HANDWRITING ANALYSIS AND PERSONALITY: A COMPUTERIZED STUDY ON THE VALIDITY OF GRAPHOLOGY

AFNAN GAROOT

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE & SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

OCTOBER 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Mrs. Afnan Garoot**

Entitled: **Handwriting Analysis and Personality:**
A Computerized Study on the Validity of Graphology

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Doctor of Philosophy (Computer Science))

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Chun Wang ________ Chair
Dr. Mounim El-Yacoubi ________ External Examiner
Dr. Nawwaf Kharma ________ External to Program
Dr. Weiyi(lan) Shang ________ Examiner
Dr. Olga Ormandjieva ________ Examiner
Dr. Ching Y. Suen ________ Supervisor

Approved ________
Chair of Department or Graduate Program Director

October 1, 2021 ________

Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Handwriting Analysis and Personality:
A Computerized Study on the Validity of Graphology

Afnan Garoot, Ph.D.
Concordia University, 2021

Handwriting analysis, also known as graphology, is defined as an analysis of a psychological structure of a human subject through his/her handwriting. It has been applied recently in different fields including areas where making a crucial decision is highly desirable such as forensic evidence, criminology, and disease analysis. However, making a crucial decision based on the results of handwriting analysis is a controversial scientific topic because the validity of graphology rules is still in question.

A few validity studies on handwriting analysis have already been conducted earlier using the evaluation of correlation between psychological questionnaires and manual handwriting analysis and they ended up with conflicting results. Manual handwriting analysis suffers from some issues which could be the reasons of the early inconsistent results. Therefore, this study conducted an empirical study that investigates the validation of graphology rules by evaluating the correlation between one of psychological tests named Big Five Factor Markers Test and our proposed automated handwriting analysis system that measures the level of the same big five personality traits based on graphological rules.

Our automated BFFM system is called Averaging of SMOTE multi-label SVM-CNN (AvgMlSC). It constructs synthetic samples using Synthetic Minority Oversampling Technique (SMOTE) and averages two learning-based classifiers i.e. Multi-label Support Vector Machine and Multi-label Convolutional Neural Network based on offline handwriting recognition to produce one optimal predictive model. The model is trained using 1066 handwriting samples written in English, French, Chinese, Arabic, and Spanish. The results reveal that our proposed model outperformed the overall performance of five traditional

models with 93% predictive accuracy, 0.94 AUC, and 90% F-Score.

The statistical test of Spearman's correlation reports that there is a statistically significant relationship between the score of the big five factors questionnaire and the graphologist's evaluation for the Big Five Factors with a different strength of relationship. A weak positive relationship is found for Extraversion. However, a moderate positive relationship is reported for Conscientiousness and Open to Experience. On the other hand, a strong positive relationship is indicated for Agreeableness, whilst a very weak positive relationship has been found for the last factor which is Emotional Stability.

Acknowledgments

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Dr. Ching Y. Suen, for supervising me with patience, understanding and providing moral and financial support. Thanks to our graphologist, Graziella Pettinati, who has dedicated her time to provide valuable workshops and resources on graphology while working hard on labelling our dataset. Thanks to Dr. Andrew Ryder, Dr. Norman Segalowitz, and Mr. Bruce Wong for sharing their knowledge on psychology test. Thanks to Dr. Muna Khayyat, Dr. Jehan Janbi, and Mr. Nicola Nobile for all their cooperation and help to facilitate what my study required from documents, software, devices, and tools.

Thanks to Um Al-Qura university for giving me the opportunity to study abroad and get my Ph.D. degree from Concordia university in Canada. Thanks to Saudi Cultural Bureau, International Student office, and Concordia Student Union for their support in any difficulties I faced during the study and the expatriation.

Special thanks for my dear husband, Ayman Garout, for his endless support and help. He was always there for me. I could not have completed this dissertation without his understanding and cooperation. Thanks to my lovely sons Mohab, Shihab, and Awwab for bringing me joy and happiness during this hard journey. Thanks to my mother and siblings for their caring, encouraging and supporting me despite the distance between us. Thanks to my friends, Mrouj and Alaa, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

Contents

List of Figures

List of Tables

Chapter 1

Study Overview

1.1 Introduction

Graphology, also known as handwriting analysis, is a technique used to evaluate and interpret the character of the writer from his or her handwriting. Nowadays, there are different psychological tests on the market used to discover human personality such as aptitude studies, psychometric tests, and other long questionnaires that require time to answer. By comparing the practicality of these different diagnostic methods, it has been found that graphology is the fastest way. It is a simple, thorough, and quick test since it only requires a sample of handwriting for assessing the personality traits of the writer.

Graphology started manually by extracting specific handwriting features from handwriting sample. Then, interpreting the extracted features into personality traits based on graphological rules. However, manual graphology has a number of issues. It is a tedious, subjective, error prone task, and sometimes leads to inconsistent results between graphologists. Therefore, computerized handwriting analysis systems have been developed in order to help graphologists to extract and analyze handwriting features faster and more precisely using computers. It takes a handwriting sample as an input and produces a personality description of the writer as an output.

There are many uses and applications for graphology. The most popular applications used today are dating and socializing, roommates and landlords, entertainment at parties and conventions, business and professional, employee hiring and human resources, police profiling, self-improvements and professional speakers, counsellors, therapists, and coaching applications. Moreover, it has been applied recently in different major areas such as forensic evidence, criminology, disease analysis which makes handwriting analysis a controversial scientific study.

1.2 Problem Statement

Graphology is considered as a controversial scientific study since the validity of the analysis rules is still in question. Few and old studies have been conducted previously to investigate the validity of handwriting analysis. All of those studies evaluated the correlation between psychological tests and manual handwriting analysis to examine the validity of graphology, and they generated inconsistent results. The reasons of the inconsistency could be the following issues associated with manual handwriting analysis. Manual graphology is a tedious, subjective, and error prone task. It gives different prediction results between graphologists. In addition, it might be influenced by the content of handwriting [45]. For this, a reliable validation method is required in order to evaluate graphological rules more accurately. Therefore, using an automated handwriting analysis system instead of manual one could be the ideal method for this purpose. See Figure 1 for comparing the early validation method and the proposed validation method.

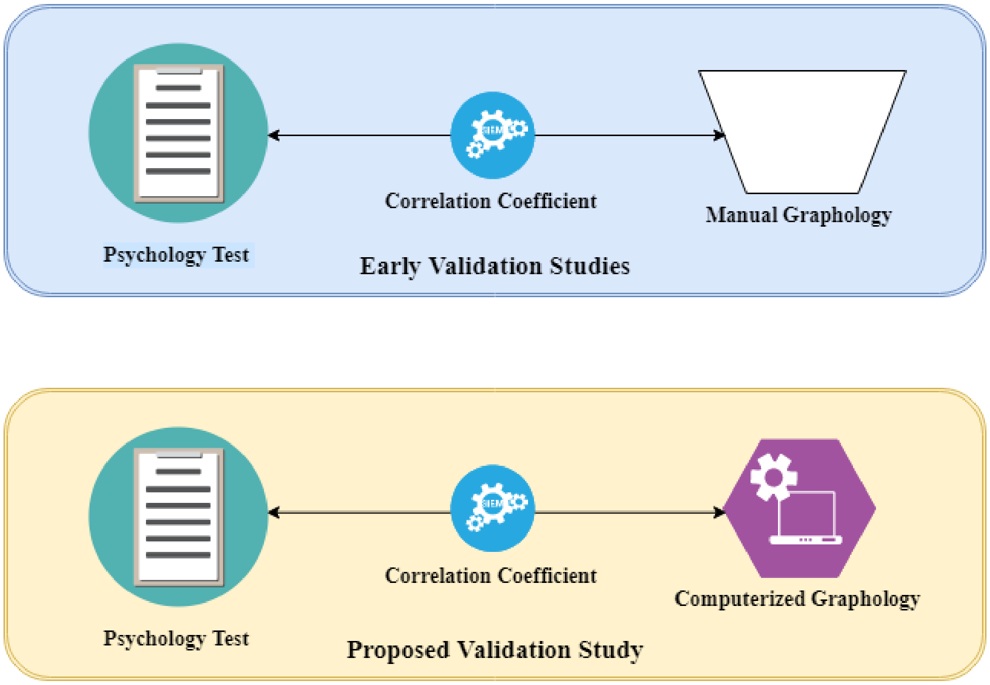


Figure 1: Early Validation Studies vs Proposed Validation Study

1.3 Motivation

The importance of investigating the validity of handwriting analysis came from its uses and applications. It has been applied in different major areas such as forensic evidence, criminology, disease analysis, psychiatry, therapists, employee hiring, human resources, personality prediction, and self-improvement. It can be observed that making a crucial decision is considered as an essential requirement in these fields especially for forensic evidence, criminology, and disease analysis. Making a crucial decision based on the results of handwriting analysis is controversial since the validity of graphology rules is still in question. Therefore, the validation of handwriting analysis must be proven firstly in order to be considered as a reliable and valid study that can be applied in these areas.

1.4 Objectives

In this study, we aim to examine the validity of handwriting analysis as a personality assessment tool by evaluating the Spearman's correlation between the scores of a psychological test named Big Five Factor Markers Test (BFFMT) which is a self-report test that measures the big five personality traits such as Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Open to Experience based on the international personality item pool-big five factor markers and our proposed automated handwriting analysis system that measures the level of the same big five personality traits based on graphological rules. Figure 2 shows the processes followed in this study for investigating the validity of graphology.

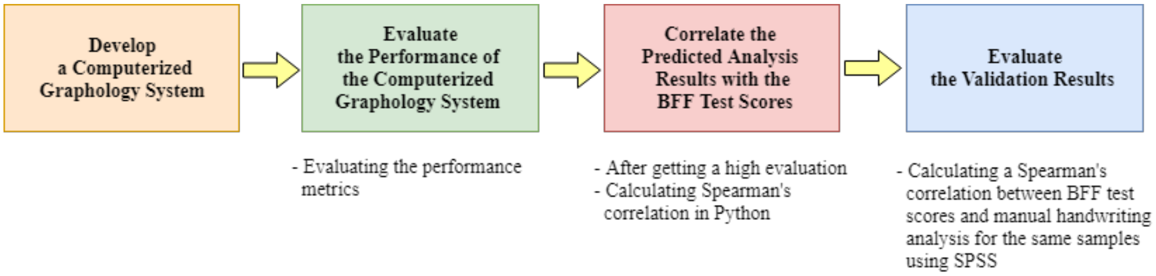

Develop a Computerized Graphology System
Evaluate the Performance of the Computerized Graphology System
Correlate the Predicted Analysis Results with the BFF Test Scores
Evaluate the Validation Results
- Evaluating the performance metrics
- After getting a high evaluation
- Calculating Spearman's correlation in Python
- Calculating a Spearman's correlation between BFF test scores and manual handwriting analysis for the same samples using SPSS

Figure 2: Diagram for Examining the Validity of Graphology

We propose an automated graphology system that implements multi-label algorithm on an ensemble learning model named AvgMlSC. AvgMlSC deploys two learning-based classifiers which are Multi-label Support Vector Machine (MLSVM) and Multi-label Convolutional Neural Network (MLCNN) based on off-line handwriting recognition. It employs the averaging ensemble method along with SMOTE resample technique in order to handle the issue of imbalanced dataset.

The study tries to answer the following research question:
RQ: Is handwriting analysis a valid tool to predict human personality traits?
The hypothesis of our study is stated based on the conclusion of the majority of early validation studies that have been conducted on handwriting analysis:
H: There is no correlation between the results of handwriting analysis and the scores of BFF test.

1.5 Limitations

These are the limitations that we encountered during our implementation:

1. Data Collection:

    - Since we are applying deep learning in our proposed system, we need to collect a big set of training data in order to get a high accuracy of results. Our aim is to collect at least 2000 handwriting samples. However, our data size has been limited to 1108 handwriting samples because of the pandemic.
    - Our questionnaire should be filled in a specific environment because handwriting could be influenced by the mental and environmental conditions of the writer. For this, we have dedicated one laboratory at Centre for Pattern Recognition and Machine Intelligence (CENPARMI) at Concordia University for collecting our data. However, waiting for participants to come over to the lab was another challenge that slowed down the process of data collection.
    - Imbalanced dataset is another challenge for our study that is caused by dataset size and data distribution.

2. Model Training:

    - Training the model with big set of data requires a significant amount of time and memory.

- To overcome the issue of time and memory, a graphical processing unit (GPU) with high capacity is needed which costs a lot of money.

1.6 Thesis Outline

The thesis is organized as follows. The second chapter reviews the definition of handwriting analysis and explains how it works. It reviews the three basic kinds of measurements for handwriting features used by graphologists to assess the personality traits. It also reviews some related works on the validity of graphology. The third chapter is intended to provide information about definition, advantages, and applications of computerized handwriting analysis. It explains the four main modules of the computerized graphology system. It also reviews some related works on computerized graphology system. The fourth chapter reviews the history of the Big Five Factors Model (BFFM) and defines each factor. It also reviews some related works on computerized prediction of BFF. The fifth chapter describes the process of data collection, data digitization, and data preprocessing. The sixth chapter describes the materials and methods. The results of the experiment are presented in chapter 7. It concludes with a discussion and a summary of the results.

Chapter 2

Graphology/Handwriting Analysis

2.1 Definition and how it Works

Graphology is considered as a modern form of psychology that reveals personality traits, including emotional outlay, fears, honesty, defences and others, from individual's handwriting but not identifying the writer's age, race, religion, or nationality. In other words, it is a technique used to evaluate and interpret the character of the writer from handwriting [37].

As it is mentioned in [44], there are two major schools of handwriting analysis which are Graphoanalysis and Gestalt graphology. Graphoanalysis is the most widely used in the United States. This is in which the graphoanalyst looks at the page as a collection of symbols where each symbol is evaluated independently from the whole [42]. However, Gestalt graphology is the school of handwriting study in Europe, particularly in Germany. It was developed alongside psychiatry and psychoanalysis. This is in which the Gestalt graphologist evaluates the symbols as these are related to the whole, to uncover the pattern or "gestalt" that indicates various aspects of the writer's personality [42]. In other words, Gestalt graphology is always a combination of features related to form, movement and space that are used to establish a personality trait. It follows the following steps mentioned in [13]:

1. Assessing the Gestalt (over-all view) of the writing.

2. Assessing the form standard of the writing; be it low or high.

3. Identifying dominant; sub dominant and counter dominant features in the writing.

4. Interpreting the co-existing features and compiling the synthesis.

In this work, we applied Gestalt graphology concept for analysing and labelling our handwriting samples used for training our computerized graphology system.

2.2 Handwriting Features

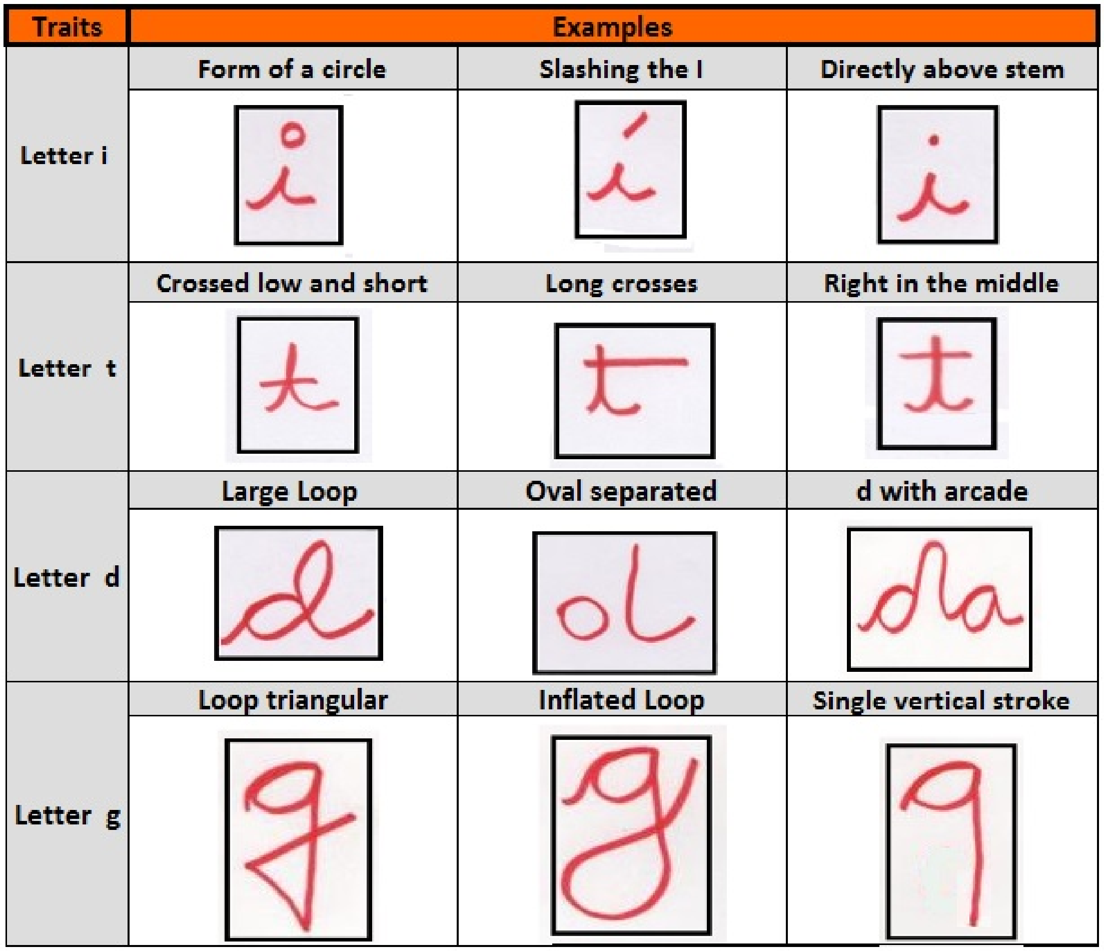
As stated in [2], there are three kinds of measurements for handwriting features used together by graphologists to assess the personality traits:

1. General Measurements give an overall impression of handwriting. Graphologists associate personality traits with the stroke quality. For example, a small stroke in superior handwriting is a measure of good concentration, while in inferior handwriting it measures pettiness.

2. Fundamental Measurements produce a primary classification of handwriting patterns. For this purpose, eight basic features are used. These features are: slant which defines emotions, baseline direction which indicates the writer's mood, letter size which indicates power, continuity which indicates mode, handwriting form (shape) which indicates natural impulse and free choice, arrangement of lines on a page or words in lines which indicate sense of organization and adaptability, pen pressure which signifies the intensity, strength, and appetites, and writing speed which depicts the rhythm of physical and mental activity of the writers.

3. Accessories Measurements are graphic symbols that appear in 't' bars, 'i' dots, punctuation marks, capital letters, signatures, numerals, initial and terminal strokes, covering strokes, flourishes, upper and lower extensions or loops and alphabet. The letter 't' is the most important graphical symbol that signifies the writer's well power and activeness. The letter 'i' is the second most important graphic symbol which signifies the speed, sphere of the imagination, intellect, ideals and aspiration. In addition, the size of a capital letter represents the pride, vanity and desire to impress others. The beginning strokes signify the consciousness, immaturity, conventionality, love of gain, selfishness, gaiety, while the ending strokes signify practicality, aspirations, and courage of the writer. Moreover, the loops occurring in letters also contribute personality related information like intellectual and physical activities, moderation,

exaggeration, fluency of thoughts, idealism, aspirations, ambition and spirituality, little vision of imaginations, and realistic practical viewpoints, egoism, boastfulness, vanity, timidity and inhibition, fear and apprehension for future etc. Table 1 shows examples of handwriting features for letter i, t, d, and g.



Traits	Examples		
Letter i	Form of a circle	Slashing the I	Directly above stem
Letter t	Crossed low and short	Long crosses	Right in the middle
Letter d	Large Loop	Oval separated	d with arcade
Letter g	Loop triangular	Inflated Loop	Single vertical stroke

Table 1: Examples of Handwriting Features for Letter i, t, d, and g [2]

Table 2 lists the handwriting traits that used together to reveal some of personal characteristics such as activity, adaptability, aggressiveness, ambition, and anger.

Personal Characterstic	Handwriting Traits
Activity	Speedy, angles, high pressure, large size, rightward slant firm down stroke
Adaptability	Curved form, garlands, even pressure, moderated speed
Aggressiveness	Heavy pressure, angles, upstrokes straight and diagonal, speed, exaggerated right thrown tending t bars
Ambition	Uneven pressure, speed, large capitals, rising bars, rising lines, extenstions into upper and lower zones
Anger	Bars and terminal strokes heavy, high and pointed

Table 2: Handwriting Traits for some Personal Characteristics [2]

2.3 Related Works on the Validity of Graphology

Handwriting analysis is considered as a controversial scientific study. The majority of the empirical studies that have been conducted previously evaluated the correlation between personality questionnaires and manual handwriting analysis in order to examine the validity of graphology. Some researchers concluded with supportive results while the others were refutative.

In 1986, two empirical studies were reported by Ben-Shakhar and Bar-Hillel concerning the validity of graphological prediction [6]. Their conclusion states that graphology cannot be considered as a valid predictor of occupational success. Two years later, a rejoinder to Ben Shakhar et al.'s paper was made by Nevo [34]. In this rejoinder an attempt was made to re-analyze the data of Ben Shakhar et al. and to reinterpret their findings. The conclusion here was supportive of graphology. In 1990, another investigation was done to examine the validity of graphology as predictor of academic achievements [35]. It shed some light on the validity of graphology as a psychometric measure, and there was fairly good evidence for its predictive value. In 2000, King and Koehler examined the illusory correlation phenomenon as a possible contributor to the persistence of graphology's use to predict personality [27]. The results were partially accounted for continued use of graphology despite overwhelming evidence against its predictive validity. Adrian F. et al. reported two studies in 2003 where similar methodology has been used to examine whether graphology predicts personality and intelligent or not [17]. The results showed no robust relationship between graphology and personality. In 2009, two researchers conducted another study on validity of handwriting analysis [15]. The study evaluated the correlation between the big five questionnaire and graphological evaluations and it did not confirm the capability of handwriting analysis to measure Big Five personality traits. The results showed that no evidence was found to validate the graphological method as a measure of personality. In 2014, Gawda aimed to verify whether or not there are any specific characteristics of writing in relation to personality traits [21]. Two different studies have been conducted with a different number of subjects and two different techniques for personality assessment. However, the same set of handwriting features was analyzed in each study. The study concluded no writing characteristics were specific to personality traits. In other words, there is no evidence for assessment of personality based on handwriting.

Chapter 3

Computerized Handwriting Analysis System

3.1 Definition, Advantages, and Applications

Computerized handwriting analysis is defined as a system that helps graphologists to extract and analyze handwriting features faster and more precisely using computers. It takes a handwriting sample as an input and produces a personality description of the writer as an output. Personal handwriting analysis made by a computer is fast, accurate and it identifies the handwriting superior to visual inspection. Moreover, computer assisted handwriting analysis is automated, efficient, and devoid of human errors [26]. It is being applied in different areas such as business, psychiatry, medicine and criminology [2].

3.2 The Main Modules

1. Handwriting Digitization: it is the process of digitizing the handwriting sample into the computer. It can be on-line or off-line. In off-line, the handwriting is written on a piece of paper and then it is digitized into the computer using scanning technology or by photographic capture. However, on-line writing is digitized at the time of creation which is usually done by using tablet PC and a special stylus or pen.

2. Preprocessing: after the digitization process, a sequence of data preprocessing operations such as thresholding, normalization, smoothing are normally applied on the digitized handwriting to put it in a suitable format ready for feature extraction [11].

3. Handwriting feature extraction: handwriting features are considered as measurements applied to a word or a character and combined together to produce a measurement vector/feature vector [1]. Any character, word, digit, or stroke has its own feature vector to distinguish it from any other characters, words, or strokes. Extracting an appropriate set of features and applying an efficient extraction method are considered as the most important factors in achieving high recognition performance for pattern recognition in general and handwriting analysis in particular. Choosing the proper type of features depends on the nature of the text, the type of handwriting recognition which might be on-line or off-line, and the script types that can be handwritten or printed.

4. Handwriting feature analysis: the goal of the final module is to find the personality traits of the writer from extracted handwriting feature using some classification methods such as template matching, support vector machine (SVM), artificial neural network (ANN), or others.

3.3 Related Works on Computerized Graphology System

In 1997, a handwriting analyzer software has been developed by Shiela Lowe [36]. It follows German's theory of handwriting analysis called Gestalt or holistic graphology concept. There are 65 personality traits that are important and identifiable in handwriting. For every trait, there is a list of handwriting characteristics to match. The software can describe up to 65 personality traits and uses up to 5000 signs to do this. There are two versions in use around the world, i.e. professional version and personal version. The professional version provides a number of features such as comprehensive report, key phrase report, results review, personality trait graph, interest indicator chart, job success potential indicators, vocational interest graph, job matching profile manager, applicant/job ranking, and ability to edit reports. The software asks the user many questions about the handwriting style. It does not do any image analysis and requires the user to provide the information such as degree of slant and margin size. At least 10 of the trait categories with ranges of low, moderate, average, high, and extreme should be completed in order for a report to be generated. The user can choose how many more he wants to choose. However, the more the user chooses, the more accurate the report should be. Unfortunately, the references do not provide technical details about the software other than that it runs on Windows. Examples of personality traits analyzed by the software are shown in Table 3 along with their list of characteristics and definitions.

1. Motivating Forces	
Emotional Independence	Sensitivity to external stimuli
Need for Freedom	Refusal to be bound by rules and convention
Need for Harmony	Preference for an agreeable environment
Need to Achieve	Restlessness; desire for physical activity
2. Personal Dynamics	
Reliability	Consistent in keeping promises and behaving dependably
3. Ego Strength	
Integrity	Moral soundness
Pride	Self-respect and self-satisfaction
Self-assertiveness	Ability to make one's presence felt; refusal to accept opposition
Self-confidence	Reliance on one's own capacities
Self-esteem	Sense of personal worth and dignity
Willpower	Power to accomplish difficult tasks by means of one's own strength of will
4. Defenses & Controls	
Acquisitive	Desire to collect and hold on to money, goods, or relationships
Impulse Control	Ability to reflect before acting
Inhibitions	Restraint of unacceptable impulses
Perfectionist	Effort to make everything the best it can be
5. Intellectual Style	
Analytical Thinking	Need to analyze problems in minute detail
Imagination	Power to form mental images of objects not physically
Objectivity	Perception of an external event without involving one's emotions
6. Communication Style	
Frankness	Lack of guile
Need for Privacy	Preference not to share personal matters with others
Sense of Humor	Capacity to see the lighter side of life
Sincerity	Completely honest and genuine
Tact & Diplomacy	Ability to refrain from giving offense when giving

	unfavorable comment
7. Interpersonal Style	
Sensitivity	Keen awareness of the moods and feelings of others
Sociability	Desire for the company of friends and/or acquaintances
8. Work Style	
Attention to Details	Deals with the finer elements of a matter
Goal-directedness	Ability to develop and focus on future objectives
Openness to Change	Readiness to consider alternatives without prejudice
Team Player	Ability to work in a group setting on a common task
Work Ethic	Belief that work is good and recreation is to be earned
9. Sales Style	
Initiative	Capacity to conceive and act on one's own ideas
Showmanship	Flair for putting on an entertaining presentation
10. Management Style	
Conflict Management	Inability to stand firm for established principles
Entrepreneurship	Degree of initiative
Leadership Aptitude	Ability to direct others
Organizational Aptitude	Ability to keep things in their proper place
11. Red Flags	
Argumentative	Need to be right at all costs
Pessimistic	Tendency to look on the downside of every situation

Table 3: Examples of Personality Traits analyzed by Sheila's Handwriting Analyzer Software

In 2008, the construction of an automated graphology system was presented by Ahmend and Mathkour [2]. The system deployed a rule-based method which was equipped with the conventional pattern recognition techniques and an inference engine to make decisions along with an explanation. The architecture of the rule-based system consists of three modules: feature extraction, inference engine, and knowledgebase. Eight features were used by the system, which are: slant, base line, speed, size, continuity, form, arrangement, and pressure. However, six personality attributes were computed by the system, which are: emotions, mood, self-confidence, coherence of thoughts, strength and organization. 35 students' handwriting samples were used by the system as a test data set. Since the system presented in this study was in a development stage, the results needed to be more accurate in measurements, reliability, and the features needed to be extracted in a finer manner.

In 2011, Champa and Ananda Kumar published another paper which focused particularly on the two important features to analyze and predict the writer's personality in an off-line system [8]. Baseline and writing pressure were the features considered by them in this article. To categorize personalities, they applied a rule-based approach with 9 rules formed by three types of baseline (ascending, descending and level), and also three kinds of writing pressure (light, medium and heavy).

In the following year, Grewal and Prashar worked on features like baseline, the letter slant, pen pressure same as the other researchers as the main features which can contain accurate information about the personality trait [24]. Furthermore, they did research on 6 common formations of the letter 'i' and the letter 'f' which show the ability of planning and organization of writer. After feature extraction, all these features were used as inputs of ANN to predict personality of the writer.

In 2013, according to [40] another system was proposed by Abdul Rahiman et al. to predict the personal behavior of individuals from their digital handwriting samples. The behavioral analysis was done based on the following features: the pen pressure, the slant of letters and the slant of baseline, the size of letters, and spacing between words. To implement the system, a simple linear regression method was used, which is an approach to model the relationship between a scalar dependent variable y and an explanatory variable denoted by x. The implemented method proved successful in analyzing handwriting regardless of the language used.

In 2014, Raut and Bobade extended another research that was conducted by Prasad et al. in 2010 [39] [41]. They added new features, margins and speed of the writing, to the six features, (size of letters, slant of letters and words, baseline, pen pressure, spacing between letters, and spacing between words), and all these eight parameters were given to the SVM as an input. A variety of kernels such as linear and polynomial were tested with SVM; however, the RBF produced a better accuracy, near 90%. The original research was about developing an automated method to predict the personal behavior of individuals from the digital form of their handwriting. They used segmentation to calculate the features from digital handwriting. Once the features were extracted, they were trained with Support Vector Machine (SVM) which produced the behavior of the writer. Handwriting image samples were collected from 100 different writers, and digitized using the scanner. The proposed method gave about 94% of accuracy rate with a Radial Basis Function (RBF) kernel.

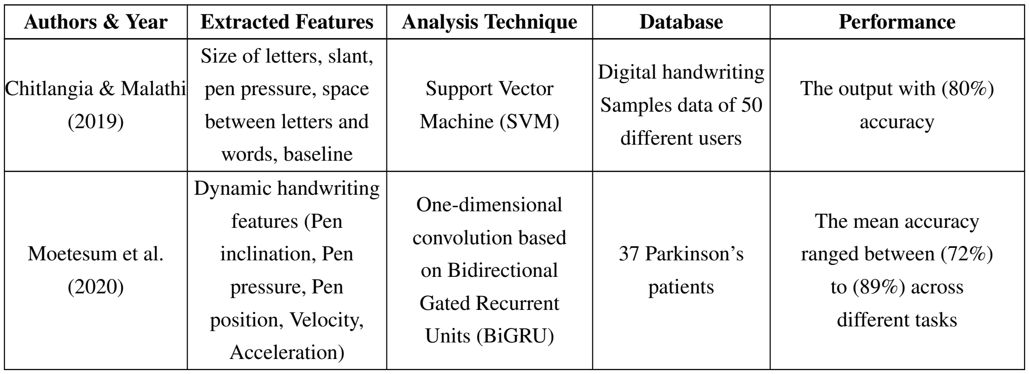
In 2015, another study was done by Gavrilescu to analyze the link between personality types and handwriting, through correlating the handwriting features with the personality primitives in a neural network with a 3-level architecture [19]. The following features were extracted: baseline, writing pressure, word slant, connecting strokes, lowercase letter 't', and lowercase letter 'f'. The database contained 64 subjects and the results showed an accuracy of 86.7% in determining the personality type, with highest accuracies for Extravert vs. Introvert and Thinking vs. Feeling personality primitives.

The following year (2016), an automated graphology system was reported on a method developed for segmentation, baseline recognition and pressure of the writing. Bal and Saha improved techniques of horizontal and vertical projections to reduce incorrect line segmentation due to overlapping [5]. Thus line and word segmentation and also skew normalization methods were developed in this research. Furthermore, the proposed methods are able to predict personality of the writer through baseline and pressure of the writing.

Three years later, Chitlangia and Malathi computerized the extraction of several features of handwriting samples (i.e. size of letters, slant, pen pressure, space between letters and words, and baseline) [12]. They classified the writer into five personality traits namely Energetic, Extrovert, Introvert, Sloppy and Optimistic. Histogram of oriented gradient (HOG) was used to extract the features from the handwriting sample of the writer which was fed to the Support Vector Machine model as an input in order to give the personality trait of the person as an output. Digital handwriting sample data of 50 different users were collected. The proposed system predicts the personality with the output with 80% accuracy.

In a recent research in 2020, a model based on Bidirectional Gated Recurrent Units (BiGRU) was proposed by Moetesum et al. to assess the potential of handwriting based sequential information in the identification of Parkinsonian symptoms [33]. One-dimensional convolution is applied to raw sequences and the resulting feature sequences are employed to train the BiGRU model for prediction. They extracted dynamic handwriting features which are pen inclination, pen pressure, pen position, velocity, and acceleration from 37 Parkinson’s patients. The proposed model achieved between 72% to 89% of accuracy across different tasks. See Table 4 for a summary of the related works on computerized graphology system.

Authors & Year	Extracted Features	Analysis Technique	Database	Performance
Shiela Lowe (1997)	such as degree of slant and margin size	Not mentioned	Not mentioned	Not mentioned
Ahmed & Mathkour (2008)	Slants, base line, speed, size, continuity, form, arrangement, and pressure	A rule-based system equipped with the conventional techniques and inference engine	35 students’ handwriting samples	Not mentioned, but in general the results were encouraging
Champa and Ananda Kumar (2011)	Size, baseline, writing pressure, slant, breaks in the writing, spacing between the words, margins, and writing speed	The inferences were made after a manual handwriting analysis	30 handwriting samples	The results were in good agreement to more than (80%) of the cases
Grewal & Prashar (2012)	Baseline, slant, pen pressure, letter ’i’ and letter ’f’	ANN	50 handwritings	Mean Square Error (MSE) reduces as number of epochs increased
Abdul Rahiman et al. (2013)	The baseline slant, the pen pressure, the slant of the writing, size of letters, and spacing between words	Simple linear regression method	Not mentioned	Proved successful in analyzing handwriting regardless of the language used
Raut & Bobade (2014)	Pen pressure, baseline, size of the letters, spacing between letters and words, slant, margins and speed of the writing	Support Vector Machine (SVM)	100 handwriting samples	Using the RBF shows a better accuracy, near (90%)
Gavrilescu (2015)	Baseline, Writing pressure, slant, Connecting strokes, Lowercase letter ’t’ and ’f”	A neural network 3-level architecture	64 subjects	Results showed an accuracy of (86%)
Bal & Saha (2016)	Baseline and writing pressure	Rule-Based system	IAM database over 550 text images containing 3800 words and some sample handwriting images	Accuracy rate of lines segmentation is (95%) and words segmentation is (92%). (96%) of lines and words were normalized perfectly with tiny error rate



Authors & Year	Extracted Features	Analysis Technique	Database	Performance
Chitlangia & Malathi (2019)	Size of letters, slant, pen pressure, space between letters and words, baseline	Support Vector Machine (SVM)	Digital handwriting Samples data of 50 different users	The output with (80%) accuracy
Moetesum et al. (2020)	Dynamic handwriting features (Pen inclination, Pen pressure, Pen position, Velocity, Acceleration)	One-dimensional convolution based on Bidirectional Gated Recurrent Units (BiGRU)	37 Parkinson's patients	The mean accuracy ranged between (72%) to (89%) across different tasks

Table 4: Related Works on Computerized Graphology System

There are other studies on computerized graphology that have been carried out and reviewed by our survey that has been published at ICDAR 2017. The most recent are reviewed above while the full list can be found in [18].

It can be observed from the prior studies that different analysis methods such as support vector machine (SVM), artificial neural networks (ANN), rule-based systems have been widely used separately in graphology. These methods are not tolerant to translation and distortion in the input image. In addition, they would have a large amount of input parameters which could add more noise during the training process. Moreover, it can be seen from Table 4 that small data was used for training the model in the early studies. The current trend of computerized graphology shows that applying a combination of analysis methods results in remarkable accuracy [16]. Moreover, the quantity of training data has an impact on achieving better results [41]. Therefore, applying ensemble methods that deploys deep learning for designing an automated graphology system with a high accuracy rate is considered in this study.

Chapter 4

The Big Five Factors Model (BFFM)

4.1 History

The Big Five Factors Model arises from the lexical hypothesis which was first proposed in the 1800s by Francis Galton. It states that every natural language contains all the personality descriptions that are relevant and important to the speakers of that language. Several researchers have explored this lexical hypothesis.

In 1936, pioneering psychologist Gordon Allport and his colleague Henry Odbert explored this hypothesis by going through an unshortened English dictionary and creating a list of 18,000 words related to individual differences. Approximately 4,500 of those terms reflected personality traits but it was not useful for research, so other scholars attempted to narrow the set of words down.

Eventually, in the 1940s, the list was reduced to 16 traits by Raymond Cattell and his colleagues using statistical methods. Several scholars, including Donald Fiske in 1949, analyzed Cattell's work. They all concluded that the data contained a strong and stable set of five traits which are Conscientiousness, Agreeableness, Emotional Stability, Openness to Experience, and Extraversion.

In the 1980s the BFFM began to receive wider scholarly attention. Today, it is a ubiquitous part of psychology research, and psychologists largely agree that personality can be grouped into the five basic traits of the BFFM.

4.2 Definitions

Each factor of the BFFM has a definition as follows:

1. Extraversion is characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness. People who are high in extraversion are outgoing and tend to gain energy in social situations. Being around other people helps them feel energized and excited [38].

2. Agreeableness includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors. People who are high in agreeableness tend to be more cooperative, friendly, and optimistic [38].

3. Conscientiousness includes high levels of thoughtfulness, good impulse control, and goal-directed behaviors. Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines. They are careful and diligent [38].

4. Emotional Stability refers to a person's ability to remain stable and balanced. A person who is high in emotional stability has a tendency to easily experience negative emotions. [38].

5. Open to Experience features characteristics such as imagination and insight. People who are high in this trait also tend to have a broad range of interests. They are curious about the world and other people and eager to learn new things and enjoy new experiences. People who are high in Open to Experience tend to be more adventurous and creative [38].

Self-report tests used for measuring each factor. They are a questionnaire that contains sets of big five factor markers that vary in their length. The scale of each test composed of a number of items on a 5-point scale ranging from complete disagreement (1: Very false for me) to complete agreement (5: Very true for me).

4.3 Related Works on Computerized Prediction of BFF

In 2016, Liu and Zhu investigated the correlations between users' personality traits and their social network linguistic behaviors [29]. For this, they built a personality prediction model based on linguistic behavior feature vectors. They constructed the personality prediction model by implementing a linear regression algorithm. For each volunteer, five linguistic behavior feature vectors corresponding to the big five factors are obtained by feature learning models, respectively. The training process of the personality prediction model is supervised, so users' five scores of five personality traits in the Big Five questionnaire are taken as their labels of the corresponding linguistic behavior feature vectors. 1,552 users' Sina microblog data have been used to train the model. Root Mean Square Error (RMSE) is used to measure the quality of different behavior feature representation methods. They obtained the following (RMSE) scores (4.80, 5.62, 5.34, 5.16, and 4.78) for Extraversion, Emotional Stability, Conscientiousness, Open to Experience, and Agreeableness, respectively.

In the following year, Majumder et al. presented a method to extract personality traits from stream-of- consciousness essays using a convolutional neural network (CNN) [31]. They trained five different networks, all with the same architecture, for the five personality traits. Each network was a binary classifier that predicted the corresponding trait to be positive or negative. They used James Pennebaker and Laura King's stream-of-consciousness essay dataset. It contains 2,468 anonymous essays tagged with the authors' personality traits based on the Big Five factors. They evaluated the model performance by measuring the accuracy obtained with different configurations. The accuracy ranged between 50% to 62% across the five factors.

In 2018, Gavrilescu and Vizireanu proposeed the non-invasive three-layer architecture based on neural networks that aims to determine the Big Five personality traits of an individual by analyzing off-line handwriting [20]. They used their own database that links the Big Five personality type with the handwriting features collected from 128 subjects containing both predefined and random texts. The main handwriting features used are the following: baseline, word slant, writing pressure, connecting strokes, space between lines, lowercase letter 't', and lowercase letter 'f'. They obtained the highest prediction accuracy for Openness to Experience, Extraversion, and Emotional Stability at 84%, while for Conscientiousness and Agreeableness, the prediction accuracy is around 77%.

In the same year, another research was conducted by Xue et al. [47]. They proposed a deep learning based method for personality recognition from text posts of on-line social network users. They first utilized a hierarchical deep neural network composed of their newly designed AttRCNN ,which introducing the attention mechanism and batch normalization (BN) technique to the recurrent-conventional neural network (RCNN), and a variant of the Inception structure to learn the deep semantic features of each user's text posts. Then, they concatenated the deep semantic features with the statistical linguistic features obtained directly from the text posts, and fed them into traditional regression algorithms to predict the real-valued Big Five personality scores. The utilized dataset involves 115,864 Facebook users, 11,494,862 text posts and 3,055,272 unique word tokens. They obtained no lower than 0.53 of MAEs average as a result of their experiments.

In 2019, Akrami et al. created a model to extract Big Five personality traits from a text using machine learning techniques [3]. They created an extensive dataset by having experts annotate personality traits in a large number of texts from multiple on-line sources. From these annotated texts, they selected a sample and made further annotations ending up in a large low-reliability dataset and a small high-reliability dataset. The results show that the models based on the small high-reliability dataset performed better than models based on large low-reliability dataset.

In 2020, Salminen et al. combined automatic personality detection (APD) and data-driven personas (DDPs) to design personas (i.e. a fictitious person that describes user or customer segments of a software system, product, or service) with personality traits that could be automatically generated using numerical and textual social media data [43]. They developed a neural network with two major sub-architectures: a single dimensional convolutional neural network since there is a spatial structure in the input text, and an long short-term memory network since there is also a temporal correlation between the words in the input text. They used the F1 macro score for evaluating the model, F1 scores obtained for each BF trait using dataset of 2,467 essays are as follows:(0.541, 0.529, 0.538, 0.553, and 0.484) for Extraversion, Openness to Experience, Conscientiousness, Agreeableness, and Emotional Stability, respectively. Table 5 summarizes the reviewed papers briefly.

Authors & Year	Analysis Technique	Database	Evaluation Measures	Performance
Liu & Zhu (2016)	Deep Learning with linear regression algorithm	1,552 Users' blogs	Root Mean Square Error (RMSE)	RMSE scores for each BF are: EXT: 4.80, NEU: 5.62, CON: 5.34, OPE: 5.16, AGR: 4.78
Majumder et al. (2017)	Convolutional neural network (CNN)	2,468 Handwriting Essays	Accuracy	The accuracy ranged between (50%) to (62%) across the five factors
Gavrilescu & Vizireanu (2018)	Three-layer architecture based on neural networks	128 subjects containing predefined and random English handwriting	Accuracy	They obtained highest prediction accuracy for OPE, EXT, and NEU (84%), while for CON and AGR, the prediction accuracy is around (77%)
Xue et al. (2018)	A deep learning based approach for personality recognition from text posts of online social network users	The final dataset involves 115,864 Facebook users, 11,494,862 text posts and 3,055,272 unique word tokens	Mean Absolute Error (MAE)	The average of MAEs is no lower than (0.53)
Akrami et al. (2019)	Machine learning method	A large dataset with lower reliability 39,370, and a smaller dataset with higher reliability 2,772	Mean Absolute Error (MAE) Mean Square Error (MSE)	For the lower reliability dataset the MAE ranged between 0.60 to 1.84, and between 0.53 to 3.85 for MSE. For the higher reliability dataset the MAE ranged between 1.04 to 1.26, and between 1.64 to 2.20 for MSE across the five factors
Salminen et al. (2020)	Deep learning classifier	2,467 Essays	F1 scores	F1 scores for each BF trait using the essays dataset are: EXT (0.541), OPE (0.529), CON (0.538), AGR (0.553), and NEU (0.484)

Table 5: Related Works on Computerized BFF Test Based on Handwriting Analysis

Chapter 5

Data Collection

As we aim to evaluate the correlation between the scores of the Big Five Factor Model (BFFM) test and the results of handwriting analysis, we need to collect the BFFM test and handwriting samples from subjects. Moreover, our model is intended to predict the level of each trait of the big five factor model, accordingly to the BFFM test, on a scale from low to high from handwriting samples written in different languages. Therefore, we required training data that contained samples representing the whole data range for each trait. Given that no such dataset was available, we set up our own large-scale collection and annotation operation.

A survey approved by University Human Research Ethics Committee at Concordia University is used for collecting our required data. The survey takes at least 40 min to complete and is composed of three sections. The first section contains demographic questions such as age, gender, occupation, education, and nationality. The second section includes the psychology test which is the International Personality Item Pool-Big Five Factor Markers Test (IPIP-BFFMT). The last is the graphology test based on handwriting and drawing but we focus on handwriting for this study. The guidelines of certification of ethical acceptability for research involving human subjects given by University Human Research Ethics Committee at Concordia University were followed. Each copy of the survey is attached with the consent form where the participant is informed about the research objectives and the confidentiality concerning identity.

For the research purposes, the survey should be filled in a specific environment because handwriting could be influenced by the mental and environmental conditions of the writer. For this, we have dedicated one laboratory at Centre for Pattern Recognition and Machine Intelligence (CENPARMI) at Concordia University for collecting our data. However, waiting for participants to come over the lab could slow down the process of data collection. Therefore, we invited the participants to our survey by posting 50 copies of letter size invitation poster that were approved and stamped by Concordia Student Union (CSU) and distributed between the campus of Loyola and Sir George Williams (SGW). In addition, email invitation is sent by the advisor of graduate students in the department of computer science and software engineering at Concordia University to Concordia faculties, staff, and students. In order to motivate the subjects to participate in our study, each participant was given $10 compensation for doing the questionnaire and a chance to win $20 Amazon gift card.

For handwriting samples, we gathered more data from our graphologist in order to enlarge the training dataset for our model. The following sections describe data analysis for the BFFM test and handwriting samples.

5.1 The Big Five Factor Model Test

5.1.1 Participants

234 participants have responded to the BFFM questionnaire. However, 43 subjects were removed from the test sample as well as the experiment of the validity of graphology because of the incomplete answers. The sample of complete BFFM test is composed of 191 subjects (48.69% male and 50.79% female). Their ages start from 18 years old in which the majority ranged between 18 and 35 years old. In terms of the level of education, 19.37% reported having started or finished high school education level, 27.23% the bachelor degree, 39.28% graduate school i.e. master or doctoral degrees, and 2.09% tertiary education i.e. diploma. Regarding the occupation, 72.77% student and 23.04% work at different places. An item of the survey asking about the country of origin of the respondents revealed that 30.37% of the participants were originally from Canada, 16.23% from Iran, 14.66% from India, 8.38% from Korea, 7.85% from China, and 21.99% from 23 different geographical countries of the world, see Table 6.

Variable	Response Category	N	%
Gender	Male	93	48.69
	Female	97	50.79
	Not informed	1	0.52
Age(years)	18-35	158	82.72
	36-55	25	13.10
	>55	8	4.19
Level of education	High school	37	19.37
	Bachelor degree	52	27.23
	Graduate school (Master & PhD)	75	39.28
	Tertiary education (Diploma)	4	2.09
	Not informed	23	12.04
Occupation	Student	139	72.77
	Worker	44	23.04
	Not informed	8	4.19
Originary geographic country	Canada	58	30.37
	Iran	31	16.23
	India	28	14.66
	Korea	16	8.38
	China	15	7.85
	Other 23 countries	42	21.99
	Not informed	1	0.52

Table 6: Demographic information of the participants (N = 191) for the (IPIP-BFFMT)

5.1.2 Instruments

International Personality Item Pool-Big Five Factor Markers Test (IPIP-BFFMT)

There are different versions of big five factor model questionnaire such as [7] and [4]. For this work, the IPIP-BFFMT developed by Goldberg and Lewis R. in 1992 [23] is used as an instrument to collect the BFF data. It is a self-report test that measures the big five personality traits i.e. (1) Extraversion, (2) Agreeableness, (3) Conscientiousness, (4) Emotional Stability, and (5) Open to Experience using the international personality item pool-big five factor markers. It was developed to represent an alternative sets of big five factor markers that vary in their length and thus in their demands on subject testing time from the previous big five factor representation. The scale composed of 50 items on a 5-point scale ranging from complete disagreement (1: Very false for me) to complete agreement (5: Very true for me), see Table 7.

	Very Inaccurate	**Moderately Inaccurate**	**Accurate Inaccurate**	**Moderately Accurate**	**Very Accurate**	
1. Am the life of the party.						(1+)
2. Feel little concern for others.						(2-)
3. Am always prepared.						(3+)
4. Get stressed out easily.						(4-)
5. Have a rich vocabulary.						(5+)
6. Don't talk a lot.						(1-)
7. Am interested in people.						(2+)
8. Leave my belongings around.						(3-)
9. Am relaxed most of the time.						(4+)
10. Have difficulty understanding abstract ideas.						(5-)
11. Feel comfortable around people.						(1+)
12. Insult people.						(2-)
13. Pay attention to details.						(3+)
14. Worry about things.						(4-)
15. Have a vivid imagination.						(5+)
16. Keep in the background.						(1-)
17. Sympathize with others' feelings.						(2+)
18. Make a mess of things.						(3-)
19. Seldom feel blue.						(4+)
20. Am not interested in abstract ideas.						(5-)
21. Start conversations.						(1+)
22. Am not interested in other people's problems.						(2-)
23. Get chores done right away.						(3+)
24. Am easily disturbed.						(4-)
25. Have excellent ideas.						(5+)
26. Have little to say.						(1-)
27. Have a soft heart.						(2+)
28. Often forget to put things back in their proper place.						(3-)
29. Get upset easily.						(4-)
30. Do not have a good imagination.						(5-)
31. Talk to a lot of different people at parties.						(1+)
32. Am not really interested in others.						(2-)
33. Like order.						(3+)
34. Change my mood a lot.						(4-)
35. Am quick to understand things.						(5+)
36. Don't like to draw attention to myself.						(1-)
37. Take time out for others.						(2+)
38. Shirk my duties.						(3-)
39. Have frequent mood swings.						(4-)
40. Use difficult words.						(5+)
41. Don't mind being the center of attention.						(1+)

42. Feel others' emotions.						(2+)
43. Follow a schedule.						(3+)
44. Get irritated easily.						(4-)
45. Spend time reflecting on things.						(5+)
46. Am quiet around strangers.						(1-)
47. Make people feel at ease.						(2+)
48. Am exacting in my work.						(3+)
49. Often feel blue.						(4-)
50. Am full of ideas.						(5+)

Table 7: International Personality Item Pool-Big Five Factor Markers Test [23]

The numbers in parentheses after each item indicate the scale on which that item is scored i.e. (1) for Extraversion, (2) for Agreeableness, (3) for Conscientiousness, (4) for Emotional Stability, and (5) for Open to Experience, and its direction of scoring (+ or -). These numbers should not be included in the actual survey questionnaire.

For + keyed items, the response "Very Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a value of 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Very Accurate" a value of 5. However, for - keyed items, the response "Very Inaccurate" is assigned a value of 5, "Moderately Inaccurate" a value of 4, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 2, and "Very Accurate" a value of 1. Once numbers are assigned for all of the items in the scale, just sum all the values to obtain a total scale score.

The reliability coefficients of the five factor scores found in Goldberg and Lewis study were considered acceptable. Their estimates ranged from .90 to .92 for Extraversion, from .84 to .97 for Agreeableness, from 88 to .94 for Conscientiousness, from .82 to .88 for Emotional Stability and from .82 to .94 for Open to Experience.

5.1.3 Instructions and Restrictions

As an instruction for answering the BFFM test, the participants were asked to describe themselves honestly as they generally are now, not as they wish to be in the future. They were asked to indicate for each statement whether it is 1. Very Inaccurate, 2. Moderately

Inaccurate, 3. Neither Accurate nor Inaccurate, 4. Moderately Accurate, or 5. Very Accurate as a description of themselves. They individually answered the test with completion time ranging from 10 to 20 min.

5.1.4 Data Analysis

5.1.4.1 Cleaning Dataset

The data analysis occurred in two stages. The first stage involved cleaning the dataset from the BFFM tests that were not completely answered by the participants. As a result, 18.38% of the cases were excluded, leaving 191 valid cases remaining in BFFMT dataset.

5.1.4.2 Estimating Normative Classifications

In the second stage, norms were estimated. National norms for the IPIP-50 to use in assignment of scores as high, medium, and low are not available [25]. Goldberg's stated position on national norms [23] is that most such norms are misleading, and he suggests that users needing norms on IPIP scales should develop local norms based on their own samples. For this study, data from the 191 participants were used to create high, medium, and low normative categories for each of the five personality traits on the IPIP-50. The three category classifications were based on a traditional normalized stanine scale with the usual interpretation of stanines 1 to 3 as low, stanines 4 to 6 as medium, and stanines of 7 to 9 as high [32]. The resulting norms for the five factors are displayed in Table 8.

	Extraversion **(n = 191)** **M (SD)**	**Conscientiousness** **(n = 191)** **M (SD)**	**Emotional Stability** **(n = 191)** **M (SD)**	**Agreeableness** **(n = 191)** **M (SD)**	**Open to Experience** **(n = 191)** **M (SD)**
	29.49 (7.58)	35.40 (6.44)	29.46 (8.09)	38.62 (5.68)	37.24 (5.70)
High	35 - 50	41 - 50	36 - 50	43 - 50	42 - 50
Medium	23 - 34	31 - 40	23 - 35	34 - 42	33 - 41
Low	10 - 22	10 - 30	10 - 22	10 - 33	10 - 32

Table 8: Normative classifications for the Big Five Factor Model from the (IPIP-BFFMT) samples (n = 191)

5.2 Handwriting Samples

5.2.1 Participants

The handwriting data consist of 1108 samples. 234 samples have been collected from our survey and 874 samples have been collected by our graphologist for her business purposes under the same condition and environment followed in this research.

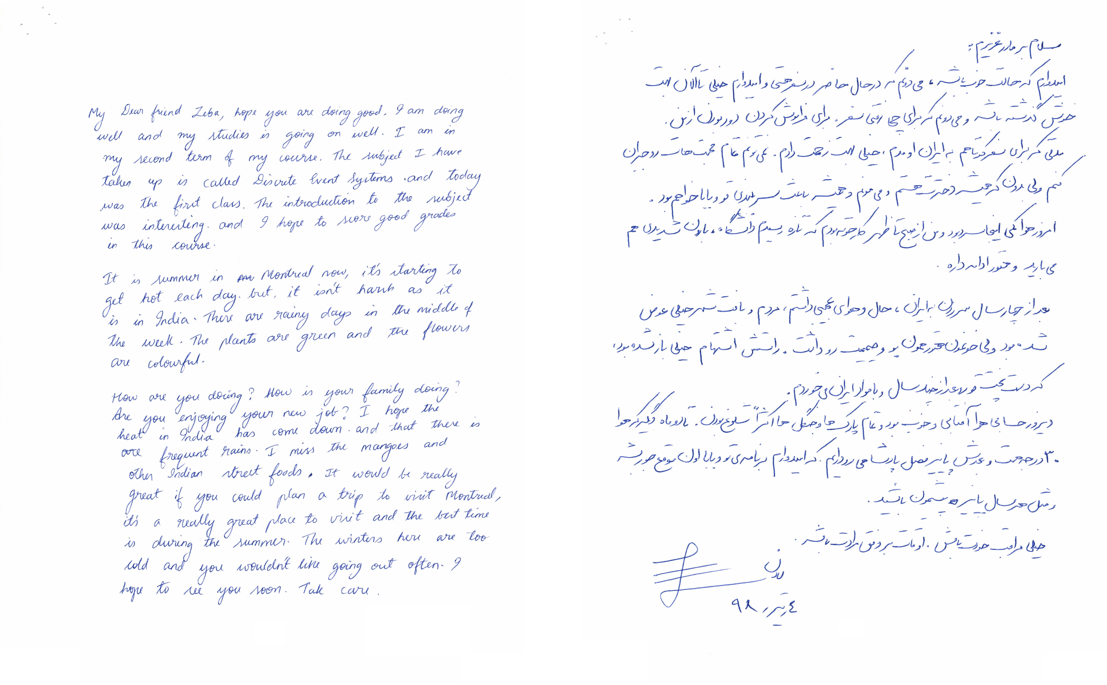
For the survey, the same individuals who responded to the (IPIP-BFFMT) were asked to write at least one page of letter on unlined letter size paper. The samples are given by 234 subjects (49.57% male and 50% female), without any exclusion. Their ages start from 18 years old in which the majority ranged between 18 and 35 years old. In terms of the level of education, 17.52% reported having started or finished the high school education level, 27.35% the bachelor degree, 41.88% graduate school i.e. master or doctoral degrees, and 1.71% tertiary education i.e. diploma. Regarding the occupation, 72.22% student and 22.65% work at different places. An item of the survey asking about the country of origin of the respondents revealed that 27.35% of the participants were originally from Canada, 18.38% from Iran, 14.53% from India, 8.10% from Korea, 7.69% from China, and 23.08% from 25 geographical countries of the world. The handwriting samples were written in 11 different languages in which the majority was written in English, see Table 9.

Variable	Response Category	N	%
Gender	Male	116	49.57
	Female	117	50
	Not Informed	1	0.43
Age (Years)	18-35	191	81.62
	36-55	32	13.68
	>55	11	4.70
Level of education	High school	41	17.52
	Bachelor degree	64	27.35
	Graduate school (Master & PhD)	98	41.88
	Tertiary education (Diploma)	4	1.71
	Not informed	27	11.54

Variable	Response Category	N	%
Occupation	Student	169	72.22
	Worker	53	22.65
	Not informed	12	5.13
Originary geographic country	Canada	64	27.35
	Iran	43	18.38
	India	34	14.53
	Korea	19	8.10
	China	18	7.69
	Other 25 countries	54	23.08
	Not informed	2	0.85
Handwriting language	English	170	72.65
	Farsi	17	7.26
	Korean	13	5.56
	French	9	3.85
	Chinese	6	2.56
	Arabic	5	2.14
	Farsi & English	4	1.71
	Spanish	2	0.85
	Chinese & English	2	0.85
	Korean & English	2	0.85
	Bengali	1	0.43
	Dutch	1	0.43
	Urdu	1	0.43
	Russian	1	0.43

Table 9: Demographic information of the participants (N = 234) for handwriting samples collected from the survey

Figure 3 shows two handwriting samples collected from the survey.



My Dear friend Zeba, hope you are doing good. I am doing well and my studies is going on well. I am in my second term of my course. The subject I have taken up is called Discrete Event Systems. and today was the first class. The introduction to the subject was interesting. and I hope to score good grades in this course.

It is summer in ~~an~~ Montreal now, it's starting to get hot each day. but, it isn't harsh as it is in India. There are rainy days in the middle of the week. The plants are green and the flowers are colourful.

How are you doing? How is your family doing? Are you enjoying your new job? I hope the heat in India has come down. and that there is ~~are~~ frequent rains. I miss the mangoes and other Indian street foods. It would be really great if you could plan a trip to visit Montreal, it's a really great place to visit and the best time is during the summer. The winters here are too cold and you wouldn't like going out often. I hope to see you soon. Take care.

سلام پرمادر عزیزم :
امیدوارم که حالت خوب باشه، می دونم که در حال حاضر در سفر هستی و امیدوارم خیلی تا الان بهت
خوش گذشته باشه و می دونم که برای چه راهی سفر. برای فراموش کردن دور بودن ازش.
مدتی که برای سفر رفتی من به ایران اومدم، خیلی ات زحمت دارم. می دونم تمام زحمت هات رو جبران
کنم ولی بدون که همیشه دخترت هستم و می مونم و همیشه باعث سربلندی تو و بابا خواهم بود.
امروز هوا کمی اینجا سرد بود و من از صبح تا ظهر کار خونه بودم که تازه رسیدم دانشگاه، بارون شدیدی هم
می بارید و هنوز ادامه داره.

بعد از چهار سال سفر رفتن به ایران، حال و هوای عجیبی داشتم، مردم و بافت شهر خیلی عوض
شده بود ولی خونمون همون بو و صمیمیت رو داشت. راستش اشتهام خیلی باز شده بود،
که دست پخت تو رو بعد از چند سال و با هوای ایران می خوردم.
دیروز حسابی هوا آفتابی و خوب بود و تمام پارک ها و جنگل ها اکثراً شلوغ بودن. تا دو ماه دیگه که هوا
۳۰ درجه است و بعدش پاییز فصل پادشاهی رو داریم. که امیدوارم برنامه ی تو و بابا اول موقع خوب شه
و مثل هر سال پاییز هم همون باشد.
خیلی مراقب خودت باش. اوقات بر وفق مرادت باشه.

لادن
٤ تیر ۹۸

(a) English Handwritten Sample

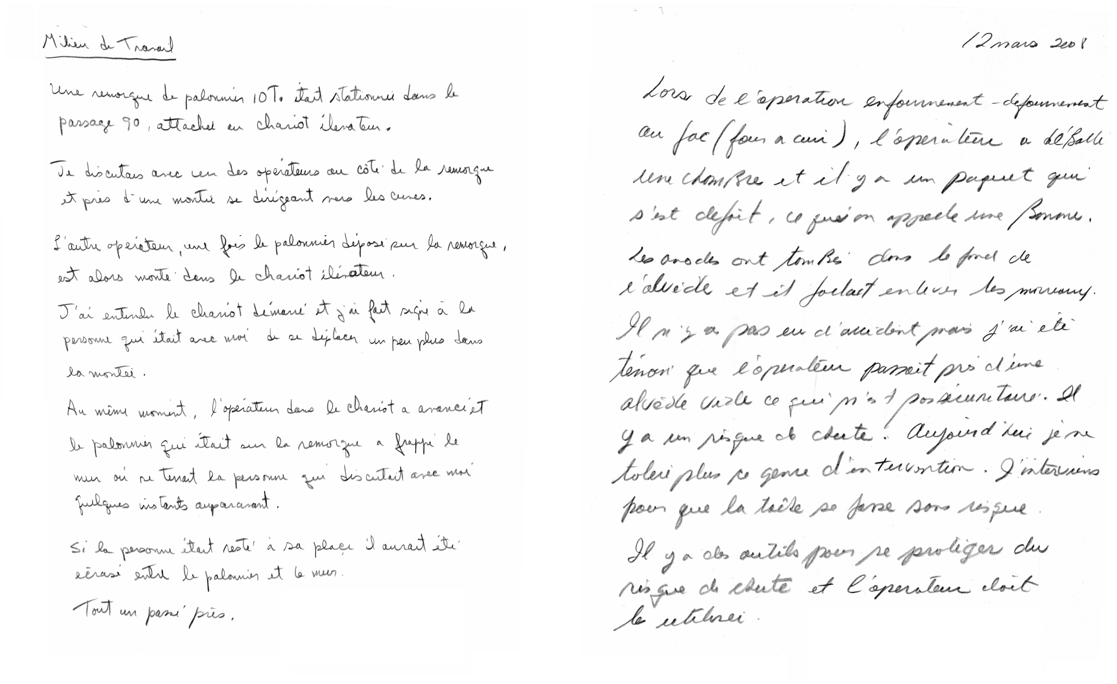
(b) Farsi Handwritten Sample

Figure 3: Two handwriting samples collected from the survey

One graphologist, well-known professional and author of various publications, was involved. 874 handwriting samples are given by 672 subjects (79.32% male and 13.79% female) collected by the graphologist for her business purposes. We ensured that these samples are collected under the same condition and environment followed in our research. The ages of the subjects start from 18 years old in which the majority ranged between 36 and 55 years old. 94.49% of subjects are from Canada and 5.51% not informed. Regarding the language of handwriting, 98.66% written in French and 1.34% written in English, see Table 10. Two examples of French handwriting samples collected by the graphologist are shown in Figure 4.

Variable	Response Category	N	%
Gender	Male	533	79.32
	Female	92	13.79
	Not Informed	47	6.99
Age(Years)	18-35	169	25.15
	36-55	403	59.97
	>55	50	7.44
	Not informed	50	7.44
Originary geographic country	Canada	635	94.49
	Not informed	37	5.51
Handwriting language	French	663	98.66
	English	9	1.34

Table 10: Demographic information of the participants (N = 672) for handwriting samples collected by graphologist



Milieu de Travail

Une remorque de palonnier 10T. était stationnée dans le passage 90, attachée au chariot élévateur.

Je discutais avec un des opérateurs au côté de la remorque et près d'une montée se dirigeant vers les cuves.

L'autre opérateur, une fois le palonnier déposé sur la remorque, est alors monté dans le chariot élévateur.

J'ai entendu le chariot démarré et j'ai fait signe à la personne qui était avec moi de se déplacer un peu plus dans la montée.

Au même moment, l'opérateur dans le chariot a avancé et le palonnier qui était sur la remorque a frappé le mur où se tenait la personne qui discutait avec moi quelques instants auparavant.

Si la personne était resté à sa place il aurait été écrasé entre le palonnier et le mur.

Tout un passé près.

12 mars 2008

Lors de l'opération enfournement-défournement au fac (four à cuire), l'opératrice a déballé une chambre et il y a un paquet qui s'est défait, ce qu'on appelle une Bonome. Les anodes ont tombé dans le fond de l'alvéole et il fallait enlever les morceaux. Il n'y a pas eu d'accident mais j'ai été témoin que l'opérateur passait près d'une alvéole vide ce qui n'est pas sécuritaire. Il y a un risque de chute. Aujourd'hui je ne tolère plus ce genre d'intervention. J'interviens pour que la tâche se fasse sans risque.

Il y a des outils pour se protéger du risque de chute et l'opérateur doit les utiliser.

Figure 4: Examples of French handwriting samples collected by the graphologist

5.2.2 Instruments

In order to predict the big five factors from handwriting, the graphologist specified handwriting features corresponding to each factor based on its definition and graphology rules. Table 11 shows the definitions of the handwriting features corresponding to each of the big five traits i.e. Extraversion, Conscientiousness, Emotional Stability, Agreeableness, and Open to Experience.

For labelling our handwriting samples, we evaluated each sample manually by scaling each feature. The scale composed of five handwriting features for Extraversion, Emotional Stability, Agreeableness, and Open to Experience on a 5-point scale where 1=None or very low, 2=Low, 3=Average, 4=High, and 5=Very high. However, for Conscientiousness, the scale composed of four features with the same 5-point scale. Once numbers are assigned for all of the features in the scale for each factor, we averaged the values to obtain the final scale score. Table 12 shows the evaluation chart that was used for labelling handwriting samples.

5.2.3 Instructions and Restrictions

The participants of the survey were asked to start with graphology part after answering the (IPIP-BFFMT). As an instruction for writing the letter, the following restrictions should be followed:

1. They must write on an unlined letter size paper
2. They must write using their ordinary handwriting
3. They should not alter or improve their writing
4. They should write with patience and calmness since the handwriting of the individual is changed by his/her mental and environmental conditions

There is no restriction on the language and the instrument. They can write in any language that they are fluent in. Moreover, they can use any instrument that they feel comfortable with in writing. The handwriting samples are written using pencil, fountain pen, and ballpoint pen in which the majority (57.02%) written with pencil.

For the handwriting samples collected by the graphologist, they are collected in her professional practice. The clients have agreed to use their samples for research or teaching purposes.

Factor	Handwriting features
Extraversion	**1. Middle zone more than 2,5 mm** The middle zone measured from baseline to top of letter includes most lowercase letters such as a, e, i, o, and u. **2. Narrow ending margin** Margin is spacing around the text page, and indentations for paragraphs **3. Dominance of garlands** Garlands: unlike the taught model, the letters "m" and "n" have a shape similar to a cup, or like a "u". Angles: replacement of curves with more or less sharp angles, whether in the letter forms and/or the inter-letter connections. **4. Progressive movement** Movement dominates, carrying the forms along with it. Often right-slanted, with a high degree of connection. **5. Slanted in the direction of the writing** Downstrokes from an angle to the baseline of between: 85 75 slight right slant 75 60 slight to moderate right-slant 60 45 strong right-slant below 45 exaggerated right slant

Factor	Handwriting features
Conscientiousness	**1. Regularity (slant, dimension, space, etc.)** Consistency in middle zone height, in the distance between downstrokes, and the slant of the writing. **2. Precision of placement of free strokes** The ’t’ bars are well centered at 2/3 of the stem. The ’I’ dots are near the letter in the same axis of the letter. **3. Legibility** Readable, clear writing even when parts are taken out of context. **4. Controlled movement** Well-structured forms. Disciplined progression to the right, resting firmly on the line.
Emotional Stability	**1. Regularity without rigidity** **2. Baseline horizontal and flexible** **3. Slightly slanted** **4. Good balance between white space and ink space** **5. Good pressure and quality of the stroke**
Agreeableness	**1. Dominance of curves versus angles** **2. Good space between letters, words, and lines** **3. Letter width >5** **4. Round letters without loops and slightly open** **5. Nourished Stroke**
Open to Experience	**1. Good openness in loops** **2. Good speed and movement** **3. Slight angles in letters** **4. Slanted in the direction of handwriting** **5. Narrow ending margin**

Table 11: Handwriting Features Corresponding to each of the Big Five Factors

Evaluator					
Sample number					
Evaluation Scale **1 = None or very low** **2 = Low** **3 = Average** **4 = High** **5 = Very high**					
Extraversion: is characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness.					
Handwriting Feature	**1**	**2**	**3**	**4**	**5**
1. Middle zone more than 2,5 mm					
2. Narrow ending margin					
3. Dominance of garlands					
4. Progressive movement					
5. Slanted in the direction of the writing					
Global Evaluation					
Conscientiousness: standard features of this dimension include high levels of thoughtfulness, good impulse control, and goal-directed behaviors.					
Handwriting Feature	**1**	**2**	**3**	**4**	**5**
1. Regularity (slant, dimension, space, etc.)					
2. Precision of placement of free strokes					
3. Legibility					
4. Controlled movement					
Global Evaluation					
Emotional Stability: refers to a person's ability to remain stable and balanced.					
Handwriting Feature	**1**	**2**	**3**	**4**	**5**
1. Regularity without rigidity					
2. Baseline horizontal and flexible					
3. Slightly Slanted					
4. Good balance between white space and ink space					
5. Good pressure and quality of the stroke					
Global Evaluation					

Agreeableness: includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors.					
Handwriting Feature	**1**	**2**	**3**	**4**	**5**
1. Dominance of curves versus angles					
2. Good space between letters, words, and lines					
3. Letter width >5					
4. Round letters without loops and slightly open					
5. Nourished stroke					
Global Evaluation					

Open to Experience: characteristics such as imagination and insight.					
Handwriting Feature	**1**	**2**	**3**	**4**	**5**
1. Good openness in loops					
2. Good speed and movement					
3. Slight angles in letters					
4. Slanted in the direction of handwriting					
5. Narrow ending margins					
Global Evaluation					

Table 12: The Evaluation Chart for Handwriting Samples

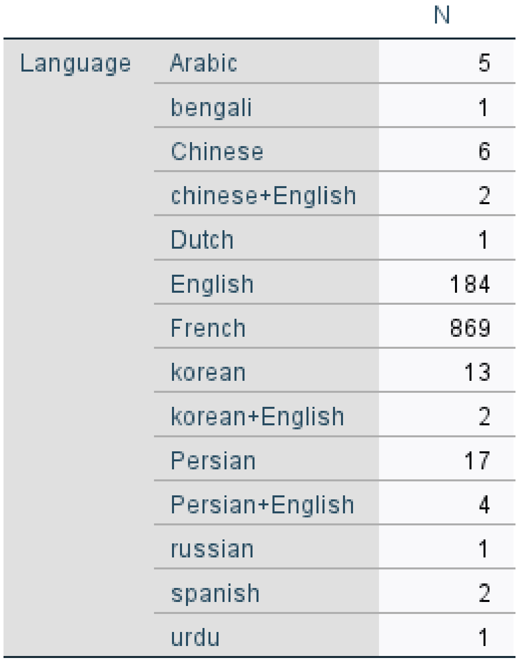
5.2.4 Data Analysis

5.2.4.1 Confounding Variables Analysis

The total of collected handwriting samples is 1108. However, some samples were excluded from the dataset after testing the confounding variables. A confounder is a variable whose presence affects the variables being studied so that the results do not reflect the actual relationship. Therefore, these variables must be excluded or controlled. In our study, we have the variable of language used for writing the sample which is considered as a confounder and the dependent variable which is the result of handwriting analysis for the big five factors individually. We need to test the impact of the language on handwriting analysis in order to control its effect and increase the power of our correlation analysis results. For this, we used two statistical methods in the Statistical Package for the Social Sciences (SPSS) named Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA). We conducted the two methods on 1108 observations of handwriting samples with no missing values and both led to the same results approximately. The following sections explain each method in details.

A. Analysis of Variance (ANOVA)

ANOVA is used to test the impact of the independent variable on the dependent variable by testing the differences between groups. It has several assumptions that need to be fulfilled but the most important ones are normality and homogeneity of variances. So, firstly, we need to test these two assumptions for each factor across different languages shown in Table 13.



		N
Language	Arabic	5
	bengali	1
	Chinese	6
	chinese+English	2
	Dutch	1
	English	184
	French	869
	korean	13
	korean+English	2
	Persian	17
	Persian+English	4
	russian	1
	spanish	2
	urdu	1

Table 13: Languages used for Writing the Samples (N=number of samples)

Normality Test

Normality means the responses for each factor level have a normal population distribution. In this section we test the normality for each of the big five factors across the languages mentioned in Figure 13 using the residual. A residual is a deviation from the sample mean which is calculated by subtracting the observed value from the expected value. It is stated by the following equation:

$$r = x - x_0 \tag{1}$$

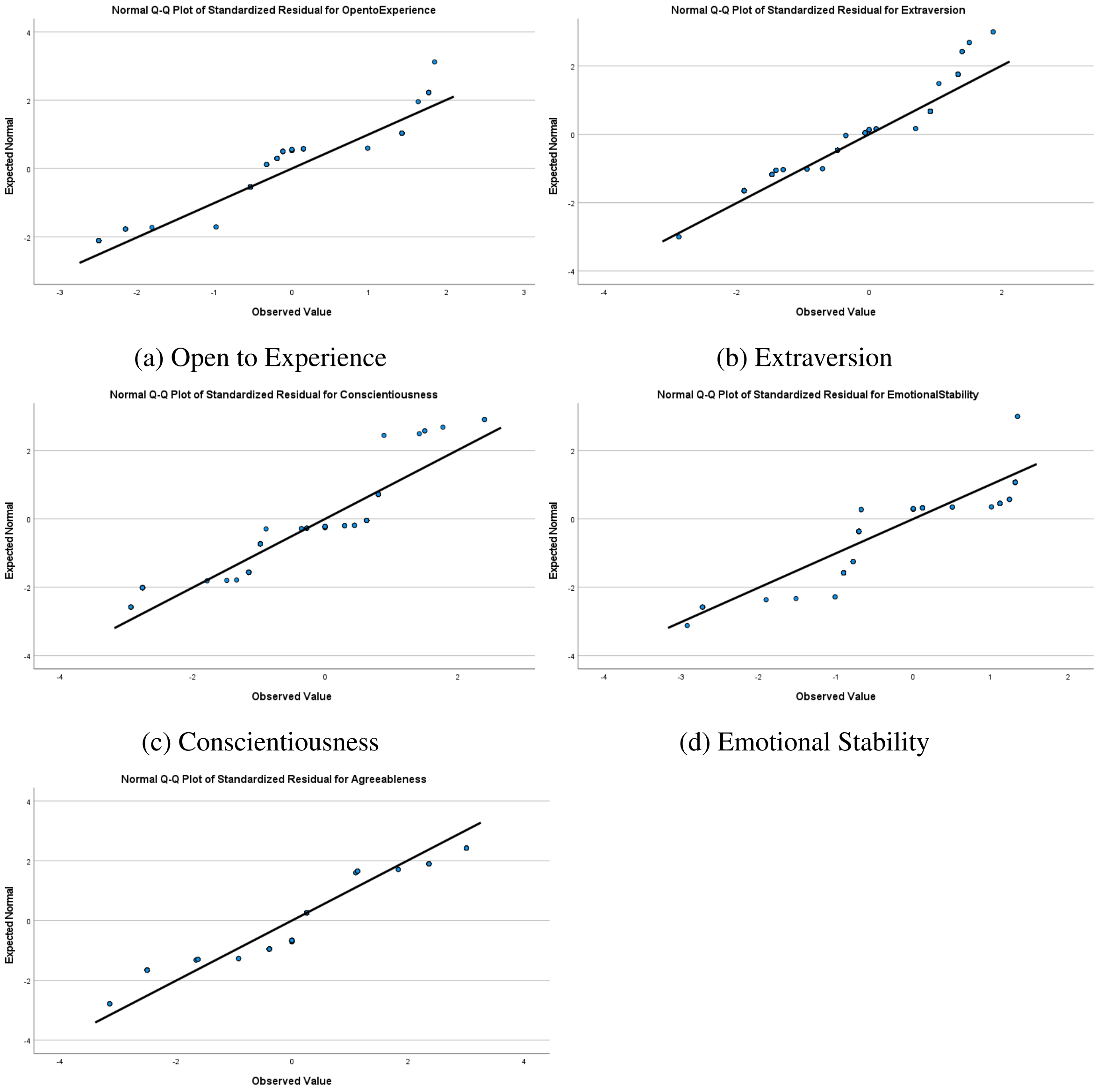
where:
r = residual
x = expected variable
x_0 = observed variable

We test the normality by assessing the residual values from the normal probability plot for each factor. Figure 5 plots the standardized residual values for Open to Experience, Extraversion, Conscientiousness, Emotional Stability, and Agreeableness, respectively. In a normal probability plot, the data are plotted in such a way that the points should form an approximate straight line. However, in the probability plots for the big five factors, the data depart from this straight line indicating departures from normality.

Homogeneity Test

In order to test the homogeneity for the big five factors, we used Levene's test which is an inferential statistic used to assess the equality of variances for a variable calculated for two or more groups. It tests the hypothesis that the population variances are equal. The test rejects the hypothesis when the p-value is less than or equal to the significant level (α = 0.05). Failing the homogeneity test means that there is a difference between the variances in the population. So, it can be observed from Table 14 that the p-value for the all five factors is less than 0.05 which means that all factors are not homogeneous.

After testing the assumptions of ANOVA, we found that the two assumptions are violated by our data. Therefore, the Kruskal-Wallis test will be used instead of ANOVA. Kruskal-Wallis test is a nonparametric test, and is used when the assumptions of ANOVA are not met.



(a) Open to Experience

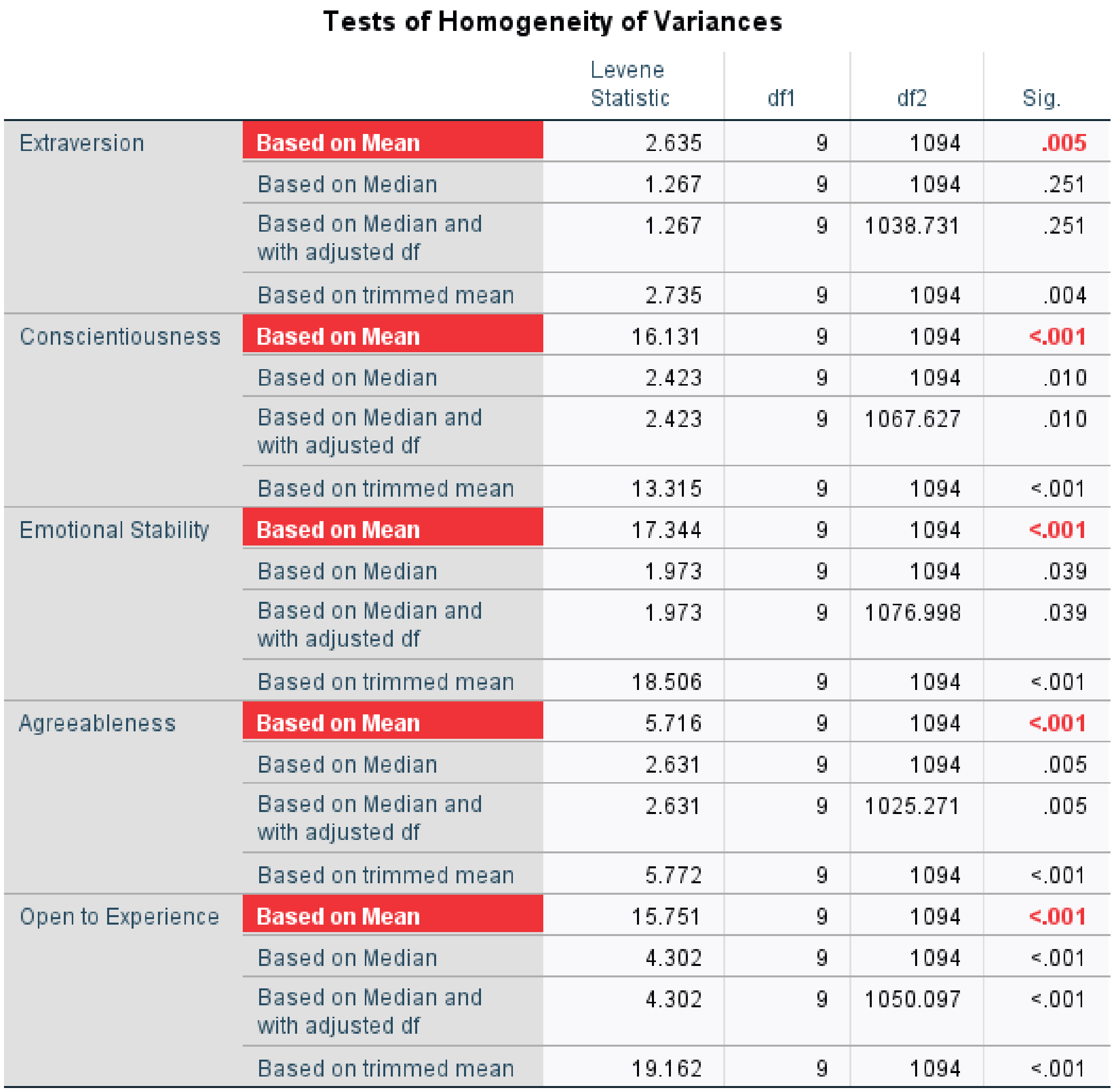
(b) Extraversion

(c) Conscientiousness

(d) Emotional Stability

(e) Agreeableness

Figure 5: Normal Probability Plot for the Standardized Residual for the Big Five Factors



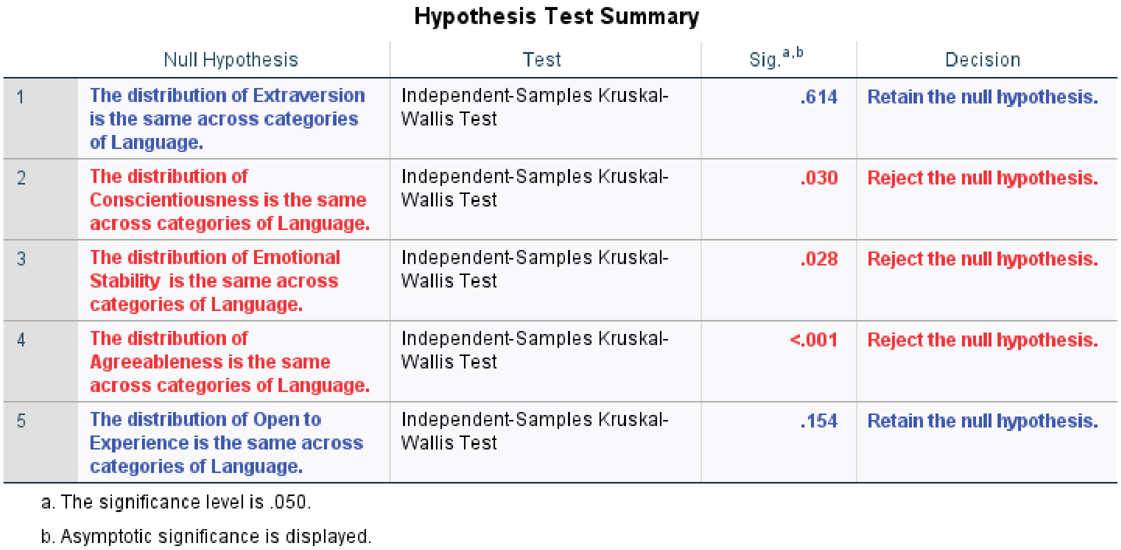
Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Extraversion	**Based on Mean**	2.635	9	1094	**.005**
	Based on Median	1.267	9	1094	.251
	Based on Median and with adjusted df	1.267	9	1038.731	.251
	Based on trimmed mean	2.735	9	1094	.004
Conscientiousness	**Based on Mean**	16.131	9	1094	**<.001**
	Based on Median	2.423	9	1094	.010
	Based on Median and with adjusted df	2.423	9	1067.627	.010
	Based on trimmed mean	13.315	9	1094	<.001
Emotional Stability	**Based on Mean**	17.344	9	1094	**<.001**
	Based on Median	1.973	9	1094	.039
	Based on Median and with adjusted df	1.973	9	1076.998	.039
	Based on trimmed mean	18.506	9	1094	<.001
Agreeableness	**Based on Mean**	5.716	9	1094	**<.001**
	Based on Median	2.631	9	1094	.005
	Based on Median and with adjusted df	2.631	9	1025.271	.005
	Based on trimmed mean	5.772	9	1094	<.001
Open to Experience	**Based on Mean**	15.751	9	1094	**<.001**
	Based on Median	4.302	9	1094	<.001
	Based on Median and with adjusted df	4.302	9	1050.097	<.001
	Based on trimmed mean	19.162	9	1094	<.001

Table 14: Test of Homogeneity of Variance for the Big Five Factors

Kruskal-Wallis Test

We conducted the Kruskal-Wallis test on 1108 handwriting samples including all languages mentioned in Figure 13 and we got the results shown in Table 15.



Hypothesis Test Summary

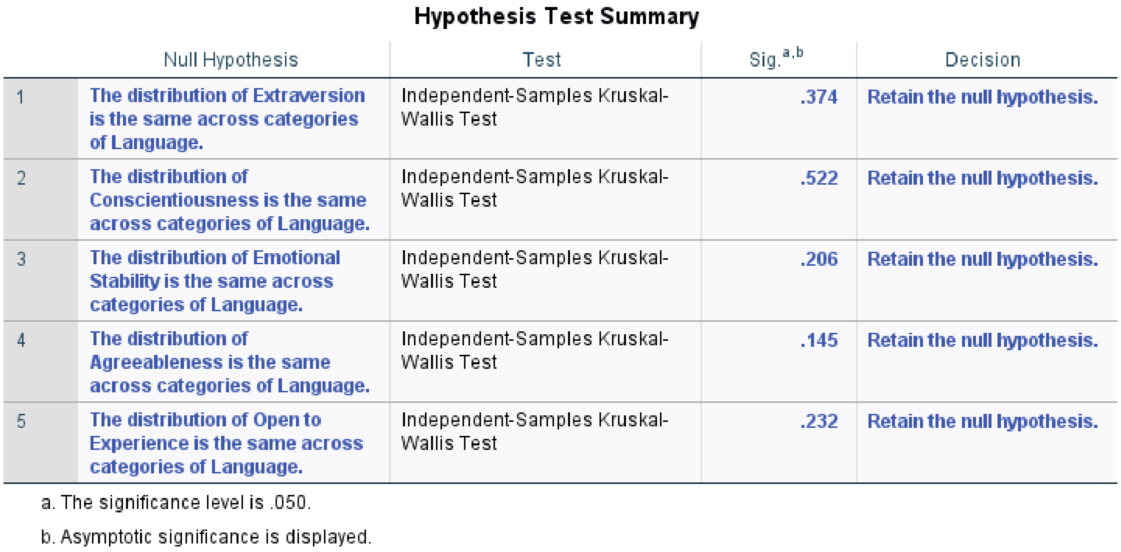
	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The distribution of Extraversion is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.614	Retain the null hypothesis.
2	The distribution of Conscientiousness is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.030	Reject the null hypothesis.
3	The distribution of Emotional Stability is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.028	Reject the null hypothesis.
4	The distribution of Agreeableness is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	<.001	Reject the null hypothesis.
5	The distribution of Open to Experience is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.154	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Table 15: Hypothesis Test Summary for Kruskal-Wallis Test for the Big Five Factors

As it can be observed the p-values for Extraversion and Open to Experience are 0.614 and 0.154, respectively, which are greater than the significant level (α = 0.05). That means there are no differences in the results of handwriting analysis for Extraversion and Open to Experience between the 14 groups of languages, therefore language has no significant effect on handwriting analysis for the two factors. However, for the other three factors i.e. Conscientiousness, Emotional Stability, and Agreeableness, the p-value is less than 0.05. That means the handwriting analysis for the three factors is influenced by the language. In order to prevent the effect of the language, we checked the pairwise comparisons of languages and we found that the p-value between six languages which are Korean, Persian, Russian, Bengali, Urdu, and Dutch is less than 0.05. That means these languages have an effect on the result of handwriting analysis. Therefore, we excluded the samples written in these languages and we re-conducted Kruskal-Wallis test with only five languages i.e. English, French, Chinese, Arabic, and Spanish. The results after re-conducting the test are shown in Table 16.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Extraversion is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.374	Retain the null hypothesis.
2	The distribution of Conscientiousness is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.522	Retain the null hypothesis.
3	The distribution of Emotional Stability is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.206	Retain the null hypothesis.
4	The distribution of Agreeableness is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.145	Retain the null hypothesis.
5	The distribution of Open to Experience is the same across categories of Language.	Independent-Samples Kruskal-Wallis Test	.232	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Table 16: Hypothesis Test Summary for Kruskal-Wallis Test for the Big Five Factors after Removing the Six Languages

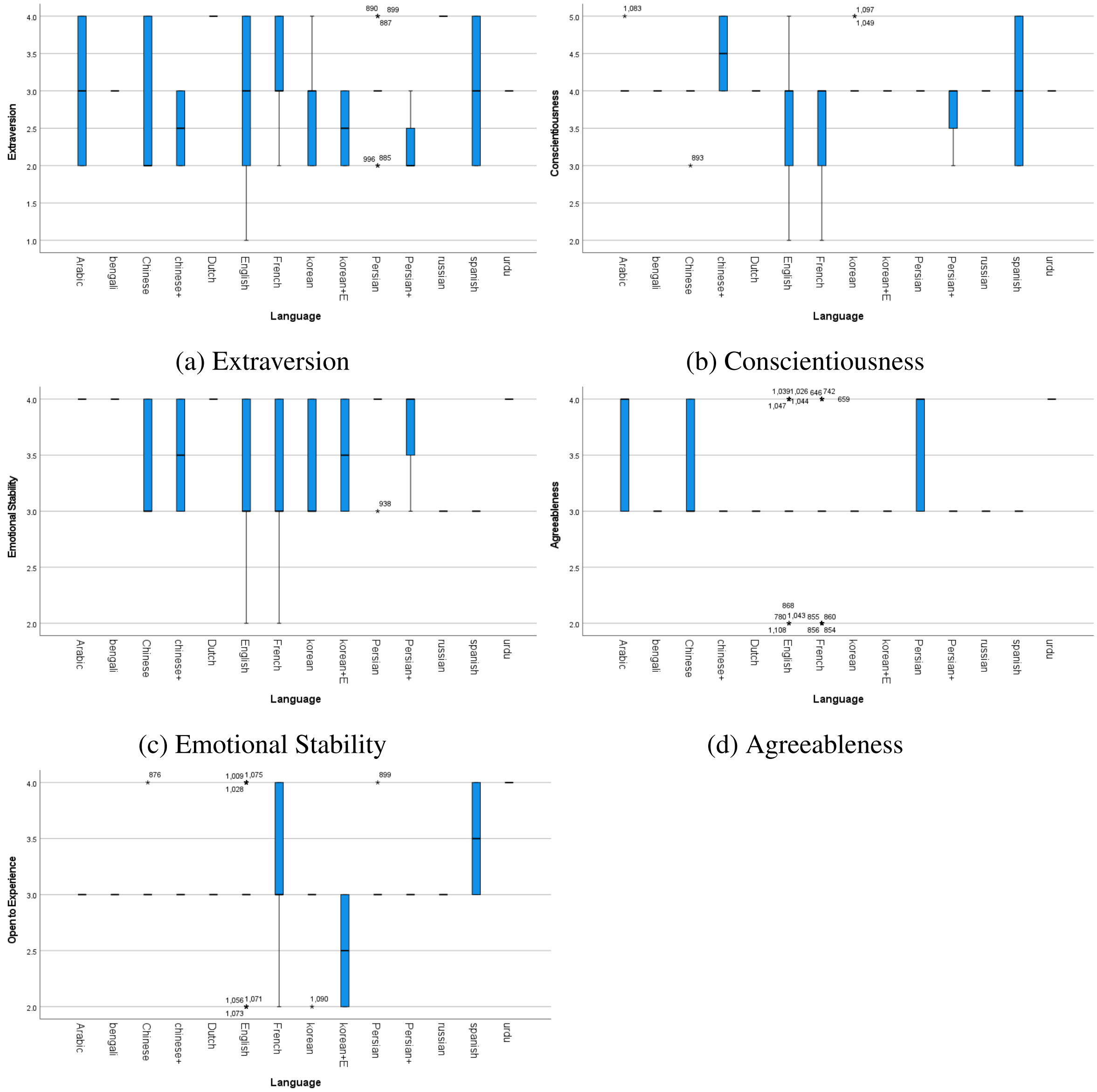
As it can be seen the p-value for the five factors is greater than 0.05. It indicates that handwriting analysis for the big five factors is not influenced by the remaining languages which are English, French, Chinese, Arabic, and Spanish.

B. Analysis of Covariance (ANCOVA)

ANCOVA is similar to ANOVA but is used to detect a difference in means of 3 or more independent groups, whilst controlling for scale covariates. In this study, we test the influence of our confounder (the language) on the dependent variable (the result of handwriting analysis) whilst controlling for scale covariate which is the gender. The same as ANOVA, ANCOVA has several assumptions that need to be met. We test the most important ones which are normality and homogeneity of variances for each factor across different languages.

Normality Test

The normality is tested by assessing the outliers graphs for each factor. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Figure 6 shows the outliers for Extraversion, Conscientiousness, Emotional Stability, Agreeableness, and Open to Experience, respectively. As it is shown on the five graphs there are outliers in the all five factors. These outliers indicate that the five factors are not normally distributed across the languages.



(a) Extraversion

(b) Conscientiousness

(c) Emotional Stability

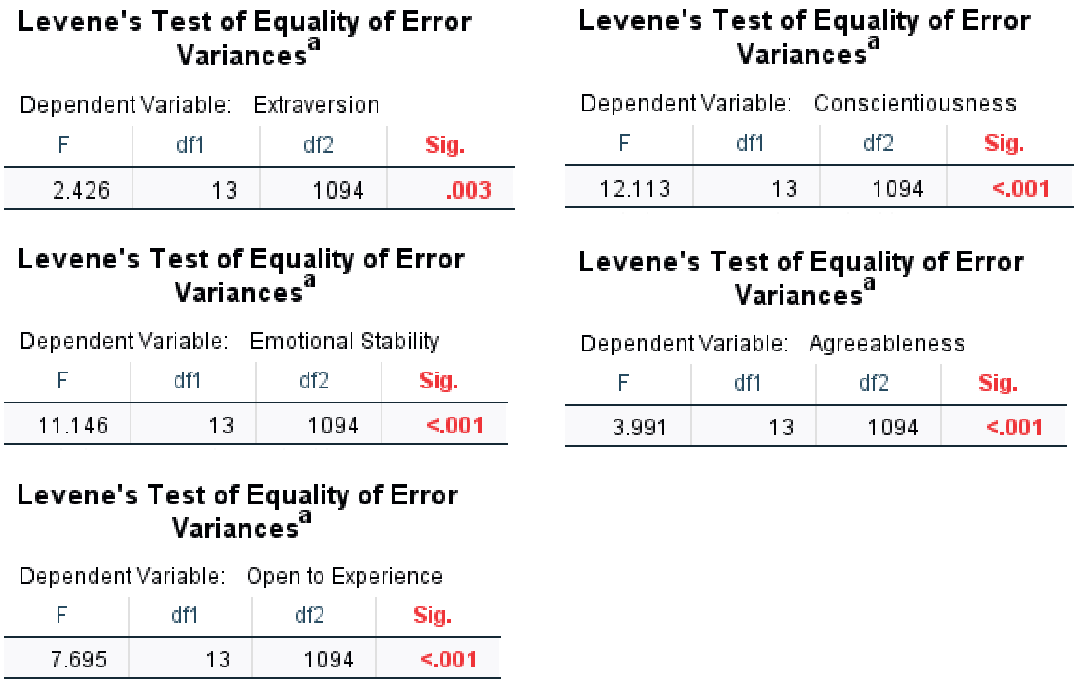
(d) Agreeableness

(e) Open to Experience

Figure 6: The Outliers for the Big Five Factors across the Gender Groups

Homogeneity Test

Levene's test is used to test the homogeneity for the big five factors. It can be observed from Table 17 that the p-value for the all five factors is less than 0.05 which means that all factors are not homogeneous across the groups of language.



Levene's Test of Equality of Error Variances[a]

Dependent Variable: Extraversion

F	df1	df2	Sig.
2.426	13	1094	.003

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Conscientiousness

F	df1	df2	Sig.
12.113	13	1094	<.001

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Emotional Stability

F	df1	df2	Sig.
11.146	13	1094	<.001

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Agreeableness

F	df1	df2	Sig.
3.991	13	1094	<.001

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Open to Experience

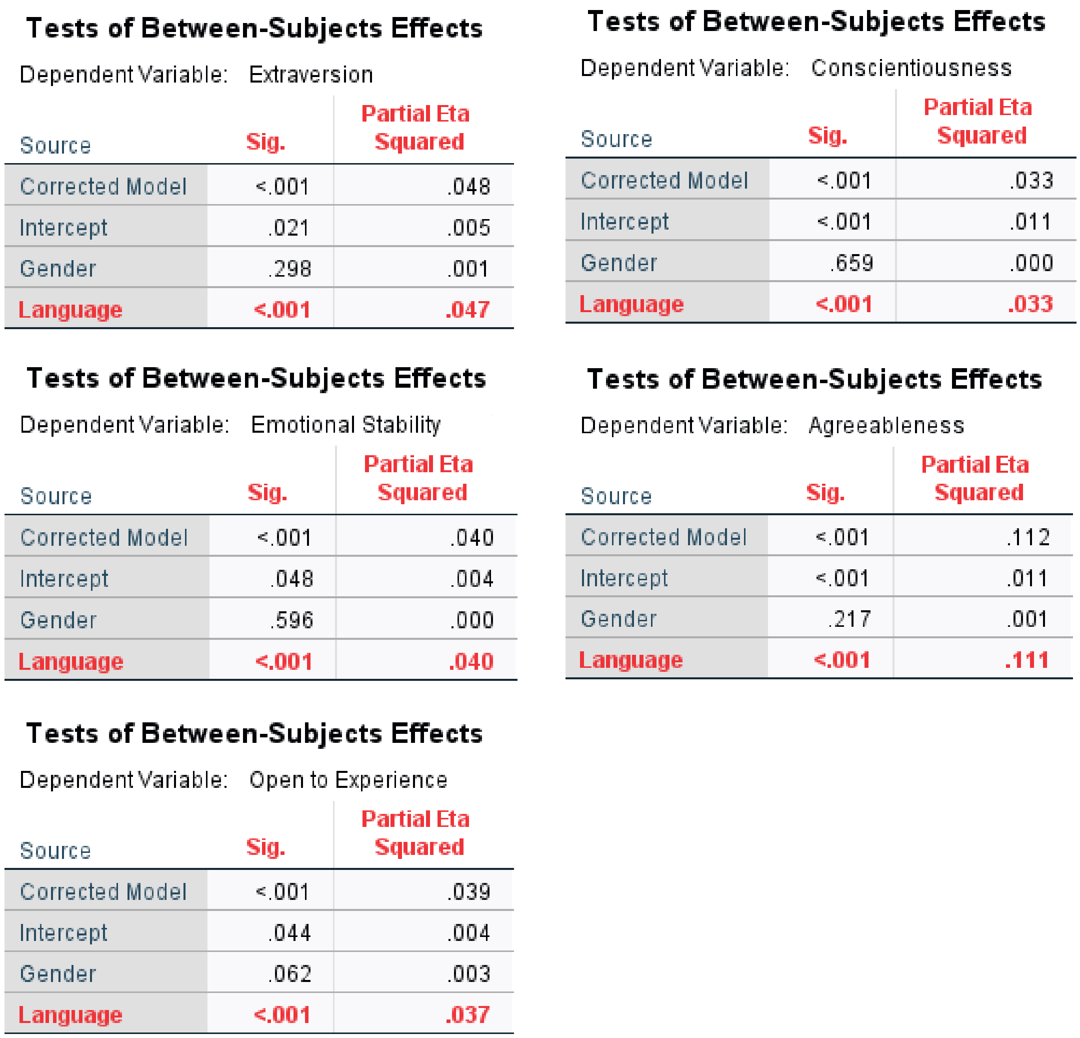
F	df1	df2	Sig.
7.695	13	1094	<.001

Table 17: Test of Homogeneity of Variance for the Big Five Factors across the Gender Groups

After testing the assumptions of ANCOVA, we found that the two assumptions are not fulfilled by our data. Therefore, the nonparametric ANCOVA named Quade's test will be used instead of traditional ANCOVA.

Quade's ANCOVA Test

Table 18 shows the influence of the confounding variable (the language) after the effect of the covariate variable (the gender) has been accounted for in each of the big five factors.



Tests of Between-Subjects Effects

Dependent Variable: Extraversion

Source	Sig.	Partial Eta Squared
Corrected Model	<.001	.048
Intercept	.021	.005
Gender	.298	.001
Language	**<.001**	**.047**

Tests of Between-Subjects Effects

Dependent Variable: Conscientiousness

Source	Sig.	Partial Eta Squared
Corrected Model	<.001	.033
Intercept	<.001	.011
Gender	.659	.000
Language	**<.001**	**.033**

Tests of Between-Subjects Effects

Dependent Variable: Emotional Stability

Source	Sig.	Partial Eta Squared
Corrected Model	<.001	.040
Intercept	.048	.004
Gender	.596	.000
Language	**<.001**	**.040**

Tests of Between-Subjects Effects

Dependent Variable: Agreeableness

Source	Sig.	Partial Eta Squared
Corrected Model	<.001	.112
Intercept	<.001	.011
Gender	.217	.001
Language	**<.001**	**.111**

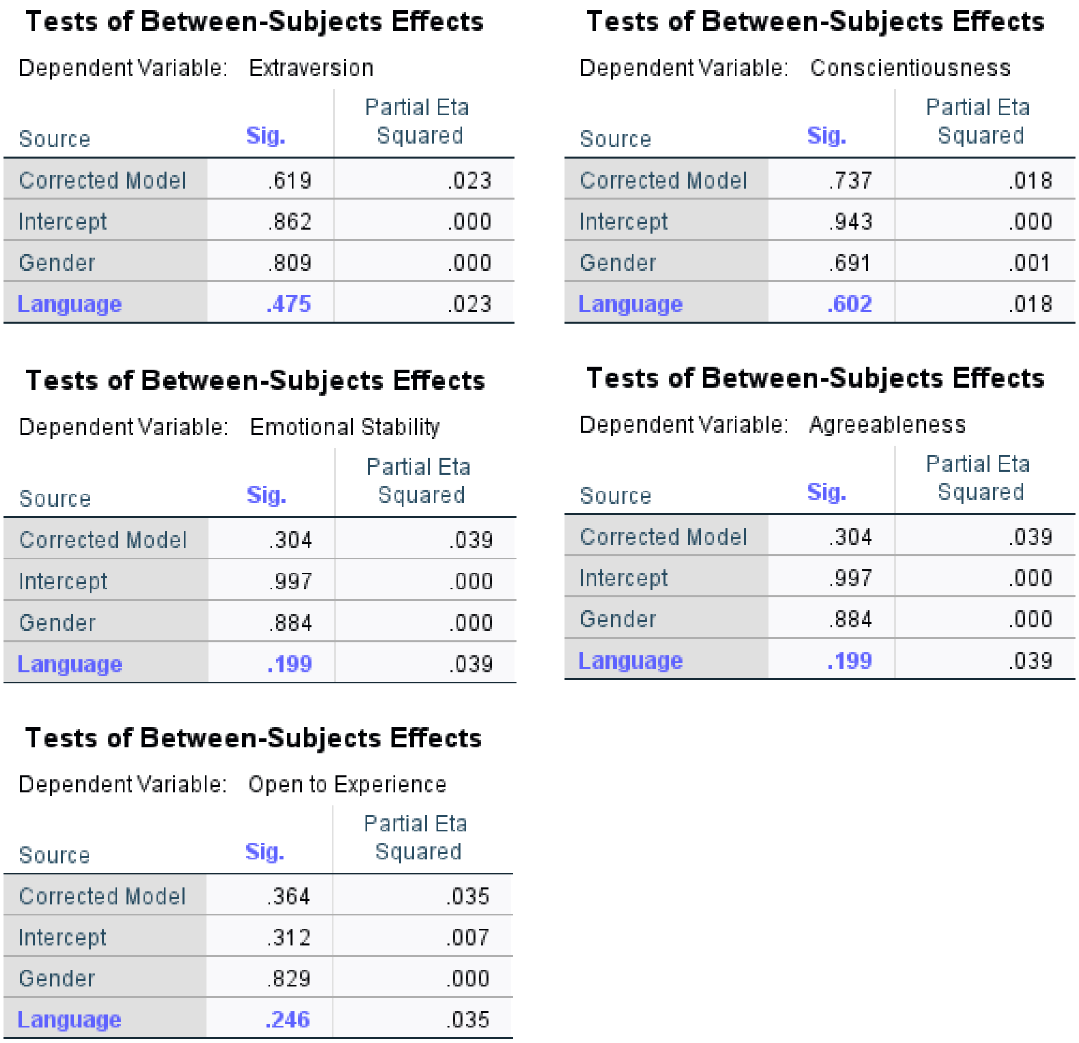
Tests of Between-Subjects Effects

Dependent Variable: Open to Experience

Source	Sig.	Partial Eta Squared
Corrected Model	<.001	.039
Intercept	.044	.004
Gender	.062	.003
Language	**<.001**	**.037**

Table 18: The Results of the Quade's ANCOVA test for the Big Five Factors before Removing the Six Languages

According to the p-value for the big five factors there is a significant difference between the language groups whilst adjusting for the gender since the p-value is less than 0.05. However, the partial Eta Squared values indicate a small effect size since they are less than 0.2 based on Cohen's guidelines. After checking the pairwise comparisons of languages in each factor we found that the p-value between the same six languages observed in ANOVA is less than 0.05. Therefore, we removed those languages and re-conducted the Quade's ANCOVA test with only five languages which are English, French, Chinese, Arabic, and Spanish. The results after re-conducting the test are shown in Table 19.



Tests of Between-Subjects Effects

Dependent Variable: Extraversion

Source	Sig.	Partial Eta Squared
Corrected Model	.619	.023
Intercept	.862	.000
Gender	.809	.000
Language	.475	.023

Tests of Between-Subjects Effects

Dependent Variable: Conscientiousness

Source	Sig.	Partial Eta Squared
Corrected Model	.737	.018
Intercept	.943	.000
Gender	.691	.001
Language	.602	.018

Tests of Between-Subjects Effects

Dependent Variable: Emotional Stability

Source	Sig.	Partial Eta Squared
Corrected Model	.304	.039
Intercept	.997	.000
Gender	.884	.000
Language	.199	.039

Tests of Between-Subjects Effects

Dependent Variable: Agreeableness

Source	Sig.	Partial Eta Squared
Corrected Model	.304	.039
Intercept	.997	.000
Gender	.884	.000
Language	.199	.039

Tests of Between-Subjects Effects

Dependent Variable: Open to Experience

Source	Sig.	Partial Eta Squared
Corrected Model	.364	.035
Intercept	.312	.007
Gender	.829	.000
Language	.246	.035

Table 19: The Results of the Quade's ANCOVA test for the Big Five Factors after Removing the Six Languages

As it can be seen the p-value for the five factors now is greater than 0.05. It indicates that the remaining languages which are English, French, Chinese, Arabic, and Spanish do not impact handwriting analysis for the big five factors. As a result of testing the confounding variables, the final total of handwriting samples in our dataset named (HWBFF) after removing the six languages from our dataset is 1066 samples. Consequently, the BFF tests that are corresponding to the removed languages are excluded from the validity test leaving 156 samples of the BFF test for the correlation. Figures 7 and 8 show the total number of samples for handwriting and BFF test before and after conducting the confounding variable test.

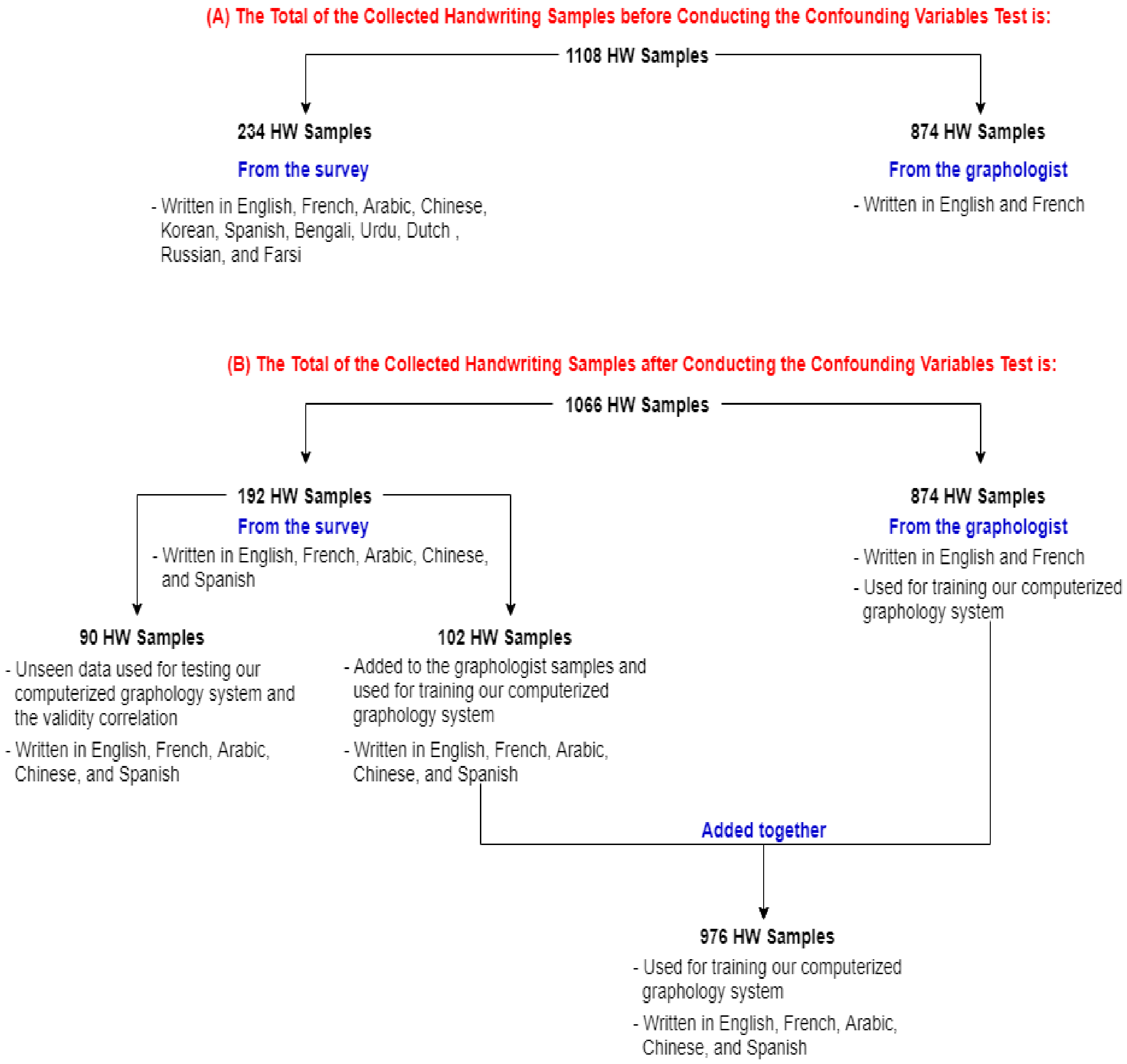
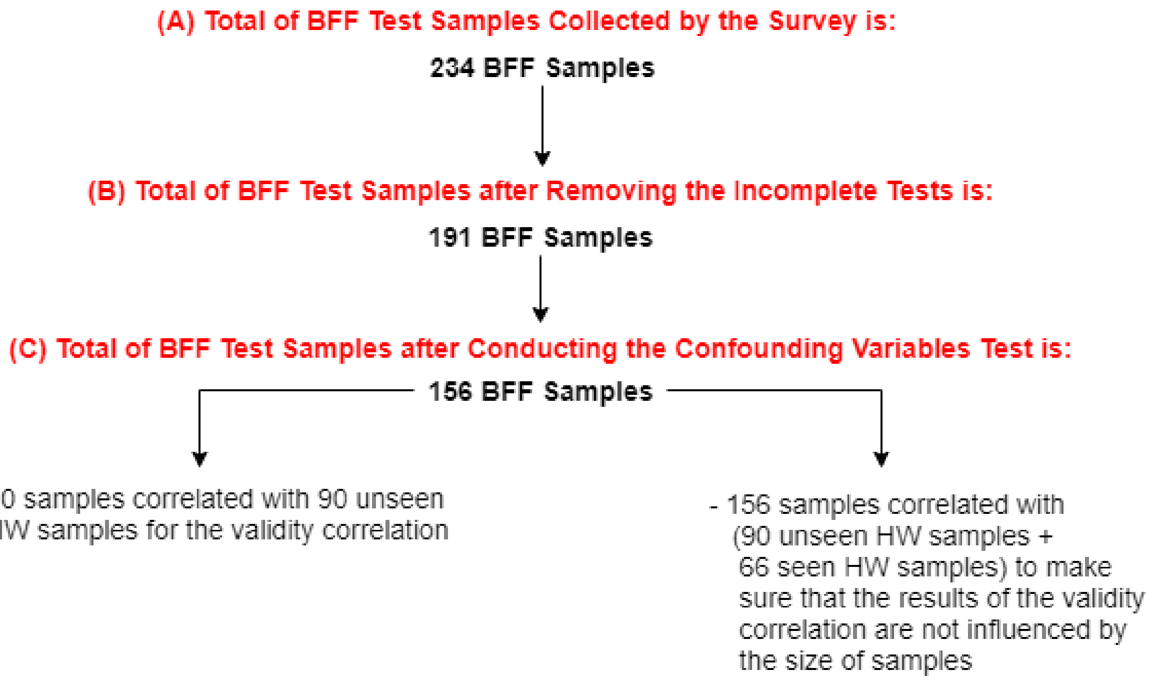

(A) The Total of the Collected Handwriting Samples before Conducting the Confounding Variables Test is:
1108 HW Samples
234 HW Samples
From the survey
- Written in English, French, Arabic, Chinese, Korean, Spanish, Bengali, Urdu, Dutch , Russian, and Farsi
874 HW Samples
From the graphologist
- Written in English and French
(B) The Total of the Collected Handwriting Samples after Conducting the Confounding Variables Test is:
1066 HW Samples
192 HW Samples
From the survey
- Written in English, French, Arabic, Chinese, and Spanish
874 HW Samples
From the graphologist
- Written in English and French
- Used for training our computerized graphology system
90 HW Samples
- Unseen data used for testing our computerized graphology system and the validity correlation
- Written in English, French, Arabic, Chinese, and Spanish
102 HW Samples
- Added to the graphologist samples and used for training our computerized graphology system
- Written in English, French, Arabic, Chinese, and Spanish
Added together
976 HW Samples
- Used for training our computerized graphology system
- Written in English, French, Arabic, Chinese, and Spanish

Figure 7: The Total of Collected Handwriting Samples before and after Conducting the Confounding Variables Test


(A) Total of BFF Test Samples Collected by the Survey is:

234 BFF Samples

(B) Total of BFF Test Samples after Removing the Incomplete Tests is:

191 BFF Samples

(C) Total of BFF Test Samples after Conducting the Confounding Variables Test is:

156 BFF Samples

- 90 samples correlated with 90 unseen HW samples for the validity correlation

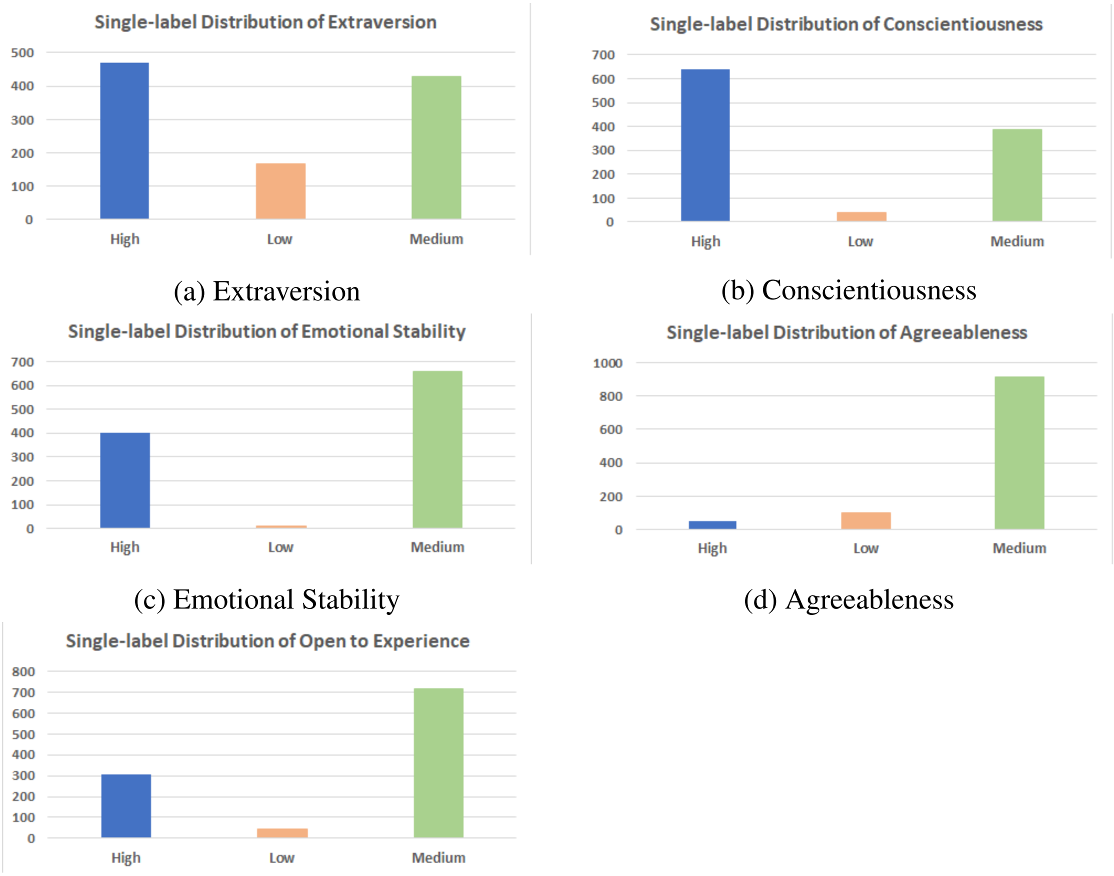
- 156 samples correlated with (90 unseen HW samples + 66 seen HW samples) to make sure that the results of the validity correlation are not influenced by the size of samples

Figure 8: The Total of Collected BFF Test Samples before and after Conducting the Confounding Variables Test

5.2.4.2 Data Distribution Analysis

In psychology filed, the goal of the big five factor model test is to predict the measurement level into (low, medium, or high) for Extraversion, Conscientiousness, Emotional Stability, Agreeableness, and Open to Experience at the same time. Thus, in computer science, it can be formulated into a multi-label classication problem with 15 single labels. Multi-label classification problem is one of the supervised learning problems where an instance may be associated with multiple labels simultaneously.

In order to understand our dataset better and receive expected results, we need to do some analysis for data distribution in order to see whether our data is balanced or not. Having imbalanced data causes the machine learning classifier tends to be more biased towards the majority class and resulting in bad classification of the minority class. So, we aim to avoid the imbalanced data in order to get a high performance evaluation for our classifier. For this, single-label distribution analysis for the five factors has been done on 1066 handwriting samples, see Figures 9 and 10.



(a) Extraversion

(b) Conscientiousness

(c) Emotional Stability

(d) Agreeableness

(e) Open to Experience

Figure 9: Single-label Distribution for the Five Factors Separately in HWBFF Dataset

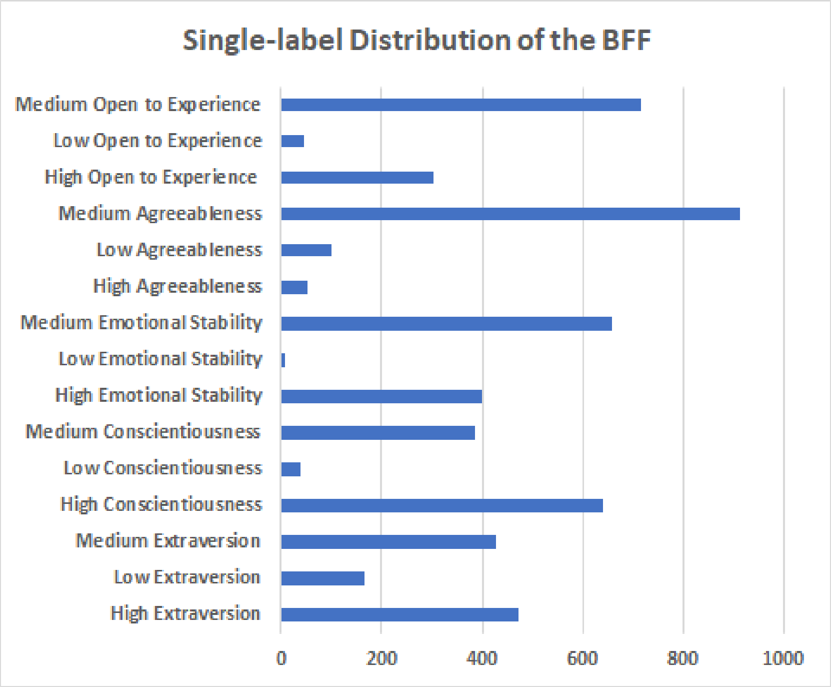


Figure 10: Single-label Distribution for the Five Factors Jointly in HWBFF Dataset

The figures demonstrate that distribution of the single-labels for the big five factors is highly skewed, 85.55% of the dataset is occupied by medium agreeableness and while low emotional stability only holds 0.85% of the dataset. Table 20 shows the number of samples per label.

Single Label	**Number of Samples**	
High Extraversion	471	1066
Low Extraversion	166	
Medium Extraversion	429	
High Conscientiousness	641	1066
Low Conscientiousness	38	
Medium Conscientiousness	387	
High Emotional Stability	400	1066
Low Emotional Stability	9	
Medium Emotional Stability	657	
High Agreeableness	52	1066
Low Agreeableness	102	
Medium Agreeableness	912	

Single Label	Number of Samples	
High Open to Experience	304	1066
Low Open to Experience	46	
Medium Open to Experience	716	

Table 20: Number of Samples of each label in the HWBFF Dataset

Our multi-label classification problem is transformed at the end into a set of independent binary classification problems by fitting one classifier per label following (one-vs-all) scheme. For this, all the samples which belong to a certain label are marked as positive represented by (1), while the reminder ones will be marked with (0) which is negative no matter what labels they contain. So, we have 15 datasets at the end, one set for each single label.Therefore, distribution analysis of positive and negative instances for each Single-label has been done on 1066 handwriting samples, see Figure 11.

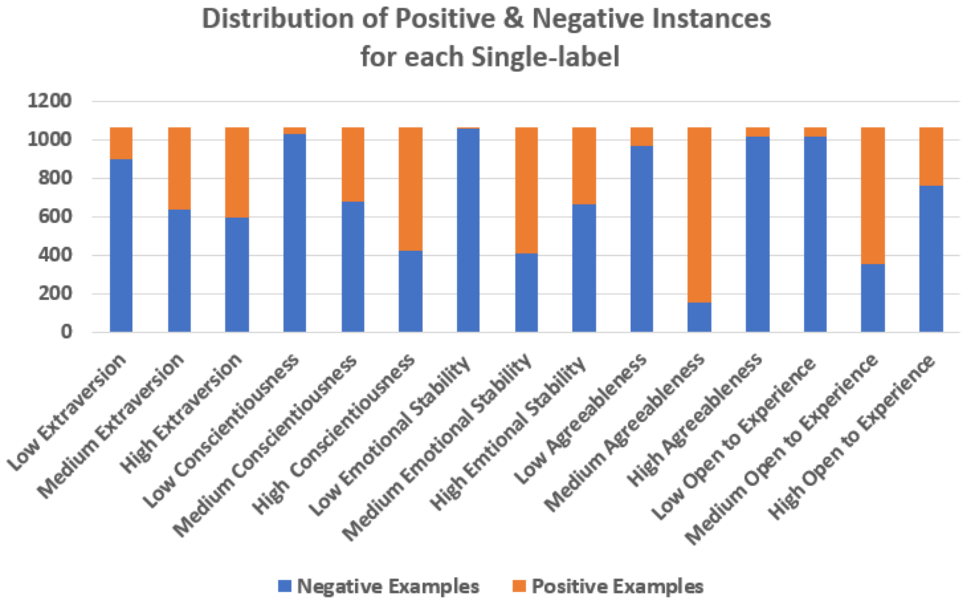


Figure 11: Distribution of Positive and Negative Instances for each Single-label in the HWBFF Dataset

It can been seen from Figure 11 that the positive and negative distribution of High Conscientiousness and Medium Emotional Stability is approximately balanced proportion, while for the other single-labels is imbalanced. Table 21 shows the number of positive and negative instances for single-labels in the big five factors.

Single-label	Negative Instances	Positive Instances	Total
Low Extraversion	898	168	1066
Medium Extraversion	639	427	
High Extraversion	593	473	
Low Conscientiousness	1028	38	
Medium Conscientiousness	678	388	
High Conscientiousness	424	642	
Low Emotional Stability	1056	10	
Medium Emotional Stability	408	658	
High Emtional Stability	666	400	
Low Agreeableness	966	100	
Medium Agreeableness	150	916	
High Agreeableness	1015	51	
Low Open to Experience	1019	47	
Medium Open to Experience	351	715	
High Open to Experience	761	305	

Table 21: Number of Positive and Negative Instances for Single-labels in the HWBFF Dataset

5.2.4.3 Imbalanced Level Assessment

The measurement of the imbalance level in a dataset is obtained as the ratio of the number of samples of the majority class and the number associated with the minority class, being known as Imbalance Ratio (IR). In binary classification, IR is calculated by dividing the frequencies of the majority class by the minority class. The higher the IR, the larger the imbalance level. However, in multi-label dataset, we use three measures in order to define the imbalance level i.e. Imbalance Ratio per Label (IRLbl), Average Imbalance Ratio per label (AvgIR), and Coefficient of Variation of IRLbl (CVIR). The higher AvgIR and CVIR, the larger the imbalance level [9].

A. Imbalance Ratio per Label (IRLbl)

With D being a multi-label dataset with a set of labels Y, and Y_i the i-th label, IRLbl is calculated for the label y as the ratio between the majority label and the label y, as shown in Equation (2). The larger the IRLbl is, the higher would be the imbalance level for the considered label [9].

$$IRLbl(y) = \frac{argmax_{y'}^{Y_{|Y|}} = Y_1(\sum_{i=1}^{|D|} h(y', Y_i))}{\sum_{i=1}^{|D|} h(y, Y_i)}, h(y, Y_i) = \begin{cases} 1, y \in Y_i \\ 0, y \notin Y_i \end{cases} \tag{2}$$

B. Average Imbalance Ratio per Label (AvgIR)

It represents the average level of imbalance in an multi-label dataset, computed as shown in Equation (3). Since different label distributions can produce the same AvgIR value, this measure should always be used jointly with measure (C) [9]:

$$AvgIR = \frac{1}{|Y|} \sum_{y=Y_1}^{Y_{|Y|}} (IRLbl(y)) \tag{3}$$

C. Coefficient of Variation of IRLbl (CVIR)

This is the coefficient of variation of IRLbl, and is obtained as shown in Equation (4). It indicates if all labels suffer from a similar level of imbalance or there are big differences in them. The larger the CVIR value, the higher would be this difference. Table 22 shows IRLbl, AvgIR, and CVIR in the big five factors. It can be observed from the value of AvgIR (14.13) and CVIR (7.12) in the table, there is a high imbalance in our multi-label dataset [9].

$$CVIR = \frac{IRLbl\sigma}{AvrIR}, IRLbl\sigma = \sqrt{\sum_{y=Y_1}^{Y_{|Y|}} \frac{(IRLbl(y) - AvgIR)^2}{|Y| - 1}} \tag{4}$$

Single-label	Negative Instances	Positive Instances	Imbalance Ratio per Label (IRLbl)
Low Extraversion	898	168	5
Medium Extraversion	639	427	2
High Extraversion	593	473	1
Low Conscientiousness	1028	38	27
Medium Conscientiousness	678	388	2
High Conscientiousness	424	642	2
Low Emotional Stability	1056	10	106
Medium Emotional Stability	408	658	2

Single-label	**Negative Instances**	**Positive Instances**	**Imbalance Ratio per Label (IRLbl)**
High Emtional Stability	666	400	2
Low Agreeableness	966	100	10
Medium Agreeableness	150	916	6
High Agreeableness	1015	51	20
Low Open to Experience	1019	47	22
Medium Open to Experience	351	715	2
High Open to Experience	761	305	3
Average Imbalance Ratio (AvgIR)			14.1333
Coefficient of Variation of IRLbl (CVIR)			7.1269

Table 22: Imbalance Ratio per Label (IRLbl), Average Imbalance Ratio (AvgIR), and Coefficient of Variation of IRLbl (CVIR) for HWBFF Dataset

Table 23 shows some basic characteristics such as number of instances, number of features, and number of labels along with imbalance measures, i.e. AvgIR and CVIR of HWBFF dataset. In addition, there are two columns for cardinality and density. Cardinality measures the average number of labels associated with each instance, as shown in Equation (5), and density is defined as cardinality divided by the number of labels, as shown in Equation (6).

$$Card(D) = \sum_{i=1}^{|D|} \frac{|Y_i|}{|D|} \tag{5}$$

$$Dens(D) = \frac{Card(D)}{|Y|} \tag{6}$$

Dataset	**Instances**	**Features**	**Labels**	**Card**	**Dens**	**AvgIR**	**CVIR**
HWBFF	1066	24	5	5	1	14.1333	7.1269

Table 23: Basic characteristics and imbalance measures of HWBFF dataset

5.3 Data Digitization

In data digitization we converted a paper based handwriting document into an electronic form. Electronic conversion is carried out using a process wherein a document is scanned and then a bitmap image of the original document is produced. The handwriting samples were scanned in a color scale at the resolution of 600 dpi using HP Color LaserJet Enterprise M553 series scanner with feature of automatic document feeder.

5.4 Data Preprocessing

5.4.1 Data Cleaning

Removing Unwanted Data
A few handwriting samples contain unwanted data such as crossed out writing, signs, scribble and drawing, signature, and number of pages. We removed all these unwanted data manually using Adobe Photoshop in order to keep only the handwriting on the page.

Image Denoising
Noise reduction was done to remove the noise caused by the scanning process and the noise of the paper. For this, the following OpenCV function was used:

cv2.fastNlMeansDenoisingColored()

Background Extraction
Some handwriting samples given by the graphologist were written on lined paper. Therefore, we extracted these lines by extracting the background using the following approach:

1. Convert image to Hue-Saturation-Value (HSV) format and color threshold with the following OpenCV function:

   cv2.inRange( )

2. Perform morphological transformations to smooth image using the following OpenCV function:

cv2.morphologyEx( )

3. Isolate characters by masking with the original image.

4. Recolour characters.

5.4.2 Data Augmentation

Our model includes data augmentation in order to enlarge the training dataset. The augmented data are generated before training the classifier. We augmented our data using techniques that do not change or alter the handwriting features that the five factors are revealed from. Table 11 in chapter 5 shows the definitions of the handwriting features corresponding to each of the five factors. Based on this table, the following four augmentation techniques were used which are rescaling, height shifting wherein the image is shifted vertically, vertical flip, and brightness.

$$rescale = 1./255$$

$$height\_shift\_range = 0.5$$

$$vertical\_flip = True$$

$$brightness\_range = [0.1, 0.9]$$

Chapter 6

Method

As mentioned in Chapter 5, the Big Five Factors test in psychology is formulated into a multi-label classification in computer science, since it predicts the measurement level (low, medium, or high) of the the five factors simultaneously. Based on the analysis of data distribution carried out in Chapter 5, our HWBFF dataset is considered as an imbalanced dataset as shown in Figure 12. Class imbalance fails to properly represent the distributive characteristics of the data and provide unsatisfying accuracy. Therefore, there is a need to handle our imbalanced dataset properly in order to get favorable results.

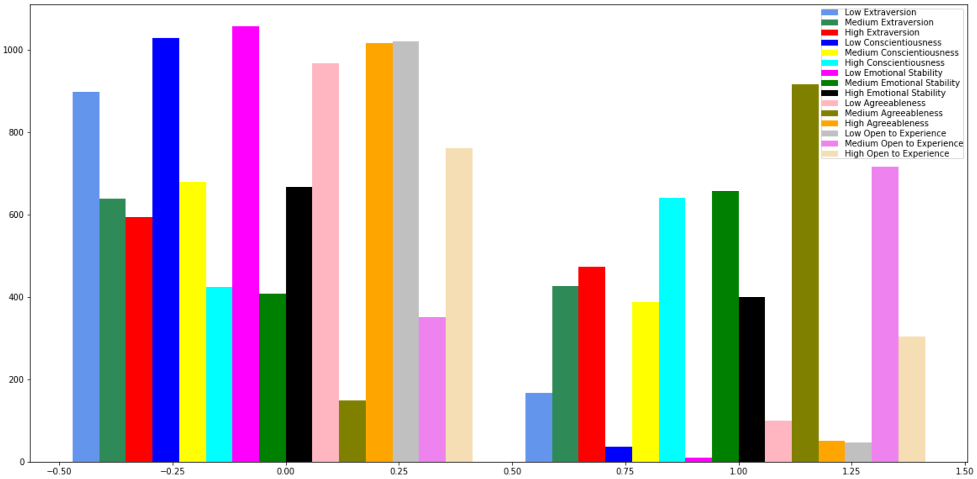


Figure 12: Number of Positive and Negative Instances for Single-labels in the HWBFF Dataset

In handling imbalanced dataset, the samples near the decision boundary should be valued since they contain more discriminative information and the skew of the boundary would be corrected by constructing synthetic samples. In addition to that, an ensemble model always tends to capture more complicated and robust decision boundary in practice. Taking these two factors into consideration, we proposed an ensemble method called Averaging of SMOTE Multi-label SVM-CNN (AvgMlSC). AvgMlSC constructs synthetic samples using Synthetic Minority Over-sampling Technique (SMOTE) to incorporate the borderline information and averaging the two classifiers i.e. Multi-label Support Vector Machine (MLSVM) and Multi-label Convolutional Neural Network (MLCNN) to produce one optimal predictive model. The following sections describe the proposed framework in detail.

6.1 Materials and Methods

Our multi-label classification problem is transformed firstly into five independent multi-class classification problems, one associated with each big five factor (Extraversion, Conscientiousness, Emotional Stability, Agreeableness, and Open to Experience). Then, each multi-class classification is transformed into three of independent binary classification problems by fitting one binary classifier for each single class which is the measurement level (low, medium, and high) following (one-vs-all) scheme. So, at the end we have 15 binary classifiers, 3 classifiers for each big five factor, see Figure 13.

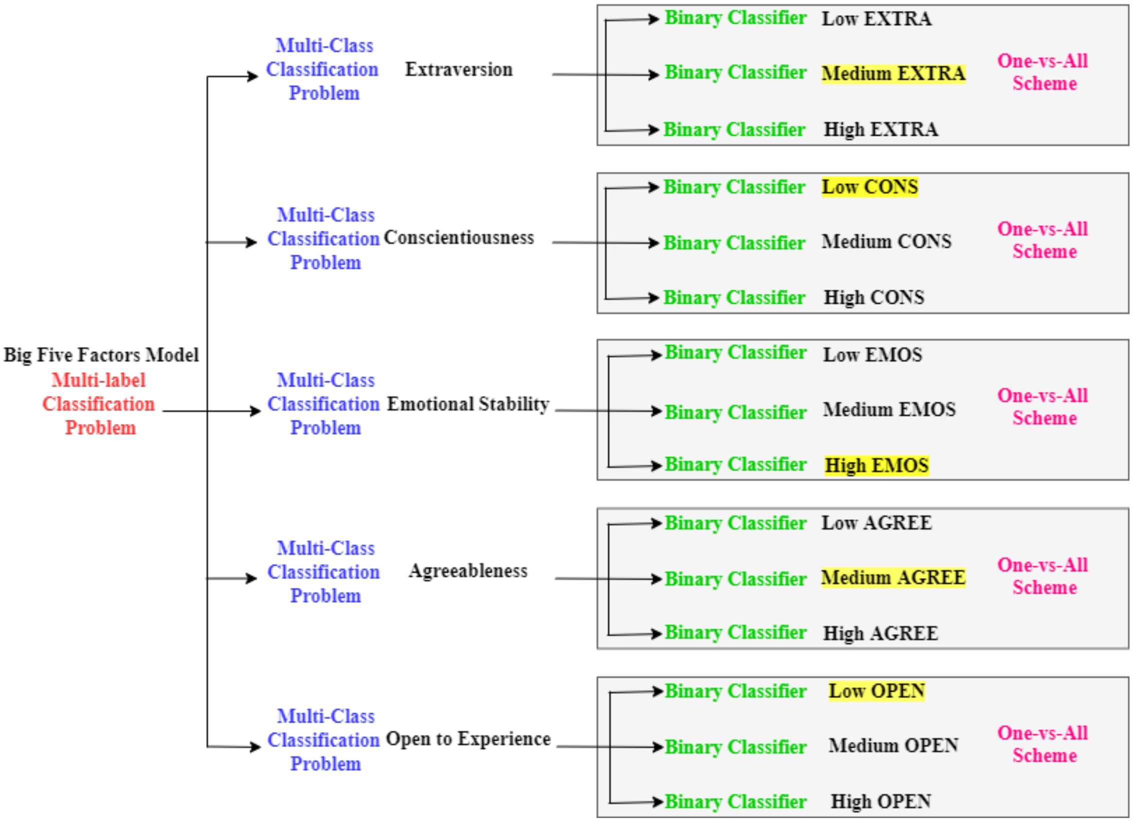


Figure 13: Transforming Multi-label BFF Classification Problem into Five Multi-class Classification Problems following (One-vs-All) Scheme

(one-vs-all) scheme deals with mutually exclusive choices, that means we should see only one binary classifier light up for each multi-class classifier. Therefore, we cannot transform our multi-label problem into one multi-class problem with following (one-vs-all) scheme because at the end we will have only one measurement level for one factor is predicted, see Figure 14.

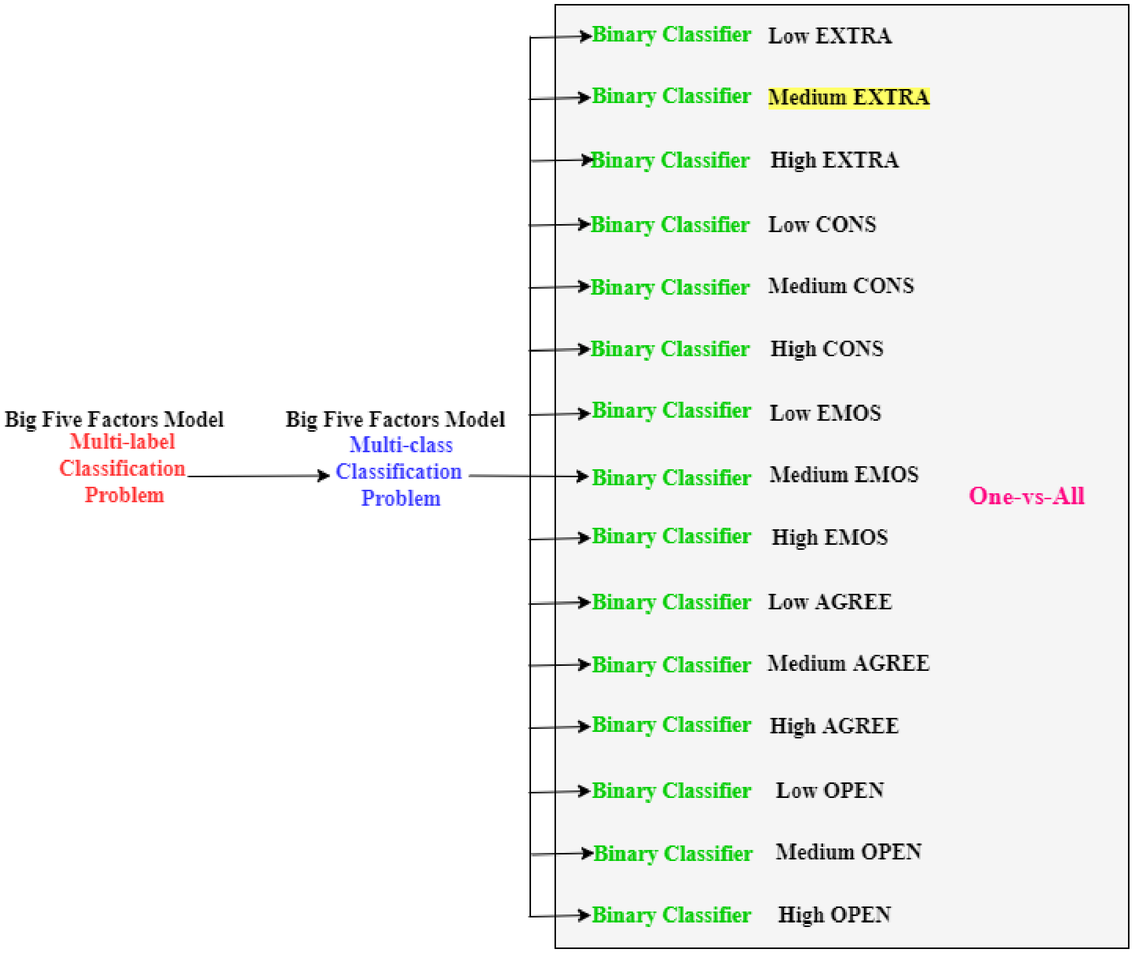


Figure 14: Transforming Multi-label BFF Classification Problem into One Multi-class Classification Problems following (One-vs-All) Scheme

Moreover, we cannot transform our multi-label problem directly into 15 of independent binary relevance problem associated with each measurement level in each factor, because in binary relevance more than one classifier light up at the same time, that means there would be one big five factor that is measured as high and low simultaneously, see Figure 15.

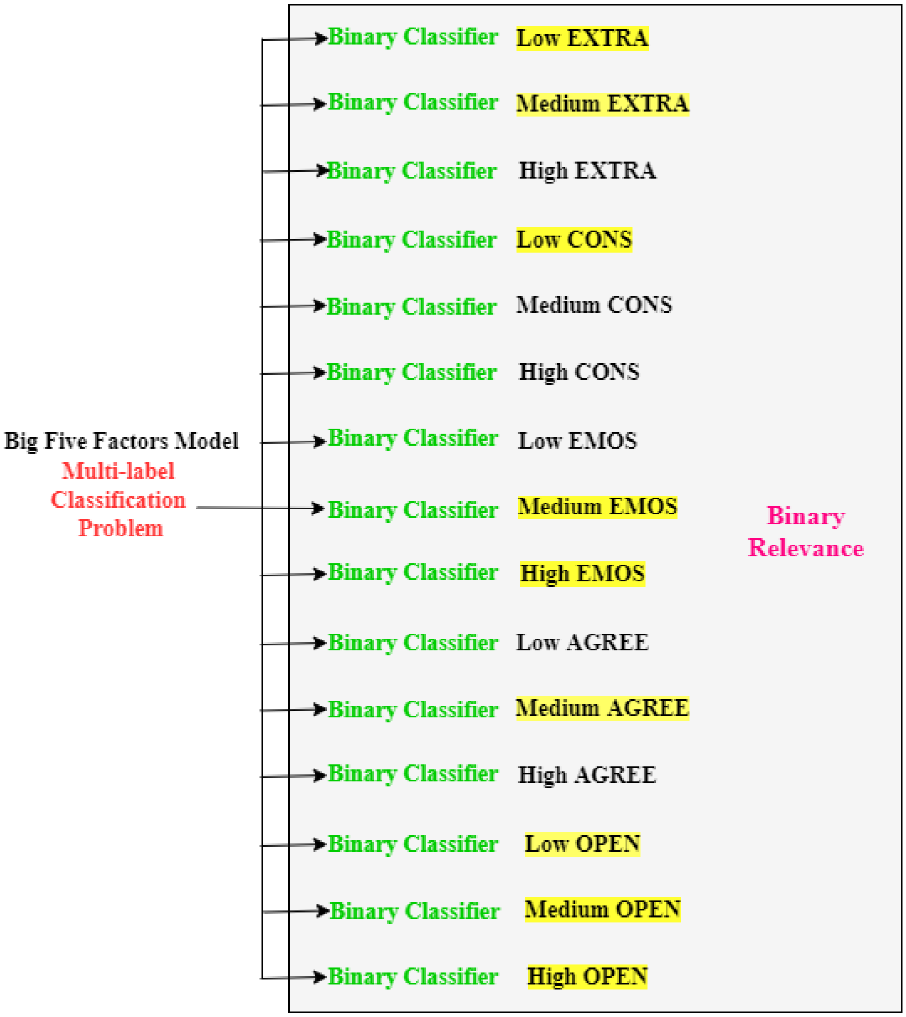

Big Five Factors Model
Multi-label
Classification
Problem
Binary Classifier Low EXTRA
Binary Classifier Medium EXTRA
Binary Classifier High EXTRA
Binary Classifier Low CONS
Binary Classifier Medium CONS
Binary Classifier High CONS
Binary Classifier Low EMOS
Binary Classifier Medium EMOS
Binary Classifier High EMOS
Binary Classifier Low AGREE
Binary Classifier Medium AGREE
Binary Classifier High AGREE
Binary Classifier Low OPEN
Binary Classifier Medium OPEN
Binary Classifier High OPEN
Binary
Relevance

Figure 15: Transforming Multi-label BFF Classification Problem into 15 Binary Classification Problems following Binary Relevance

The two classifiers (MLSVM and MLCNN) of the ensemble learning are trained and evaluated individually and sequentially. In each binary classifier inside each multi-class SVM and CNN, SMOTE which is an oversampling strategy is applied within 10-fold cross validation on each original single-label training set to construct synthetic samples from the minority classes. That means, before starting the oversampling process, Original Single-label Training Set is splitted into 10 folds. After that, the SMOTE is applied on each fold, then the resampled fold used for the training. After training, the model is evaluated using the validation set generated in the cross validation. The process of training and evaluation are repeated 10 times. At the end, the outputs of the five multi-class classifiers are joined together. Each model tested using unseen data and the predicted results of each classifier are averaged to produce one optimal output, see Figure 16 and Figure 17. The following sub sections illustrate some basic knowledge and architecture of MLSVM, MLCNN, and the algorithm of SMOTE.

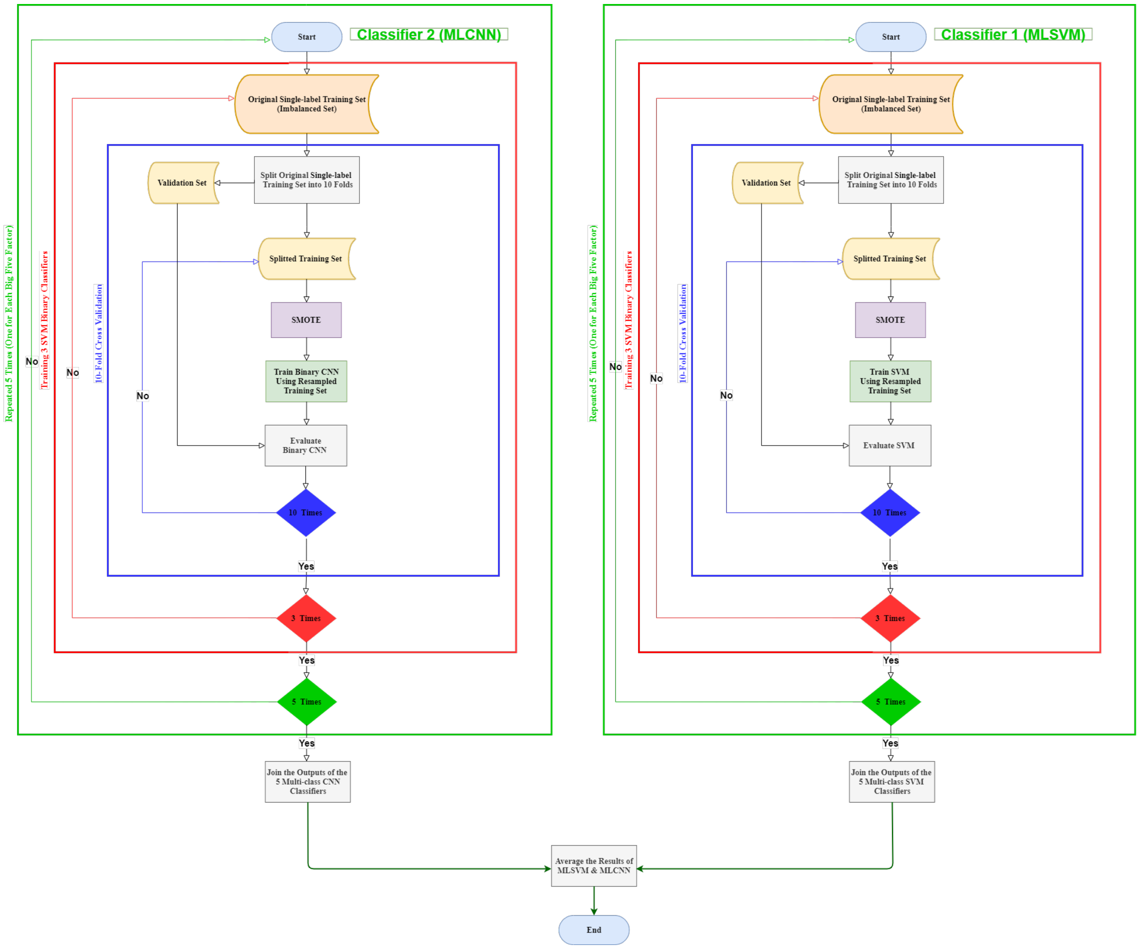

Repeated 5 Times (One for Each Big Five Factor)
Training 3 SVM Binary Classifiers
10-Fold Cross Validation
Start
Classifier 2 (MLCNN)
Original Single-label Training Set (Imbalanced Set)
Validation Set
Split Original Single-label Training Set into 10 Folds
Splitted Training Set
SMOTE
Train Binary CNN Using Resampled Training Set
Evaluate Binary CNN
10 Times
Yes
3 Times
Yes
5 Times
Yes
No
No
No
Join the Outputs of the 5 Multi-class CNN Classifiers
Repeated 5 Times (One for Each Big Five Factor)
Training 3 SVM Binary Classifiers
10-Fold Cross Validation
Start
Classifier 1 (MLSVM)
Original Single-label Training Set (Imbalanced Set)
Validation Set
Split Original Single-label Training Set into 10 Folds
Splitted Training Set
SMOTE
Train SVM Using Resampled Training Set
Evaluate SVM
10 Times
Yes
3 Times
Yes
5 Times
Yes
No
No
No
Join the Outputs of the 5 Multi-class SVM Classifiers
Average the Results of MLSVM & MLCNN
End

Figure 16: Flowchart of AvgMlSC Framework

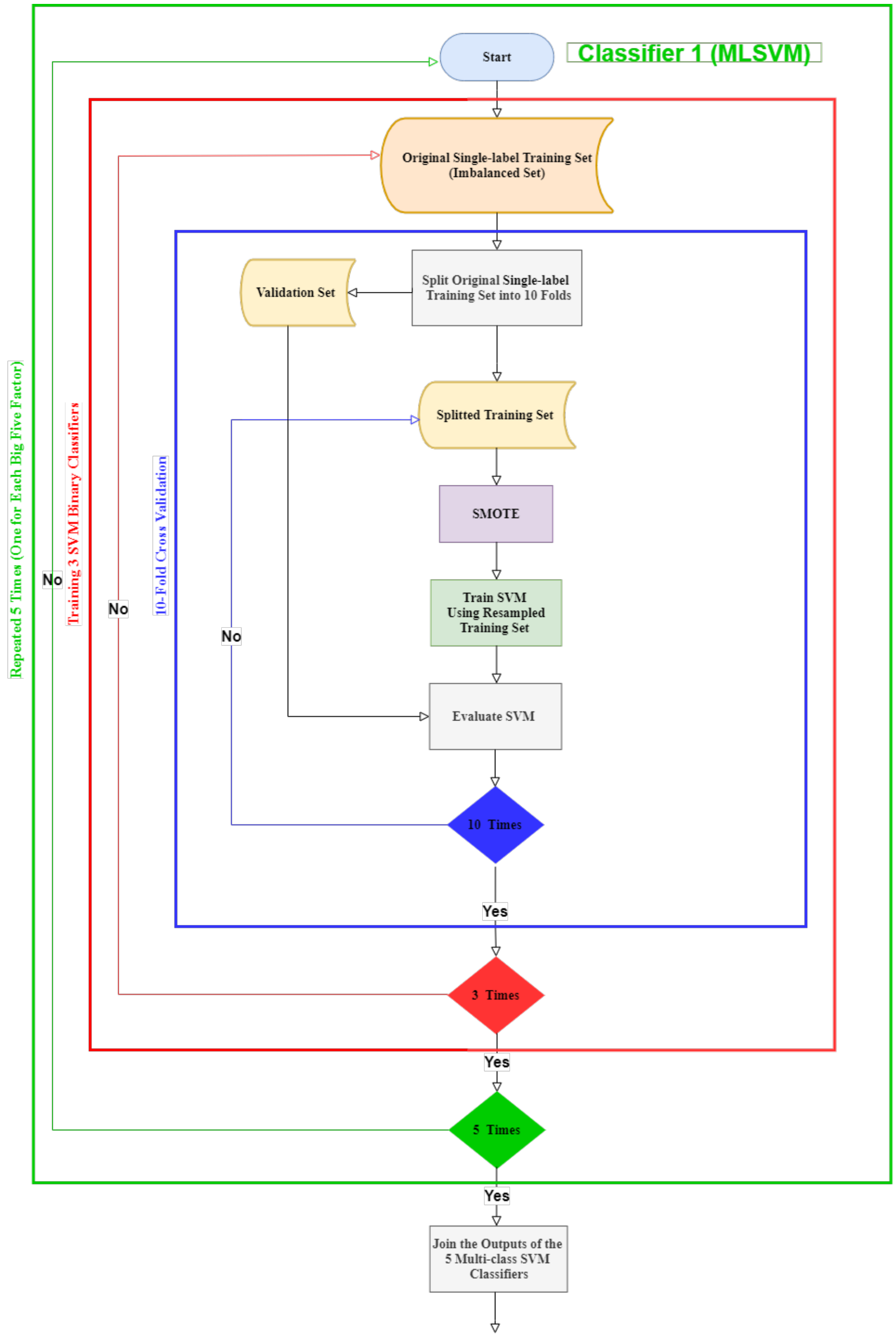
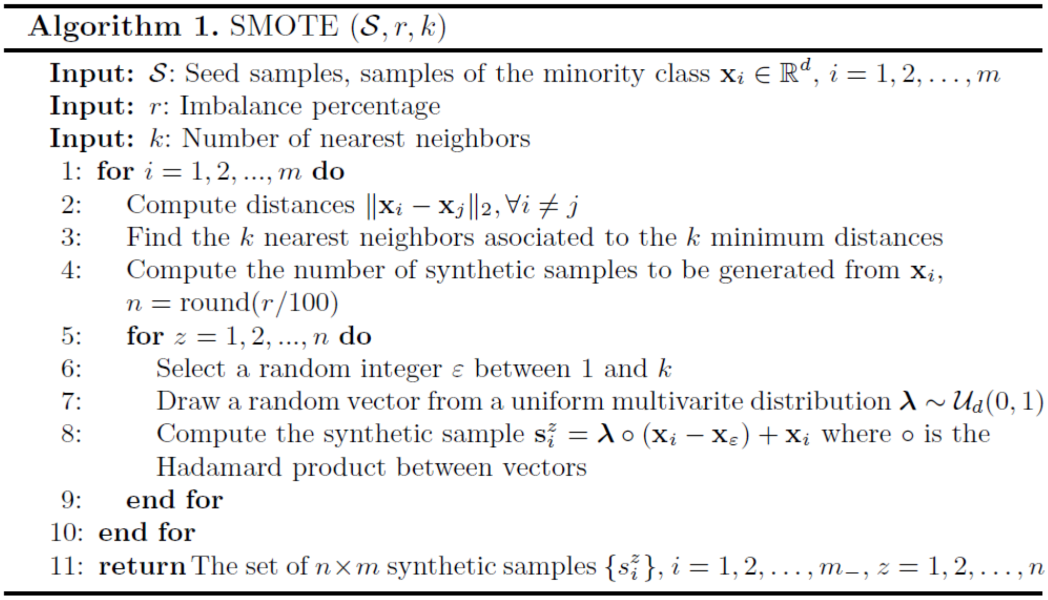


Figure 17: Flowchart of one classifier (MLSVM)

6.1.1 Synthetic Minority Over-sampling Technique (SMOTE)

There are many techniques used for dealing with an imbalanced dataset. One of the most commonly preferred approaches is resampling which consists of two types of methods i.e. random undersampling and random oversampling. In random undersampling, samples are removed from the majority class, while in random oversampling, more examples are added from the minority class. However, these two techniques provide undesiarable results in several cases because the former causes a loss of potentially useful information, while the latter induces overfitting due to the exact replication of samples. In order to improve these limitations, Synthetic Minority Over-sampling Technique (SMOTE) [10] used in this work. SMOTE is an oversampling strategy that helps to overcome overfitting by focusing on the feature space rather than data space and interpolating synthetic samples along the line segments connecting seed samples and forcing the decision region of the minority class to become more general. Thus, in SMOTE, synthetic samples are not exact copies of the original ones. Figure 18 shows the algorithm that describes the procedure of SMOTE [22].



```
Algorithm 1. SMOTE $(\mathcal{S}, r, k)$
Input: $\mathcal{S}$: Seed samples, samples of the minority class $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, m$
Input: $r$: Imbalance percentage
Input: $k$: Number of nearest neighbors
for $i = 1, 2, ..., m$ do
    Compute distances $\|\mathbf{x}_i - \mathbf{x}_j\|_2, \forall i \neq j$
    Find the $k$ nearest neighbors associated to the $k$ minimum distances
    Compute the number of synthetic samples to be generated from $\mathbf{x}_i$,
    $n = \text{round}(r/100)$
    for $z = 1, 2, ..., n$ do
        Select a random integer $\varepsilon$ between 1 and $k$
        Draw a random vector from a uniform multivarite distribution $\boldsymbol{\lambda} \sim \mathcal{U}_d(0, 1)$
        Compute the synthetic sample $\mathbf{s}_i^z = \boldsymbol{\lambda} \circ (\mathbf{x}_i - \mathbf{x}_\varepsilon) + \mathbf{x}_i$ where $\circ$ is the
        Hadamard product between vectors
    end for
end for
return The set of $n \times m$ synthetic samples $\{s_i^z\}$, $i = 1, 2, \ldots, m_-$, $z = 1, 2, \ldots, n$
```

Figure 18: Synthetic Minority Over-sampling Technique (SMOTE) Algorithm [22]

As illustrated in Algorithm 1, SMOTE oversamples the minority class by taking each minority class sample and introducing synthetic examples in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. Depending upon the amount of oversampling required, neighbours from the k nearest neighbours are randomly chosen and joined to the synthetic examples.

6.1.2 Multi-label Support Vector Machine (MLSVM)

Support vector machine (SVM) is one of the popular classifiers in binary classification. In this work, we transformed our multi-label classification problem at the end into a set of independent binary classification problems by fitting one classifier per class. This mechanism named (one-vs-all) scheme which is a conceptually simple and computationally efficient solution for multi-label classification. Therefore, as a first classifier in our ensemble method, multi-label learning using support vector machine (MLSVM) for the binary classification problem associated with each class is conducted in this study.

Standard SVM was defined formally by Li and Guo as follows [28]. Given a labeled multi-label training set $D = \{(x_{i,i})\}_{i=1}^{N}$ where x_i is the input feature vector for the i-th instance, and its label vector y_i is a $\{+1, -1\}$-valued vector with length K such as $K = |y|$. If $y_{ik} = 1$, it indicates that the instance x_i is assigned into the $k - th$ class; otherwise, the instance does not belong to the $k - th$ class. For the $k - th$ class $(k = 1, ..., K)$, the binary SVM training is a standard quadratic optimization problem:

$$min_{W_k,b_k,\{\xi_{ik}\}}\frac{1}{2}\left\|W_k\right\|^2 + C\sum_{i=1}^{N}\xi_{ik} \tag{7}$$

Subject to $y_{ik}\left(w_k^T x_i + b_k\right) \geq 1 - \xi_{ik}, \xi_{ik} \geq 0, \forall_i$, where $\{\xi_{ik}\}$ are the slack variables and C is the trade-off parameter. It maximizes the soft class separation margin. The model parameters w_k and b_k returned by this binary learning problem define a binary classifier associated with the $k - th$ class: $f_k\left(x_i\right) = w_k^T x_i + b_k$. The set of binary classifiers from all classes can be used independently to predict the label vector $\hat{y}$ for an unlabeled instance $\hat{x}$ The $k - th$ component of the label vector $\hat{y}_k$ has value 1 if $f_k\left(\hat{x}\right) > 0$, and has value -1 otherwise. The absolute value $\left|f_k\left(\hat{x}\right)\right|$ can be viewed as a $confidence$ value for its prediction $\hat{y}_k$ on instance $\hat{x}$.

6.1.3 Multi-label Convolutional Neural Network (MLCNN)

The second classifier in our ensemble method is Multi-label Convolutional Neural Network (MLCNN). For this, we conduct multi-label learning under (one-vs-all) scheme by using Convolutional Neural Networks (CNNs) in order to transform our multi-label classification problem into a set of independent binary classification problems. Trucco et al. defined standard Convolutional Neural Network (CNN) as a particular kind of neural network where the weights are learned for the application of a series of convolutions on the input image, being the filter weights shared across the same convolutional layer [46]. CNN replaces fully connected layers in neural network by operators ℓ defined by small convolution kernels. This localizes computations, effectively reducing the number of parameters in $\cup_\Theta$. The resulting network is defined as:

$$\cup_\Theta\left(X\right)=\left(\ell^{f^L}\circ ... \circ \ell^{f^j}\circ ...\ell^{f^2}\circ \ell^{f^1}\right)\left(X\right) \tag{8}$$

Convolutional layer j is determined by a set $f^j=\left\{f_1^j,...,f_1^{j+1}\right\}$ of such kernels, and accepts as input a tensor x^j of dimension $h_j \times w_j \times c_j$. Convolving x^j with each of these $j+1$ filters and stacking the output results in a tensor x^{j+1} of dimension $h_j \times w_j \times c_{j+1}$. Each of these convolutional layers is followed by a nonlinear pointwise function, and the spatial size $h_j \times w_j$ of the output tensor is decreased by means of a pooling operator $p^j : \mathbb{R}^{h_j \times w_j} \rightarrow \mathbb{R}^{h_{j+1} \times w_{j+1}}$ In a CNN, learnable weights lie in convolution kernels, and the training process leads to finding the optimal way of filtering the training data so that irrelevant information is discarded and the error (loss) in the training set is decreased as much as possible.

Figure 19 shows the architecture implemented in each binary CNN classifier for each single label. As we can see, the input of our MLCNN combines two types of data which are structured and unstructured data. Structured data is organized and fits tidily into spreadsheets and relational databases such as names, dates, addresses, and credit card numbers. On the other hand, unstructured data has no predefined construction or systemization such as text form, audio, images, and videos. Therefore, our MLCNN consists of two neural networks for inputs. The first one named ImageCNN which is a convolutional neural network used for unstructured data i.e. images of handwriting samples. The second network called FeatureFCNN which is a fully connected neural network used for structured data i.e. the values of handwriting features. Then, the outputs from the two networks are concatenated and passed to ClassifierFCNN which is a fully connected neural network that classifies the handwriting samples into one class.

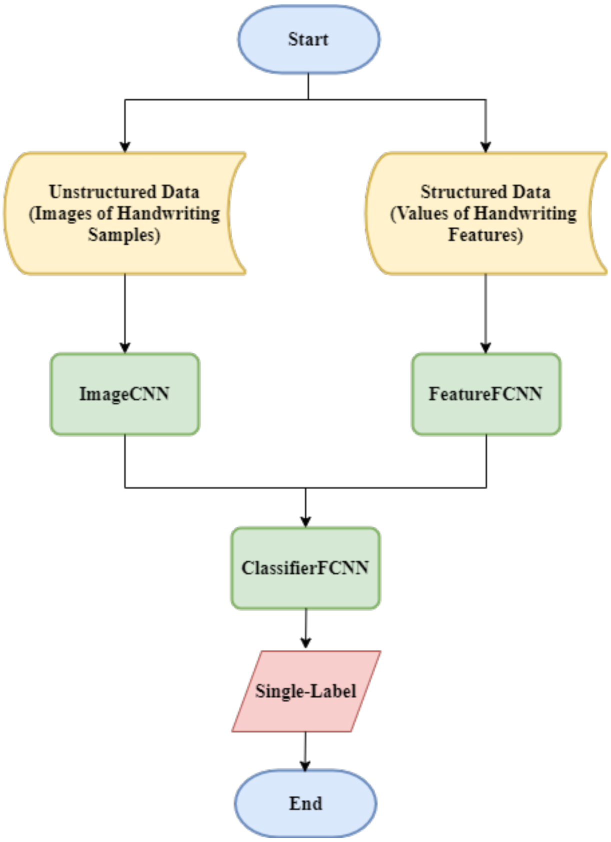
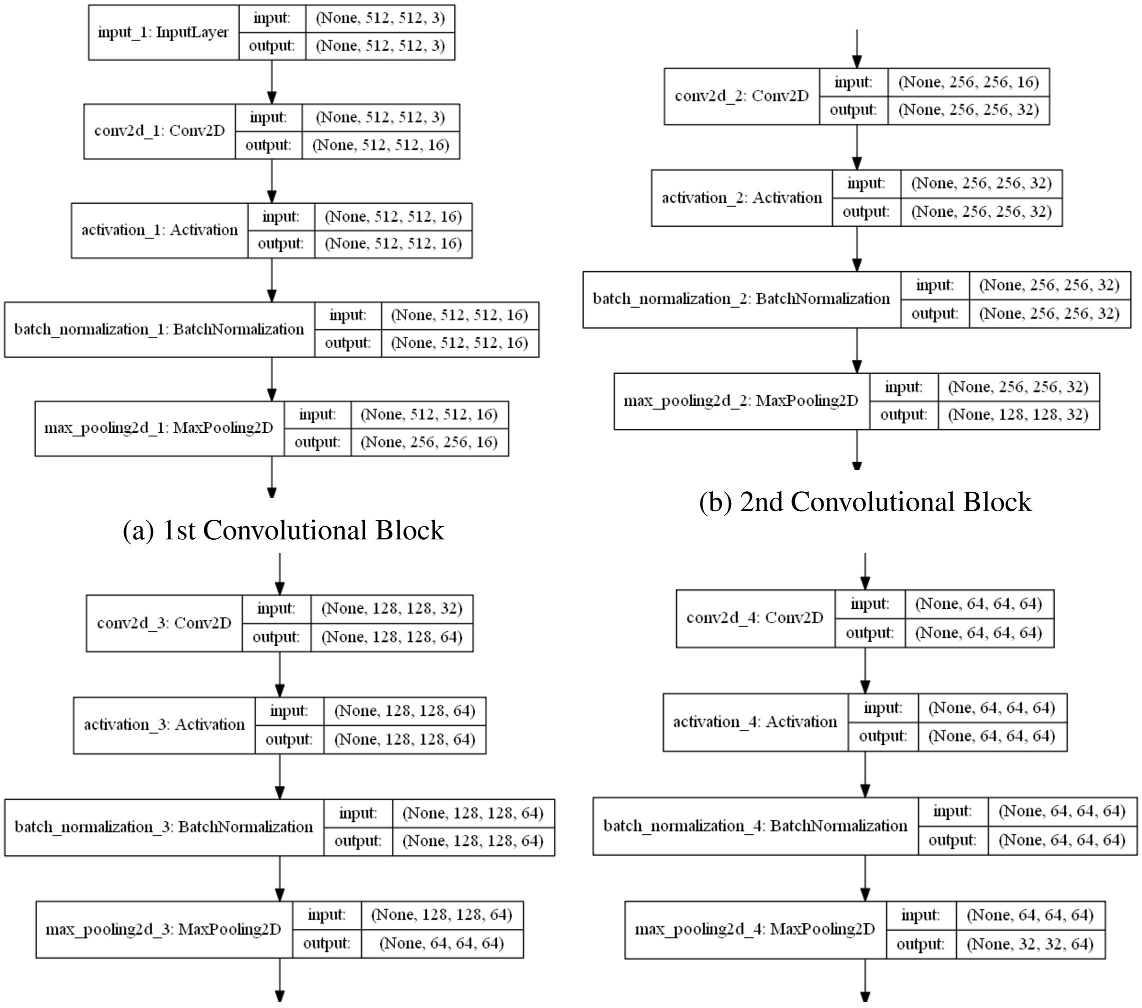

Start
Unstructured Data (Images of Handwriting Samples)
Structured Data (Values of Handwriting Features)
ImageCNN
FeatureFCNN
ClassifierFCNN
Single-Label
End

Figure 19: The Architecture of the Binary CNN

6.1.3.1 Image Convolutional Neural Network (ImageCNN)

To input the images of handwriting samples into MLCNN, we create a convolutional neural network that consists of one input layer which accepts a three dimensional color image of a fixed-size (512×512). Then, the input image is passed through 8 convolutional blocks. Each block consists of one convolution layer with filters of (3×3) pixel window, one rectified linear unit (ReLU) activation function layer for reducing the effect of gradient vanishing during backpropagation, and one batch normalization layer, then the layers are followed by Max-pooling layer which is performed over a (2×2) pixel window. The number of filters in each block is: 16, 32, 64, 64, 64, 128, 256, and 512, respectively. Once the filtration process is applied on the input image, it is passed through a layer to be flattened out to two fully connected hidden layers. The first layer contains 16 nodes followed by (ReLU) activation function layer, batch normalization layer, and finally dropout with a rate of 0.5. The second hidden layer contains 4 nodes followed by (ReLU) activation function layer, the number of nodes in this layer should match the number of nodes coming out from

FeatureFCNN. Max Pooling, Dropout and Batch Normalization layers are added to prevent overfitting and control the number of parameters in the network, see Figure 20 and 21, and 22.


input_1: InputLayer
input: (None, 512, 512, 3)
output: (None, 512, 512, 3)
conv2d_1: Conv2D
input: (None, 512, 512, 3)
output: (None, 512, 512, 16)
activation_1: Activation
input: (None, 512, 512, 16)
output: (None, 512, 512, 16)
batch_normalization_1: BatchNormalization
input: (None, 512, 512, 16)
output: (None, 512, 512, 16)
max_pooling2d_1: MaxPooling2D
input: (None, 512, 512, 16)
output: (None, 256, 256, 16)
conv2d_2: Conv2D
input: (None, 256, 256, 16)
output: (None, 256, 256, 32)
activation_2: Activation
input: (None, 256, 256, 32)
output: (None, 256, 256, 32)
batch_normalization_2: BatchNormalization
input: (None, 256, 256, 32)
output: (None, 256, 256, 32)
max_pooling2d_2: MaxPooling2D
input: (None, 256, 256, 32)
output: (None, 128, 128, 32)
conv2d_3: Conv2D
input: (None, 128, 128, 32)
output: (None, 128, 128, 64)
activation_3: Activation
input: (None, 128, 128, 64)
output: (None, 128, 128, 64)
batch_normalization_3: BatchNormalization
input: (None, 128, 128, 64)
output: (None, 128, 128, 64)
max_pooling2d_3: MaxPooling2D
input: (None, 128, 128, 64)
output: (None, 64, 64, 64)
conv2d_4: Conv2D
input: (None, 64, 64, 64)
output: (None, 64, 64, 64)
activation_4: Activation
input: (None, 64, 64, 64)
output: (None, 64, 64, 64)
batch_normalization_4: BatchNormalization
input: (None, 64, 64, 64)
output: (None, 64, 64, 64)
max_pooling2d_4: MaxPooling2D
input: (None, 64, 64, 64)
output: (None, 32, 32, 64)

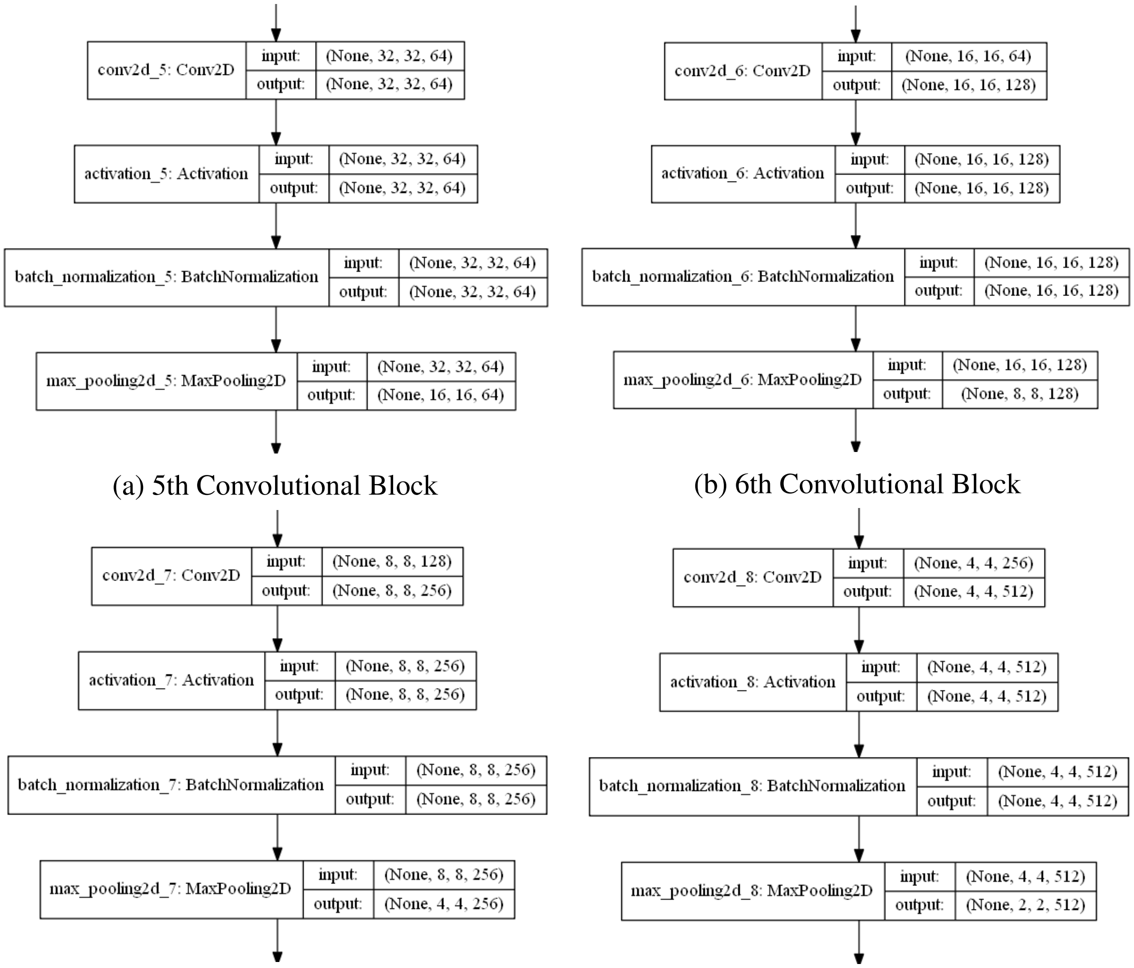
(a) 1st Convolutional Block

(b) 2nd Convolutional Block

(c) 3ed Convolutional Block

(d) 4th Convolutional Block

Figure 20: The First Four Convolutional Blocks in ImageCNN


conv2d_5: Conv2D
input: (None, 32, 32, 64)
output: (None, 32, 32, 64)
activation_5: Activation
input: (None, 32, 32, 64)
output: (None, 32, 32, 64)
batch_normalization_5: BatchNormalization
input: (None, 32, 32, 64)
output: (None, 32, 32, 64)
max_pooling2d_5: MaxPooling2D
input: (None, 32, 32, 64)
output: (None, 16, 16, 64)
conv2d_6: Conv2D
input: (None, 16, 16, 64)
output: (None, 16, 16, 128)
activation_6: Activation
input: (None, 16, 16, 128)
output: (None, 16, 16, 128)
batch_normalization_6: BatchNormalization
input: (None, 16, 16, 128)
output: (None, 16, 16, 128)
max_pooling2d_6: MaxPooling2D
input: (None, 16, 16, 128)
output: (None, 8, 8, 128)
conv2d_7: Conv2D
input: (None, 8, 8, 128)
output: (None, 8, 8, 256)
activation_7: Activation
input: (None, 8, 8, 256)
output: (None, 8, 8, 256)
batch_normalization_7: BatchNormalization
input: (None, 8, 8, 256)
output: (None, 8, 8, 256)
max_pooling2d_7: MaxPooling2D
input: (None, 8, 8, 256)
output: (None, 4, 4, 256)
conv2d_8: Conv2D
input: (None, 4, 4, 256)
output: (None, 4, 4, 512)
activation_8: Activation
input: (None, 4, 4, 512)
output: (None, 4, 4, 512)
batch_normalization_8: BatchNormalization
input: (None, 4, 4, 512)
output: (None, 4, 4, 512)
max_pooling2d_8: MaxPooling2D
input: (None, 4, 4, 512)
output: (None, 2, 2, 512)

(a) 5th Convolutional Block

(b) 6th Convolutional Block

(c) 7th Convolutional Block

(d) 8th Convolutional Block

Figure 21: The Last Four Convolutional Blocks in ImageCNN

6.1.3.2 Feature Fully Connected Neural Network (FeatureFCNN)

In order to input the structured data into MLCNN, a sequential model named FeatureFCNN is created for accepting the values of 24 handwriting features selected by the graphologist based on graphological rules. These structured data contain 5 features for Extraversion (i.e. middle zone more than 2,5 mm, narrow ending margin, dominance of garlands, progressive movement, and slanted in the direction of the writing), 4 features for Conscientiousness (i.e. regularity, legibility, Controlled movement, and Precision of placement of free strokes), 5 features for Emotional Stability (i.e. regularity without rigidity, baseline horizontal and flexible, slightly slanted, good balance between white space and ink space, and good pressure and quality of the stroke), 5 features for Agreeableness (i.e. dominance of curves

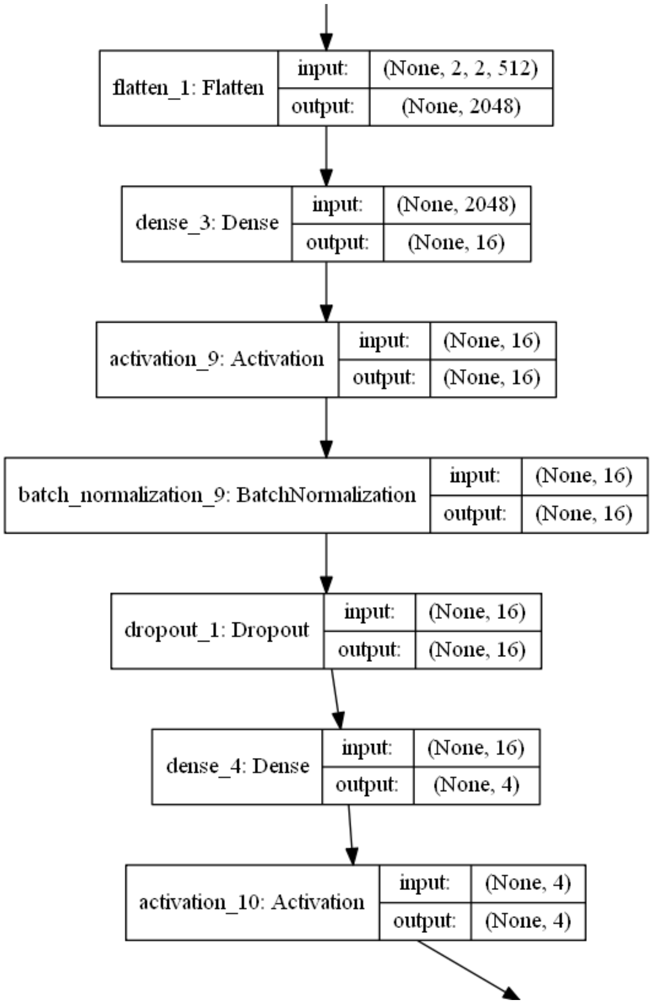


Figure 22: Architecture of fully connected layers in ImageCNN

versus angles, good space between letters, words, and lines, letter width >5, round letters without loops and slightly open, and nourished strokes), and finally 5 features for Open to Experience (i.e. good openness in loops, good speed and movement, slight angles in letters, slanted in the direction of handwriting, and narrow ending margin). FeatureFCNN is a fully connected neural network that consists of three layers. The first is the input layer which consists of the input shape as (None, 24). Then, the input features are passed through two fully connected hidden Layers. The first hidden layer contains 2 input vectors and 8 output vectors followed by rectified linear unit (ReLU) activation function and the second one contains 8 input vectors and 4 output vectors followed by rectified linear unit (ReLU) activation function. Figure 23 shows the architecture of the feature neural network.

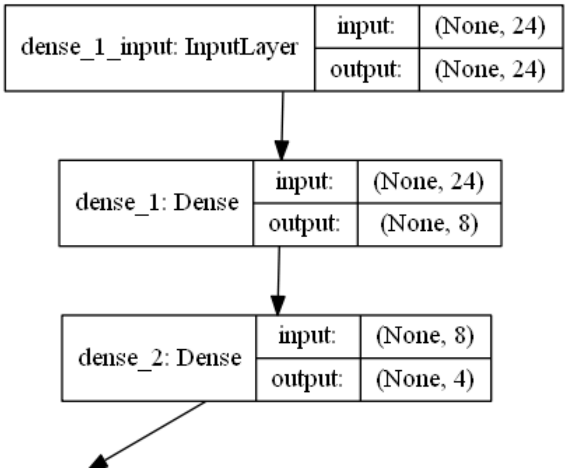


Figure 23: Architecture of Feature Fully Connected Neural Network (FeatureFCNN)

6.1.3.3 Classifier Fully Connected Neural Network (ClassifierFCNN)

The outputs of ImageCNN and FeatureFCNN are passed through Keras concatenation function to be concatenated and passed to the ClassifierFCNN which is a fully connected neural network that outputs multiple values. The ClassifierFCNN consists of two fully connected dense layers. The first one contains 4 nodes with (ReLU) activation function and the second layer contains one output class with sigmoid activation function. Figure 24 shows the architecture of ClassifierFCNN.

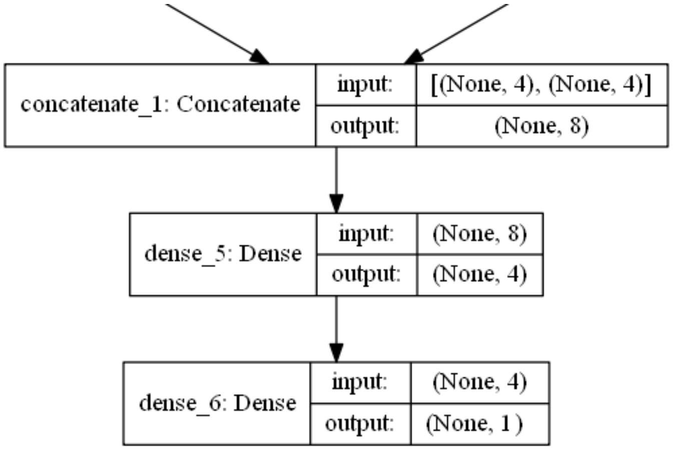


Figure 24: Architecture of Classifier Fully Connected Neural Network (ClassifierFCNN)

6.1.3.4 Model Optimization

To optimize the model, we use binary cross-entropy loss, since the output for each personality trait is either 0 or 1. Cross-entropy loss measures the performance of a model that outputs probabilities between 0 and 1. The loss increases when the prediction diverges from the actual label, so the goal of the network is to learn weights that minimize the loss. We use Adam with Weight Decay (AdamW) optimizer [30] in order to improve model generalization. AdamW uses two parameters which are weight_decay with rate of $1e-5$ and AMSGrad which is a stochastic optimization method that seeks to fix a convergence issue with Adam based optimizers. L2 regularization with rate of 0.0001 was used to avoid overfitting. In addition, class_weight parameter with 'balanced' value was used at fitting in order to handle imbalanced dataset.

6.1.4 Ensemble method

Since our HWBFF dataset is imbalanced, an ensemble method is used to improve the performance of the overall system. Model averaging is used for this work. In averaging approach each ensemble member contributes an equal amount to the final prediction. In the case of regression, the ensemble prediction is calculated as the average of the member predictions. In the case of predicting a class label, the prediction is calculated as the mode of the member predictions. In the case of predicting a class probability, the prediction can be calculated as the argmax of the summed probabilities for each class label. Argmax is an operation that finds the argument that gives the maximum value from a target function. Figure 25 shows the ensemble method for AvgMlSC.

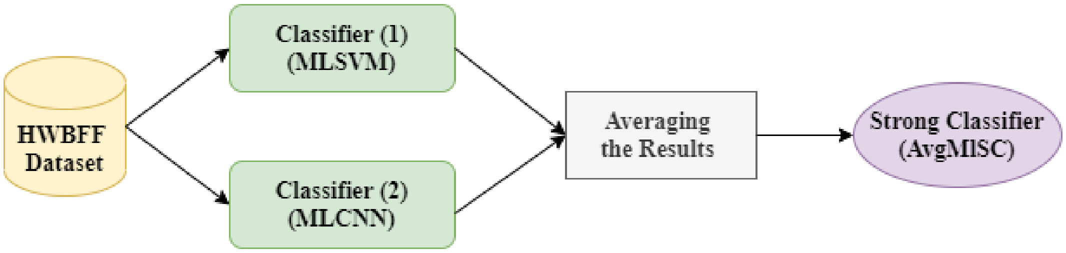
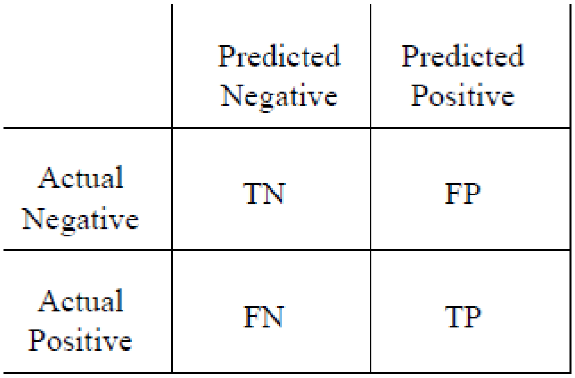


Figure 25: Ensemble Method for AvgMlSC

6.2 Performance Measures

A confusion matrix, as illustrated in Figure 24, is a typical measure for evaluating the performance of machine learning algorithms in a binary class problem. The columns are the Predicted class and the rows are the Actual class. In the confusion matrix, TN is the number of negative examples correctly classified (True Negatives), FP is the number of negative examples incorrectly classified as positive (False Positives), FN is the number of positive examples incorrectly classified as negative (False Negatives) and TP is the number of positive examples correctly classified (True Positives).



	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 24: Confusion Matrix

Predictive Accuracy is the performance measure generally associated with machine learning algorithms and is defined by Equation (9).

$$PredectiveAccuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

Error rate which is calculated by $(1 - PredictiveAccuracy)$ is used in the context of balanced datasets and equal error costs as a performance metric. However, for imbalanced datasets with unequal error costs, it is more appropriate to use the ROC curve. It stands for Receiver Operating Characteristic Curve. ROC curves represent the family of best decision boundaries for relative costs of TP and FP. On an ROC curve the X-axis represents $(\%FP = FP/(TN + FP))$ and the Y-axis represents $(\%TP = TP/(TP + FN))$. The ideal point on the ROC curve would be (1.00), that is all positive examples are classified correctly and no negative examples are misclassified as positive. There is a need for manipulating the balance of training samples for each class in the training set in order to increase the ROC curve. Figure 26 shows an example of the ROC curve.

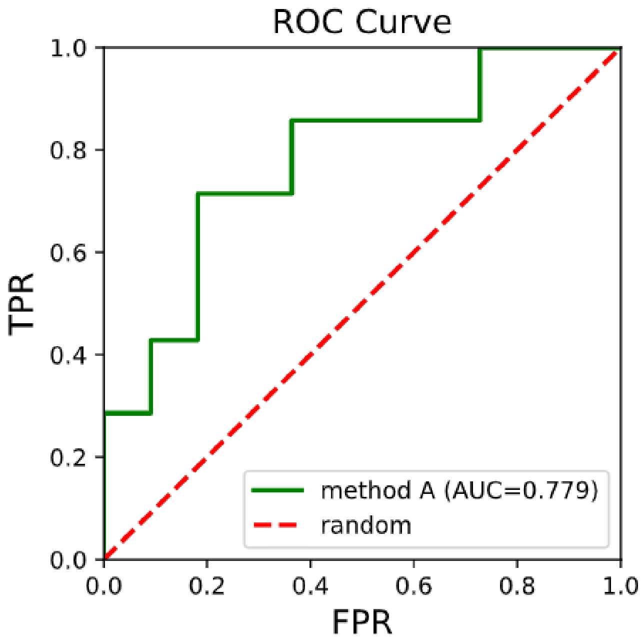


Figure 26: Example of ROC Curve

The line $y = x$ represents the scenario of randomly guessing the class. Area Under the ROC Curve (AUC) is a useful metric for classifier performance as it is independent of the decision criterion selected and prior probabilities. The AUC comparison can establish a dominance relationship between classifiers. F-Score is used as another performance measure since it combines between Recall which is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive, and Precision which is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly. F-Score is defined by the following equation.

$$F - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{10}$$

Chapter 7

Experimental Study

7.1 SMOTE influence on the imbalance level

The imbalance level for single-labels was reassessed after applying SMOTE on HWBFF dataset in order to analyze how SMOTE has influenced the label distributions in our dataset after adding 100, 200, and 300 synthetic samples. Since SMOTE is applied only to training partitions, Table 27 was obtained from the training set used in experimentation. However, the imbalance data previously shown in Table 22 correspond to the whole datasets. The training set was generated using approximately 90% of 1066 handwriting samples while the other 10% was used for testing. We have approximately 9761 examples in the majority class and 4879 examples in the minority class for the training set. Table 25 shows the number of positive and negative instances for each single-label in the training dataset.

Single-label	**Negative Instances**	**Positive Instances**	**IRLbl**	Total
Low Extraversion	838	138	6.07	976
Medium Extraversion	580	396	1.47	
High Extraversion	535	441	1.21	
Low Conscientiousness	948	28	33.86	
Medium Conscientiousness	623	353	1.16	
High Conscientiousness	381	595	1.56	
Low Emotional Stability	968	8	121	
Medium Emotional Stability	372	604	1.62	
High Emotional Stability	612	364	1.68	
Low Agreeableness	879	97	9.06	
Medium Agreeableness	143	833	5.83	

Single-label	**Negative Instances**	**Positive Instances**	**IRLbl**	Total
High Agreeableness	930	46	20.22	
Low Open to Experience	938	38	24.68	
Medium Open to Experience	332	644	1.94	
High Open to Experience	682	294	1.32	

Table 25: Number of Positive and Negative Instances for Single-labels in the training set (n = 976)

SMOTE was done within 10-fold cross-validation for each binary classifier, that means after splitting the original training set into 10 folds. The splitted training set in each fold consists of 90% of the original training set selected at random, with the remaining 10% used as a hold out set for validation. For this experiment, one splitted training set for one fold for each binary classifier was selected randomly for assessing the imbalance level before and after applying SMOTE. Table 26 shows IRLbl measures before applying SMOTE on one splitted training set for one fold generated within the cross-validation. As it is shown by the table, the imbalance ratio per label ranges from small to very high imbalanced dataset, the highest ones are 34, 124, 19, and 24 which belong to low Conscientiousness, low Emotional Stability, high Agreeableness, and low Open to Experience, respectively. However, the low Extraveraion, low Agreeableness, and medium Agreeableness represent the moderate level while the remaining labels are small imbalance level.

Before SMOTE				
Single-label	**Negative Instances**	**Positive Instances**	**IRLbl**	**Total**
Low Extraversion	753	125	6.024	878
Medium Extraversion	522	356	1.466	
High Extraversion	481	397	1.212	
Low Conscientiousness	853	25	34.12	
Medium Conscientiousness	560	318	1.761	
High Conscientiousness	343	535	1.559	
Low Emotional Stability	871	7	124.429	
Medium Emotional Stability	335	543	1.621	
High Emotional Stability	550	328	1.677	
Low Agreeableness	791	87	9.092	
Medium Agreeableness	129	749	5.806	

Before SMOTE				
High Agreeableness	836	42	19.905	
Low Open to Experience	844	34	24.824	
Medium Open to Experience	299	579	1.936	
High Open to Experience	613	265	2.313	

Table 26: IRLbl before Applying SMOTE on one splitted training set for one fold generated within the 10-Fold Cross-Validation for Each Single-label

Table 27 shows IRLbl after applying SMOTE and adding 100, 200, and 300 synthetic samples with 5-nearest neighbour to the the splitted training set. As it can be seen from the table, after adding 300 samples, all single-labels are changed to small imbalance level. From these results it can be drawn that SMOTE produces an improvement on imbalance level. For this work, SMOTE with 300 synthetic samples added to the training set was applied since it achieved the least imbalance level than others. Nevertheless, the change in imbalance level will not necessarily imply better classification results. The crucial factor for obtaining better predictions will be how these new instances change the model built by the classifier. Therefore, the following section will compare the classification results produced before and after applying SMOTE with adding 300 synthetic samples on the training set to one of the two classifiers which is MLCNN.

7.2 The Results Before and After SMOTE-MLCNN

This section analyzes the results produced by one of the two classifiers i.e. MLCNN before and after SMOTE is applied. Therefore, in this case there are only two sets of results, one produced by MLCNN from the HWBFF dataset without resampling and another one obtained from the same classifier using the same dataset after being processed by SMOTE. Table 28 shows the Predictive Accuracy and AUC for the single-labels before and after applying SMOTE, while Table 29 presents the average of the two measures for the big five factors before and after SMOTE is applied. However, Table 30 shows the overall performance for the MLCNN before and after SMOTE. It can be observed from the highlighted values in the three tables that the two performance measures after applying SMOTE obtained higher values in the most cases in each table than the base results. See Figures 27, 28, 29, 30, and 31 for comparing the AUC for the BFF before and after applying SMOTE on MLCNN.

	After SMOTE											
	(100 Synthetic Samples)				(200 Synthetic Samples)				(300 Synthetic Samples)			
Single-label	Negative Instances	Positive Instances	IRLbl	#Samples	Negative Instances	Positive Instances	IRLbl	#Samples	Negative Instances	Positive Instances	IRLbl	#Samples
Low Extraversion	753	225	3.35	978	753	325	2.32	1078	753	425	1.77	1178
Medium Extraversion	522	456	1.14		522	556	1.07		522	656	1.26	
High Extraversion	481	497	1.03		481	597	1.24		481	697	1.45	
Low Conscientiousness	853	125	6.82		853	225	3.79		853	325	2.62	
Medium Conscientiousness	560	418	1.34		560	518	1.08		560	618	1.10	
High Conscientiousness	443	535	1.21		543	535	1.01		643	535	1.20	
Low Emotional Stability	871	107	8.14		871	207	4.21		871	307	2.84	
Medium Emotional Stability	435	543	1.25		535	543	1.01		635	543	1.17	
High Emotional Stability	550	428	1.29		550	528	1.04		550	628	1.14	
Low Agreeableness	791	187	4.23		791	287	2.76		791	387	2.04	
Medium Agreeableness	229	749	3.27		329	749	2.28		429	749	1.75	
High Agreeableness	836	142	5.89		836	242	3.45		836	342	2.44	
Low Open to Experience	844	134	6.30		844	234	3.61		844	334	2.53	
Medium Open to Experience	399	579	1.45		499	579	1.16		599	579	1.03	
High Open to Experience	613	365	1.68		613	465	1.32		613	565	1.08	

Table 27: IRLbl after Applying SMOTE on one splitted training set for one fold generated within the 10-Fold Cross-Validation for Each Single-label in MLCNN

	Low EXTRA		Medium EXTRA		High EXTRA		Low CONS		Medium CONS		High CONS		Low EMOS		Medium EMOS		High EMOS		Low AGREE		Medium AGREE		High AGREE		Low OPEN		Medium OPEN		High OPEN	
Measure	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**
Predictive Accuracy	**0.76**	0.70	**0.60**	0.50	**0.84**	0.78	**0.90**	**0.90**	**0.53**	**0.53**	0.59	**0.62**	**0.99**	**0.99**	0.62	**0.66**	0.61	**0.66**	0.96	**0.99**	0.91	**0.93**	0.94	**0.96**	0.90	**0.91**	**0.82**	0.77	**0.93**	0.88
AUC	**0.86**	0.79	**0.61**	0.56	**0.93**	0.83	0.62	**0.93**	0.53	**0.54**	0.61	**0.74**	0.67	**0.73**	0.66	**0.80**	0.65	**0.80**	0.25	**0.91**	0.39	**0.77**	0.55	**0.94**	**0.84**	**0.84**	**0.69**	0.56	**0.96**	0.85

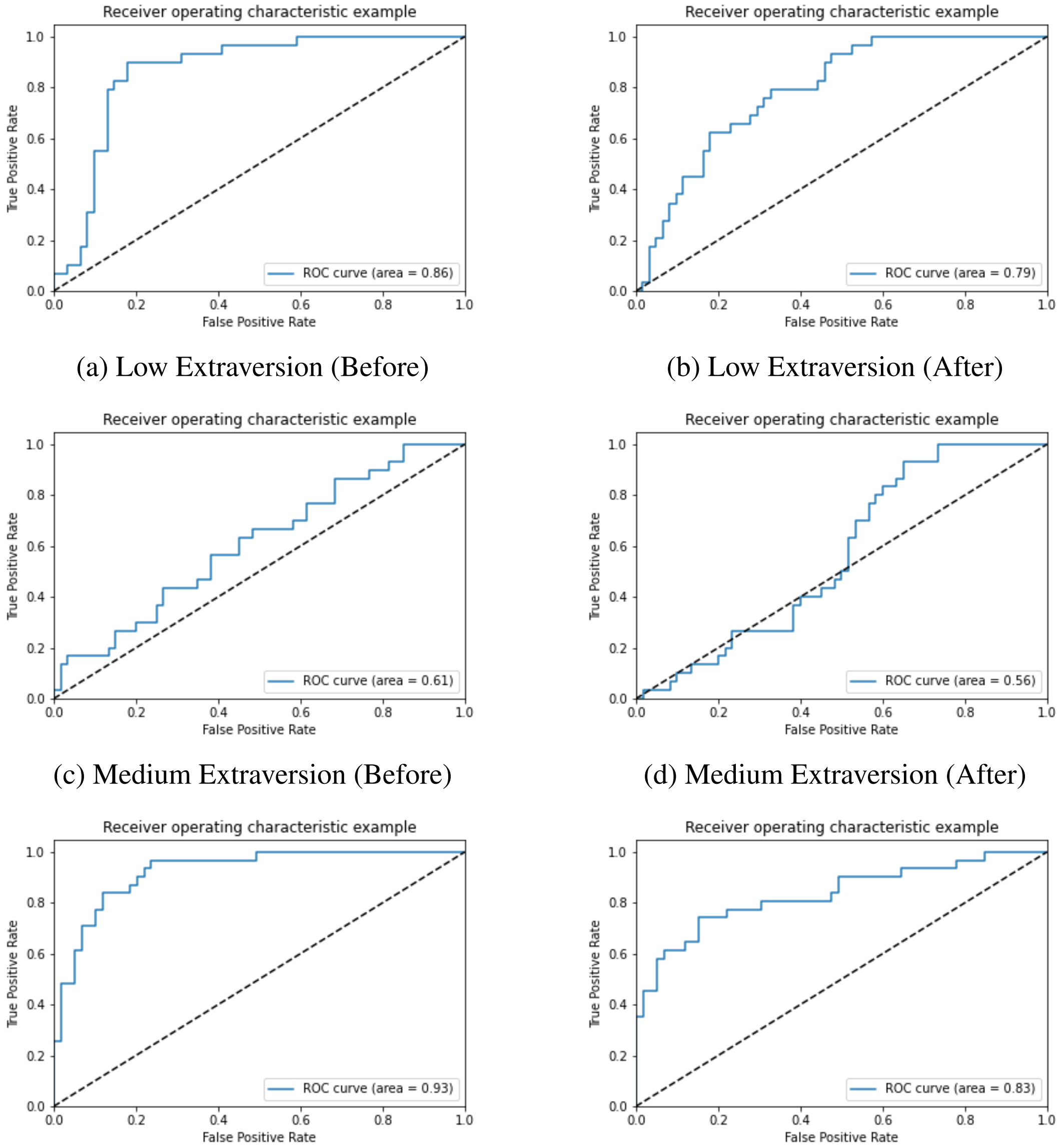
Table 28: Predictive Accuracy and AUC for the single-labels before and after Applying SMOTE on MLCNN

	Extraversion		Conscientiousness		Emotional Stability		Agreeableness		Open to Experience	
Measure	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Predictive Accuracy	**0.73**	0.66	0.67	**0.68**	0.74	**0.77**	0.94	**0.96**	**0.88**	0.85
AUC	**0.80**	0.73	0.59	**0.74**	0.66	**0.78**	0.40	**0.87**	**0.83**	0.75

Table 29: The Average of Predictive Accuracy and AUC for the Big Five Factors Before and after Applying SMOTE on MLCNN

	Predictive Accuracy	AUC
Before	**0.79**	0.65
After	**0.79**	**0.78**

Table 30: The Overall Performance of MLCNN Before and after Applying SMOTE
Best Values are Highlighted in Bold



(a) Low Extraversion (Before)

(b) Low Extraversion (After)

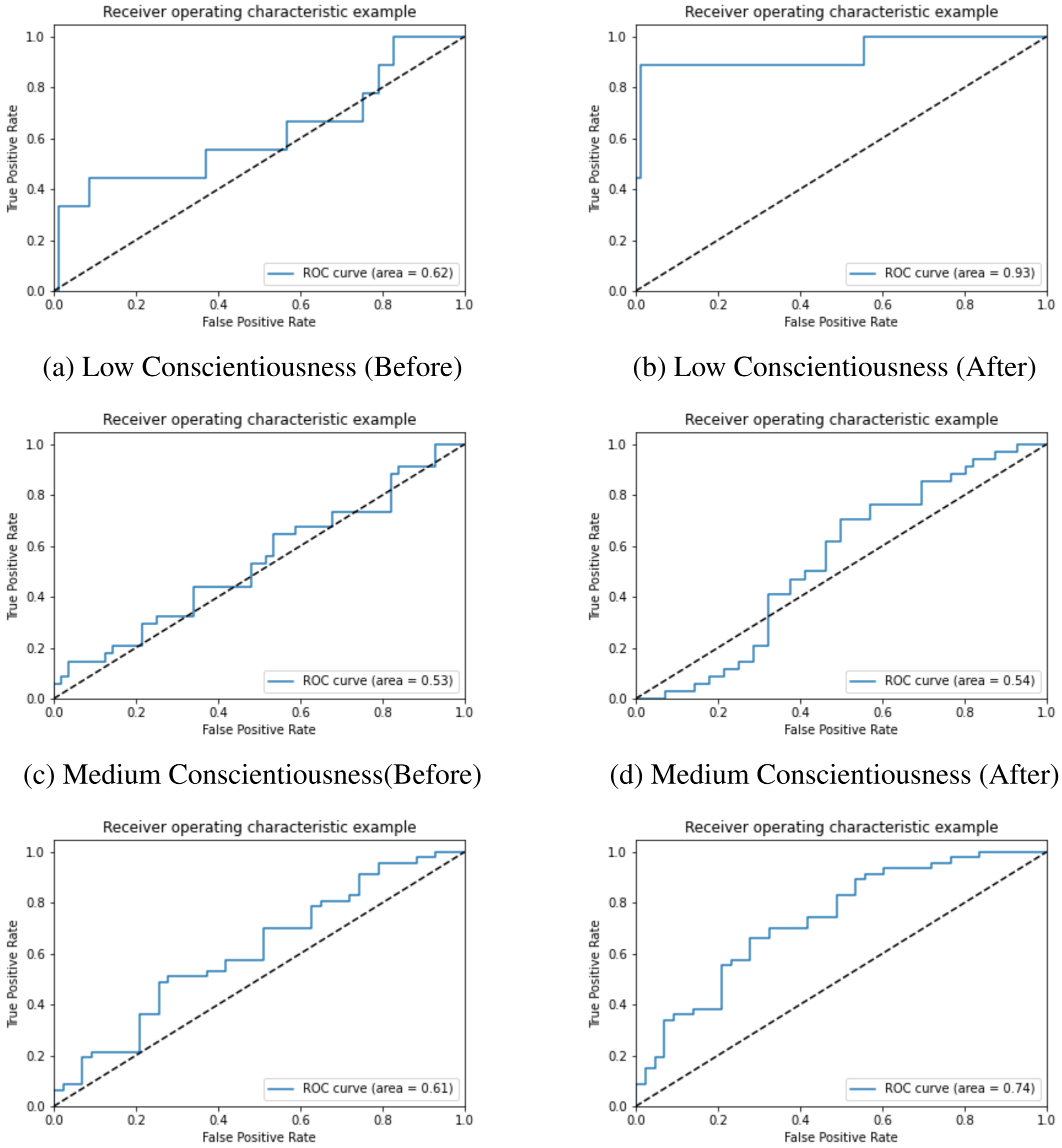
(c) Medium Extraversion (Before)

(d) Medium Extraversion (After)

(e) High Extraversion (Before)

(f) High Extraversion (After)

Figure 27: Area Under the Curve (AUC) for Extraversion Before and After Applying SMOTE on MLCNN



(a) Low Conscientiousness (Before)

(b) Low Conscientiousness (After)

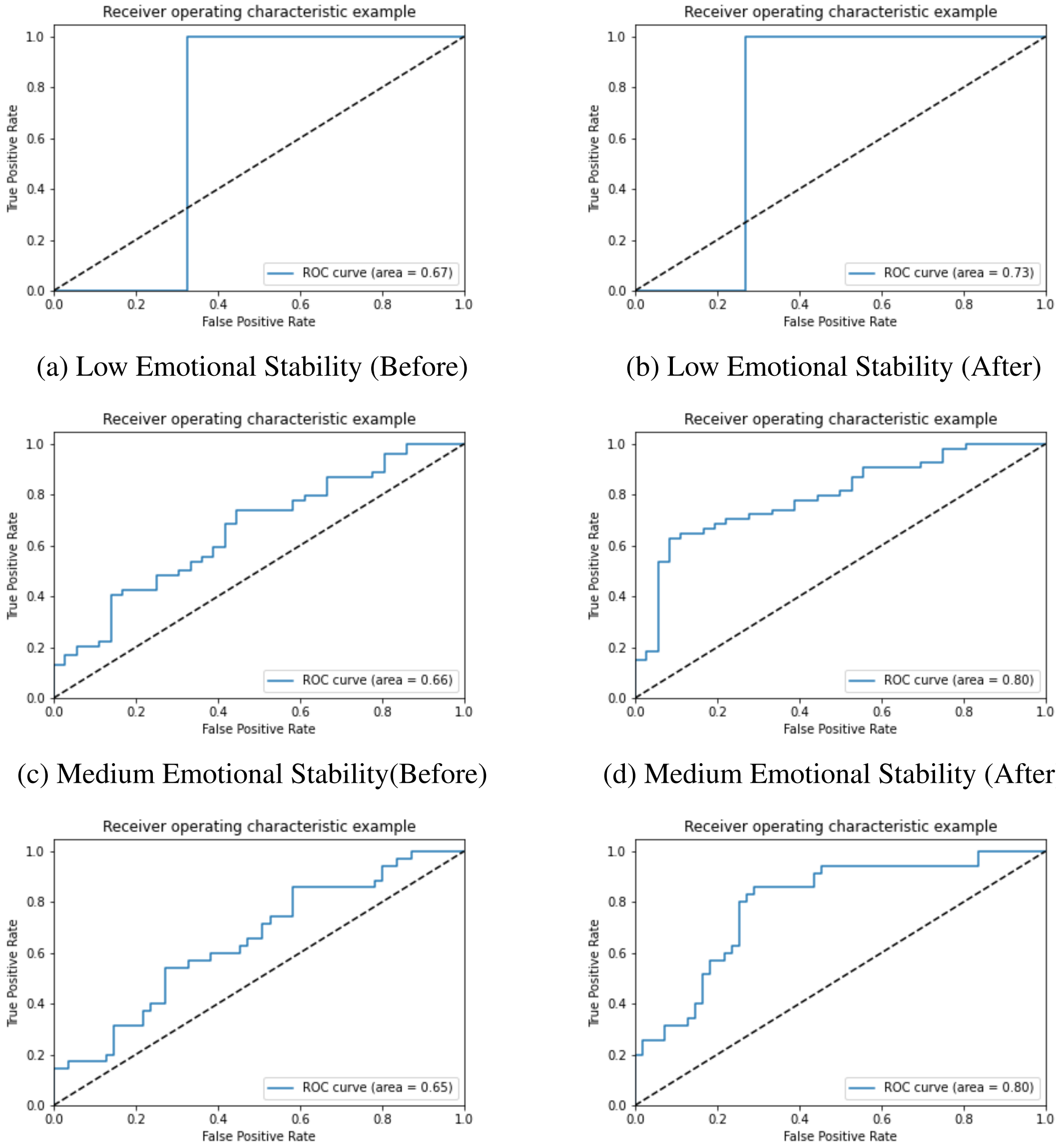
(c) Medium Conscientiousness(Before)

(d) Medium Conscientiousness (After)

(e) High Conscientiousness (Before)

(f) High Conscientiousness (After)

Figure 28: Area Under the Curve (AUC) for Conscientiousness Before and After Applying SMOTE on MLCNN



(a) Low Emotional Stability (Before)

(b) Low Emotional Stability (After)

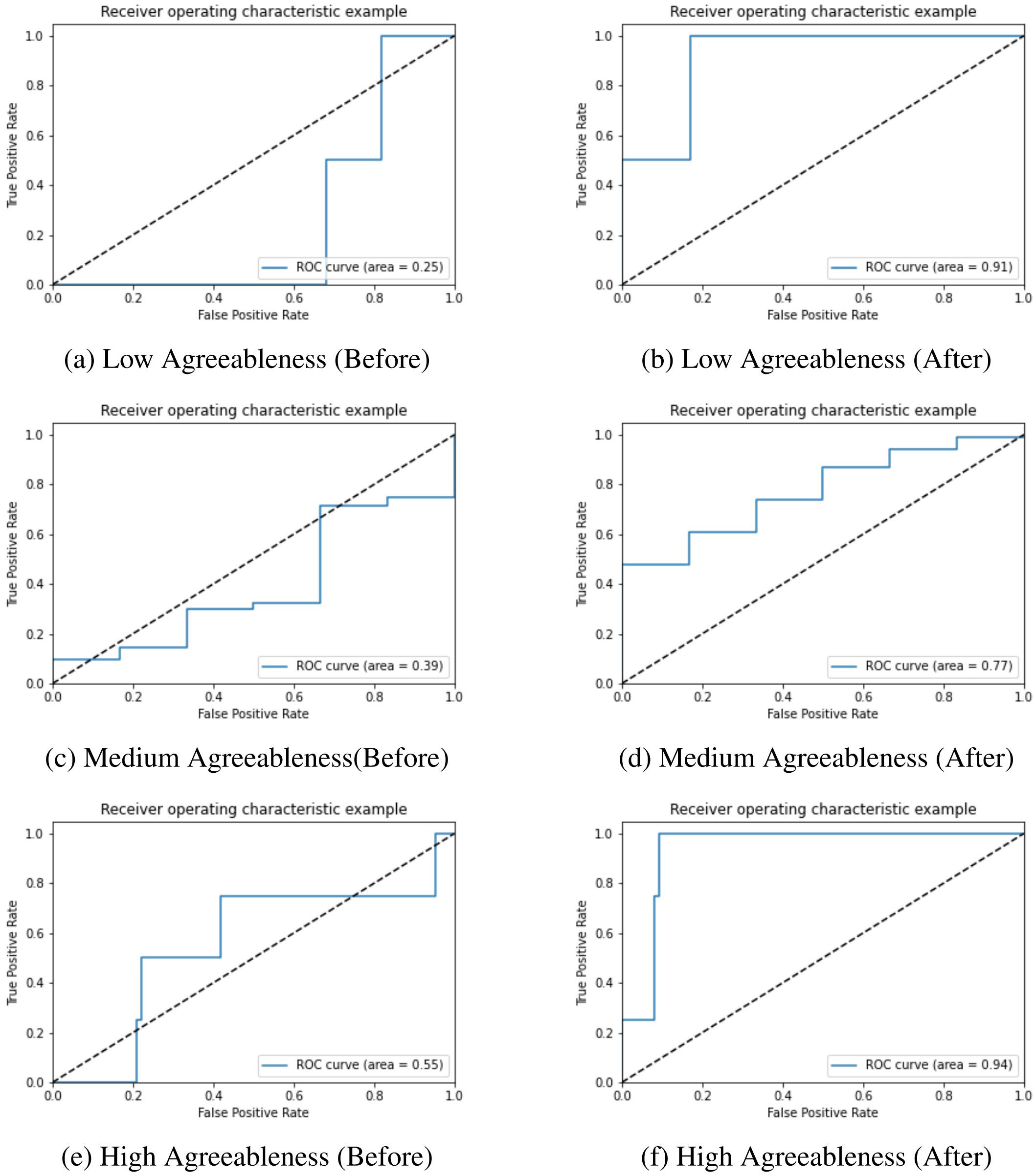
(c) Medium Emotional Stability(Before)

(d) Medium Emotional Stability (After)

(e) High Emotional Stability (Before)

(f) High Emotional Stability (After)

Figure 29: Area Under the Curve (AUC) for Emotional Stability Before and After Applying SMOTE on MLCNN



(a) Low Agreeableness (Before)

(b) Low Agreeableness (After)

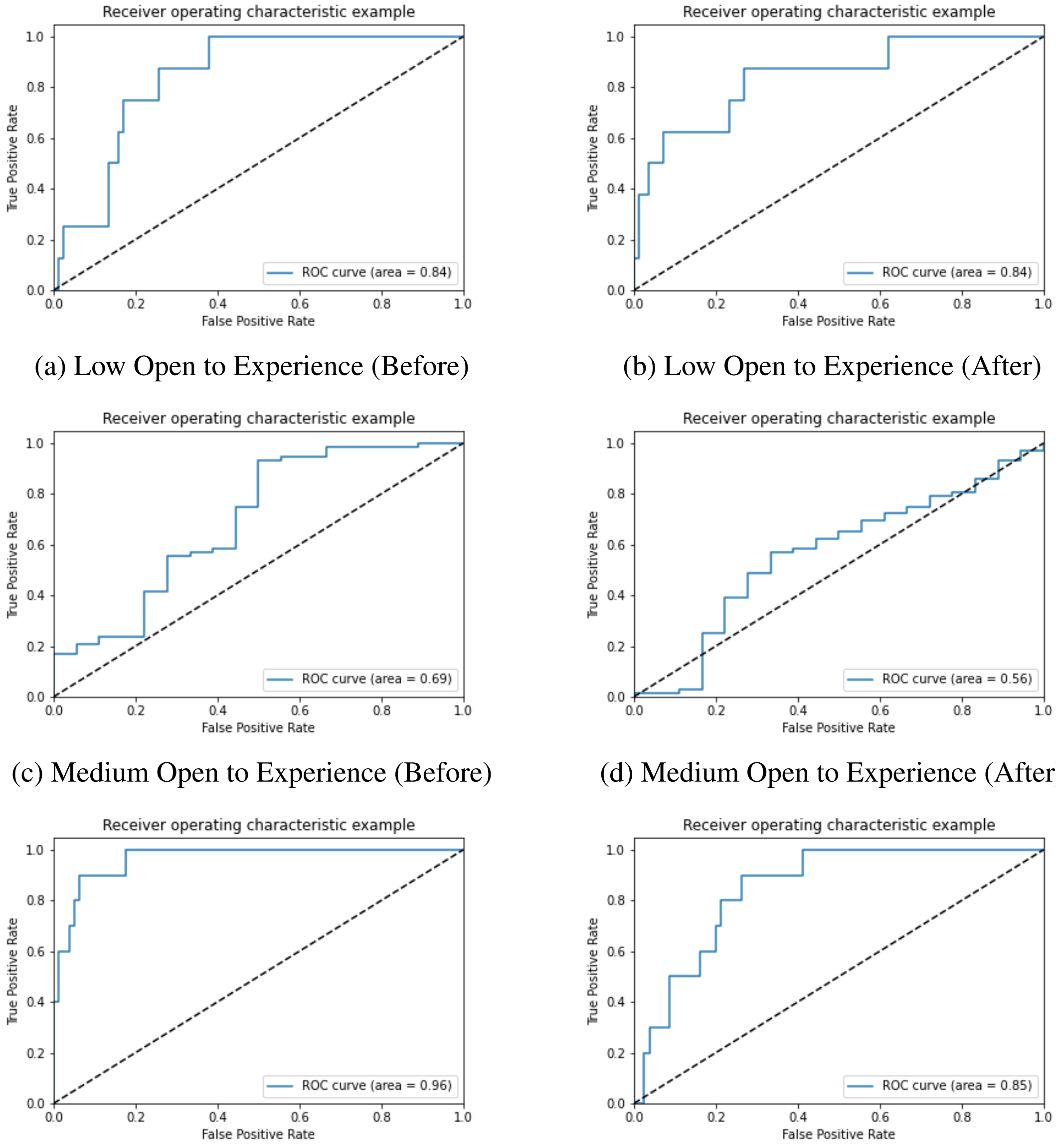
(c) Medium Agreeableness(Before)

(d) Medium Agreeableness (After)

(e) High Agreeableness (Before)

(f) High Agreeableness (After)

Figure 30: Area Under the Curve (AUC) for Agreeableness Before and After Applying SMOTE on MLCNN



(a) Low Open to Experience (Before)

(b) Low Open to Experience (After)

(c) Medium Open to Experience (Before)

(d) Medium Open to Experience (After)

(e) High Open to Experience (Before)

(f) High Open to Experience (After)

Figure 31: Area Under the Curve (AUC) for Open to Experience Before and After Applying SMOTE on MLCNN

7.3 The Results After Ensembling

Although the performance measures have been improved after applying SMOTE on ML-CNN as it is shown in Tables 28, 29, and 30, there is a need for more improvement in the BFF classification results in order to get a reliable result for the Spearman's rank correlation coefficient ρ done for investigating the validity of handwriting analysis. For this, averaging ensemble method is applied combining the two classifiers SMOTE-MLSVM and SMOTE-MLCNN generating our proposed strong classifier AvgMLSC. The following sub section compares between the results of single classifiers (i.e. MLSVM and MLCNN), and ensemble classifier (i.e. AvgMLSC) in order to show the influence of applying the ensemble method.

7.3.1 SMOTE-MLSVM vs SMOTE-MLCNN vs AvgMLSC Results

For demonstrating the performance of our proposed method, the performance of the ensemble1 classifier (AvgMlSC) with the two base classifiers (SMOTE-MLSVM and SMOTE-MLCNN) predicting the measurement level of the big five factors for 90 unseen handwriting samples are compared. Tables 31, 32, and 33 report the predicted results of accuracy, AUC, and F-Score for ensemble1 classifier and the two based classifiers for each single-label calculated based on the confusion matrix. The results of AvgMlSC for the 15 single-labels (i.e. Low Extraversion, Medium Extraversion, High Extraversion, Low Conscientiousness, Medium Conscientiousness, High Conscientiousness, Low Emotional Stability, Medium Emotional Stability, High Emotional Stability, Low Agreeableness, Medium Agreeableness, High Agreeableness, Low Open to Experience, Medium Open to Experience, and High Open to Experience) are reported respectively to achieve (92%, 71%, 90%, 100%, 83%, 92%, 100%, 92%, 96%, 100%, 99%, 100%, 100%, 88%, 100%) Accuracy, (0.95, 0.67, 0.97, 1.00, 0.83, 0.93, 1.00, 0.95, 0.94, 1.00, 0.95, 1.00, 1.00, 0.86, and 1.00) AUC, and (87%, 32%, 85%, 100%, 77%, 93%, 100%, 93%, 94%, 100%, 99%, 100%, 100%, 92%, and 100%) F-Score. While the results of SMOTE-MLSVM and SMOTE-MLCNN are reported to achieve lower values than the ensemble method. Figures 33, 34, 35, 36, and 37 show the ROC Curve for each single-label obtained by the three classifiers. Tables 34, 35, and 36 report the average of predicted results of accuracy, AUC, and F-Score for AvgMlSC and the two based classifiers for each factor. The results of AvgMlSC for the five factors (i.e. Extraversion, Conscientiousness, Emotional Stability, Agreeableness, Open to Experience) are reported respectively to achieve (84%, 92%, 96%, 99%, and 96%) Accuracy, (0.86, 0.92, 0.97, 0.98, and 0.95) AUC, and (68%, 90%, 97%, 99%, and 97%) F-Score. Whereas the results of the two classifiers separately obtained lower values than the ensemble method for the three performance measures. The overall average of the results of the three performance measures obtained by the three classifiers are shown in Table 37 and

Figure 32. As it can be seen from the table and the Figure, AvgMlSC achieved 93% Accuracy, 0.94 AUC, and 90% F-Score. Whilst SMOTE-MLSVM obtained 89% Accuracy, 0.77 AUC, and 72% F-Score, and SMOTE-MLCNN reported 79% Accuracy, 0.78 AUC, and 40% F-Score. Based on the obtained values of the results, we can draw the following conclusion about the performance of the AvgMlSC. The ensemble method produces more successful results than single classifiers. As it can be seen from Table 37 and Figure 32 that the performance of ensemble method are better than any other based classifier, especially for the F-Score value which increased remarkably. The comparison results confirm that averaging ensemble method with applying SMOTE technique on MLSVM and MLCNN can effectively deal with imbalanced data and obviously improve prediction performance.

Classifier	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
	EXTRA	EXTRA	EXTRA	CONS	CONS	CONS	EMOS	EMOS	EMOS	AGREE	AGREE	AGREE	OPEN	OPEN	OPEN
SMOTE-MLSVM	0.83	0.58	0.88	0.94	0.72	0.87	1.00	0.94	0.98	0.98	0.92	0.97	0.98	0.86	0.98
SMOTE-MLCNN	0.70	0.50	0.78	0.90	0.53	0.62	0.99	0.66	0.66	0.99	0.93	0.96	0.91	0.77	0.88
AvgMLSC	0.92	0.71	0.90	1.00	0.83	0.92	1.00	0.92	0.96	1.00	0.99	1.00	1.00	0.88	1.00

Table 31: The Predictive Accuracy for the Single-Labels Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

Classifier	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
	EXTRA	EXTRA	EXTRA	CONS	CONS	CONS	EMOS	EMOS	EMOS	AGREE	AGREE	AGREE	OPEN	OPEN	OPEN
SMOTE-MLSVM	0.83	0.45	0.86	0.82	0.68	0.86	1.00	0.94	0.98	0.50	0.49	0.62	0.93	0.70	0.94
SMOTE-MLCNN	0.79	0.56	0.83	0.93	0.54	0.74	0.73	0.80	0.80	0.91	0.77	0.94	0.84	0.56	0.85
AvgMLSC	0.95	0.67	0.97	1.00	0.83	0.93	1.00	0.95	0.94	1.00	0.95	1.00	1.00	0.86	1.00

Table 32: AUC for the Single-Labels Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

Classifier	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
	EXTRA	EXTRA	EXTRA	CONS	CONS	CONS	EMOS	EMOS	EMOS	AGREE	AGREE	AGREE	OPEN	OPEN	OPEN
SMOTE-MLSVM	0.75	0.10	0.82	0.71	0.58	0.89	1.00	0.95	0.97	0	0.96	0.40	0.88	0.91	0.90
SMOTE-MLCNN	0.23	0.35	0.67	0	0.09	0.73	0	0.77	0.41	0.67	0.96	0	0	0.86	0.27
AvgMLSC	0.87	0.32	0.85	1.00	0.77	0.93	1.00	0.93	0.94	1.00	0.99	1.00	1.00	0.92	1.00

Table 33: F-Score for the Single-Labels Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

Classifier	Extraversion	Conscientiousness	Emotional Stability	Agreeableness	Open to Experience
SMOTE-MLSVM	0.76	0.84	0.97	0.96	0.94
SMOTE-MLCNN	0.66	0.68	0.77	0.96	0.85
AvgMLSC	0.84	0.92	0.96	0.99	0.96

Table 34: The Average of Predictive Accuracy for the Big Five Factors Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

Classifier	Extraversion	Conscientiousness	Emotional Stability	Agreeableness	Open to Experience
SMOTE-MLSVM	0.71	0.79	0.97	0.54	0.86
SMOTE-MLCNN	0.73	0.74	0.78	0.87	0.75
AvgMLSC	0.86	0.92	0.97	0.98	0.95

Table 35: The Average of AUC for the Big Five Factors Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

Classifier	Extraversion	Conscientiousness	Emotional Stability	Agreeableness	Open to Experience
SMOTE-MLSVM	0.56	0.73	0.97	0.45	0.90
SMOTE-MLCNN	0.42	0.27	0.39	0.54	0.38
AvgMLSC	0.68	0.90	0.97	0.99	0.97

Table 36: The Average of F-Score for the Big Five Factors Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

Classifier	**Predictive Accuracy**	**AUC**	**F-Score**
SMOTE-MLSVM	0.89	0.77	0.72
SMOTE-MLCNN	0.79	0.78	0.40
AvgMLSC	0.93	0.94	0.90

Table 37: The Overall Performance of SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

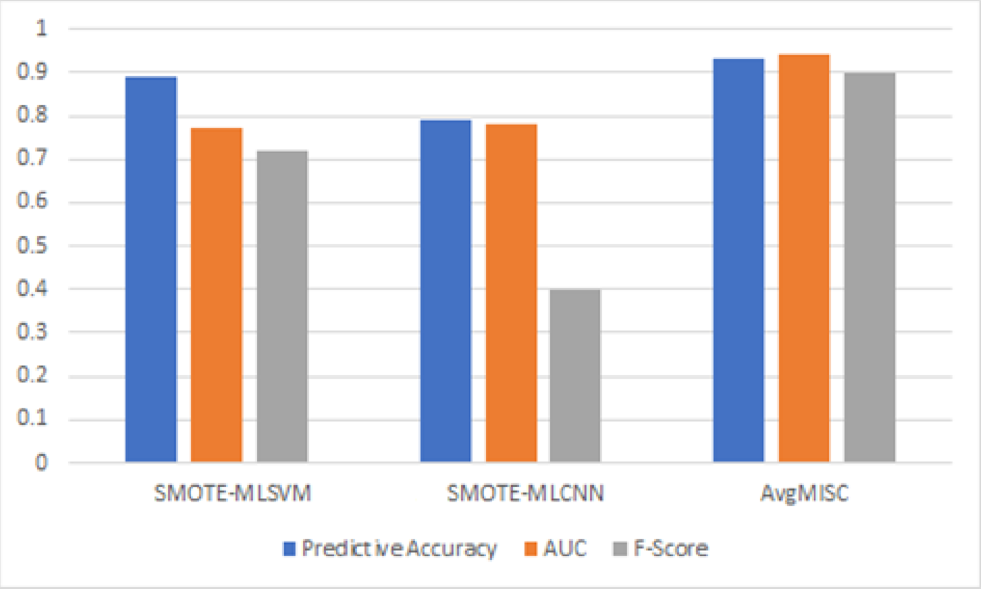
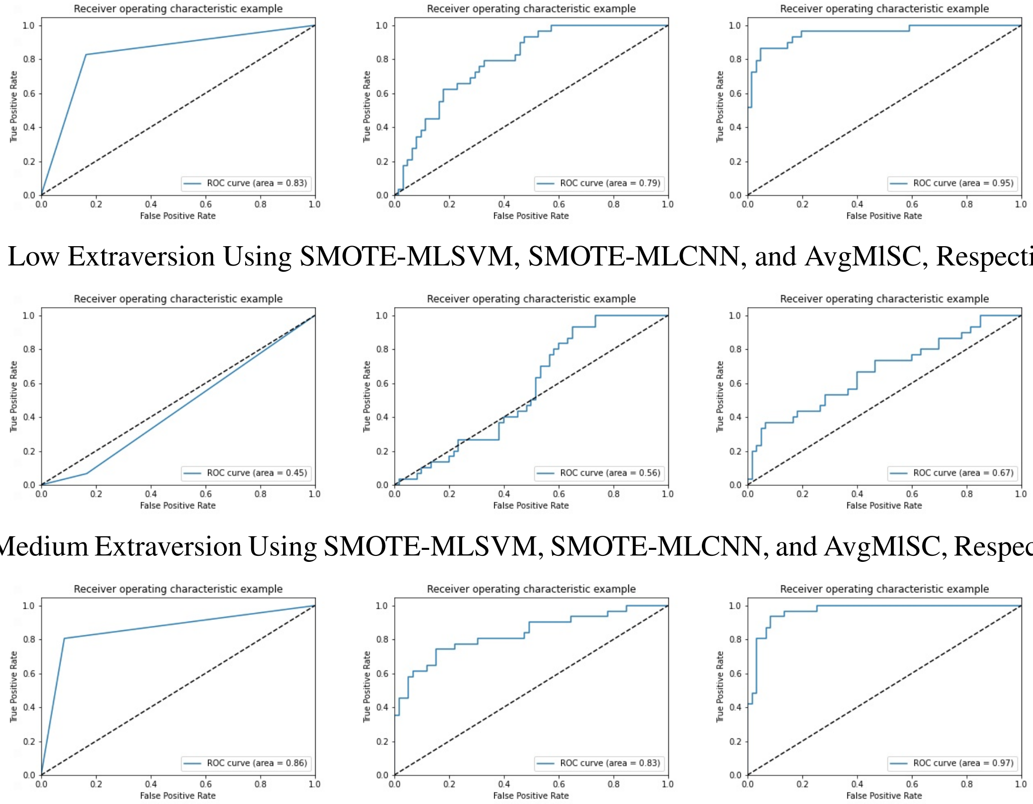


Figure 32: The Overall Average of Predictive Accuracy, AUC, and F-Score for SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC

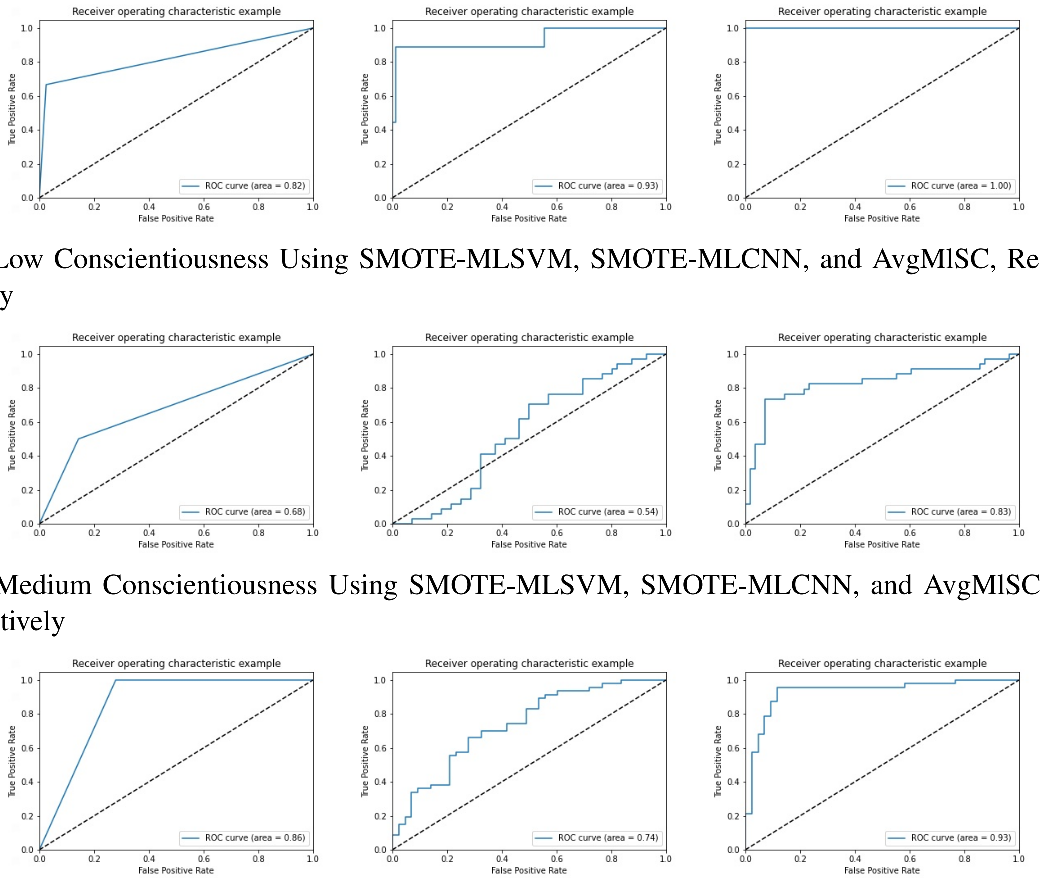


(a) Low Extraversion Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(b) Medium Extraversion Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(c) High Extraversion Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

Figure 33: Area Under the Curve (AUC) for Extraversion Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

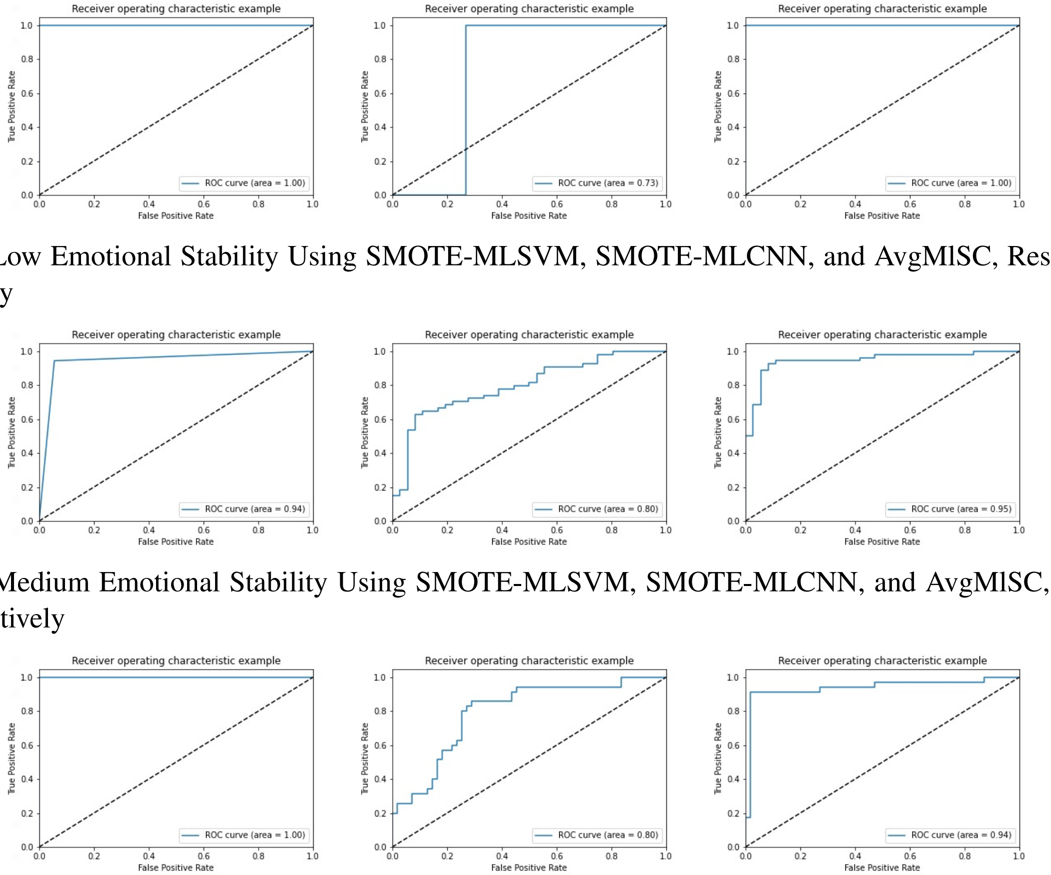


(a) Low Conscientiousness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(b) Medium Conscientiousness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(c) High Conscientiousness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

Figure 34: Area Under the Curve (AUC) for Conscientiousness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

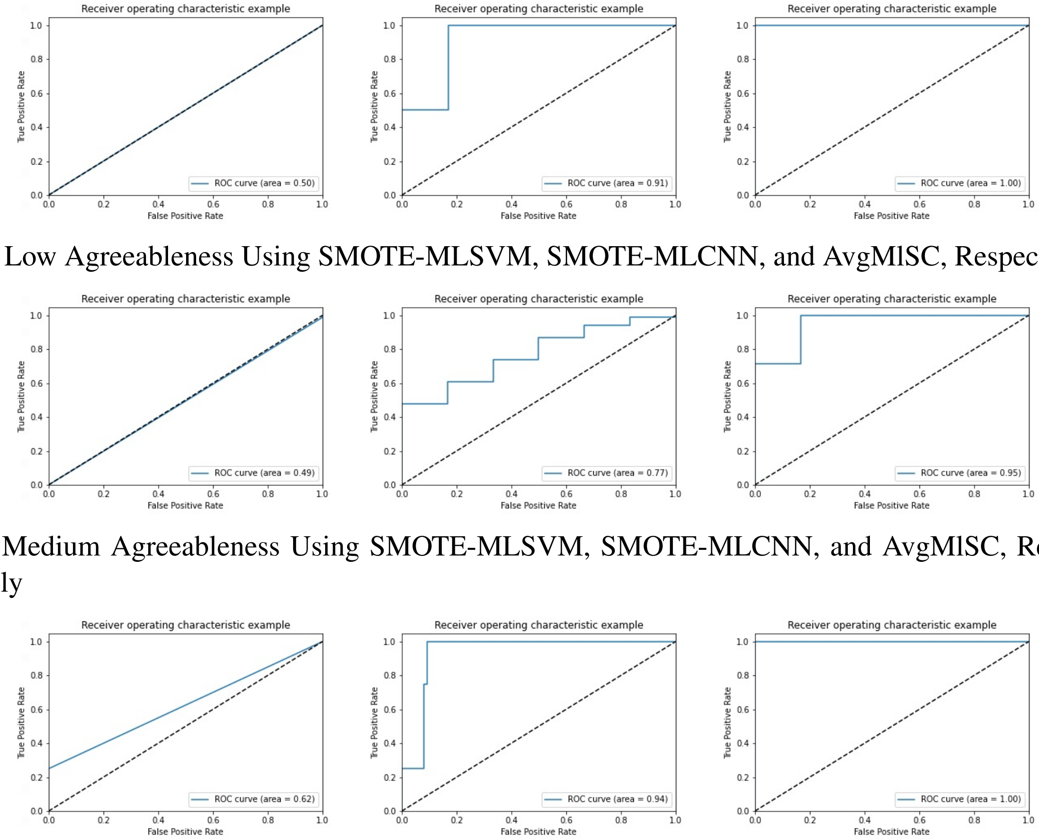


(a) Low Emotional Stability Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(b) Medium Emotional Stability Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(c) High Emotional Stability Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

Figure 35: Area Under the Curve (AUC) for Emotional Stability Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

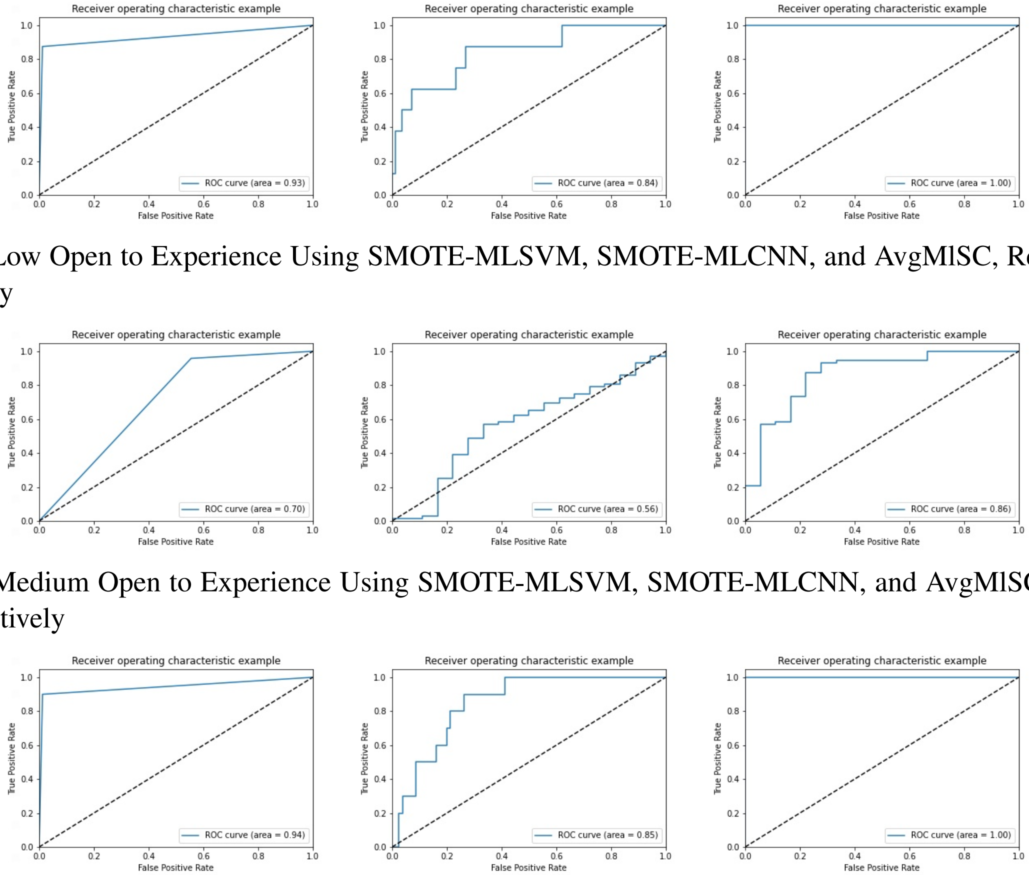


(a) Low Agreeableness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(b) Medium Agreeableness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(c) High Agreeableness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

Figure 36: Area Under the Curve (AUC) for Agreeableness Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively



(a) Low Open to Experience Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(b) Medium Open to Experience Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

(c) High Open to Experience Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

Figure 37: Area Under the Curve (AUC) for Open to Experience Using SMOTE-MLSVM, SMOTE-MLCNN, and AvgMlSC, Respectively

7.4 A Comparative Analysis with the Baseline Classifiers

To further establish the effectiveness of the proposed model, a comparative analysis with five popular baseline classifiers, i.e. Logistic Regression (LR), Naïve Bayes (NB), K-Neighbors (KN) Support Vector Machine (SVM), and Convolutional Neural Network (CNN) is presented in this section. For this, all classifiers have employed the same resampled HWBFF dataset for training and multi-label learning under "one-vs-all" scheme with the same experimental protocol (10-fold cross validation) was considered. Tables 38, 39, and 40 present the overall average of predictive accuracy, AUC, and F-Score, respectively, for the big five fators. For the predictive accuracy, the proposed model achieved the highest value for Extraversion, Conscientiousness, and Agreeableness. While for Emotional Stability LR obtained the highest value. However, for Open to Experience, LR and AvgMLSC were the best classifiers. For AUC, AvgMlSC obtained the highest numbers for Extraversion, Conscientiousness, Agreeableness, and Open to Experience. Whilst LR achieved the highest value for Emotional Stability. For the last measure which is F-Score, AvgMlSC produced the best result for Conscientiousness, Agreeableness, and Open to Experience. However, KN was the best classifier for Extraversion while LR was the best for Emotional Stability.

Classifier	EXTRA	CONS	EMOS	AGREE	OPEN
SMOTE-MLLR	0.79	0.86	**0.98**	0.97	**0.96**
SMOTE-MLNB	0.69	0.79	0.89	0.93	0.88
SMOTE-MLKN	0.82	0.88	0.95	0.97	0.95
SMOTE-MLSVM	0.76	0.84	0.97	0.96	0.94
SMOTE-MLCNN	0.66	0.68	0.77	0.96	0.85
AvgMLSC	**0.84**	**0.92**	0.96	**0.99**	**0.96**

Table 38: The Average of Predictive Accuracy for each Factor using the Five Baseline Classifiers and AvgMlSC

Classifier	EXTRA	CONS	EMOS	AGREE	OPEN
SMOTE-MLLR	0.74	0.82	**0.98**	0.83	0.91
SMOTE-MLNB	0.70	0.79	0.87	0.82	0.83
SMOTE-MLKN	0.80	0.80	0.79	0.64	0.82
SMOTE-MLSVM	0.71	0.79	0.97	0.54	0.86
SMOTE-MLCNN	0.73	0.74	0.78	0.87	0.75
AvgMLSC	**0.86**	**0.92**	0.97	**0.98**	**0.95**

Table 39: The Average of AUC for each Factor using the Five Baseline Classifiers and AvgMlSC

Classifier	EXTRA	CONS	EMOS	AGREE	OPEN
SMOTE-MLLR	0.63	0.79	**0.98**	0.98	0.96
SMOTE-MLNB	0.61	0.69	0.66	0.69	0.75
SMOTE-MLKN	**0.73**	0.75	0.62	0.55	0.77
SMOTE-MLSVM	0.56	0.73	0.97	0.45	0.9
SMOTE-MLCNN	0.42	0.27	0.39	0.54	0.38
AvgMLSC	0.68	**0.9**	0.97	**0.99**	**0.97**

Table 40: The Average of F-Score for each Factor using the Five Baseline Classifiers and AvgMlSC

Table 41 and Figure 38 compare the overall performance for the six classifiers in terms of the three measures. The table and the figure reveal that the overall performance of AvgMlSC which is our proposed ensemble learning is better than the individual learners with 93% predictive accuracy, 0.94 AUC, and 90% F-Score.

All experiments were conducted in Spyder (Python 3.8) programming environment with Anaconda Navigator. They are performed on TensorBook that is manufactured and configured by Lambda. It is a GPU laptop with RTX 2070, Intel i7-9750H Processor (6 Cores), 32 GB DDR4 Memory, and 1 TB SSD (NVMe) running on Linux. Max-Q. Ubuntu, TensorFlow, PyTorch, Keras, CUDA, and cuDNN are pre-installed.

Classifier	Predictive Accuracy	AUC	F-Score
SMOTE-MLLR	0.91	0.86	0.88
SMOTE-MLNB	0.84	0.80	0.68
SMOTE-MLKN	0.91	0.77	0.68
SMOTE-MLSVM	0.89	0.77	0.72
SMOTE-MLCNN	0.79	0.78	0.40
AvgMLSC	**0.93**	**0.94**	**0.90**

Table 41: The Overall performance for the Five Baseline Classifiers and AvgMlSC

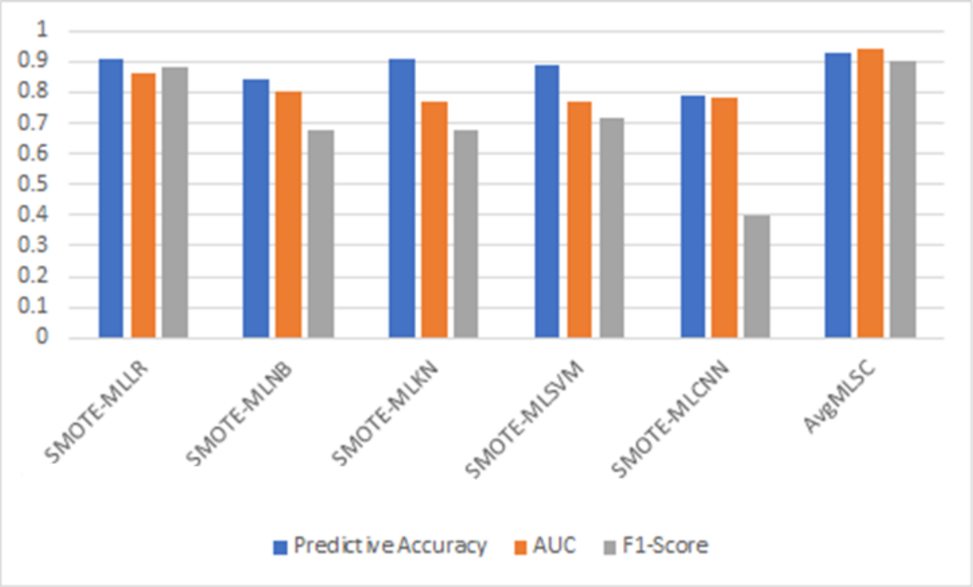


Figure 38: The Overall Average of Predictive Accuracy, AUC, and F-Score for the Five Baseline Classifiers and AvgMlSC

7.5 A Comparative Analysis with the State-of-the-Art

The results of three early computerized BFF model from the state-of-the-art have been chosen to be compared with the results of our proposed model. We choose these two studies because they have used the same form of data used in our model and the same performance measures to evaluate their proposed models.

The first study is published in 2018 [20], the authors proposed the non-invasive three-layer architecture based on neural networks that aims to determine the Big Five personality traits of an individual by analyzing off-line handwriting. They used their own database that links the Big Five personality type with the handwriting features containing both predefined and random text. The main handwriting features used are the following: baseline, word slant, writing pressure, connecting strokes, space between lines, lowercase letter ’t’, and lowercase letter ’f’. They measured the model performance by calculating the predictive accuracy, see Table 42.

	Predictive Accuracy				
Automated BFF Model	**EXTRA**	**CONS**	**EMOS**	**AGREE**	**OPEN**
Gavrilescu & Vizireanu (2018)	84.00	77.00	84.00	77.00	84.00
AvgMlSC	**84.00**	**92.00**	**96.00**	**99.00**	**96.00**

Table 42: A Comparison between AvgMlSC and Gavrilescu & Vizireanu (2018)

The second work is [31] that presented a method to extract personality traits from stream of-consciousness essays using a convolutional neural network (CNN). They trained five different networks, all with the same architecture, for the five personality traits. Each network was a binary classifier that predicted the corresponding trait to be positive or negative. They used James Pennebaker and Laura King's stream-of-consciousness essay dataset. It contains 2,468 anonymous essays tagged with the authors' personality traits based on the Big Five factors. They evaluated the model performance by measuring the predictive accuracy, see Table 43.

	Predictive Accuracy				
Automated BFF Model	**EXTRA**	**CONS**	**EMOS**	**AGREE**	**OPEN**
Majumder et al. (2017)	58.09	57.30	59.38	56.71	62.68
AvgMlSC	**84.00**	**92.00**	**96.00**	**99.00**	**96.00**

Table 43: A Comparison between AvgMlSC and Majumder et al. (2017)

The third work is [43] that combined automatic personality detection (APD) and data-driven personas (DDPs) to design personas with personality traits that could be automatically generated using numerical and textual social media data. They developed a neural network with two major sub-architectures: a single dimensional convolutional neural network since there is a spatial structure in the input text, and a long short-term memory network since there is also a temporal correlation between the words in the input text. They used the F-score for evaluating their model, F-score obtained for each BF trait using the same dataset used in the first work, see Table 44.

	F-Score				
Automated BFF Model	**EXTRA**	**CONS**	**EMOS**	**AGREE**	**OPEN**
Salminen et al. (2020)	0.54	0.53	0.48	0.55	0.52
AvgMlSC	**0.68**	**0.90**	**0.97**	**0.99**	**0.97**

Table 44: A Comparison between AvgMlSC and Salminen et al. (2020)

As it can be observed from Tables 42, 43, and 44 that our proposed model achieved a remarkable improvement in accuracy and F-Score than the two models.

7.6 Validation Coefficients for the Big Five Factors using Handwriting Analysis

7.6.1 Spearman's Correlation Between the Results of the Computerized Handwriting Analysis and the Scores of the BFF Test using Python

In order to assess the validity of the BFF evaluation assigned by the graphologist, Spearman's rho (ρ) correlation coefficients are conducted on the testing set after getting a high performance evaluation for AvgMlSC. The testing set consists of 90 handwriting samples written in five languages i.e. English, French, Chinese, Arabic, and Spanish. The Spearman's correlation coefficients calculated in Python using the spearmanr() SciPy function between the handwriting analysis results predicted by AvgMlSC and the scores of the Big Five Factor Markers Test.

Before conducting the correlation test, the BFF probabilities predicted by AvgMlSC of each class are converted firstly into binary labels using a threshold of (≥ 0.5). If the probability is ≥ 0.5, it is converted to 1 otherwise it is 0, see Tables 45 and 46.

Low EXTRA	**Medium EXTRA**	**High EXTRA**	**Low CONS**	**Medium CONS**	**High CONS**
0.06	0.89	0.05	0.76	0.04	0.20
0.65	0.30	0.05	0.05	0.89	0.05
0.01	0.04	0.95	0.90	0.07	0.03

Table 45: Examples of the Predicted Probabilities for Extraversion and Conscientiousness

Low EXTRA	**Medium EXTRA**	**High EXTRA**	**Low CONS**	**Medium CONS**	**High CONS**
0	1	0	1	0	0
1	0	0	0	1	0
0	0	1	1	0	0

Table 46: Binary Labels for Extraversion and Conscientiousness

After that, for each factor, the binary labels under each class were converted into scale of 3. Those ones under Low classes were converted to 1, ones under Medium classes converted to 2, and ones under High classes converted to 3, see Table 47.

EXTRA	**CONS**
2	1
1	2
3	1

Table 47: The Predicted Analysis for Extraversion and Conscientiousness after Converting them to the Scale of 3

Then, the values in Table 47 are correlated with the scores of the Big Five Factor Markers Test after converting them into scale of 3 using the normative classification shown in Table 8.

The statistical test reports there is a sufficient evidence to conclude that there is a statistically significant relationship between the score of the Big Five Factor Markers Test (BFFMT) and the graphologist's evaluation at the 0.01 level of significance for the big five factors. Therefore, the null hypothesis that states there is no correlation between the results of handwriting analysis and the scores of BFF test is rejected. Based on a standard interpretation table of Spearman's correlation coefficients [14], shown in Table 48, the strength of the correlation is varied among the five factors. For Extraversion, a weak positive relationship is found with $(\rho = 0.220)$. However, a moderate positive relationship is reported for Conscientiousness and Open to Experience with $(\rho = 0.340)$ and $(\rho = 0.356)$, respectively. On the other hand, a strong positive relationship is indicated for Agreeableness with $(\rho = 0.445)$. For the last factor which is Emotional Stability, a very weak positive relationship is found with $(\rho = 0.032)$.

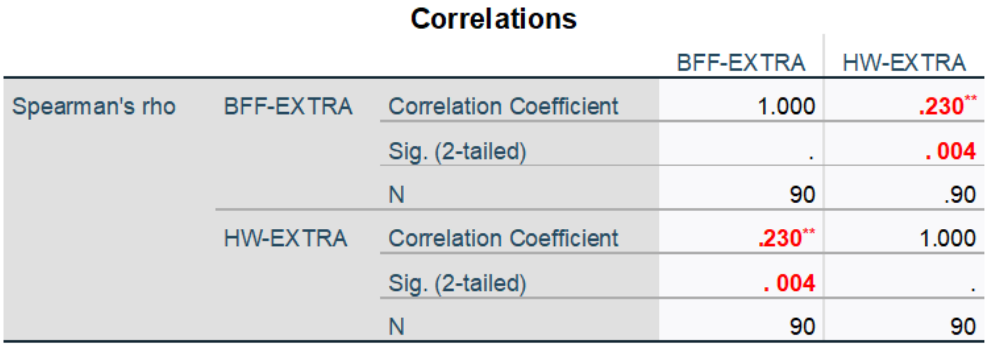
Spearman rho	**Correlation**
>= 0.70	Very strong relationship
0.40 - 0.69	Strong relationship
0.30 - 0.39	Moderate relationship
0.20 - 0.29	Weak relationship
0.01 - 0.19	Very weak relationship

Table 48: Interpretation Table of Spearman Correlation Coefficients (Adapted from Dancey and Reidy, 2004)

In order to make sure that the results of Spearman's correlation is not influenced by the small size of samples, the correlation coefficients were re-performed on a large size of samples in Python using the spearmanr() SciPy function. They were re-carried out on 156 handwriting samples written in the same five languages and some of them used for the model training. As a result, the re-conducted test reports there is a statistically significant relationship between the score of the Big Five Factor Markers Test (BFFMT) and the graphologist's evaluation for the big five factors with the same strength of correlation mentioned above for each factor.

7.6.2 Spearman's Correlation Between the Manual Handwriting Analysis Evaluation and the Scores of the BFF Test using SPSS

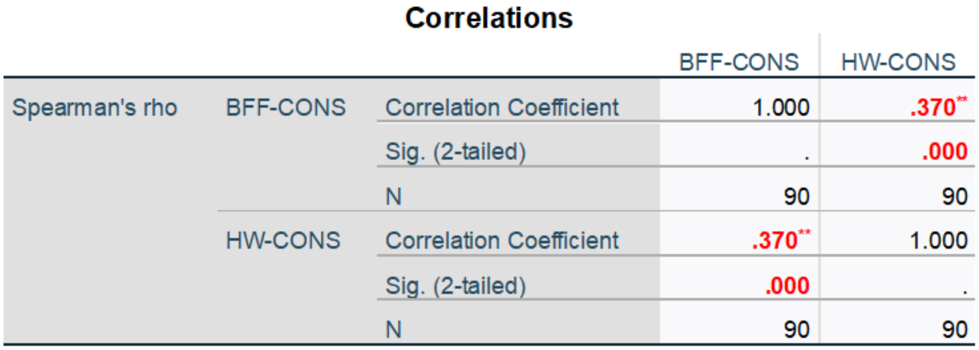
In order to evaluate the results of the validation coefficients produced based on the handwriting analysis results of AvgMlSC in Python, we carried out Spearman's rho (ρ) correlation coefficients between the scores of the Big Five Factor Markers Test (BFFMT) collected by the survey and the handwriting analysis scores calculated manually by the graphologist on the same 90 handwriting samples using SPSS. Tables 49, 50, 51,52, and 53 show the results of the statistical test for Extraversion, Conscientiousness, Agreeableness, Open to Experience, and Emotional Stability, respectively.



Correlations

			BFF-EXTRA	HW-EXTRA
Spearman's rho	BFF-EXTRA	Correlation Coefficient	1.000	.230**
		Sig. (2-tailed)	.	. 004
		N	90	.90
	HW-EXTRA	Correlation Coefficient	.230**	1.000
		Sig. (2-tailed)	. 004	.
		N	90	90

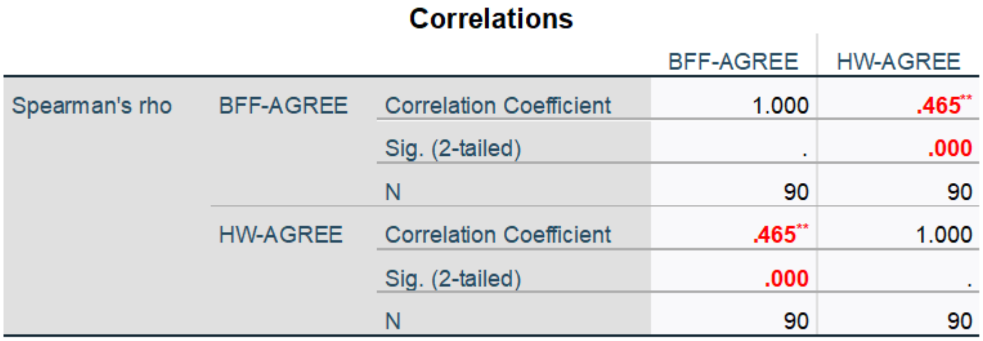
Table 49: Spearman's rho (ρ) Correlation Coefficients between BFFM Test and Handwriting Analysis for Extraversion



Correlations

			BFF-CONS	HW-CONS
Spearman's rho	BFF-CONS	Correlation Coefficient	1.000	.370**
		Sig. (2-tailed)	.	.000
		N	90	90
	HW-CONS	Correlation Coefficient	.370**	1.000
		Sig. (2-tailed)	.000	.
		N	90	90

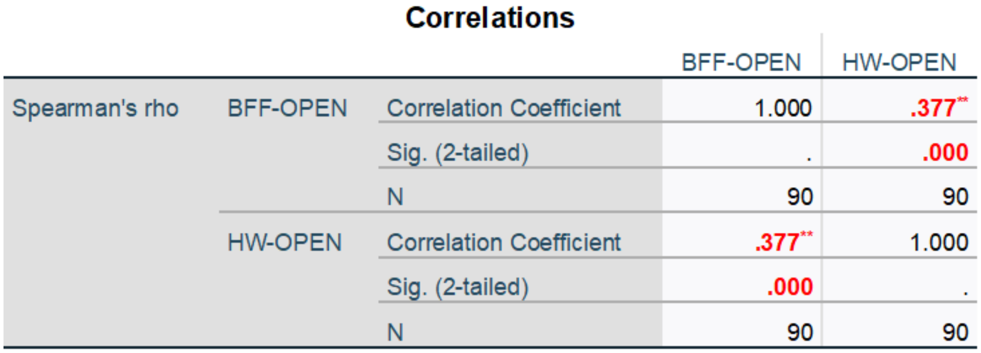
Table 50: Spearman's rho (ρ) Correlation Coefficients between BFFM Test and Handwriting Analysis for Conscientiousness



Correlations

			BFF-AGREE	HW-AGREE
Spearman's rho	BFF-AGREE	Correlation Coefficient	1.000	.465**
		Sig. (2-tailed)	.	.000
		N	90	90
	HW-AGREE	Correlation Coefficient	.465**	1.000
		Sig. (2-tailed)	.000	.
		N	90	90

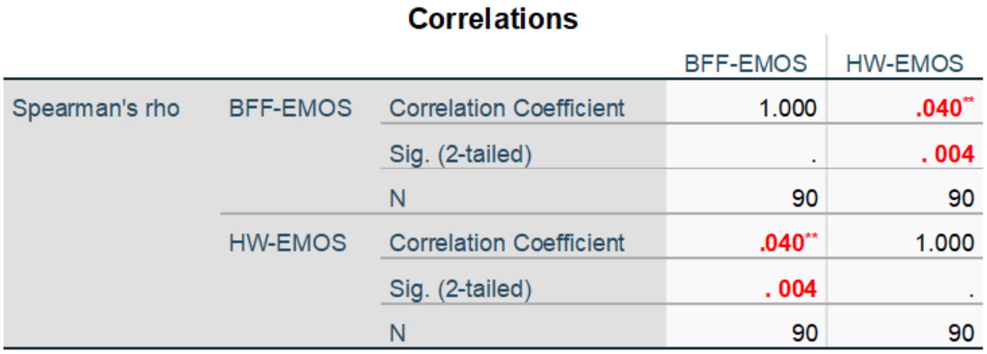
Table 51: Spearman's rho (ρ) Correlation Coefficients between BFFM Test and Handwriting Analysis for Agreeableness



Correlations

			BFF-OPEN	HW-OPEN
Spearman's rho	BFF-OPEN	Correlation Coefficient	1.000	.377**
		Sig. (2-tailed)	.	.000
		N	90	90
	HW-OPEN	Correlation Coefficient	.377**	1.000
		Sig. (2-tailed)	.000	.
		N	90	90

Table 52: Spearman's rho (ρ) Correlation Coefficients between BFFM Test and Handwriting Analysis for Open to Experience



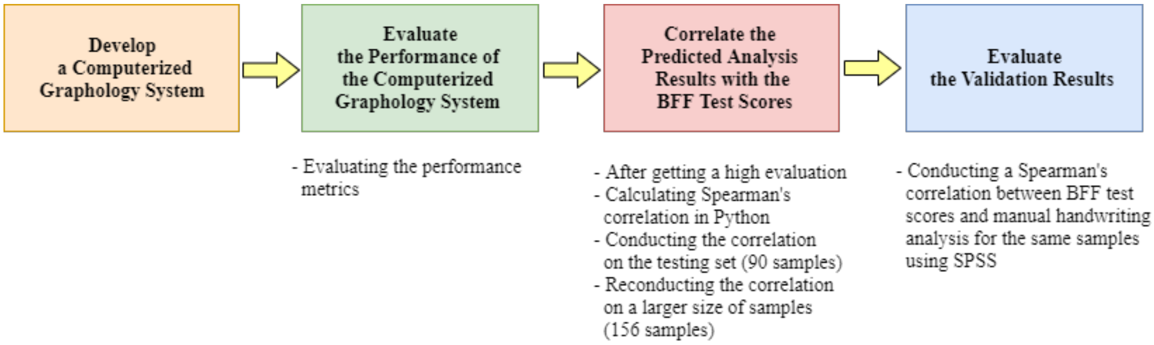
Correlations

			BFF-EMOS	HW-EMOS
Spearman's rho	BFF-EMOS	Correlation Coefficient	1.000	.040**
		Sig. (2-tailed)	.	. 004
		N	90	90
	HW-EMOS	Correlation Coefficient	.040**	1.000
		Sig. (2-tailed)	. 004	.
		N	90	90

Table 53: Spearman's rho (ρ) Correlation Coefficients between BFFM Test and Handwriting Analysis for Emotional Stability

As it can be seen from the Figures above, the Spearman's correlation coefficient between the scores of the Big Five Factors Questionnaire and the graphologist's evaluation for Extraversion, Conscientiousness, Agreeableness, Open to Experience, and Emotional Stability are 0.230, 0.370, 0.465, 0.377, and 0.040, respectively. Based on Table 48, the value of (ρ) for Extraversion indicates a weak positive correlation, while it is a moderate positive correlation for Conscientiousness and Open to Experience. However, it is a strong positive correlation for Agreeableness while it is a very weak positive correlation for Emotional Stability. By applying the statistically significant manner, we found that the p-values for the five factors are less than level of significance ($\alpha = 0.01$). Therefore, we can conclude that there is a statistically significant relationship between the score of Big Five Factors questionnaire and the graphologist's evaluation for the Big Five Factors which means that the null hypothesis is rejected.

It can be observed from the results of the validation coefficients obtained based on the handwriting analysis results of AvgMlSC and handwriting analysis results given manually by graphologist that both led to the same conclusion. The conclusion states there is a statistically significant relationship between the score of Big Five Factors questionnaire and the graphologist's evaluation for the Big Five Factors. The following diagram illustrates the processes followed by this study for investigating the validity of handwriting analysis.


Develop a Computerized Graphology System
Evaluate the Performance of the Computerized Graphology System
Correlate the Predicted Analysis Results with the BFF Test Scores
Evaluate the Validation Results

- Evaluating the performance metrics

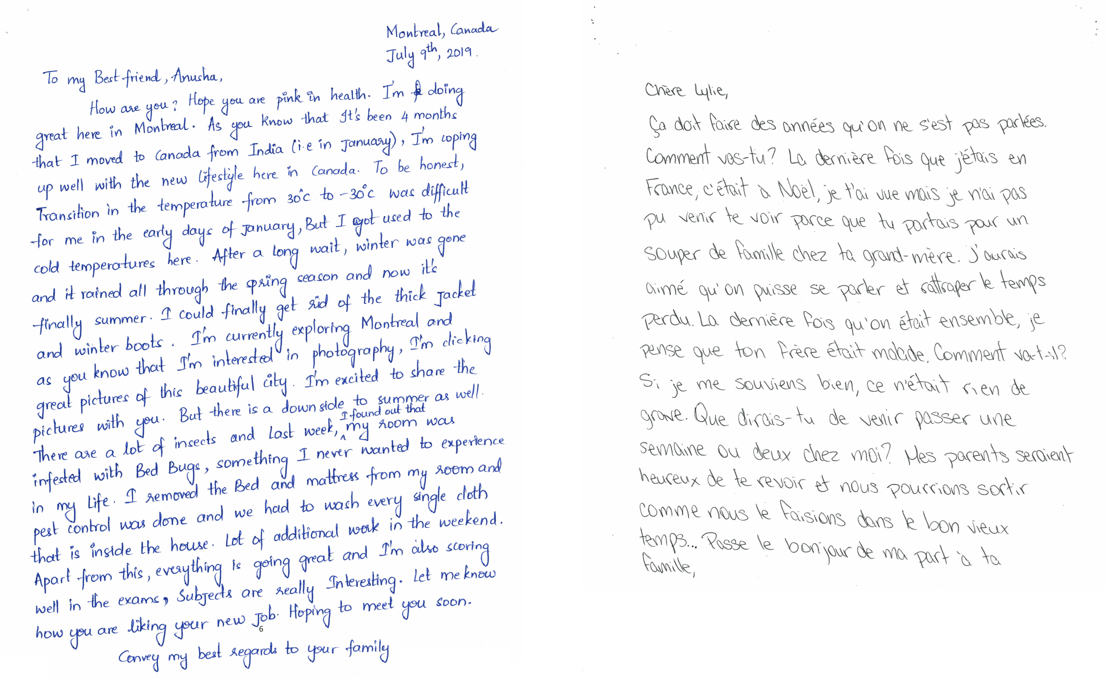
- After getting a high evaluation
- Calculating Spearman's correlation in Python
- Conducting the correlation on the testing set (90 samples)
- Reconducting the correlation on a larger size of samples (156 samples)

- Conducting a Spearman's correlation between BFF test scores and manual handwriting analysis for the same samples using SPSS

Figure 39: Diagram for the Process of Examining the Validity of Graphology

7.7 A comparison between the BFF Correlation of Printed and Cursive Writing

In this section, the BFF correlation of printed and cursive writing for Latin languages is compared using SPSS. Printed writing is a style of handwriting in which the letters are written separately while forming a word. However, cursive writing is a style of penmanship in which some characters are written joined together in a flowing manner, generally for the purpose of making writing faster. The printed handwritings with a complete BFF psychology test that are collected by our survey consist of 39 samples, see Figure 40. While we have 40 cursive handwriting samples with a complete BFF psychology test, see Figure 41.



Montreal, Canada
July 9th, 2019.

To my Best-friend, Anusha,

How are you? Hope you are pink in health. I'm ~~f~~ doing
great here in Montreal. As you know that It's been 4 months
that I moved to Canada from India (i.e in January), I'm coping
up well with the new lifestyle here in Canada. To be honest,
Transition in the temperature from 30°c to -30°c was difficult
for me in the early days of January, But I got used to the
cold temperatures here. After a long wait, winter was gone
and it rained all through the spring season and now it's
finally summer. I could finally get rid of the thick Jacket
and winter boots. I'm currently exploring Montreal and
as you know that I'm interested in photography, I'm clicking
great pictures of this beautiful city. I'm excited to share the
pictures with you. But there is a down side to summer as well.
I found out that
There are a lot of insects and Last week, my room was
infested with Bed Bugs, something I never wanted to experience
in my Life. I removed the Bed and mattress from my room and
pest control was done and we had to wash every single cloth
that is inside the house. Lot of additional work in the weekend.
Apart from this, everything is going great and I'm also scoring
well in the exams, Subjects are really Interesting. Let me know
how you are liking your new Job. Hoping to meet you soon.
Convey my best regards to your family

Chère Lylie,

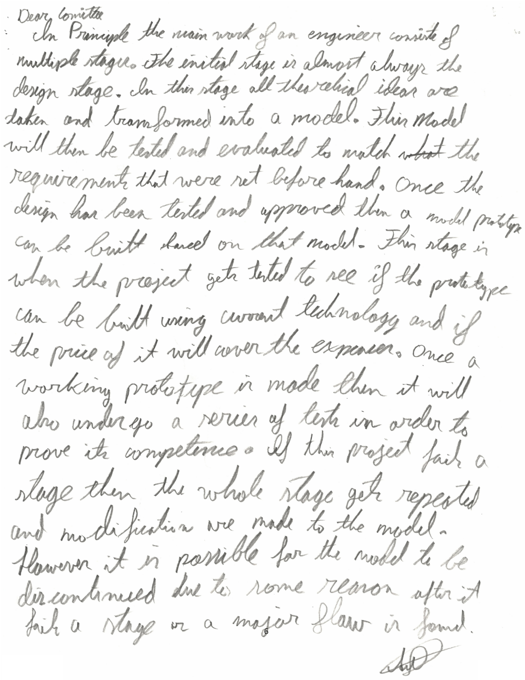
Ça doit faire des années qu'on ne s'est pas parlées.
Comment vas-tu? La dernière fois que j'étais en
France, c'était à Noël, je t'ai vue mais je n'ai pas
pu venir te voir parce que tu partais pour un
souper de famille chez ta grand-mère. J'aurais
aimé qu'on puisse se parler et rattraper le temps
perdu. La dernière fois qu'on était ensemble, je
pense que ton frère était malade. Comment va-t-il?
Si je me souviens bien, ce n'était rien de
grave. Que dirais-tu de venir passer une
semaine ou deux chez moi? Mes parents seraient
heureux de te revoir et nous pourrions sortir
comme nous le faisions dans le bon vieux
temps... Passe le bonjour de ma part à ta
famille,

(a) Printed English Handwritten Sample

(b) Printed French Handwritten Sample

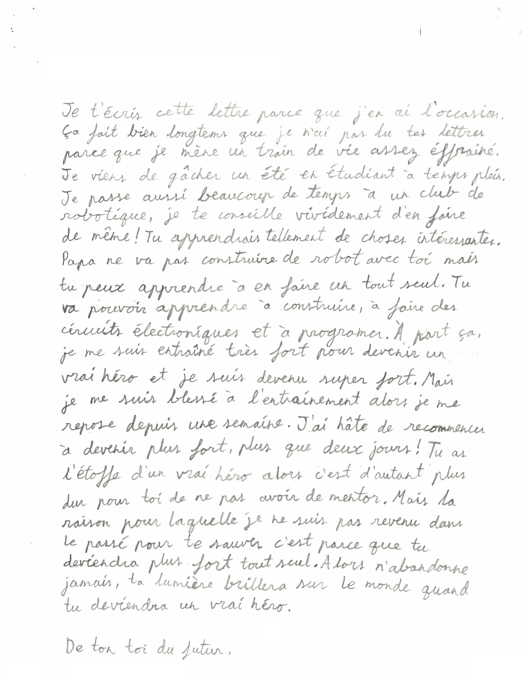
Figure 40: Two Printed Handwriting Samples Collected by our Survey

Spearman's correlation coefficients are conducted between the scores of the BFF psychology test and the BFF evaluation using handwriting analysis for printed and cursive writing. The results of the correlations are the same as expected by our graphologist. They indicate that the strength of the BFF correlation for printed writing is lower than cursive writing. This difference is explained by the fact that printed writings are often artificial and are often associated with a persona personality.



Dear Comittee
In Principle the main work of an engineer consists of
multiple stages. The initial stage is almost always the
design stage. In this stage all theoretical ideas are
taken and transformed into a model. This Model
will then be tested and evaluated to match ~~what~~ the
requirements that were set before hand. Once the
design has been tested and approved then a model prototype
can be built based on that model. This stage is
when the project gets tested to see if the prototype
can be built using current technology and if
the price of it will cover the expenses. Once a
working prototype is made then it will
also undergo a series of tests in order to
prove its competence. If this project fails a
stage then the whole stage gets repeated
and modification are made to the model.
However it is possible for the model to be
discontinued due to some reason after it
fails a stage or a major flaw is found.

(a) Cursive English Handwritten Sample



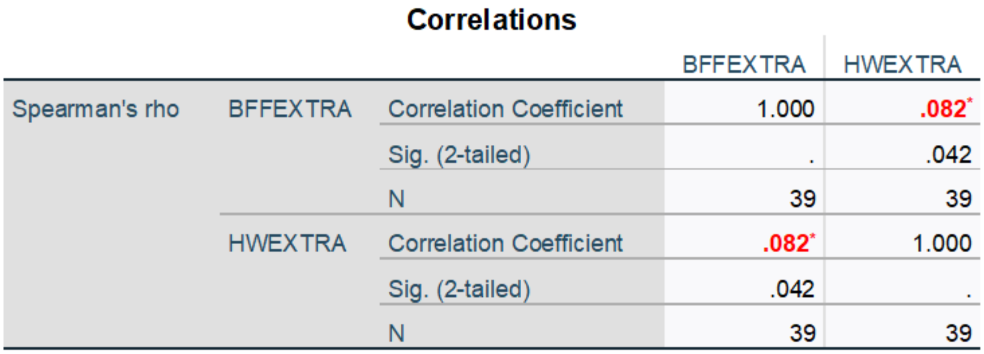
Je t'écris cette lettre parce que j'en ai l'occasion.
Ça fait bien longtems que je n'ai pas lu tes lettres
parce que je mène un train de vie assez éffrainé.
Je viens de gâcher un été en étudiant à temps plein.
Je passe aussi beaucoup de temps à un club de
robotique, je te conseille vividement d'en faire
de même! Tu apprendrais tellement de choses intéressantes.
Papa ne va pas construire de robot avec toi mais
tu peux apprendre à en faire un tout seul. Tu
va pouvoir apprendre à construire, à faire des
circuits électroniques et à programer. À part ça,
je me suis entraîné très fort pour devenir un
vrai héro et je suis devenu super fort. Mais
je me suis blessé à l'entrainement alors je me
repose depuis une semaine. J'ai hâte de recommencer
à devenir plus fort, plus que deux jours! Tu as
l'étoffe d'un vrai héro alors c'est d'autant plus
dur pour toi de ne pas avoir de mentor. Mais la
raison pour laquelle je ne suis pas revenu dans
le passé pour te sauver c'est parce que tu
deviendra plus fort tout seul. Alors n'abandonne
jamais, ta lumière brillera sur le monde quand
tu deviendra un vrai héro.

De ton toi du futur.

(b) Cursive French Handwritten Sample

Figure 41: Two Cursive Handwriting Samples Collected by our Survey

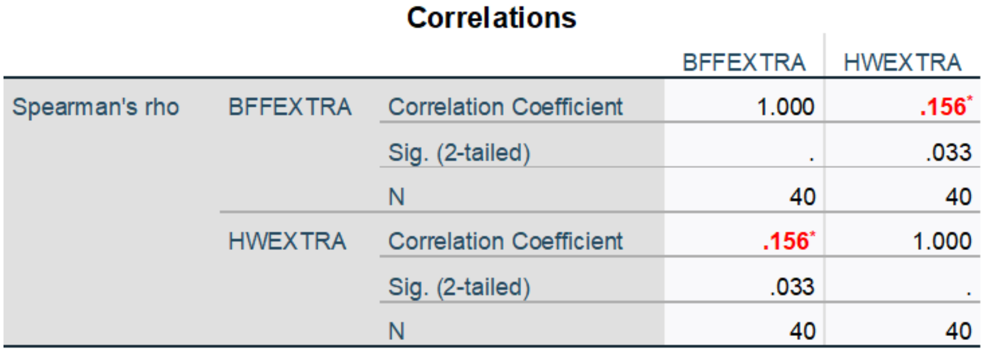
As it can be seen from Tables 54 and 55, the strength of Extraversion correlation for printed writings is $(\rho = 0.082)$ while it is $(\rho = 0.156)$ for cursive writings with a statistically significant correlations for both. Even though the two results are interpreted as a very weak relationship based on the interpretation table of Spearman's correlation coefficients, still the correlation of printed writing is weaker than cursive writing.



Correlations

			BFFEXTRA	HWEXTRA
Spearman's rho	BFFEXTRA	Correlation Coefficient	1.000	.082*
		Sig. (2-tailed)	.	.042
		N	39	39
	HWEXTRA	Correlation Coefficient	.082*	1.000
		Sig. (2-tailed)	.042	.
		N	39	39

Table 54: The Correlation of Extraversion for Printed Handwritings

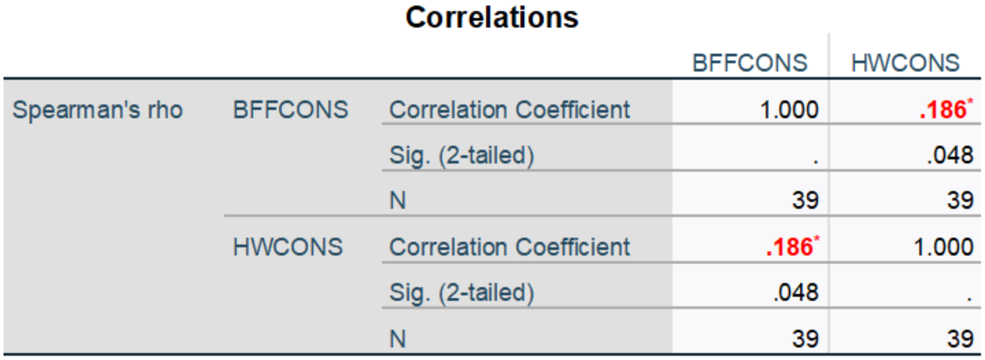


Correlations

			BFFEXTRA	HWEXTRA
Spearman's rho	BFFEXTRA	Correlation Coefficient	1.000	.156*
		Sig. (2-tailed)	.	.033
		N	40	40
	HWEXTRA	Correlation Coefficient	.156*	1.000
		Sig. (2-tailed)	.033	.
		N	40	40

Table 55: The Correlation of Extraversion for Cursive Handwritings

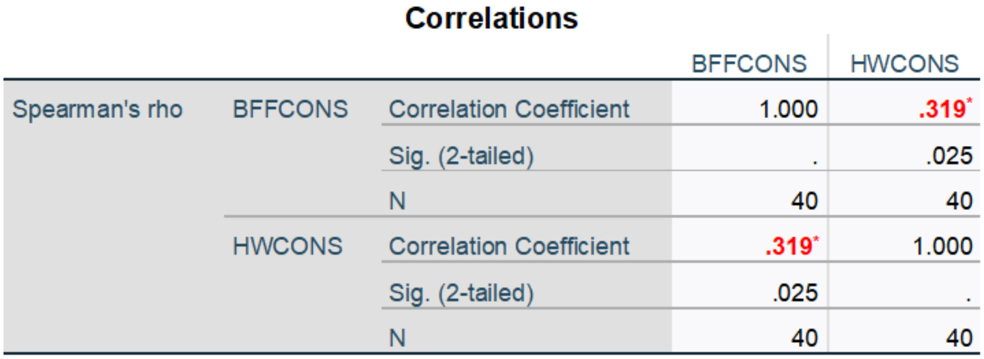
For Conscientiousness correlation, Table 56 shows that the strength of correlation for printed writings is very weak $(\rho = 0.186)$ while it is moderate $(\rho = 0.319)$ for cursive writings as it is shown in Table 57. The two p-values which are 0.048 and 0.025 indicate that there is a significant correlation between the score of BFF test and the BFF evaluation using graphology for both writings.



Correlations

			BFFCONS	HWCONS
Spearman's rho	BFFCONS	Correlation Coefficient	1.000	.186*
		Sig. (2-tailed)	.	.048
		N	39	39
	HWCONS	Correlation Coefficient	.186*	1.000
		Sig. (2-tailed)	.048	.
		N	39	39

Table 56: The Correlation of Conscientiousness for Printed Handwritings

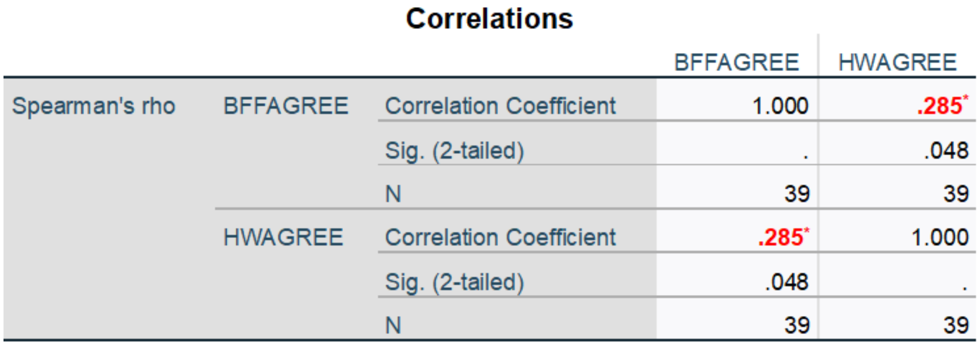


Correlations

			BFFCONS	HWCONS
Spearman's rho	BFFCONS	Correlation Coefficient	1.000	.319*
		Sig. (2-tailed)	.	.025
		N	40	40
	HWCONS	Correlation Coefficient	.319*	1.000
		Sig. (2-tailed)	.025	.
		N	40	40

Table 57: The Correlation of Conscientiousness for Cursive Handwritings

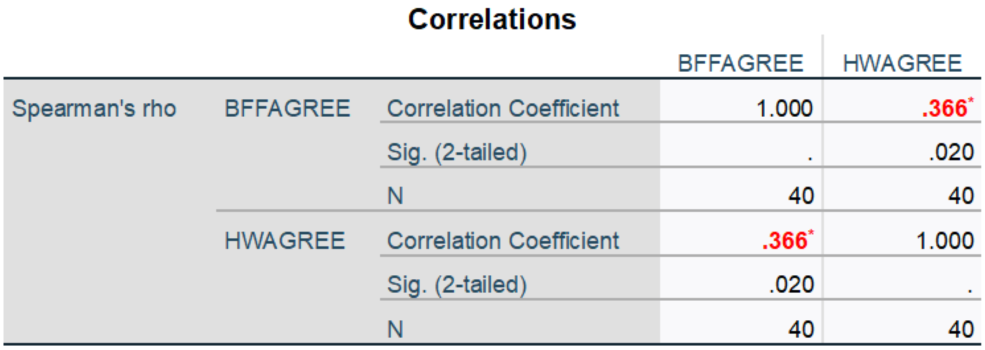
The same findings are indicated by Tables 58 and 59 for Agreeableness correlation. There is a significant correlation with a strength of weak relationship ($\rho = 0.285$) for printed writings and moderate relationship ($\rho = 0.366$) for cursive writings.



Correlations

			BFFAGREE	HWAGREE
Spearman's rho	BFFAGREE	Correlation Coefficient	1.000	.285*
		Sig. (2-tailed)	.	.048
		N	39	39
	HWAGREE	Correlation Coefficient	.285*	1.000
		Sig. (2-tailed)	.048	.
		N	39	39

Table 58: The Correlation of Agreeableness for Printed Handwritings

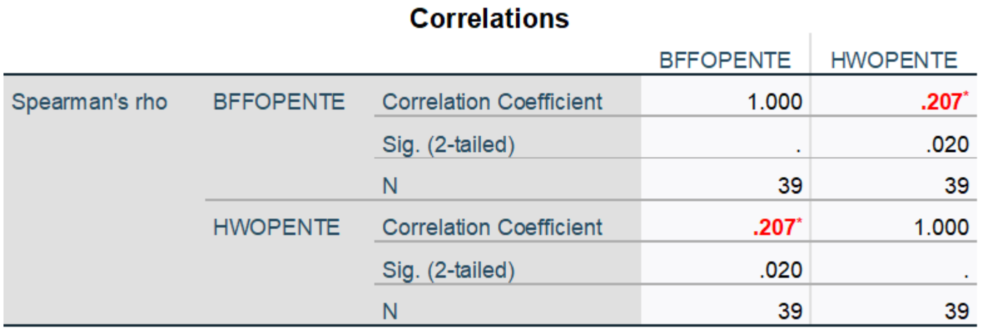


Correlations

			BFFAGREE	HWAGREE
Spearman's rho	BFFAGREE	Correlation Coefficient	1.000	.366*
		Sig. (2-tailed)	.	.020
		N	40	40
	HWAGREE	Correlation Coefficient	.366*	1.000
		Sig. (2-tailed)	.020	.
		N	40	40

Table 59: The Correlation of Agreeableness for Cursive Handwritings

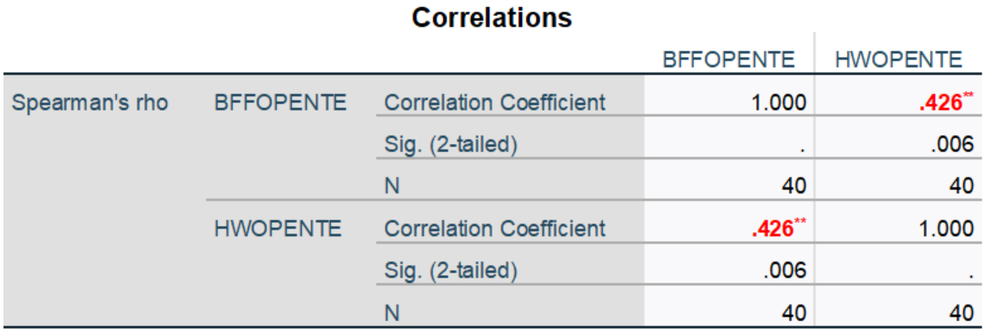
The following two tables display Open to Experience correlation which indicate that the printed writings have lower correlation than cursive writings with a significant correlation for both. Printed writings had a weak relationship with $(\rho = 0.207)$ while cursive had a strong relationship with $(\rho = 0.426)$.



Correlations

			BFFOPENTE	HWOPENTE
Spearman's rho	BFFOPENTE	Correlation Coefficient	1.000	.207*
		Sig. (2-tailed)	.	.020
		N	39	39
	HWOPENTE	Correlation Coefficient	.207*	1.000
		Sig. (2-tailed)	.020	.
		N	39	39

Table 60: The Correlation of Open to Experience for Printed Handwritings

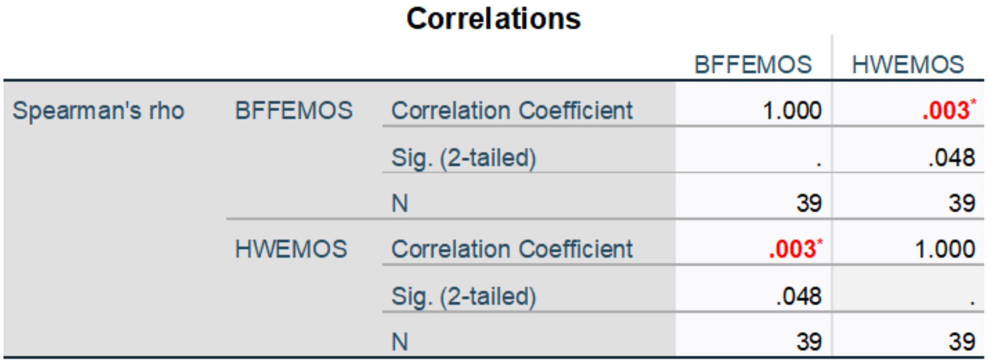


Correlations

			BFFOPENTE	HWOPENTE
Spearman's rho	BFFOPENTE	Correlation Coefficient	1.000	.426**
		Sig. (2-tailed)	.	.006
		N	40	40
	HWOPENTE	Correlation Coefficient	.426**	1.000
		Sig. (2-tailed)	.006	.
		N	40	40

Table 61: The Correlation of Open to Experience for Cursive Handwritings

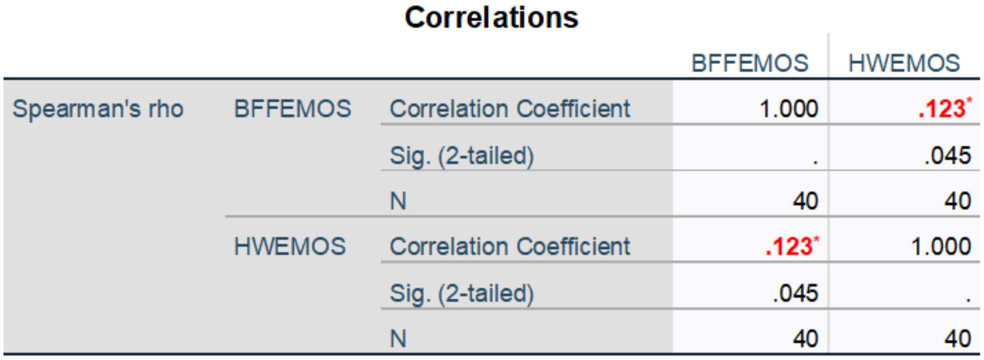
The same conclusion has been found for the last factor which is Emotional Stability. Both style of writings had a very weak relationship with a statistically significant correlation. However, the printed writings with $(\rho = 0.003)$ is weaker than the cursive ones with $(\rho = 0.123)$, see Tables 62 and 63.



Correlations

			BFFEMOS	HWEMOS
Spearman's rho	BFFEMOS	Correlation Coefficient	1.000	.003*
		Sig. (2-tailed)	.	.048
		N	39	39
	HWEMOS	Correlation Coefficient	.003*	1.000
		Sig. (2-tailed)	.048	.
		N	39	39

Table 62: The Correlation of Emotional Stability for Printed Handwritings



Correlations

			BFFEMOS	HWEMOS
Spearman's rho	BFFEMOS	Correlation Coefficient	1.000	.123*
		Sig. (2-tailed)	.	.045
		N	40	40
	HWEMOS	Correlation Coefficient	.123*	1.000
		Sig. (2-tailed)	.045	.
		N	40	40

Table 63: The Correlation of Emotional Stability for Cursive Handwritings

Chapter 8

Conclusion

This work conducted an empirical study for evaluating the Spearman's correlation between a psychological test named big five factor markers test and our automated handwriting analysis system named AvgMlSC. AvgMlSC is based on ensemble learning that was employed along with SMOTE resample technique in order to handle the issue of imbalanced dataset. The prediction results of AvgMlSC were compared to five baseline classifiers and outperformed their results with 93% predictive accuracy, 0.94 AUC, and 90% F-Score. Moreover, it achieved higher values of accuracy and F-Score than three of early computerized BFF models. The contributions of the study can be summarized as follows:

- To the best of our knowledge, this is the first study that investigates the validity of graphology by evaluating the correlation between a psychological test and a computerized graphology system.
- We introduced a robust yet simple framework to address imbalance problem in the big five factors classification.
- To the best of our knowledge, this is the first study to systematically investigate data imbalance issue in handwriting analysis in general and the big five factor classification in particular.
- To the best of our knowledge, this is the first study that predicts the big five factors model based on handwriting analysis using both ensemble and resample methods.
- Based on the literature review, imbalance in multilabel classification has been faced mainly through algorithmic adaptation and the use of ensemble, while the resampling

approach that is used in our study is the least examined part until now.

- We collected a handwriting sample dataset that is labeled manually with the measurement level of the big five factors model based on graphological rules.

For the key findings of the work, they can be summarized as follows:

- Both validation results obtained based on our computerized graphology and manual graphology led to the some conclusion.
- The validation study concludes that there is a statistically significant relationship between the score of BFF questionnaire and the BFF graphologist evaluation with different strength of relationship for each factor.
- Since the strength of the correlation between the BFF test and our computerized BFF is ranged from very weak to strong across the big five actors, big five factor questionnaire is considered more accurate than the computerized BFF. Therefore, handwriting analysis is usually applied as a complement tool. In other words, it does not replace traditional tool and direct evidence in some areas such as employee hiring or forensics.
- The results indicate that the strength of the BFF correlation for printed writing is lower than cursive writing because the former is often mechanically artificial and associated with a persona personality.
- The results of our proposed model for the computerized graphology significantly outperform the results which are reported by the baseline models.
- The results show the potential of ensembling and SMOTE oversampling for predicting the measurement level of BFF using an imbalance handwriting analysis dataset.
- The study shows the potential of machine learning methods for predicting the measurement level of BFF using graphology data.

A number of assumptions can be considered as future work for researchers which are as follows:

- Adding features of signature and drawing to develop an automated graphology system for investigating the validity of graphology.

- Validating handwriting analysis using the five languges (English, French, Chinese, Arabic, and Spanish) separately.
- Validating handwriting analysis using balanced data for the training process.
- Study the validity of graphology for a specific age group such as children or teenager.

Bibliography

[1] Mustafa Ali Abuzaraida, Akram M Zeki, and Ahmed M Zeki. Feature extraction techniques of online handwriting arabic text recognition. In *Proc. 2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, pages 1–7. IEEE, 2013.

[2] Pervez Ahmed and Hassan Mathkour. On the development of an automated graphology system. In *Proc. Ic-ai*, pages 897–901, 2008.

[3] Nazar Akrami, Johan Fernquist, Tim Isbister, Lisa Kaati, and Björn Pelzer. Automatic extraction of personality from text: Challenges and opportunities. In *Proc. 2019 IEEE International Conference on Big Data (Big Data)*, pages 3156–3164. IEEE, 2019.

[4] Monica Albu. Cp5f: A new questionnaire for the evaluation of the big five superfactors. *Cognition, Brain, Behavior*, 13(1):79, 2009.

[5] Abhishek Bal and Rajib Saha. An improved method for handwritten document analysis using segmentation, baseline recognition and writing pressure detection. *Procedia Computer Science*, 93:403–415, 2016.

[6] Gershon Ben-Shakhar, Maya Bar-Hillel, Yoram Bilu, Edor Ben-Abba, and Anat Flug. Can graphology predict occupational success? two empirical studies and some methodological ruminations. *Journal of Applied Psychology*, 71(4):645, 1986.

[7] Gian Vittorio Caprara, Claudio Barbaranelli, Laura Borgogni, and Marco Perugini. The "big five questionnaire": A new questionnaire to assess the five factor model. *Personality and individual Differences*, 15(3):281–288, 1993.

[8] HN Champa and KR Ananda Kumar. Rule based approach for personality prediction through handwriting analysis. In *Proc. 2nd International Conference on Biomedical*

Informatics and Signal processing, organized by Sai's BioSciences Research Institute Pvt. Ltd., 2009, 2009.

[9] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. A first approach to deal with imbalance in multi-label datasets. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 150–160. Springer, 2013.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[11] Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, and Ching Suen. *Character recognition systems: a guide for students and practitioners*. John Wiley & Sons, 2007.

[12] Aditya Chitlangia and G Malathi. Handwriting analysis based on histogram of oriented gradient for predicting personality traits using svm. *Procedia Computer Science*, 165:384–390, 2019.

[13] Pierre Etienne Cronje. *The viability of graphology in psycho-educational assessment*. PhD thesis, University of South Africa, 2009.

[14] Christine P Dancey and John Reidy. *Statistics without maths for psychology*. Pearson education, 2007.

[15] Carla Dazzi and Luigi Pedrabissi. Graphology and personality: an empirical study on validity of handwriting analysis. *Psychological reports*, 105(3_suppl):1255–1268, 2009.

[16] Esmeralda Contessa Djamal and Risna Darmawati. Recognition of human personality trait based on features of handwriting analysis using multi structural algorithm and artificial neural networks. In *Proc. 2013 IEEE Conference on Control, Systems & Industrial Informatics (ICCSII)*, pages 22–24, 2013.

[17] Adrian Furnham, Tomas Chamorro-Premuzic, and Ines Callahn. Does graphology predict personality and intelligence? *Individual Differences Research*, 1(2), 2003.

[18] Afnan H Garoot, Maedeh Safar, and Ching Y Suen. A comprehensive survey on handwriting and computerized graphology. In *Proc. 2017 14th IAPR International*

Conference on Document Analysis and Recognition (ICDAR), volume 1, pages 621–626. IEEE, 2017.

[19] Mihai Gavrilescu. Study on determining the myers-briggs personality type based on individual's handwriting. In *Proc. 2015 E-Health and Bioengineering Conference (EHB)*, pages 1–6. IEEE, 2015.

[20] Mihai Gavrilescu and Nicolae Vizireanu. Predicting the big five personality traits from handwriting. *EURASIP Journal on Image and Video Processing*, 2018(1):1–17, 2018.

[21] Barbara Gawda. Lack of evidence for the assessment of personality traits using handwriting analysis. *Polish Psychological Bulletin*, 45(1):73–79, 2014.

[22] Andrés Felipe Giraldo-Forero, Jorge Alberto Jaramillo-Garzón, José Francisco Ruiz-Muñoz, and César Germán Castellanos-Domínguez. Managing imbalanced data sets in multi-label problems: a case study with the smote algorithm. In *Iberoamerican Congress on Pattern Recognition*, pages 334–342. Springer, 2013.

[23] Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96, 2006.

[24] Parmeet Kaur Grewal and Deepak Prashar. Behavior prediction through handwriting analysis. *IJCST*, 3(2):520–523, 2012.

[25] W Paul Jones. Enhancing a short measure of big five personality traits with bayesian scaling. *Educational and Psychological Measurement*, 74(6):1049–1066, 2014.

[26] Vikram Kamath, Nikhil Ramaswamy, P Navin Karanth, Vijay Desai, and SM Kulkarni. Development of an automated handwriting analysis system. *ARPN Journal of Engineering and Applied Sciences*, 6(9):1819–660, 2011.

[27] Roy N King and Derek J Koehler. Illusory correlations in graphological inference. *Journal of Experimental Psychology: Applied*, 6(4):336, 2000.

[28] Xin Li and Yuhong Guo. Active learning with multi-label svm classification. In *Proc. IjCAI*, pages 1479–1485. Citeseer, 2013.

[29] Xiaoqian Liu and Tingshao Zhu. Deep learning for constructing microblog behavior representation to identify social media user's personality. *PeerJ Computer Science*, 2:e81, 2016.

[30] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[31] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.

[32] Leslie A Miller and Robert L Lovler. *Foundations of psychological testing: A practical approach.* Sage publications, 2018.

[33] Momina Moetesum, Imran Siddiqi, Farah Javed, and Uzma Masroor. Dynamic handwriting analysis for parkinson's disease identification using c-bigru model. In *Proc. 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 115–120. IEEE, 2020.

[34] Baruch Nevo. Yes, graphology can predict occupational success: Rejoinder to ben-shakhar, et al. *Perceptual and Motor Skills*, 66(1):92–94, 1988.

[35] Stanley Oosthuizen. Graphology as predictor of academic achievement. *Perceptual and Motor Skills*, 71(3):715–721, 1990.

[36] Siew Hock Ow, Kean Siang Teh, and Li Yi Yee. An overview on the use of graphology as a tool for career guidance. department of software engineering, faculty of computer science and information technology university of malaysia, kuala lumpur, malaysia. *Journal CMU*, 4(1), 2005.

[37] Helmut Ploog. *Handwriting Psychology: Personality reflected in handwriting*. iUniverse, 2013.

[38] Robert A Power and Michael Pluess. Heritability estimates of the big five personality traits based on common genetic variants. *Translational psychiatry*, 5(7):e604–e604, 2015.

[39] Shitala Prasad, Vivek Kumar Singh, and Akshay Sapre. Handwriting analysis based on segmentation method for prediction of human personality using support vector machine. *International Journal of Computer Applications*, 8(12):25–29, 2010.

[40] Abdul Rahiman, Diana Varghese, and Manoj Kumar. Habit: Handwritten analysis based individualistic traits prediction. *International Journal of Image Processing (IJIP)*, 7(2):209, 2013.

[41] Anike A Raut and Ankur M Bobade. Prediction of human personality by handwriting analysis based on segmentation method using support vector machine. *International journal of pure and applied research in engineering and technology*, 2014.

[42] Klara Goldzieher Roman. Handwriting: A key to personality. 1954.

[43] Joni Salminen, Rohan Gurunandan Rao, Soon-gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. Enriching social media personas with personality traits: A deep learning approach using the big five classes. In *Proc. International Conference on Human-Computer Interaction*, pages 101–120. Springer, 2020.

[44] Roberta Satow and Jacqueline Rector. Using gestalt graphology to identify entrepreneurial leadership. *Perceptual and motor skills*, 81(1):263–270, 1995.

[45] Gholamhosein Sheikholeslami, SN Srihari, and V Govindaraju. Computer aided graphology. Master's thesis, State University of New York at Buffalo, Citeseer, 1995.

[46] Emanuele Trucco, Tom MacGillivray, and Yanwu Xu. *Computational Retinal Image Analysis: Tools, Applications and Perspectives*. Academic Press, 2019.

[47] Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48(11):4232–4246, 2018.