORK

Though the study has significant contribution in its field there are some limitations that I will try to overcome in future. Firstly in future I would like to focus on more consistent dataset collection to improve the degree of accuracy. My plans also include the expansion of the proposed framework for identifying personalities or emotions. This could be done in real-time, like in a banking system's real-time user authentication. The more reliable and accurate data is collected, the more precise the outcome would be.

5.3 CONCLUSION

This paper aims to build a framework capable of identifying a person's gender and identity from their handwriting using machine learning. In this paper, I propose a user authentication scheme based on pen tablet data and handwriting. The accuracy levels for SVM, LR, LDA, and RF are 87 per cent, 85 per cent, 73 per cent, and 77 per cent, respectively, based on the review of the experimental results. SVM and LR both have an accuracy rating of greater than 80 per cent.

BIBLIOGRAPHY

[1] M. Al Emran, S. Naief, M. Hossain, et al., Handwritten character recognition and prediction of age, gender and handedness using machine learning. PhD thesis, BRAC University, 2018.

[2] S. Gupta, "Automatic person identification and verification using online handwriting," International Institute of Information Technology, vol. 408, 2008.

[3] Á. Morera, Á. Sánchez, J. F. Vélez, and A. B. Moreno, "Gender and handedness prediction from offline handwriting using convolutional neural networks," Complexity, vol. 2018, 2018.

[4] V. Pervouchine and G. Leedham, "Extraction and analysis of forensic document examiner features used for writer identification," Pattern Recognition, vol. 40, no. 3, pp. 1004–1013, 2007.

[5] S.-H. Cha and S. Srihari, "Multiple feature integration for writer verification," in Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition, pp. 333–342, 2000.

[6] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, "Individuality of handwriting: a validation study," in Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 106–109, IEEE, 2001.

[7] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biological cybernetics, vol. 36, no. 4, pp. 193–202, 1980.

[8] J. Fierrez, A. Pozo, M. Martinez-Diaz, J. Galbally, and A. Morales, "Benchmarking touchscreen biometrics for mobile authentication," IEEE Transactions

on Information Forensics and Security, vol. 13, no. 11, pp. 2720–2733, 2018.

[9] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello, “Continuous user authentication on mobile devices: Recent progress and remaining challenges,” IEEESignal Processing Magazine, vol. 33, no. 4, pp. 49–61, 2016.

[10] N. Sae-Bae, N. Memon, K. Isbister, and K. Ahmed, “Multitouch gesture-based authentication,” IEEE transactions on information forensics and security, vol. 9, no. 4, pp. 568–582, 2014.

[11] A. Rahiman, D. Varghese, and M. Kumar, “Habit: Handwritten analysis based individualistic traits prediction,” International Journal of Image Processing (IJIP), vol. 7, no. 2, p. 209, 2013.

[12] S. Z. Mishu and S. Rafiuddin, “Performance analysis of supervised machine learning algorithms for text classification,” in 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 409–413, IEEE, 2016.

[13] X. Wang, X. Ding, and H. Liu, “Writer identification using directional element features and linear transform,” in Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 942–945, Citeseer, 2003.

[14] S. N. Srihari, S.-H. Cha, and S. Lee, “Establishing handwriting individuality using pattern recognition techniques,” in Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 1195–1204, IEEE, 2001.

[15] R. Plamondon and S. N. Srihari, “Online and off-line handwriting recognition: a comprehensive survey,” IEEE transactions on pattern analysis and machine intelligence, vol. 22, no. 1, pp. 63–84, 2000.

[16] H. Bunke, S. Bengio, and A. Vinciarelli, “Offline recognition of unconstrained handwritten texts using hmms and statistical language models,” IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 6, pp. 709–720, 2004.

[17] M. T. F. Rabbi, S. T. Rahman, P. Biswash, J. Kim, A. Sheikh, A. K. Saha, and M. S. Uddin, “Handwritten signature verification using CNN with data augmentation,” The Journal of Contents Computing, vol. 1, no. 1, pp. 25–37, 2019.

[18] A. Kumar, D. C. Wong, H. C. Shen, and A. K. Jain, “Personal authentication using hand

[19] images,” Pattern Recognition Letters, vol. 27, no. 13, pp. 1478–1486, 2006.

[20] D. Ritter, F. Schaub, M. Walch, and M. Weber, “Miba: Multitouch image-based authentication on smartphones,” in CHI’13 Extended Abstracts on Human Factors in Computing Systems, pp. 787–792, 2013.

[21] X. Zhao, T. Feng, W. Shi, and I. A. Kakadiaris, “Mobile user authentication using statistical touch dynamics images,” IEEE Transactions on Information Forensics and Security, vol. 9, no. 11, pp. 1780–1789, 2014.

[22] J. L. Traenkenschuh and D. W. Gilles, “User authentication with image password,” Feb. 5 2009. US Patent App. 11/882,553.

[23] M. Kosugi, T. Suzuki, O. Uchida, and H. Kikuchi, “Swipass: Image-based user authentication for touch screen devices,” Journal of Information Processing, vol. 24, no. 2, pp. 227–236, 2016.

[24] W. Yicong and H. Xu, “Image analysis for user authentication,” Dec. 1 2015. US Patent 9,202,105.

[25] N. O'donoghue, "Method and device for customized picture-based user identification and authentication," Mar. 17 2005. US Patent App. 10/930,741.

[26] P. Samangouei, V. M. Patel, and R. Chellappa, "Attribute-based continuous user authentication on mobile devices," in 2015 IEEE 7th international conference on biometrics theory, applications and systems (BTAS), pp. 1–8, IEEE, 2015.

[27] S. Bandyopadhyay and V.Y. Law, "Image-based unlock functionality on a computing device," June 4 2013. US Patent 8,458,485.

[28] S. V. Raghavan, "Methods and devices for pattern-based user authentication," May 29 2012. US Patent 8,191,126.

[29] S. Cho and M. Jang, "System and method for performing user authentication based on user behavior patterns," Oct. 11 2007. US Patent App. 11/651,132.

[30] B. Fallah and H. Khotanlou, "Identify human personality parameters based on handwriting using neural network," 2016.

[31] T. Masui, "Image processing apparatus, user authentication method and storage medium storing program for user authentication," May 3 2007. US Patent App. 11/403,084.

[32] M. Savvides and B. V. Kumar, "Illumination normalization using logarithm transform for face authentication," in International Conference on Audio-and Video-Based Biometric Person Authentication, pp. 549–556, Springer, 2003.

[33] H. Atobe, "Communication system, image processing apparatus, image processing method, authentication server, image managing method, image managing program, and image processing system," Nov. 20 2012. US Patent 8,314,958.

PERSON AND GENDER IDENTIFICATION FROM HANDWRITING USING MACHINE LEARNING

ORIGINALITY REPORT

13%	3%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Md. Azim Hossain Akash, Nasima Begum, Sayma Rahman, Jungpil Shin, Md Amiruzzaman, Md Rashedul Islam. "User Authentication Through Pen Tablet Data Using Imputation and Flatten Function", 202020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), 2020 Publication	7%
2	Submitted to University of Asia Pacific Student Paper	1%
3	Refat Khan Pathan, Mohammad Amaz Uddin, Nazmun Nahar, Ferdous Ara, Mohammad Shahadat Hossain, Karl Andersson. "Chapter 51 Gender Classification from Inertial Sensor-Based Gait Dataset", Springer Science and Business Media LLC, 2021 Publication	<1%
4	Jana Dittmann, Arnd Steinmetz, Ralf Steinmetz. "Chapter 21 Content-Fragile	<1%