Automatic Handwriting Analysis for Classifying Multi-Label Personality Traits using Transformer OCR

Marzieh Adeli Shamsabad

A Thesis

in

The Department

of

Computer Science & Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

April 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Marzieh Adeli Shamsabad**

Entitled: **Automatic Handwriting Analysis for Classifying Multi-Label Personality Traits using Transformer OCR**

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

Master of Applied Science (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 Committee:

Chair
Dr. Adam Krzyzak

Examiner
Dr. Adam Krzyzak

Examiner
Dr. Muna Khayyat

Thesis Supervisor
Dr. Ching Yee Suen

Approved by

Dr. Charalambos Poullis, Chair
Department of Computer Science & Software Engineering

2025

Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Automatic Handwriting Analysis for Classifying Multi-Label Personality Traits using Transformer OCR

Marzieh Adeli Shamsabad

Handwriting analysis, or graphology, studies an individual's psychological traits through handwriting patterns and features. It is used in forensic science, criminology, and disease diagnosis.

Previous studies have evaluated the correlation between psychological questionnaires and manual handwriting analysis, but results were inconsistent due to its limitations and human error. This research addresses these challenges by developing an automated handwriting analysis system using deep learning to predict multi-label personality traits based on the Big Five Factor Model (BFFM).

The proposed model is built on the Transformer OCR (TrOCR) architecture, pre-trained on diverse datasets, including handwritten texts like IAM. In this study, the text generation function is replaced with a classification approach to predict levels (Low, Average, High) of BFFM traits from handwriting samples. The model uses Focal Loss to handle class imbalance and Binary Cross-Entropy with Logits for accurate classification.

The dataset includes 873 French and 181 English handwriting samples from CENPARMI, originally labeled for Extraversion and Conscientiousness. It has been expanded to cover all five BFFM traits: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience, totaling 1,054 samples. Each sample is segmented into individual lines to improve generalization.

The model's performance is compared with ResNet50 and Vision Transformers (ViT Base 16 - 224 and 384). Results show that TrOCR outperforms them in accuracy and overall performance. For two personality traits, it achieves 90.05% accuracy, AUROC of 0.97, and F-Score of 89%. For all five traits, it reaches 89.01% accuracy, AUROC of 0.95, and F-Score of 87%. Extraversion shows the weakest performance (AUROC of 91), while Agreeableness performs best (AUROC of 97). These results highlight the model's effectiveness in classifying BFFM traits despite class imbalance.

Acknowledgments

I would like to sincerely thank all those who have played an essential role in the completion of my thesis.

First and foremost, I am deeply grateful to my supervisor, Prof. Ching Yee Suen, for his unwavering support, guidance, and mentorship. His expertise and insightful feedback have been invaluable, and his encouragement has greatly influenced my academic growth. I also appreciate the personal care he has shown, always offering his advice and assistance whenever I needed it.

I would like to thank the staff at CENPARMI, Concordia University, for their help and support during my studies. Special thanks to Nicola Nobile, the research manager at CENPARMI, for his technical assistance and collaborative approach, which have been incredibly helpful throughout my research.

To my family, I owe a debt of gratitude for their love, encouragement, and constant belief in me. Their sacrifices and unwavering support have been the foundation of my academic achievements, and I am forever thankful for their presence and support in my life.

I would also like to express my deepest appreciation to my partner, Sina, for his constant support and encouragement. His belief in my abilities and his support during this journey have been a constant source of strength and motivation, and I am truly fortunate to have him by my side.

Contents

List of Figures **viii**

List of Tables **ix**

1 Study Overview **1**

1.1 Introduction . . . 1

1.2 Problem Statement . . . 2

1.3 Motivation . . . 3

1.4 Objectives . . . 3

1.5 Limitations . . . 4

1.6 Thesis Outline . . . 6

2 Literature Review **7**

2.1 Handwriting Analysis and Graphology . . . 7

2.1.1 Definition and Concept . . . 7

2.1.2 Historical Development . . . 7

2.1.3 Handwriting Features . . . 8

2.1.4 Validity of Graphology and Related Works . . . 11

2.2 The Big Five Factor Model of Personality Traits (BFFM) . . . 13

2.2.1 Historical Development . . . 13

2.2.2 Definition and Explanation . . . 14

2.2.3 Related Works on Automated BFFM . . . 15

2.3 Automated Handwriting Analysis System 16

2.3.1 Concepts, Advantages, and Real-World Applications 16

2.3.2 The Core Process 17

2.3.3 Related Work 19

3 Data Collection and Analysis 24

3.1 Dataset Overview 24

3.1.1 Data Collection 25

3.1.2 Labeling Process 27

3.1.3 Digitization 27

3.2 Dataset Distribution Analysis 29

3.3 Imbalance Ratio Assessment 30

4 Methodology 33

4.1 Preprocessing Techniques 33

4.1.1 Line Segmentation 34

4.1.2 Image Processing 35

4.1.3 Data Augmentation 36

4.2 Focal Loss 37

4.3 Optimization 38

4.4 Classification 39

4.4.1 Multi-Class Classification 40

4.4.2 Multi-Label Binary Classification 41

4.5 Model Development 41

4.5.1 CNN Architecture 41

4.5.2 Vision Transformer: ViT base 16 42

4.5.3 Transformer OCR 43

4.6 Evaluation Metrics 45

5 Experimental Result and Discussion 47

5.1 Segmentation 47

5.2 Impact of Focal Loss 50

5.3 Performance of Optimizers 53

5.4 Multi-Label vs. Multi-Class Classification 55

5.5 Model Performance 57

5.5.1 English Language Dataset 57

5.5.2 French Language Dataset 59

5.5.3 Combined Dataset 62

5.6 A Comparative Analysis with Related Studies 69

6 Conclusion and Future Work 71

6.1 Summary of Findings 71

6.2 Conclusions 72

6.3 Suggestions for Future Research 73

7 Publications 74

References 84

List of Figures

Figure 3.1 Handwriting Samples from Survey Participants and Graphologist 26

Figure 3.2 Language Distribution of 1110 Handwriting Samples. 29

Figure 3.3 Overview of Five-Factor Personality Traits Distribution 30

Figure 4.1 French Handwriting Sample and Line Segmentation Results 35

Figure 4.2 Different Examples of English Handwriting Sub-samples 35

Figure 4.3 Comparison of Before and After Image Processing 36

Figure 4.4 Data Augmentation Samples 37

Figure 4.5 The Proposed TrOCR Model for Classification 43

Figure 5.1 Performance of ResNet50 Before and After Segmentation. 49

Figure 5.2 Training Curves of ResNet50: Comparison with and without Focal Loss . . 50

Figure 5.3 Impact of Focal Loss on AUROC Performance of Optimizers in the Test Phase 51

Figure 5.4 Performance of Optimizers: Training and Validation Analysis in ResNet50 . 54

Figure 5.5 Training Results for English Sub-samples 57

Figure 5.6 AUROC English Sub-samples Test results 59

Figure 5.7 Training Results for French Sub-samples 60

Figure 5.8 AUROC French Sub-samples Test results 61

Figure 5.9 Training Results for Total Sub-samples 63

Figure 5.10 AUROC Total Sub-samples Test results 64

Figure 5.11 Confusion Matrices for TrOCR by Trait on Test Sub-samples 65

Figure 5.12 AUROC Performance of TrOCR Across Traits 66

Figure 5.13 AUROC Curves of TrOCR for Personality Trait Levels (Low, Average, High) 68

List of Tables

Table 2.1 Examples of Handwriting Features for the Letters 'a', 'y', 't', 'i', and 'g'. . . 11

Table 2.2 Handwriting Features and Associated Personality Traits 12

Table 2.3 Summary of Automated Handwriting Analysis Studies 22

Table 3.1 Handwriting Features Corresponding to Each of the Big Five Factors 28

Table 3.2 Multi-Class Imbalance Ratio Calculation for the English Dataset 32

Table 3.3 Multi-Class Imbalance Ratio Calculation for the French Dataset 32

Table 5.1 Number of Class Samples and IR_{multi} for English After Segmentation 48

Table 5.2 Number of Class Samples and IR_{multi} for French After Segmentation 48

Table 5.3 Validation Performance of ResNet50 with and without Focal Loss 52

Table 5.4 Impact of Focal Loss on ResNet50 Performance in the Test Phase 52

Table 5.5 Comparative Evaluation of Classification Methods on a Validation Dataset . 55

Table 5.6 Comparison of AUROC Scores for Classification Methods Across Models . . 56

Table 5.7 Validation Performance Analysis of Models on English Subsamples 58

Table 5.8 Test Phase Performance Analysis of Models on English Subsamples 59

Table 5.9 Validation Performance Analysis of Models on French Subsamples 60

Table 5.10 Test Phase Performance Analysis of Models on French Subsamples 61

Table 5.11 Validation Performance Analysis of Models on Total Subsamples 63

Table 5.12 Test Phase Performance Analysis of Models on Total Subsamples 64

Table 5.13 Performance Metrics of TrOCR for Personality Trait Classification 66

Table 5.14 TrOCR Metrics for Personality Traits Across Levels (Low, Average, High) . 67

Table 5.15 F1-Score (%) Comparison of Personality Trait Classification Models 69

Table 5.16 Comparison of Test Accuracy (%) for ML Classifiers with SMOTE 69

Table 5.17 Test Accuracy (%) Comparison of Deep Learning Models 70

Chapter 1

Study Overview

1.1 Introduction

Handwriting is a unique form of personal expression, often considered as distinctive as a fingerprint. Graphology, the study of handwriting analysis, examines various handwriting features such as slant, margin width, size, specific letter shapes, and strokes to predict personality traits. These features can reveal numerous aspects of an individual's personality, including emotional state, self-esteem, and creativity. Unlike aptitude tests, psychometric evaluations, or lengthy questionnaires, handwriting analysis offers a faster and more accessible approach. By analyzing a simple written sample, valuable information about an individual's character can be obtained efficiently and effectively [1].

Handwriting analysis has traditionally relied on manual interpretation to assess personality traits using predefined guidelines. Although this approach provided meaningful assessments, it was often time-consuming, prone to personal bias, and lacked consistency, leading to varying outcomes among graphologists. To address these challenges, computerized handwriting analysis was developed. By using advanced algorithms, these systems process handwriting samples with enhanced speed and accuracy, ensuring consistent and dependable results while increasing efficiency [2].

Today, handwriting analysis finds applications in areas such as recruitment, personal development, forensic investigations, and healthcare, providing information about behavior, identity, and potential medical conditions [3]. Although questions about its scientific credibility remain, technological progress has enhanced its accuracy, establishing it as a practical method for exploring personality and behavioral traits.

1.2 Problem Statement

Understanding personality traits has an important role in decision-making across various fields. In psychology, it helps design therapy plans; in recruitment, it connects candidates to suitable roles. Forensic science uses it for profiling, and healthcare applies it to customize treatments and enhance patient care. Traditional methods, such as questionnaires and psychometric tests, are commonly used but have their limitations. These methods can be slow, costly, and require specialized expertise, making them less practical for large-scale or rapid evaluations. Furthermore, self-reported data is often subject to biases, such as social desirability bias, where individuals may respond in ways they believe are more socially acceptable rather than reflecting their true thoughts or behaviors, reducing the reliability of the results [4].

Handwriting analysis provides a practical way to understand personality traits using just a writing sample. It is based on the idea that handwriting reflects a person's psychological state and character. This connection comes from the fact that handwriting is guided by the brain, which controls the hand's movements, creating unique patterns in letter shapes, slant, pressure, and spacing. These patterns are shaped by a mix of physical factors, cultural influences, and personal experiences, making handwriting a reflection of an individual's personality [2].

However, manual handwriting analysis is subjective and depends on the graphologist's expertise, leading to errors like inconsistent interpretations or misjudging handwriting features such as slant or pressure [5]. Automated handwriting analysis offers a more objective and efficient solution by using algorithms to improve accuracy and consistency. However, it still faces challenges, such as handling diverse handwriting styles, managing imbalanced datasets, and interpreting complex patterns, which emphasize the need for further advancements in this field.

1.3 Motivation

Automated handwriting analysis provides a more efficient and objective way to assess personality traits compared to traditional methods. Recent advancements in machine learning and deep learning allow for precise analysis of handwriting patterns, enabling the extraction of subtle features such as stroke pressure, slant, and spacing that are linked to personality traits. These algorithms help to address challenges such as the variability in handwriting styles and imbalances in the data, using techniques like data augmentation and custom loss functions to enhance accuracy and reliability. By removing the subjectivity of manual analysis, automated systems ensure consistent results and can handle large datasets effectively.

1.4 Objectives

This study aims to develop an advance deep learning model for handwriting analysis to assess personality traits based on the Big Five Personality Traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. The dataset is labeled using graphological rules, and the validity of the labels is confirmed through the Big Five Factor Markers Test (BFFMT), a widely recognized self-report questionnaire from the International Personality Item Pool.

The study addresses the following challenges in automated handwriting-based personality assessment:

1. **Multi-Label Classification:** This approach is used to predict multiple personality traits simultaneously at different levels, by evaluating Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) and Cross Entropy with Softmax to identify the best method.

2. **Imbalanced Dataset:** Significant class imbalances in the dataset are addressed using Focal Loss, which helps focus learning on underrepresented classes.

3. **Limited Handwriting Samples:** To overcome the limited number of handwriting samples, the Line segmentation technique is applied to expand the dataset and enhance better generalization to unseen data.

4. **Feature Extraction:** TrOCR model is used for automated feature extraction as a new approach in the classification task, employing a transformer-based encoder to analyze handwriting patterns. A classification head is added to predict personality traits at different levels. The performance of TrOCR is compared with other models, such as ResNet50 (a CNN-based model) and Vision Transformer (ViT) base 16 with input sizes of 224 and 384, to assess its effectiveness in capturing handwriting patterns and accurately predicting personality traits.

To enhance model performance, state-of-the-art optimizers, including Adam, AdaBelief, and SGD with momentum, are evaluated. This combination of multi-label classification, advanced loss functions, segmentation, augmentation, and robust feature extraction aims to develop a reliable and scalable automated handwriting analysis model for personality assessment.

1.5 Limitations

During the implementation of this study, we encountered several challenges that impacted the development and evaluation process:

1. Dataset

- **Data Insufficiency:** The number of handwriting samples was insufficient for deep learning models, which require large datasets to achieve reliable generalization. The limited data made it difficult to train the model effectively and ensure that it performs well on unseen handwriting samples. This issue also restricted the diversity of handwriting styles included in the dataset, potentially affecting the model's ability to handle variations.

- **Imbalanced Dataset:** The dataset had an uneven distribution of samples across different personality trait levels, with some traits being significantly underrepresented. This imbalance led to challenges in model training, as it became more difficult for the model to learn patterns for minority classes, ultimately impacting the classification accuracy for those traits.

- **Labeling Consistency:** The labeling process involved assigning personality traits based on graphological rules and validating them through psychological questionnaires. Ensuring that these two methods aligned required significant effort and introduced complexity. Discrepancies in the interpretation of graphological features sometimes created additional challenges in maintaining consistent labels.

2. Model Training

- **Computational Resources:** Training deep learning models requires a lot of computing power and memory. To meet these needs, a powerful graphical processing unit (GPU) was used to handle time and memory limits.

- **Optimization Complexity:** Selecting and fine-tuning the most effective optimization strategies for the model required significant effort. Different optimizers performed variably with the dataset, and the need to experiment extensively with learning rates and other hyperparameters added complexity to the training process. This process consumed considerable time and required careful evaluation to ensure the model achieved its best possible performance.

3. Preprocessing

- **Segmentation:** To increase the dataset size, a segmentation method was implemented to extract meaningful sections from the handwriting samples. Developing and applying this technique required significant effort to ensure the segments are consistent and suitable for analysis. This step was essential for improving data quality but presented considerable challenges during preprocessing.

4. Validation and Evaluation

- **Comparative Model Analysis:** Multiple deep learning models were implemented to ensure the robustness and generalizability of the proposed method. The performance of different architectures was compared to highlight the strengths and limitations of each model in recognizing handwriting patterns for personality trait classification based on the BFFM. Through this approach, the effectiveness of the proposed method was validated, and valuable insights were gained into the suitability of various models for handwriting-based personality analysis.

1.6 Thesis Outline

This thesis is organized as follows: Chapter 2 provides a review of handwriting analysis, the Big Five factor measurement, and the application of computerized handwriting analysis, along with related work on these topics. Chapter 3 describes the dataset, data preprocessing, and feature extraction methods. Chapter 4 outlines the materials and methods used in the study. Chapter 5 presents the experimental results, and Chapter 6 concludes with a discussion, summarizing the findings and suggesting directions for future work.

Chapter 2

Literature Review

2.1 Handwriting Analysis and Graphology

2.1.1 Definition and Concept

Graphology, the study of handwriting to determine personality traits, has been used to analyze and understand an individual's character, emotional state, and behavior. It evaluates specific handwriting features, such as letter size, slant, spacing, pressure, and overall structure, to analyze traits like emotional stability, honesty, fears, and defenses. Unlike methods focused on demographic information such as age or nationality, graphology aims to reveal psychological characteristics unique to the individual [6].

2.1.2 Historical Development

The origins of graphology can be traced back to ancient times, with figures like Aristotle observing links between handwriting and behavior [7]. However, the field gained structure and recognition in the 19th century when Jean-Hippolyte Michon formalized it by introducing the term ”graphology” and developing systematic techniques to analyze handwriting [8]. This basis was expanded by Jules Crépieux-Jamin, who integrated psychological principles into handwriting analysis, elevating its potential as a tool for personality assessment [9].

Over time, graphology evolved into two main schools of thought: Graphoanalysis and Gestalt

Graphology [10]. Graphoanalysis, primarily used in the United States, focuses on analyzing individual symbols in handwriting independently. Each symbol is treated as a separate entity to determine its specific meaning [11]. On the other hand, Gestalt Graphology, prominent in Europe, particularly in Germany, adopts a holistic approach. This method considers handwriting as a unified whole, analyzing the interplay of form, movement, and space to uncover patterns, or ”gestalts,” that reflect various aspects of the writer’s personality [12]. Gestalt Graphology analyzes handwriting by looking at its overall visual impression, or ”Gestalt,” to capture the general style and flow. It then evaluates the handwriting’s clarity, consistency, and structure, determining whether it is of high or low quality. Key features, including dominant, secondary, and contrasting traits, are identified to understand the unique characteristics of handwriting. Finally, these elements are combined to form a detailed interpretation of the writer’s personality [13]. In this study, the principles of Gestalt Graphology are applied to analyze and label handwriting samples, which is the first step in training our models.

2.1.3 Handwriting Features

Handwriting features are fundamental to graphology, offering an understanding of an individual’s personality, emotional tendencies, and cognitive patterns. These features are typically categorized into three main types: general measurements, fundamental measurements, and accessory measurements. Together, they form a comprehensive structure that graphologists use to assess personality traits and behavioral characteristics [14].

General Measurements

General measurements provide an overall impression of the handwriting, focusing on stroke quality, consistency, and visual appearance. Graphologists often associate the general style of handwriting with the writer’s psychological state. For instance, small, precise strokes in clear handwriting are interpreted as a sign of attention to detail and good concentration. Conversely, small strokes in poorly executed handwriting may indicate traits such as pettiness or rigidity.

The overall impression also captures the balance and flow of the handwriting. A balanced and harmonious style suggests a well-organized and adaptable personality, while irregular or inconsistent handwriting may reflect emotional instability or disorganization.

Fundamental Measurements

Fundamental measurements explore deeper into specific handwriting characteristics that form the basis of personality analysis. These features include:

- **Slant:** The direction of the slant reflects emotional tendencies. A rightward slant suggests openness, expressiveness, and sociability, while a leftward slant may indicate caution, reserve, or emotional withdrawal. Neutral or upright slants are often associated with objectivity and emotional balance.
- **Baseline Direction:** The alignment of handwriting on the page signifies mood and outlook. Upward baselines indicate optimism and enthusiasm, whereas downward baselines suggest negativity, fatigue, or a lack of motivation.
- **Letter Size:** Letter size is linked to self-image and confidence. Larger letters are associated with extroversion, confidence, and assertiveness, while smaller letters suggest introversion, humility, or focus on detail.
- **Continuity:** The connection between letters reflects thinking styles. Continuous strokes indicate logical, organized thought processes, while disconnected strokes reveal spontaneity, creativity, or even impulsiveness.
- **Handwriting Form (Shape):** The overall shape of handwriting, whether rounded or angular, provides information in natural impulses and decision-making tendencies. Rounded forms suggest flexibility and adaptability, while angular forms denote determination and a strong-willed personality.
- **Spacing and Alignment:** The arrangement of words, lines, and letters highlights organizational skills and adaptability. Wide spacing between words suggests independence, while narrow spacing indicates a need for closeness or sociability.
- **Pen Pressure:** The intensity of pen pressure reflects the writer’s energy and strength. Heavy

pressure is often linked to boldness, determination, and emotional intensity, while light pressure indicates sensitivity, delicacy, or passivity.

- **Writing Speed:** Speed is a marker of mental and physical activity. Fast handwriting is associated with dynamism, decisiveness, and a quick thought process, while slow handwriting suggests caution, deliberation, and attention to detail.

Accessory Measurements

Accessory measurements analyze specific graphic symbols and unique features within handwriting. These elements provide additional depth to personality assessments:

- **Crossbars on ‘t’:** The height, alignment, and firmness of the crossbar on the letter ‘t’ indicates the writer’s ambition and self-control. High crossbars signify high aspirations, while low crossbars may reflect modesty or low self-esteem (Table 2.1).
- **Dots on ‘i’:** The placement and size of the dot above the letter ‘i’ give information in focus and imagination (Table 2.1). Precisely placed dots suggest attention to detail and ambition, while scattered dots indicate creativity and spontaneity.
- **Capital Letters:** The size and style of capital letters show self-esteem and a desire to impress. Large, ornate capitals suggest confidence and a need for recognition, while smaller capitals reflect modesty or humility.
- **Loops and Extensions:** Loops in letters like ‘g’ and ‘y’ provide clues about creativity and aspirations. Large loops may signify idealism and ambition, while small loops suggest practicality and realism (Table 2.1).
- **Initial and Terminal Strokes:** The way letters begin and end offers clues about the writer’s mindset. Strong initial strokes suggest determination and proactivity, while faint terminal strokes may indicate timidity or hesitation.

For a concise overview, Table 2.2 summarizes these features, their interpretations, and their associated personality traits.

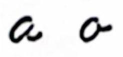
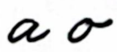
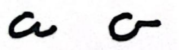
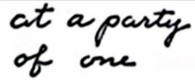
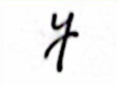
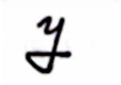
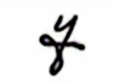
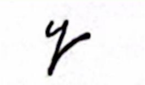
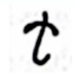
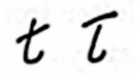
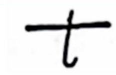
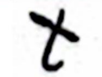
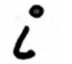
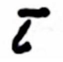
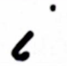
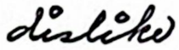
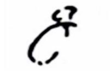
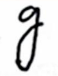
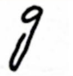
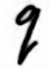
Features	Examples			
Shape	Lightly opened at the top.	Mostly closed.	Opened at the top.	Sometimes open, and closed.
Letter				
Traits	Frank, talkative, generous.	Discreet, diplomatic.	Too talkative, outspoken.	Can keep a secret, sincere.
Shape	Broken.	Triangular	Loop within a loop.	Right turned.
Letter				
Traits	Hidden anxiety about social matters.	Persistent, active, impatient.	Persistence.	Selfless and compassionate.
Shape	Arched (convex).	Rising upward.	Long heavy pressure.	Slanting down, heavy pressure.
Letter				
Traits	Excellent self-control.	Hopeful, ambitious.	Energetic, aggressive, enthusiastic.	Brutal, short temper.
Shape	Heavy, large.	Horizontal, dashed.	Faint dot.	Circle i dot.
Letter				
Traits	Materialistic and sensitive.	Energetic.	Weak will power, poor vitality.	Individualistic, interest in arts and crafts.
Shape	With light pressure.	Large, inflated in vertical slant.	Failure to close lower loop.	Upstroke to the right.
Letter				
Traits	Sensitive to music, color, and rhythm.	Self- control, strong desire.	Nervous and irritable.	Sympathetic, Volunteer.

Table 2.1: Examples of Handwriting Features for the Letters ’a’, ’y’, ’t’, ’i’, and ’g’ [15]

2.1.4 Validity of Graphology and Related Works

The validity of graphology, or handwriting analysis, has long been a controversial subject. While graphology believes that handwriting can reveal personality traits through features like slant, size, and pressure, scientific research has often questioned its reliability.

Studies have shown that handwriting analysis is prone to errors and inconsistencies. For example, in research conducted in 2000, King and Koehler found that people often see connections

Table 2.2: Handwriting Features and Associated Personality Traits [14]

Handwriting Feature	**Interpretation**	**Associated Personality Traits**
Slant	The direction of the handwriting's tilt.	Rightward: Sociable, expressive. Leftward: Reserved, cautious. Upright: Balanced, objective.
Baseline Direction	The alignment of handwriting across the page.	Upward: Optimistic, enthusiastic. Downward: Pessimistic, fatigued. Wavy: Emotional instability.
Letter Size	The height and width of letters.	Large: Confident, extroverted. Small: Introverted, humble. Medium: Balanced self-image.
Spacing Between Words	The distance between words.	Wide: Independent, distant. Narrow: Sociable, seeks closeness.
Pen Pressure	The amount of pressure applied to the writing instrument.	Heavy: Determined, intense. Light: Sensitive, gentle.
Writing Speed	The tempo of the handwriting.	Fast: Dynamic, decisive. Slow: Cautious, deliberate.
Crossbar on 't'	The height, length, and firmness of the cross stroke on the letter 't'.	High: Ambitious, goal-oriented. Low: Lacks confidence. Firm: Strong willpower. Weak: Indecisive.
Dots on 'i'	The position, size, and shape of the dots above lowercase 'i'.	Precise: Focused, attentive to detail. Scattered: Carefree, imaginative. Round: Idealistic. Slash-like: Impatient.
Loops in Letters	The size and shape of loops in letters like 'g', 'y', or 'd'.	Large: Creative, idealistic, ambitious. Small: Practical, realistic. Closed: Reserved. Open: Expressive, imaginative.
Initial Strokes	The way letters begin (e.g., bold, faint, curved).	Strong: Determined, proactive. Weak: Hesitant, cautious.
Terminal Strokes	The way letters end (e.g., upward, downward, straight).	Upward: Ambitious, aspiring. Downward: Practical, grounded. Firm: Resolute, courageous.
Capital Letters	The size and embellishment of uppercase letters.	Large: Confident, self-important, seeks attention. Small: Modest, humble, reserved.
Alignment of Lines	The straightness or curvature of lines on the page.	Straight: Organized, disciplined. Wavy: Creative, unconventional, emotional.

between handwriting and personality traits that are not real. These false connections happen because of personal biases, making handwriting analysis less trustworthy. They also pointed out that factors like mood and environment can change handwriting, adding to its unreliability [16].

In a study conducted in 2003, Adrian et al. found no strong link between handwriting and traits like intelligence or personality. This raised serious questions about whether graphology could truly measure these traits [17]. Similarly, Thiry and Rohmer examined handwriting analysis in 2007 by comparing it to psychological tools like the Rorschach test. While they found some small connections between handwriting and certain psychological traits, they concluded that these were not strong enough to make graphology a reliable tool for evaluating personality [18].

In 2009, Dazzi and Pedrabissi tested whether graphology could predict the Big Five personality traits. They found no reliable connection between handwriting features and traits like openness or neuroticism, further weakening the claims of graphology [19].

Gawda's research in 2014 identified specific handwriting features tied to personality traits but found no consistent patterns. Similarly, Harne et al., in 2018, demonstrated that cultural and individual differences greatly influence handwriting. These variations make it very difficult to create standardized methods for analyzing handwriting [20, 21].

More recent research conducted in 2021 by Garoot et al. highlighted how handwriting interpretations often vary between graphologists, underlining the need for clear and standardized methods. They concluded that, while there may be statistically significant correlations between handwriting-based evaluations and personality assessments like the Big Five Factors, the inconsistency in correlation strength across traits indicates that handwriting analysis is not yet a reliable standalone method for assessing personality. Current evidence supports its use as a complementary tool rather than a substitute for validated psychological instruments [22].

2.2 The Big Five Factor Model of Personality Traits (BFFM)

2.2.1 Historical Development

The concept of the Big Five Personality Traits came out from research aimed at identifying universal dimensions that describe human personality. The origins of this approach can be traced back to Allport and Odbert in 1936, who compiled a list of words from the English language used to describe personality. By analyzing and reducing this list, they laid the basis for a widely accepted model of personality.

In the 1940s, Cattell refined this approach by grouping these traits into 16 personality factors using factor analysis. Later, in the 1960s, Tupes and Christal re-examined earlier studies and consistently identified five core dimensions. Their findings became a significant moment in the development of the Big Five. However, it wasn't until the 1980s, through the influential work of Costa and McCrae in 1985 and Goldberg in 1990, that the Big Five gained importance. These researchers demonstrated its reliability and validity across cultures and contexts, making it one of the most robust models for understanding personality [23, 24].

2.2.2 Definition and Explanation

BFFM is a well-known model for describing an individual's personality. It is based on five basic personality traits which are grouped into sub-factors, as follows [1]:

1. **Extraversion (EX)**: Extraversion measures sociability, energy levels, and the tendency to seek stimulation from the external environment. Extroverts are typically outgoing, assertive, and thrive in social interactions, while introverts prefer solitude and self-reflection.

2. **Neuroticism (NE)**: Neuroticism is often discussed about emotional stability and reflects a tendency to experience negative emotions, such as anxiety and stress. Individuals with high neuroticism are more prone to mood swings and emotional instability, while those with low neuroticism exhibit greater emotional stability and resilience.

3. **Agreeableness (AG)**: Agreeableness reflects interpersonal traits such as empathy, kindness, and cooperation. Highly agreeable individuals are trusting and generous, whereas those with low agreeableness may exhibit competitiveness or skepticism.

4. **Conscientiousness (CO)**: This trait is associated with self-discipline, organization, and a strong sense of duty. Highly conscientious individuals are reliable, goal-oriented, and careful in their work. In contrast, those with low conscientiousness may exhibit impulsiveness and a lack of reliability.

5. **Openness to Experience (OE)**: Openness reflects curiosity, imagination, and a preference for novelty. Individuals scoring high on openness are often creative, adventurous, and reasonably curious, while those with low openness are more practical, routine-oriented, and traditional.

Self-report tests are widely used to measure each of the Big Five personality traits. These tests consist of questionnaires that include sets of markers representing each trait. The number of items in these questionnaires can vary depending on the specific version of the test. Each item is rated on a 5-point scale, ranging from 1 (Strongly Disagree or Very False for Me) to 5 (Strongly Agree or Very True for Me). Participants evaluate how accurately each statement describes their personality, providing a structured and standardized method to assess openness, conscientiousness, extraversion, agreeableness, and neuroticism.

2.2.3 Related Works on Automated BFFM

In 2018, Gavrilescu and Vizireanu developed a neural network-based system to predict the Big Five personality traits using handwriting samples. Their model achieved with notable performance exceeding 84% for Openness, Extraversion, and Neuroticism, and slightly lower accuracies of around 77% for Conscientiousness and Agreeableness [1].

In 2020, Salminen et al. introduced a deep learning model combining a one-dimensional convolutional neural network and a long short-term memory network to predict personality traits from textual data. Using a dataset of 2,467 essays, their model achieved F1 scores ranging from 0.484 for Emotional Stability to 0.553 for Agreeableness, demonstrating the potential of combining text analysis with advanced neural networks [25].

In 2022, Ramezani et al. proposed KGrAt-Net, a Knowledge Graph Attention Network designed to improve personality prediction from text. Incorporating knowledge graph embeddings, their model achieved an average accuracy of 70.26%, which increased to 72.41% when graph features were added [26]. Around the same time, Kerz et al. explored the integration of psycholinguistic features with transformer-based models, such as BERT. Their approach enhanced the detection of personality traits from verbal behavior by leveraging psycholinguistic feature analysis alongside pre-trained transformer embeddings [27].

By 2023, Sirasapalli and Malla introduced a deep learning model that used convolutional and recurrent neural networks to map Myers-Briggs Type Indicator profiles to the Big Five traits. Their methodology, which combined datasets for improved generalization, achieved an impressive accuracy of 87.89% and an F1 score of 0.924 [28].

In 2024, Yan et al. examined the performance of large language models in predicting personality traits from Chinese counseling dialogues. Their fine-tuned model significantly outperformed earlier methods, achieving a 36.94% improvement over the state-of-the-art in personality prediction accuracy [29]. That same year, Sze et al. demonstrated the feasibility of using mobile phone sensor data to assess personality traits. Their machine learning model achieved an F1 score of 0.78 in a two-class classification problem, showcasing the potential of behavioral data for personality prediction [30].

Peters et al. explored the use of GPT-4-powered chatbots for inferring the Big Five personality traits. Their chatbot inferred personality traits with moderate accuracy, outperforming earlier static-text-based methods [31].

In a recent study, Ouarka et al. proposed a multimodal fusion approach to predict personality traits using visual, audio, and text data. Their model employed pre-trained architectures such as ViT-B16 and VGG16 for visual features, VGGish for audio features, and GloVe embeddings for text. Long Short-Term Memory networks were used to capture temporal dependencies, while attention mechanisms enhanced performance. The method achieved a prediction accuracy of 91.70%, demonstrating the potential of combining multiple modalities for personality prediction [32].

2.3 Automated Handwriting Analysis System

2.3.1 Concepts, Advantages, and Real-World Applications

Automated handwriting analysis involves using computational methods to evaluate and interpret handwriting samples. Unlike traditional manual techniques, which rely on expert judgment, automated systems employ algorithms to analyze handwriting features such as shape, size, slant, and spacing. These models are designed to extract patterns and correlations, offering information about the writer's identity, cognitive abilities, or personality traits. By applying machine learning and artificial intelligence, automated handwriting analysis has transformed into a scalable and efficient tool. The shift from manual to automated handwriting analysis brings several distinct advantages:

1. **Precision and Objectivity:** Automated systems eliminate human bias, ensuring consistent results across diverse datasets [25].

2. **Speed and Efficiency:** Large-scale handwriting datasets can be processed within minutes, enabling faster decision-making in areas like forensics and education [33].

3. **Scalability:** With minimal additional resources, automated methods can scale to analyze extensive collections of handwriting samples [34].

4. **Integration with Advanced Technologies:** Combining handwriting analysis with neural networks and deep learning models significantly enhances prediction accuracy [35].

Automated handwriting analysis finds applications across various domains:

- **Forensic Science:** Identifying forgeries, verifying signatures, and authenticating documents are common use cases [36].
- **Healthcare:** Handwriting analysis aids in diagnosing motor-related disorders such as Parkinson's disease and in monitoring rehabilitation progress [37].
- **Educational Assessment:** Automated systems evaluate handwriting to assess students' cognitive and motor skills, offering valuable feedback for educators [38].
- **Psychological Profiling:** In psychology, handwriting features are correlated with personality traits, emotions, and mental states, contributing to studies on human behavior [1].

2.3.2 The Core Process

Automated handwriting analysis typically involves five main stages:

1. **Data Collection:** The first step in automated handwriting analysis is collecting samples, either by scanning documents with high-resolution scanners for clarity or capturing handwriting directly using digital devices like tablets, which also record dynamic features such as speed

and pressure. Mobile applications further simplify the process, enabling scalable and remote data collection [38]. High-quality data is important, as noise or distortion can affect analysis accuracy.

2. **Image Preprocessing:** In handwriting analysis for personality trait assessment, pen pressure is considered one of the features. However, this information is often lost during the digitization of handwriting samples. Therefore, enhancing the quality of scanned images through preprocessing becomes a critical step to preserve and highlight all meaningful features. This process includes techniques such as noise removal, binarization, and normalization. Methods like median filtering and adaptive thresholding are commonly applied to improve image clarity and ensure the handwriting is suitable for further analysis.[39].

3. **Feature Extraction:** Feature extraction identifies the specific characteristics of handwriting that can be analyzed further. Commonly extracted features include:

   - Structural Features: Letter shapes, loops, alignment, and spacing between letters and words. The steadiness and direction of handwriting along a baseline.
   - Dynamic Features: Pressure, speed, and rhythm. Algorithms like Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) are employed for robust feature extraction [40].

4. **Pattern Classification:** The extracted features are fed into classifiers to identify handwriting patterns and predict personality traits. Machine learning models such as Support Vector Machines (SVMs) and Random Forests, as well as deep learning architectures like Convolutional Neural Networks (CNNs), are frequently used for personality trait classification from handwriting [33].

5. **Results Interpretation:** The system analyzes handwriting features to assess personality traits, using models like the BFFM. The results are validated against established benchmarks or expert evaluations to ensure accuracy and are presented in clear, user-friendly formats for easy interpretation.

2.3.3 Related Work

Gahmousse et al., in 2021 used Edge Hinge (EH) for feature extraction and AdaBoost for trait classification based on the Five Factor Model. A majority voting technique with three AdaBoost models (C1, C2, C4) achieved 83.02% accuracy, outperforming previous works [41].

In 2022, Alamsyah et al. used a simple CNN with two layers to classify handwriting features such as entry strokes, slantness, and size on the AND dataset, achieving 80.88% accuracy. However, since the dataset was imbalanced, accuracy alone may not fully reflect the model's performance [42].

Mukherjee et al., also in 2022, proposed a method for predicting Big Five personality traits using handwriting features. Their study extracted specific characters ('a', 'g', 'n', 't') and the word 'of' to identify features like slants, loops, and connectivity. They utilized multi-label classification techniques such as Classifier Chains (CC), Binary Relevance (BR), and Label Powerset (LP) with KSTAR, KNN, and MLP as base classifiers. The CC-KNN combination achieved the best accuracy of 98.1% on a custom dataset of 50 participants [38].

In the same year, Yusof et al. analyzed graphological features such as slant, spacing, and baselines using Agglomerative Hierarchical Clustering (AHC) with Principal Component Analysis (PCA) for feature extraction and dimensionality reduction. They used a dataset of 70 Malaysian handwriting samples labeled according to the BFFM, and reported moderate clustering performance, achieving a silhouette score of 0.054 [43].

Durga and Deepu, in 2022, introduced a handwriting-based personality classification model that combined document and character-level handwriting features. They employed a novel Directional Movement (DM) kernel in CNNs, designed to capture the directional flow of handwriting strokes for fine-grained feature extraction, and used Hinge loss for optimization. Their model achieved an average accuracy of 86% on a private dataset of 200 samples [44].

In another study from 2022, Shree and Siddaraju developed a deep learning pipeline for handwriting-based personality analysis using a custom dataset of 3,000 images. Their approach involved two key stages: handwriting detection and personality trait classification. YOLO v5 was used to accurately detect and localize handwriting regions within the images, achieving a high F1 score of 0.95. These detected regions were then classified using ResNet-34, which achieved an F1 score of 0.91 for predicting personality traits [35].

In 2022, Garoot and Suen introduced the AvgMlSC ensemble learning model, which combines Multi-label SVM and CNN classifiers to predict the Big Five personality traits from handwriting. To address class imbalance, they applied the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic examples of minority classes by interpolating between existing samples. Their approach achieved 93% accuracy, an AUC of 0.94, and a 90% F-score on the dataset, which contains 1,108 handwriting samples in five languages [45].

In the same year, Samsuryadi et al. proposed a handwriting analysis model aimed at predicting the Big Five personality traits. They extracted a range of handcrafted features from the IAM Handwriting Database, including baselines, letter size, slant, spacing, and pen pressure, using image processing techniques implemented in OpenCV. These features were then used to train traditional machine learning classifiers, including Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The model demonstrated high predictive performance, achieving over 99% accuracy in personality trait classification [46].

In 2023, Peralta-Rodríguez et al. conducted a study to classify personality traits based on the Big Five Factor Model using handwritten images. Their approach involved a two-stage deep learning pipeline: first, a U-Net architecture was employed for image preprocessing to enhance handwriting regions and reduce noise. Then, a Convolutional Neural Network with five convolutional layers was used for personality trait classification. The model was evaluated on the HWxPI dataset, which consists of 418 handwritten essays. While the overall performance was modest, with an average AUC of 0.56, the model achieved its highest AUC score of 0.62 for the Extraversion trait [47].

Later in 2023, Dhumal et al. introduced a hybrid deep learning model that combined CNNs and Long Short-Term Memory (LSTM) networks for handwriting-based personality trait prediction. In this architecture, CNNs were used to extract spatial features from handwriting images, while LSTMs captured temporal dependencies and sequential patterns present in the handwritten strokes. The model was trained and evaluated on a custom offline handwriting dataset, achieving a high classification accuracy of 96% [48].

In 2024, Ahmed et al. utilized models like VGG16, ResNet, DenseNet201, and InceptionV3 for handwriting-based personality prediction. They extracted features such as letter size, slant, and pressure from the IAM Database and a proprietary dataset. Their VGG16 model achieved a maximum accuracy of 73.8% [33].

Yan Xu et al., in 2024, focused on handwriting-based personality prediction, emphasizing preprocessing methods like binarization and noise removal. Using the CENPARMI dataset of 234 samples, they achieved 82.90% accuracy for conscientiousness and extraversion with the ConvNeXt-Tiny model [29].

Nair et al. explored handwriting analysis in 2024 using algorithms: KNN, SVM, Naive Bayes, Decision Trees, and Random Forest. Extracting features like stroke pressure, letter size, slant, and spacing from a personal handwriting dataset, SVM achieved the highest accuracy of over 95% [49].

Chethan et al., also in 2024, combined CNNs, ANNs, and ResNets for handwriting-based personality prediction. Their dataset of 1000 samples categorized traits such as Anxious and Cooperative, achieving individual accuracy rates of up to 85% [50].

In 2024, Safar and Suen integrated traditional graphology with machine learning to predict personality traits from handwriting. They used VGG16 for feature extraction and applied machine learning algorithms like k-NN, Random Forest, and Logistic Regression, alongside SMOTE for data balancing. Ensemble methods, including Stacking and Majority Voting, were also employed, achieving over 90% accuracy for traits like Agreeableness and Openness to Experience [51].

In a recent 2025 study, Puttaswamy and Thillaiarasu employed Fine DenseNet for feature extraction and an Attention-Mechanism-based Deep LSTM with CTC loss (AMDLSTM-CTC) for classification. Their model achieved 97.6% accuracy on the Kaggle handwriting dataset [52].

See Table 2.3 for a summary of the related works on Automated Handwriting analysis systems.

The reviewed studies show progress in using handwriting to analyze personality traits, with methods ranging from traditional machine learning to deep learning models. These approaches have demonstrated the potential to extract features from handwriting and predict traits with reasonable accuracy. However, there are still challenges and limitations that need to be addressed to improve the effectiveness of these methods.

Many studies depend on small or custom datasets, such as Mukherjee et al.'s work with 50 participants or Alamsyah et al.'s dataset containing 15 handwriting feature categories. Limited datasets reduce the ability of the models to generalize across different handwriting styles and populations.

Study & Year	Features	Methodology	Dataset	Results
Gahmousse et al., 2021	Edge Hinge (EH) distributions	AdaBoost with majority voting	Custom dataset (285 samples)	Avg. Accuracy: 83.02%
Alamsyah et al., 2022	Entry stroke 'A,' size, slantness	Convolutional Neural Network (CNN)	AND dataset	Accuracy: 80.88%
Mukherjee et al., 2022	Characters ('a', 'g', 'n', 't'), word 'of'; slant, ellipticity, loops, connectivity	Multi-label classification: CC, BR, LP with KSTAR, KNN, MLP	Custom dataset (50 participants)	Best Accuracy: 98.1% (CC-KNN)
Yusof et al., 2022	Slanting, spacing, baselines	Agglomerative Hierarchical Clustering with PCA	70 handwriting samples with Big Five labels	Silhouette score: 0.054
Durga and Deepu, 2022	Document- and character-level: baselines, margins, spacing, 't', 'i'	CNN (DM kernel), self-adaptive ANN	Custom dataset (200 samples)	Avg. accuracy: 86% (Big Five traits)
Navya Shree K S & Siddaraju, 2022	Handwriting features (e.g., slant, margin)	YOLO v5 for detection, ResNet-34 for classification	Custom dataset (3000 images)	F1 scores: YOLO v5: 0.95, ResNet-34: 0.91
Afnan Garoot & Ching Y. Suen, 2022	Handwriting features: slant, size, spacing, baseline, letter curvature, pressure	AvgMlSC (MLSVM+MLCNN with SMOTE)	HWBFF dataset (1066 samples)	Predictive accuracy: 93%, AUC: 0.94, F-Score: 90%
Dhumal et al., 2023 (First Paper)	Signature strokes, structural patterns	Transformer + LSTM	Custom dataset	Accuracy: 96% (outperformed LSTM: 93%)
Samsuryadi et al., 2023	Baselines, margins, spacing, size, slant, pressure	Decision Tree, SVM, KNN (OpenCV)	IAM Handwriting Database	Accuracy: ¿99% (Big Five traits)
Peralta-Rodríguez et al., 2023	Image-based features (no explicit extraction)	U-Net for preprocessing, CNN for classification	HWxPI dataset (418 essays)	Avg. AUC: 0.56, Max: 0.62 (Extraversion)
Dhumal et al., 2023 (Second Paper)	Handwriting patterns and strokes	Hybrid CNN-LSTM, multi-task learning	Custom offline handwriting dataset	Accuracy: 96%
Ahmed et al., 2024	Letter size, slant, pressure, spacing	CNNs (VGG16, DenseNet201, ResNet, InceptionV3)	IAM Database, proprietary dataset	VGG16 accuracy: 73.8%
Xu et al., 2024	Automatic extraction (CNNs)	ConvNeXtTiny, binary cross-entropy loss, Adam optimizer	234 handwriting samples	Best accuracy: 82.90%
Nair et al., 2024	Stroke pressure, letter size, slant, spacing	KNN, SVM, Naive Bayes, Decision Trees, Random Forest	Personal handwriting dataset	SVM: 95% accuracy
Dr. H.K. Chethan et al., 2024	Handwriting attributes: slant, size, pressure	CNNs, ANNs, ResNets, ensemble methods	Kaggle, student samples (1000)	Prediction: Anxious: 85%, Cooperative: 80%, etc.
Maedeh Safar & Ching Y. Suen, 2024	Automatic extraction (VGG16)	k-NN, Random Forest, Logistic Regression, SMOTE, Stacking	1,108 samples (CENPARMI)	Over 90% accuracy for Agreeableness and Openness
Puttaswamy & Thillaiarasu, 2025	Handwriting traits: slants, spacing, font tilting	Fine DenseNet, AMDLSTM-CTC	Kaggle handwriting dataset	Accuracy: 97.6%, F1 Score: 92.67%

Table 2.3: Summary of Automated Handwriting Analysis Studies

Another common issue is class imbalance, where certain personality traits are underrepresented, leading to less accurate predictions for those traits. For example, Afnan Garoot and Ching Y. Suen addressed this problem using SMOTE; however, SMOTE has limitations when applied to high-dimensional data like images, as it may not effectively capture the complex patterns present in such data [53].

Many studies also depend on manual feature extraction, which requires researchers to identify and analyze specific handwriting features such as slants, spacing, or letter characteristics like Yusof et al. and Mukherjee et al. While this approach provides detailed analysis, it is time-consuming and not scalable for large datasets or real-world applications. Automated feature extraction methods are necessary to address these limitations and enhance the efficiency of handwriting analysis models.

Another key limitation is the focus on single-label classification, where traits are treated independently. This simplification misses the complexity of personality analysis, as a single handwriting sample can reflect multiple traits simultaneously at different levels. While studies like Afnan Garoot and Ching Y. Suen have explored multi-label classification, their approach relied on specific graphological feature extraction, which may not fully capture the nuanced variations present in handwriting features that represent all five traits of the BFFM within a single sample.

These challenges emphasize the need for more reliable, automated, and scalable models that can effectively address class imbalance and handle multi-label classification. Developing models that can predict multiple personality traits simultaneously, each with varying levels, would be a significant step forward in this field. This study aims to fill these gaps by offering a more thorough and effective approach to analyzing personality traits from handwriting.

Chapter 3

Data Collection and Analysis

This chapter provides an overview of the dataset used in this study, detailing its structure, the process of collecting handwriting samples and personality scores, and the preprocessing steps applied to prepare the data for analysis. The dataset, initially labeled for two personality traits, is expanded in this thesis to include all five traits of BFFM: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. This expansion was achieved by combining handwriting samples with graphology rules and their corresponding BFFM test scores.

3.1 Dataset Overview

The dataset consists of handwriting samples and personality trait scores collected from 1,110 participants. It includes data gathered through a structured survey conducted at Concordia University and additional samples provided by a professional graphologist. Below, we detail the dataset's structure, the participant demographics, and the collection methodology [22].

- **Survey Participants:** A total of 234 individuals participated in the survey.
- **Graphologist's Dataset:** An additional 876 handwriting samples were collected from 672 individuals by a professional graphologist. These samples maintain the same collection standards and conditions as the survey data.

3.1.1 Data Collection

Survey Design and Structure: The survey was designed to collect handwriting samples and corresponding personality trait scores. It consisted of three main sections:

(1) **Demographics:** This section collected information about participants' age, gender, education, occupation, and nationality. The participants ranged from 18 to 35 years old, with a nearly equal gender distribution (48.69% male, 50.79% female). Regarding education, 39.28% had graduate-level qualifications, 27.23% held a bachelor's degree, 19.37% were in or had completed high school, and 2.09% held a diploma. Most participants were students (72.77%), while 23.04% were employed. The group was internationally diverse, with 30.37% Canadian, 16.23% Iranian, 14.66% Indian, 8.38% Korean, 7.85% Chinese, and 21.99% from 23 other nationalities.

(2) **BFFM Test:** The International Personality Item Pool-Big Five Factor Markers Test (IPIP-BFFMT) was used to measure personality traits.

- The test comprised 50 self-report items rated on a 5-point Likert scale from 1 (Very Inaccurate) to 5 (Very Accurate).
- Participants were instructed to answer honestly based on their current state, with a completion time of 10–20 minutes.

(3) **Graphology Test:** Participants were required to write at least one page of text on unlined letter-sized paper.

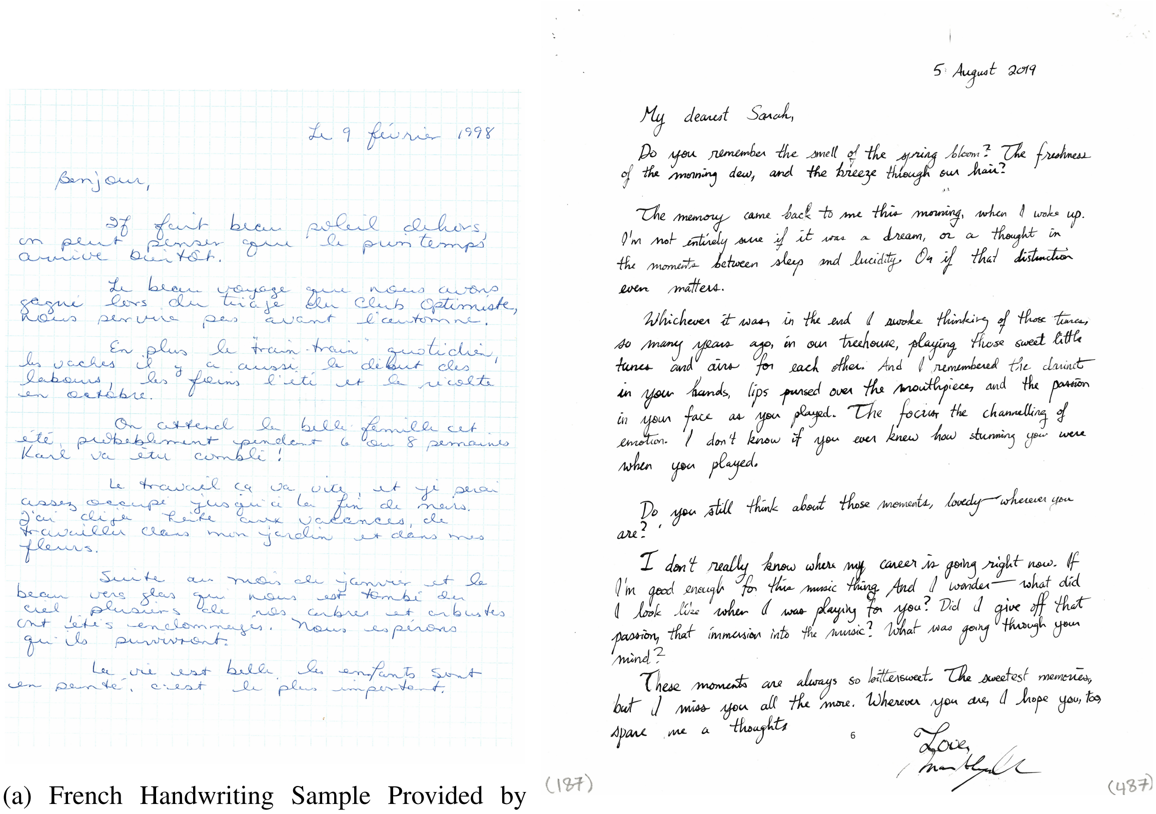
- Instructions emphasized natural handwriting without modification or enhancement.
- Writing was completed in a calm and patient manner to minimize external influences.
- Participants could use any writing tool (e.g., pencil, ballpoint pen, or fountain pen) and any language they preferred.

Recruitment and Rewards: Participants were recruited through posters on Concordia University campuses (approved by Concordia Student Union) and email invitations via the Computer Science department. They received $10 as a reward and a chance to win a $20 Amazon gift card.

Controlled Environment: Data collection occurred in a dedicated laboratory at the Centre for Pattern Recognition and Machine Intelligence (CENPARMI). This controlled environment ensured consistency in handwriting samples by minimizing external factors.

Graphologist's Dataset: To enhance the dataset, a professional graphologist contributed 876 handwriting samples. These samples were collected during her professional practice, where her clients agreed to allow their samples to be used for research and teaching purposes. These samples followed similar conditions and standards as the survey data, ensuring consistency across the dataset.

Figure 3.1 shows two examples of French and English handwriting samples collected from the survey participants and the graphologist.



Le 9 février 1998

Bonjour,

Il fait beau soleil dehors on peut penser que le printemps arrive bientôt.

Le beau voyage que nous avons gagné lors du tirage du Club Optimiste, nous servira pas avant l'automne.

En plus le "train-train" quotidien, les vaches il y a aussi le début des labours, les foins l'été et la récolte en octobre.

On attend la belle famille cet été, probablement pendant 6 ou 8 semaines Karl va être comblé !

Le travail ça va vite, et je serai assez occupé jusqu'à la fin de mars. J'ai déjà hâte aux vacances, de travailler dans mon jardin et dans mes fleurs.

Suite au mois de janvier et le beau verglas qui nous est tombé du ciel plusieurs de nos arbres et arbustes ont été endommagés. Nous espérons qu'ils survivront.

La vie est belle, les enfants sont en santé, c'est le plus important.

(187)

5 August 2019

My dearest Sarah,

Do you remember the smell of the spring bloom? The freshness of the morning dew, and the breeze through our hair?

The memory came back to me this morning, when I woke up. I'm not entirely sure if it was a dream, or a thought in the moments between sleep and lucidity. Or if that distinction even matters.

Whichever it was, in the end I awoke thinking of those times, so many years ago, in our treehouse, playing those sweet little tunes and airs for each other. And I remembered the clarinet in your hands, lips pursed over the mouthpiece, and the passion in your face as you played. The focus the channelling of emotion. I don't know if you ever knew how stunning you were when you played.

Do you still think about those moments, lovedy— wherever you are?

I don't really know where my career is going right now. If I'm good enough for this music thing. And I wonder— what did I look like when I was playing for you? Did I give off that passion, that immersion into the music? What was going through your mind?

These moments are always so bittersweet. The sweetest memories, but I miss you all the more. Wherever you are, I hope you, too spare me a thoughts

6

Love,

(487)

(a) French Handwriting Sample Provided by Graphologist

(b) English Handwriting Sample from Survey

Figure 3.1: Handwriting Samples from Survey Participants and Graphologist

3.1.2 Labeling Process

To predict the Big Five personality traits from handwriting, the graphologist specified handwriting features corresponding to each trait based on its definitions and established graphology rules. Table 3.1 provides an overview of the handwriting features used for each of the five traits: Extraversion, Conscientiousness, Neuroticism, Agreeableness, and Openness to Experience.

Each handwriting sample was manually evaluated by scaling these features. The labeling process involved:

- A 5-point scale was used for most traits: 1 = None or Very Low, 2 = Low, 3 = Average, 4 = High, 5 = Very High.
- For Extraversion, Neuroticism, Agreeableness, and Openness to Experience, each trait was assessed based on five handwriting features.
- For Conscientiousness, the scale comprised four handwriting features.
- For each trait, the individual scores for all corresponding features were averaged to compute the final trait score.

The manual labeling process ensured that the handwriting samples were accurately aligned with graphology principles, providing robust data for model training.

3.1.3 Digitization

The process of digitizing handwriting samples involved converting physical documents into electronic formats suitable for analysis [22].

- **Document Scanning:** Handwriting samples were scanned at a resolution of 600 dpi using an HP Color LaserJet Enterprise M553 series scanner. The scanner's automatic document feeder ensured efficiency and consistency during the digitization process.
- **Image Output:** Scanned documents were saved as high-resolution bitmap images, preserving the details of the original handwriting for accurate feature extraction.
- **Bias Control:** Steps were taken to assess and minimize the impact of factors such as handwriting language, ensuring fair and unbiased data analysis.

Table 3.1: Handwriting Features Corresponding to Each of the Big Five Factors [22]

Factor	**Handwriting Features**
Extraversion	• Middle zone more than 2.5 mm. The middle zone includes most lowercase letters such as a, e, i, o, and u. • Narrow ending margin. Margin refers to spacing around the text page and indentations for paragraphs. • Dominance of garlands: Letters like ”m” and ”n” have a ”u” shape instead of the taught model. • Progressive movement: Often right-slanted, with a high degree of connection. • Slanted in the direction of the writing: Downstrokes angle to the baseline between 85°-45°.
Conscientiousness	• Regularity in slant, dimension, and spacing. • Precision of free stroke placement: ’t’ bars are well-centered, and ’i’ dots align with their stems. • Legibility: This is measured by how clear and readable the handwriting is, even when taken out of context. • Controlled movement: Well-structured forms progressing firmly along the baseline.
Neuroticism	• Regularity without rigidity. • Horizontal and flexible baseline. • Slightly slanted handwriting. • Balance between white space and ink space. • Good pressure and quality of strokes.
Agreeableness	• Dominance of curves over angles. • Adequate spacing between letters, words, and lines. • Letter width greater than 5. • Rounded letters without loops, slightly open. • Nourished and smooth strokes.
Openness to Experience	• Good openness in loops. • Appropriate speed and movement in writing. • Slight angles in letters. • Slanted handwriting in the direction of writing. • Narrow ending margin.

3.2 Dataset Distribution Analysis

In psychology, the Big Five Factor Model test aims to assess the levels of Extraversion, Conscientiousness, Neuroticism, Agreeableness, and Openness to Experience, categorizing them into low, average, or high levels simultaneously. Similarly, in computer science, we aim to develop a supervised learning model capable of performing a classification task for all these traits at once. This approach involves handling instances that can be associated with multiple labels at the same time.

To ensure accurate results, it is essential to examine the dataset's distribution to determine if it is balanced. An imbalanced dataset can lead machine learning models to perform well on the majority class but poorly on the minority class. As shown in Figure 3.2, the majority of handwriting samples are in French (873) and English (181), while other languages have fewer than 20 samples each. This imbalance means the models would primarily learn patterns from French and English handwriting, offering little understanding of the underrepresented languages. Including these minority languages would add minimal value to the study and could even lead to misleading results due to their small sample sizes.

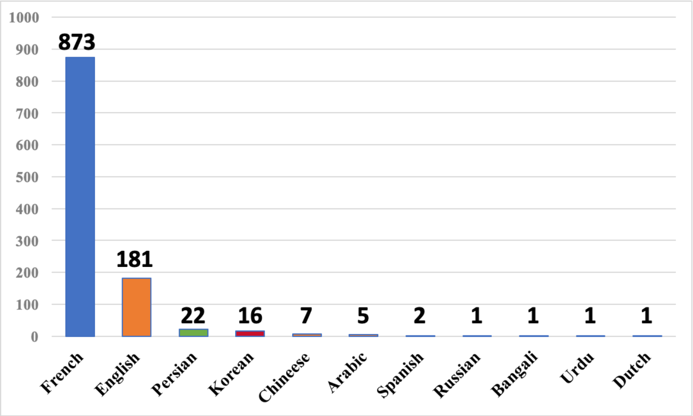
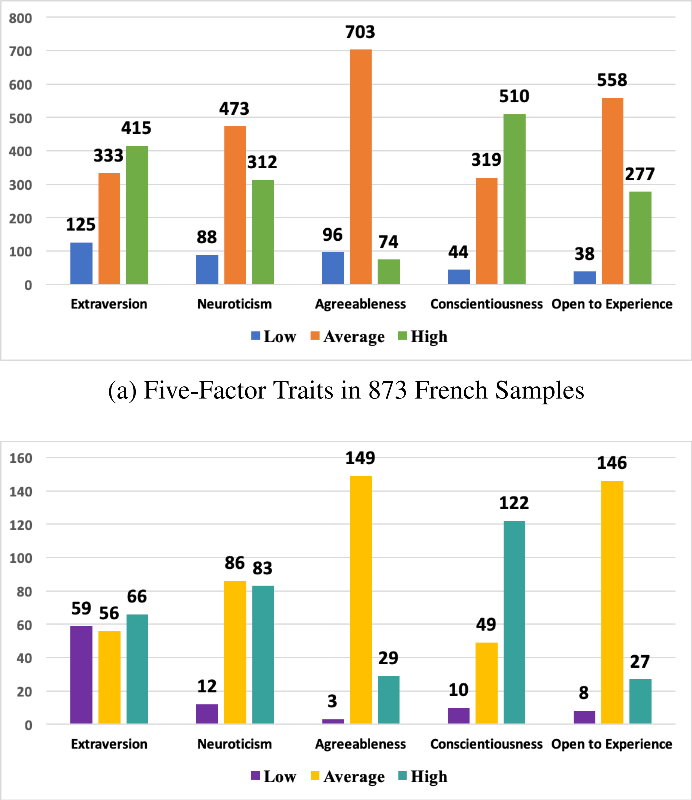


Figure 3.2: Language Distribution of 1110 Handwriting Samples.

To address this, this study focuses on French and English handwriting samples to ensure sufficient data for reliable and meaningful results. A single-label distribution analysis is conducted for each five personality traits in these datasets. This analysis, presented in Figure 3.3, helps determine whether further steps are required to address potential biases in the data.



(a) Five-Factor Traits in 873 French Samples

(b) Five-Factor Traits in 181 English Samples

Figure 3.3: Overview of Five-Factor Personality Traits Distribution

3.3 Imbalance Ratio Assessment

Figure 3.3 visually shows the personality trait distribution in French (873) and English (181) handwriting samples. This study aims to develop a model that can predict all five personality traits at different levels simultaneously. To achieve this, it is important to understand how balanced the dataset is for each trait, as an imbalanced dataset can lead to several challenges.

A common problem with imbalanced datasets is that the model predicts the majority class more often, leading to high accuracy but poor performance on the minority class. This imbalance also reduces the model's ability to generalize to unseen data, especially when the minority class has very few samples [54].

To address these challenges, it is important to first identify whether an imbalance exists and how it impacts the dataset. This involves analyzing class distributions and using statistical measures to assess the imbalance. In this study, since the dataset is multi-class, we calculate the multi-class imbalance ratio (IR_{multi}) for each trait and use the average imbalance ratio to measure the overall balance of the dataset.

The IR_{multi} measures how uneven a dataset is by comparing the proportions of samples across all classes. A perfectly balanced dataset has an IR_{multi} of 1, where all classes have the same number of samples. As the IR_{multi} increases, the smallest class contains far fewer samples than the largest class, making it harder for the model to learn from underrepresented classes. Ideally, a lower IR_{multi} is better because it indicates a more balanced dataset with better representation across all classes.

While specific threshold values for interpreting IR_{multi} can vary across different fields and applications, a general guideline is [55]:

- $\boldsymbol{IR_{\text{multi}} \leq 1.5}$: The dataset is balanced or only slightly imbalanced.
- **$1.5 < \textbf{IR}_{\textbf{multi}} \leq 3$:** The imbalance is moderate and can be addressed with techniques like class weighting or simple oversampling.
- **$\boldsymbol{IR_{\textbf{multi}} > 3}$:** The imbalance is significant and may require advanced methods to handle effectively.

The goal is to reduce IR_{multi} as much as possible, as a high imbalance ratio makes it challenging for the model to learn effectively from smaller classes. The IR_{multi} is calculated using the following formula, based on the approach outlined in [56]. For each class (Low, Average, High), the proportion of samples $P(\text{Class})$ is calculated as:

$$P(\text{Class}) = \frac{\text{Samples in Class}}{\text{Total Samples}}$$

This provides the relative proportion of each class within the total samples for each trait. The maximum and minimum class proportions ($\max$ and $\min$) are identified, and the IR_{multi} is calculated:

$$IR_{\text{multi}} = \frac{\max(P(\text{Low}), P(\text{Average}), P(\text{High}))}{\min(P(\text{Low}), P(\text{Average}), P(\text{High}))}$$

To evaluate the overall imbalance across all traits, the average multi-class imbalance ratio is calculated as:

$$\text{Average } IR_{\text{multi}} = \frac{\sum IR_{\text{multi (traits)}}}{\text{Number of Traits}}$$

The calculated IR_{multi} for the English dataset is shown in Table 3.2, and the results for the French dataset are presented in Table 3.3.

Trait	**Low (n)**	**Avg (n)**	**High (n)**	**Proportions (Max/Min)**	**$\mathbf{IR_{multi}}$**
Extraversion	59	56	66	0.3646 / 0.3094	1.18
Neuroticism	12	86	83	0.4751 / 0.0663	7.17
Agreeableness	3	149	29	0.8232 / 0.0166	49.62
Conscientiousness	10	49	122	0.6740 / 0.0552	12.22
Open to Experience	8	146	27	0.8066 / 0.0442	18.25
				Average $\mathbf{IR_{multi}}$	**17.69**

Table 3.2: Multi-Class Imbalance Ratio Calculation for the English Dataset

Trait	**Low (n)**	**Avg (n)**	**High (n)**	**Proportions (Max/Min)**	**$\mathbf{IR_{multi}}$**
Extraversion	125	333	415	0.4756 / 0.1432	3.32
Neuroticism	88	473	312	0.5421 / 0.1008	5.38
Agreeableness	96	703	74	0.8057 / 0.0848	9.50
Conscientiousness	44	319	510	0.5844 / 0.0504	11.59
Open to Experience	38	558	277	0.6395 / 0.0435	14.69
				Average $\mathbf{IR_{multi}}$	**8.89**

Table 3.3: Multi-Class Imbalance Ratio Calculation for the French Dataset

The analysis of IR_{multi} shows a clear difference in class balance between the English and French handwriting datasets. The English dataset has an average IR_{multi} of approximately 17.69, indicating a significant imbalance with some classes being heavily underrepresented. In contrast, the French dataset has a lower average IR_{multi} of about 8.89, showing a more balanced distribution of samples across the traits. Based on these results, we started our study by focusing on the French dataset because it has more samples and a lower imbalance ratio, making it a better starting point for developing and evaluating techniques. To assess the impact of language on the model's performance, we then will apply the same techniques to the English dataset. Finally, we will combine both datasets to analyze the results and understand how the model performs when using data from both languages.

The methods used in this study to address the imbalance challenge are evaluated using the same imbalance ratio formula. This ensures that the techniques are assessed consistently, determining whether they improve the dataset's balance and enhance the model's performance across all traits.

Chapter 4

Methodology

This chapter outlines the methodology used in this research to process handwriting data for personality trait prediction. It covers preprocessing techniques aimed at improving data quality and increasing the number of handwriting samples, particularly for underrepresented classes, as discussed in the previous chapter. To address class imbalance, the focal loss is employed, assigning greater importance to minority classes during training.

Additionally, the chapter introduces the optimization techniques used to enhance model performance and explains the classification approach adopted in this study. Transformer OCR is presented as the primary deep-learning model for handwriting feature extraction. Finally, the evaluation metrics such as accuracy, precision, recall, and AUROC are described to assess the effectiveness of the models.

4.1 Preprocessing Techniques

Preprocessing and feature extraction are important initial steps in preparing handwriting data for personality trait prediction, particularly in this study, where deep learning models are employed to automatically learn patterns from handwriting samples. For these models to be effective, it is essential to have a sufficient number of training samples while preserving all meaningful handwriting patterns.

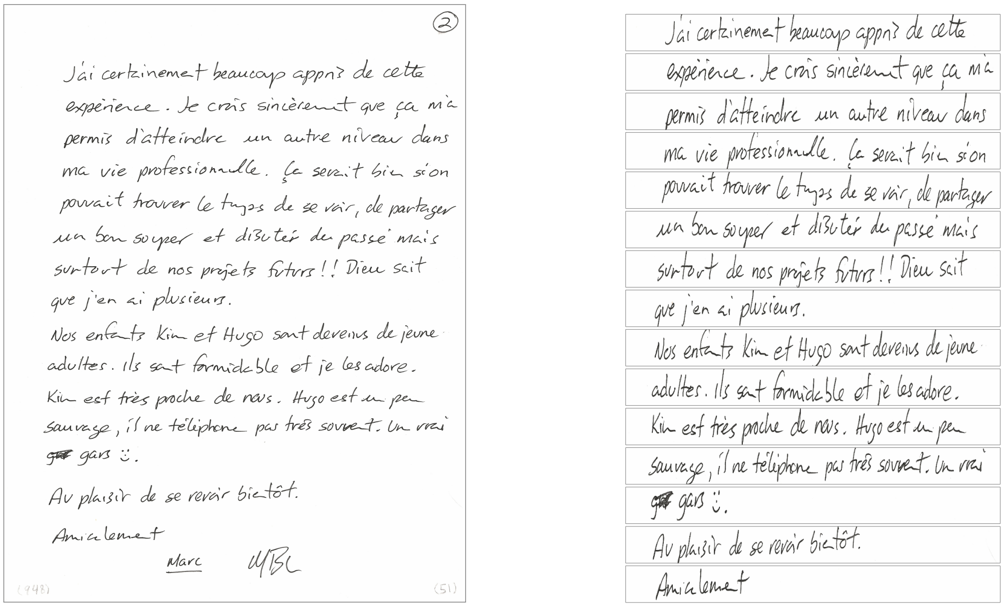
To achieve this, handwriting features are categorized into two types: global features, which capture the overall handwriting style, such as margins, line spacing, and slant; and local features, which focus on finer details, including stroke pressure, letter connections, and the shapes of individual letters. We focused on techniques that preserve both global and local features while increasing the number of samples to ensure the model comprehensively learns and identifies patterns associated with personality traits.

4.1.1 Line Segmentation

A key preprocessing technique used in this study is line segmentation, which increases the dataset size, reduces class imbalance, and ensures that both local and global handwriting features are preserved. Some advantages of this technique are as follows:

- **Increasing Dataset Size:** Splitting handwritten documents into lines significantly increases the number of samples available for training, which is especially useful when working with smaller datasets.
- **Reducing Class Imbalance:** Generating more samples from underrepresented handwriting classes helps create a more balanced dataset, improving the model's performance on less frequent traits.
- **Preserving Handwriting Features:** Ensures that both global features that represent the writer's overall and local features that capture the writer's unique style handwriting features are preserved.
- **Simplified Processing:** Analyzing individual lines simplifies the data processing pipeline and ensures uniform input dimensions, making it easier to train deep learning models effectively.

The process is implemented on both French and English datasets using OpenCV, a widely used library for computer vision tasks. Contour detection was applied to identify and isolate individual lines of text by detecting their boundaries. Bounding rectangles were applied around each line to extract them accurately. Padding with a white background was added to each image to ensure uniformity, keeping the dimensions consistent. After extraction, the dataset was manually cleaned to remove irrelevant lines, such as signatures or numbers, providing a refined and usable dataset. This simple and effective approach preserved the structure of the lines and prepared them for further processing. The results of this process are shown in Figures 4.1 and 4.2.


②
J'ai certainement beaucoup appris de cette
expérience. Je crois sincèrement que ça m'a
permis d'atteindre un autre niveau dans
ma vie professionnelle. Ce serait bien si on
pouvait trouver le temps de se voir, de partager
un bon souper et discuter du passé mais
surtout de nos projets futurs!! Dieu sait
que j'en ai plusieurs.
Nos enfants Kim et Hugo sont devenus de jeune
adultes. Ils sont formidable et je les adore.
Kim est très proche de nous. Hugo est un peu
sauvage, il ne téléphone pas très souvent. Un vrai
gars :).
Au plaisir de se revoir bientôt.
Amicalement
Marc
(51)

(a) Handwriting Sample Before Segmentation

(b) Handwriting Samples After Line Segmentation

Figure 4.1: French Handwriting Sample and Line Segmentation Results

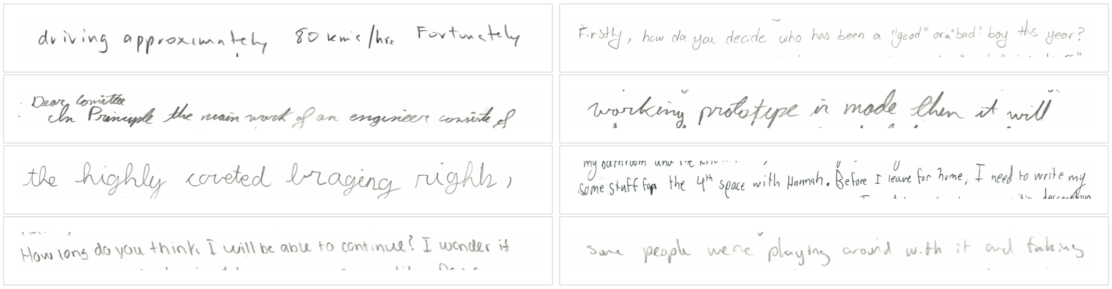
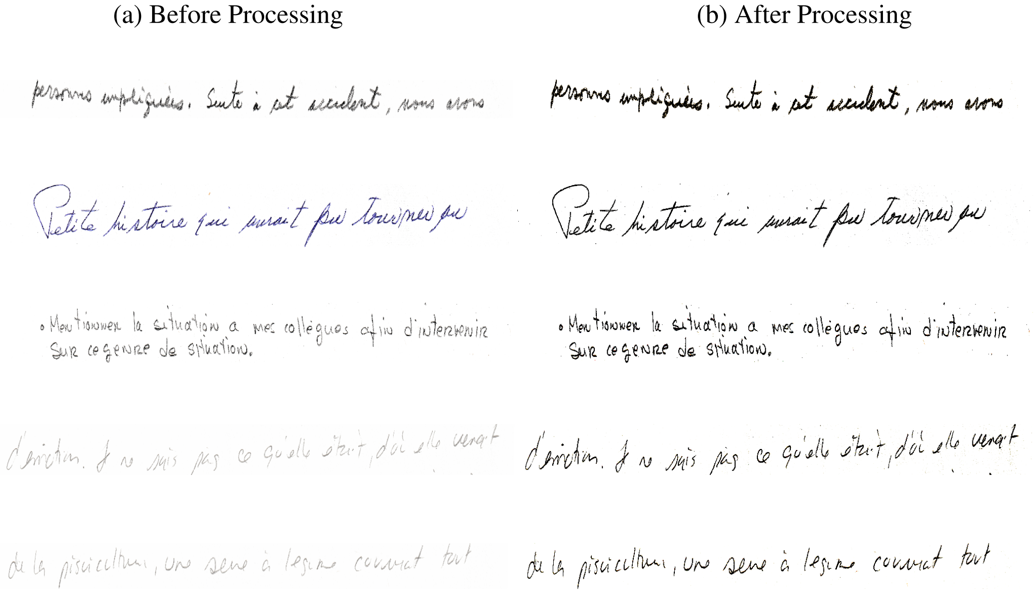

driving approximately 80 km/hr. Fortunately
Firstly, how do you decide who has been a "good" or a "bad" boy this year?
Dear Committee
In Principle the main work of an engineer consists of
working prototype is made then it will
the highly coveted bragging rights,
Some stuff for the 4th space with Hannah. Before I leave for home, I need to write my
How long do you think I will be able to continue? I wonder if
Some people were playing around with it and taking

Figure 4.2: Different Examples of English Handwriting Sub-samples

4.1.2 Image Processing

As handwriting samples are digitized, several image processing techniques are explored to optimize the quality of handwriting images for analysis. Among these, Otsu's binarization method combined with bilateral filtering proved the most effective approach. Otsu's binarization simplifies the images by converting them to black and white using an automatically determined optimal threshold, effectively separating the foreground (handwriting) from the background. This method enhances the clarity of the text, making it easier for the model to identify key features [29].

To complement this, bilateral filtering is applied to reduce noise while preserving important structural details such as edges and fine strokes [57]. This step ensures that the handwriting retains its essential characteristics, which are important for accurate classification (Figure 4.3).



(a) Before Processing

(b) After Processing

Figure 4.3: Comparison of Before and After Image Processing

These processed images are then converted into a three-channel format to align with the input requirements of neural networks, which are typically designed to process RGB images.

4.1.3 Data Augmentation

Although image segmentation helped to expand our dataset, the limited number of handwriting samples still required additional data augmentation to train neural networks effectively, a large dataset is essential for learning meaningful patterns [58]. To address this, we carefully selected augmentation techniques designed to introduce realistic variations in handwriting that the model is likely to encounter in real-world scenarios. These techniques include:

- **Random Rotations and Flips:** Simulate natural variations in how individuals might rotate or flip their writing.

- **Affine Transformations:** Imitate different handwriting scales and distortions, such as stretching or compressing, common in handwritten notes.

- **Perspective Changes and Blurring:** Adjust the viewing angle and simulate slight blurring when handwriting is photographed or scanned.

- **Random Erasing:** Introduces small areas of missing data, encouraging the model to focus on the most informative parts of handwriting.

Together, these augmentations help create a reliable dataset that trains the neural network to be adaptive and accurate in analyzing diverse handwriting styles (Figure 4.4).

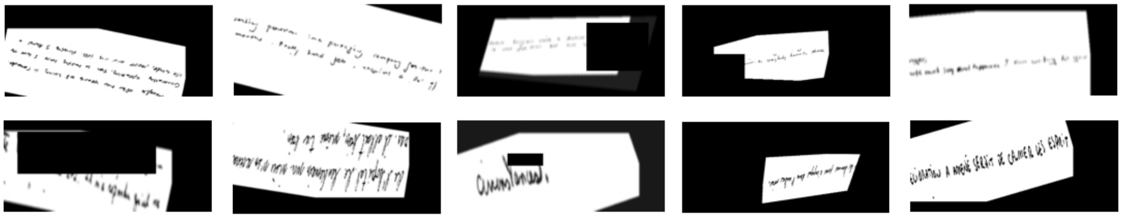


Figure 4.4: Data Augmentation Samples

4.2 Focal Loss

Focal Loss is introduced to address the class imbalance by extending the standard cross-entropy loss [59]. It achieves this by incorporating a modulating factor to reduce the contribution of well-classified examples and a balancing factor to manage differences in class frequencies [60]. The formula for Focal Loss is as follows:

$$\text{Focal Loss} = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{1}$$

In Eq. 1, p_t represents the predicted probability of the true class, defined as $p_t = p$ if the true label $t = 1$, and $p_t = 1 - p$ if $t = 0$. The parameter α_t balances the loss between classes, and γ focuses the model's attention on difficult examples by reducing the loss from samples with high p_t. A higher γ further decreases the influence of easy examples.

The algorithm for Focal Loss, outlined in Algorithm 1, processes each sample by computing the weighted loss based on p_t, the modulating factor $(1 - p_t)^\gamma$, and the balancing weight α_t. The losses for all samples are then accumulated and normalized by the total number of samples to produce the final loss value [61]. This approach allows the model to prioritize learning from challenging cases, making it highly effective for tasks with significant class imbalance [62].

Algorithm 1 Focal Loss

```
Require: Predicted probabilities $p$, true labels $y \in \{0, 1\}$, focusing parameter $\gamma \geq 0$, balancing factor $\alpha \in [0, 1]$
Ensure: Focal Loss $FL$
Initialize $FL \leftarrow 0$
for each sample $i$ in the dataset do
    Compute $p_t$:
        $$p_t = \begin{cases} p_i & \text{if } y_i = 1, \\ 1 - p_i & \text{if } y_i = 0 \end{cases}$$
    Compute the modulating factor: $modulating\_factor \leftarrow (1 - p_t)^\gamma$
    Compute the balanced cross-entropy weight: $weight \leftarrow \alpha$ if $y_i = 1$, else $1 - \alpha$
    Update focal loss for the sample:
        $$FL_i \leftarrow -weight \cdot modulating\_factor \cdot \log(p_t)$$
    Accumulate: $FL \leftarrow FL + FL_i$
end for
Normalize the loss: $FL \leftarrow \frac{FL}{N}$ where $N$ is the total number of samples
return $FL$
```

4.3 Optimization

Selecting the appropriate optimizer is an important aspect of training deep neural networks for handwriting analysis. In our observation, this factor has not been mentioned in previous studies, despite its importance in addressing challenges such as imbalanced datasets, varying writing styles, complex patterns, noise, and multiple labels per sample. To evaluate their impact on performance, three state-of-the-art optimizers are used: SGD (Stochastic Gradient Descent) with Momentum, Adabelief (Adaptive Belief), and ADAM (Adaptive Moment Estimation). The selection of each optimizer is based on its distinct advantages, and they are evaluated for their effect on the model's performance, particularly in handling the training dynamics specific to handwriting analysis.

- **SGD with Momentum**: It is selected for its ability to accelerate convergence and help the model escape local minima. The momentum term smooths out the gradient descent path, enabling more stable and faster convergence, which is useful for handling complex handwriting data [63].

- **AdaBelief**: This is used for its ability to adapt the learning rate based on the variance of the gradients, similar to Adam, but with improved stability and convergence. By incorporating a belief about the gradient's direction that computes adaptive learning rates for each parameter, AdaBelief leads to faster convergence and better generalization, making it effective for training on imbalanced and complex handwriting datasets [64].

- **Adam**: It is selected for its adaptability and efficiency in handling sparse gradients and noisy data. Additionally, it dynamically adjusts the learning rate for each parameter, which benefits datasets with varying patterns. Its initial learning rate and minimal tuning requirements make it robust for deep learning tasks [65].

These optimizers are employed to identify the most suitable ones for achieving a balance between convergence speed, stability, and accuracy. This enhances the model's capability to classify personality traits from handwriting data.

4.4 Classification

Since the dataset structure includes three classes for each personality trait: low, average, and high, two classification approaches, multi-class and multi-label binary classification, are evaluated. These approaches are analyzed to determine which method better handles imbalanced data and improves the model's ability to accurately classify traits with fewer samples. The goal is to identify the approach that provides more balanced learning and enhances classification performance across all personality traits.

4.4.1 Multi-Class Classification

The multi-class classification is used to predict each personality trait as a separate task. Each trait is classified into three mutually exclusive levels: low, average, and high. This setup requires the model to learn the distinct features associated with each level for all five traits.

To handle this classification, the model is designed with five independent classification heads, one for each personality trait. Each classification head outputs logits, raw outputs from the network before activation, (z_t) corresponding to the three levels (low, average, and high) for trait t. These logits are passed through a softmax activation function to convert them into probabilities [66]:

$$P(y_t = i) = \frac{\exp(z_{t,i})}{\sum_{j=1}^{3} \exp(z_{t,j})}, \quad i \in \{1, 2, 3\}$$

Where $z_{t,i}$ is the logit output for class i of trait t, and $P(y_t = i)$ represents the predicted probability of level i for trait t.

The model is trained using the cross-entropy loss function, which measures the difference between the predicted probabilities and the true labels. For a single sample, the loss for a given trait t is:

$$L_t = -\sum_{i=1}^{3} y_{t,i} \log(P(y_t = i))$$

where $y_{t,i}$ is the one-hot encoded true label for class i and $P(y_t = i)$ is the predicted probability for class i. The total loss for all five traits is the sum of the individual trait losses:

$$L_{\text{total}} = \sum_{t=1}^{5} L_t$$

In this classification, oversampling is implemented alongside sample weighting based on class frequencies. This approach increases the presence of underrepresented classes during training and encourages the model to learn their features effectively, enhancing its ability to differentiate among traits [67].

4.4.2 Multi-Label Binary Classification

Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) combines the sigmoid activation function with binary cross-entropy loss. This combination transforms raw logits from the neural network's final layer into probabilities, then calculates the binary cross-entropy for each label. This approach models each label as an independent Bernoulli distribution. The loss function is defined as:

$$\text{BCEWithLogitsLoss}(z, y) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log(\sigma(z_i)) + (1 - y_i) \cdot \log(1 - \sigma(z_i))\right)$$

Where z_i represents the logits, $\sigma(z_i) = \frac{1}{1+e^{-z_i}}$ is the sigmoid function applied to the logits, y_i is the target label (0 or 1), and N is the number of samples or batch size. By applying the sigmoid function, logits are converted into probabilities, allowing the model to independently predict the presence or absence of each label.

This classification model includes 15 binary classification heads, one for each level across the five personality traits, enabling the simultaneous prediction of multiple labels per instance, as instances can belong to multiple classes. The BCEWithLogitsLoss function is applied independently to each classification head, allowing the model to learn the features of each level without being influenced by the distribution of other levels.

4.5 Model Development

This section presents an overview of the deep learning models employed for automatic handwriting feature extraction. The Transformer OCR model is introduced as a new approach specifically adapted for classification tasks in this study. Its performance is evaluated against three pre-trained models: ResNet50 and Vision Transformer (ViT) base 16 at two input resolutions (224 × 224 and 384 × 384).

4.5.1 CNN Architecture

Convolutional Neural Networks (CNNs) are well-known for their success in extracting meaningful features from images due to their ability to handle spatial data processing effectively [68].

In this study, ResNet50 is chosen for its strong feature detection capabilities, making it suitable for handwriting analysis. It is pre-trained on ImageNet and uses residual blocks to help train deeper networks without running into vanishing gradient problems.

- **ResNet50:** This 50-layer network offers a good balance between accuracy and computational efficiency. It is a practical choice for handwriting feature extraction, especially in resource-limited environments, where it performs well compared to more demanding models like transformers [69].

4.5.2 Vision Transformer: ViT base 16

ViT is a transformer-based model designed specifically for computer vision tasks. Unlike CNNs, which use convolutional filters to detect local patterns, ViT divides an image into patches, treats each patch as a token, and processes these tokens using a transformer encoder. This architecture enables ViT to capture global relationships within an image, making it highly suitable for handwriting analysis [70].

To evaluate its effectiveness at different scales and examine the trade-off between computational efficiency and the ability to capture handwriting features, this study considers two configurations of ViT base 16 with input resolutions of 224×224 and 384×384. Both configurations are pre-trained on the ImageNet-21k dataset and share the same transformer-based architecture, providing consistency in feature extraction for comparison with the proposed TrOCR model [71].

- **ViT base 16-224:** This model uses an input resolution of 224×224. This configuration is chosen for its computational efficiency and ability to capture general handwriting features. It provides a good balance between processing speed and feature representation [72].
- **ViT base 16-384:** This model operates with a higher input resolution of 384×384, which enhances the model's ability to capture more detailed handwriting features. This configuration allows the analysis of finer spatial patterns, making it better suited for tasks requiring higher precision. However, it comes at the cost of increased computational demands [72].

4.5.3 Transformer OCR

TrOCR, or Transformer Optical Character Recognition, is a transformer-based model developed by Microsoft specifically for OCR applications. Unlike traditional OCR systems that rely on CNNs for image processing and RNNs for sequential text generation [73], TrOCR is designed as an end-to-end transformer model that integrates a ViT encoder, initialized with BEiT weights for image encoding, and a RoBERTa-based text decoder for autoregressive text generation.

The encoder processes images by dividing them into 16x16 fixed-size patches, embedding each patch as a sequence token, and using absolute positional embeddings to retain spatial information [74]. This architecture effectively captures local and global features within an image, demonstrating state-of-the-art performance for OCR tasks like printed and handwritten text recognition without requiring complex pre- or post-processing steps [75].

The TrOCR model is trained on ImageNet-1k for its image encoder and fine-tuned on the IAM handwriting dataset. In this study, we adapted TrOCR to our handwriting dataset and proposed a new approach that modifies the model for personality trait classification instead of text generation (Figure 4.5).

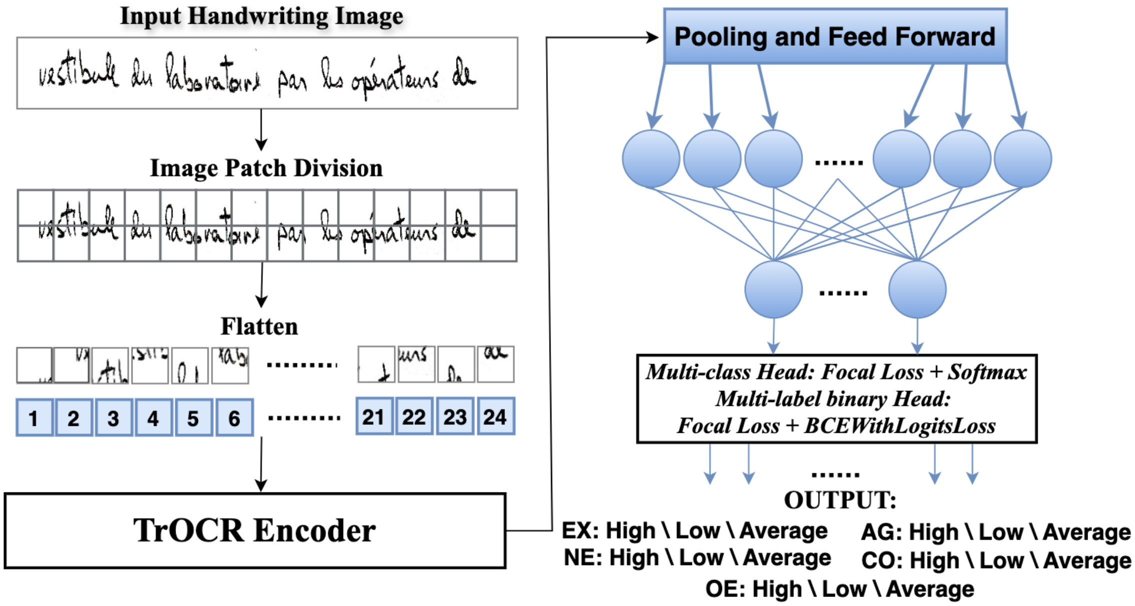


Figure 4.5: The Proposed TrOCR Model for Classification

In our approach, the handwriting image is first divided into smaller, non-overlapping patches to capture local features such as stroke patterns, letter shapes, and other distinctive handwriting characteristics. Each of these patches is then flattened into 1D vectors, converting the 2D spatial information into a sequential format that can be processed by the model. These flattened vectors are combined to form a sequence of visual tokens, which are fed into the TrOCR encoder based on the Vision Transformer (ViT) architecture.

Within the encoder, self-attention mechanisms analyze the relationships between these tokens to extract high-level representations of handwriting features. After this, a pooling layer is applied to aggregate information from these tokens, summarizing the most relevant features while reducing the dimensionality of the data. The pooled outputs are then passed through a series of feed-forward neural network layers to further transform and refine the extracted features. This step enhances the model's ability to capture complex handwriting patterns and relationships between different parts of the text.

Instead of using the original text decoder for text generation, we replace it with a custom classification head designed to predict personality traits. This classification head is divided into two parts: a multi-class head with Softmax activation for predicting mutually exclusive classes and a multi-label binary head with BCEWithLogitsLoss for independently classifying each trait as a binary task. To handle class imbalance effectively, Focal Loss is applied to both components, enabling the model to focus more on challenging and minority samples [76].

This adaptation highlights TrOCR's flexibility, showing that it can go beyond OCR tasks to handle complex classification. The model's transformer-based design captures detailed handwriting features, making it useful for analyzing personality traits from handwriting images.

4.6 Evaluation Metrics

Five fundamental metrics are used to evaluate our models' performance: accuracy, the area under the receiver operating characteristic (AUROC), F1-score, precision, and recall. These metrics are defined based on the content of the confusion matrix for each class i [77]:

- True Positives (TP$_i$): Correctly predicted instances of class i.
- True Negatives (TN$_i$): Correctly predicted instances that are not class i.
- False Positives (FP$_i$): Instances incorrectly predicted as class i.
- False Negatives (FN$_i$): Instances of class i incorrectly predicted as another class.

- **Accuracy:** It is the ratio of correct predictions to total predictions. The overall accuracy is calculated based on Eq. 2:

$$\text{Accuracy} = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{2}$$

While accuracy is useful, it can be misrepresented by class imbalances, making precision, recall, F1-score, and AUROC important for a comprehensive evaluation.

- **Precision:** It is defined as the proportion of true positive predictions out of all positive predictions by Eq. 3:

$$\text{Precision} = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

High precision indicates a low false positive rate, which is important in this context to ensure that traits are not misclassified as other traits. Precision is particularly valuable for evaluating the model's performance on minority classes, where false positives could have a more significant impact.

- **Recall, or sensitivity:** measures the proportion of true positive predictions out of all actual positives according to Eq. 4:

$$\text{Recall} = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

High recall means the model effectively identifies the target class, minimizing false negatives. This metric is essential for ensuring that the model accurately detects all personality traits, especially those with fewer samples.

- **The F1-score:** It balances recall and precision and the overall F1-score can be calculated as a weighted average of the F1-scores for each class by Eq. 5.

$$F1_i = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

- **The AUROC curve:** This curve represents a two-dimensional plot of the true positive rate (recall on the y-axis) against the false positive rate (x-axis). An AUROC close to 1 signifies a perfect model. In this study, AUROC is calculated separately for each trait to assess the model's performance in differentiating between levels(low, average, high) within each trait which allows for a detailed evaluation of the model's strengths and weaknesses in classifying handwriting features linked to different personality traits.

Chapter 5

Experimental Result and Discussion

This chapter presents and discusses the experimental results obtained in this study. The dataset is split into 60% training, 20% validation, and 20% testing, and each model is trained for 100 epochs. The effectiveness of focal loss and segmentation in addressing class imbalance is highlighted. The performance of three state-of-the-art optimizers is evaluated to improve model training, and the impact of multi-class and multi-binary classification heads is examined. The primary deep learning model, Transformer OCR, is analyzed and compared with ResNet50 and Vision Transformer using evaluation metrics such as accuracy, precision, recall, and AUROC to assess and compare model performance. The results are visualized using figures and tables for clarity.

5.1 Segmentation

The comparison of $\mathrm{IR}_{\mathrm{multi}}$ values before and after segmentation clearly shows how segmentation technique helped address the imbalance in the English and French handwriting datasets. Before segmentation, the English dataset had a very high average $\mathrm{IR}_{\mathrm{multi}}$ of 17.69, indicating severe imbalance with some classes having very few samples. The French dataset, though slightly better, still had an average $\mathrm{IR}_{\mathrm{multi}}$ of 8.8, highlighting the need for improvement.

After segmentation and manual data cleaning, which involved removing unnecessary sub-samples such as signatures and numbers, the dataset expanded to 5,765 sub-samples for French and 1,807 sub-samples for English, significantly improving the dataset size. The English dataset's average

IR$_{\text{multi}}$ dropped to 6.02, a noticeable reduction that shows better class representation (Table 5.1). Similarly, the French dataset's average IR$_{\text{multi}}$ decreased to 5.81, reflecting a more balanced distribution across classes (Table 5.2). These improvements demonstrate how segmentation effectively increased the number of samples in each class, particularly addressing the under-representation in smaller classes.

Trait	**Low (n)**	**Avg (n)**	**High (n)**	**Proportions (Max/Min)**	**IR$_{\text{multi}}$**
Extraversion	649	528	630	0.3593 / 0.2922	1.23
Neuroticism	134	903	770	0.4997 / 0.0742	6.74
Agreeableness	157	1348	302	0.7459 / 0.0869	8.58
Conscientiousness	164	362	1281	0.7091 / 0.0907	7.81
Open to Experience	228	1303	276	0.7211 / 0.1262	5.71
				Average IR$_{\text{multi}}$	**6.02**

Table 5.1: Number of Class Samples and IR$_{\text{multi}}$ for English After Segmentation

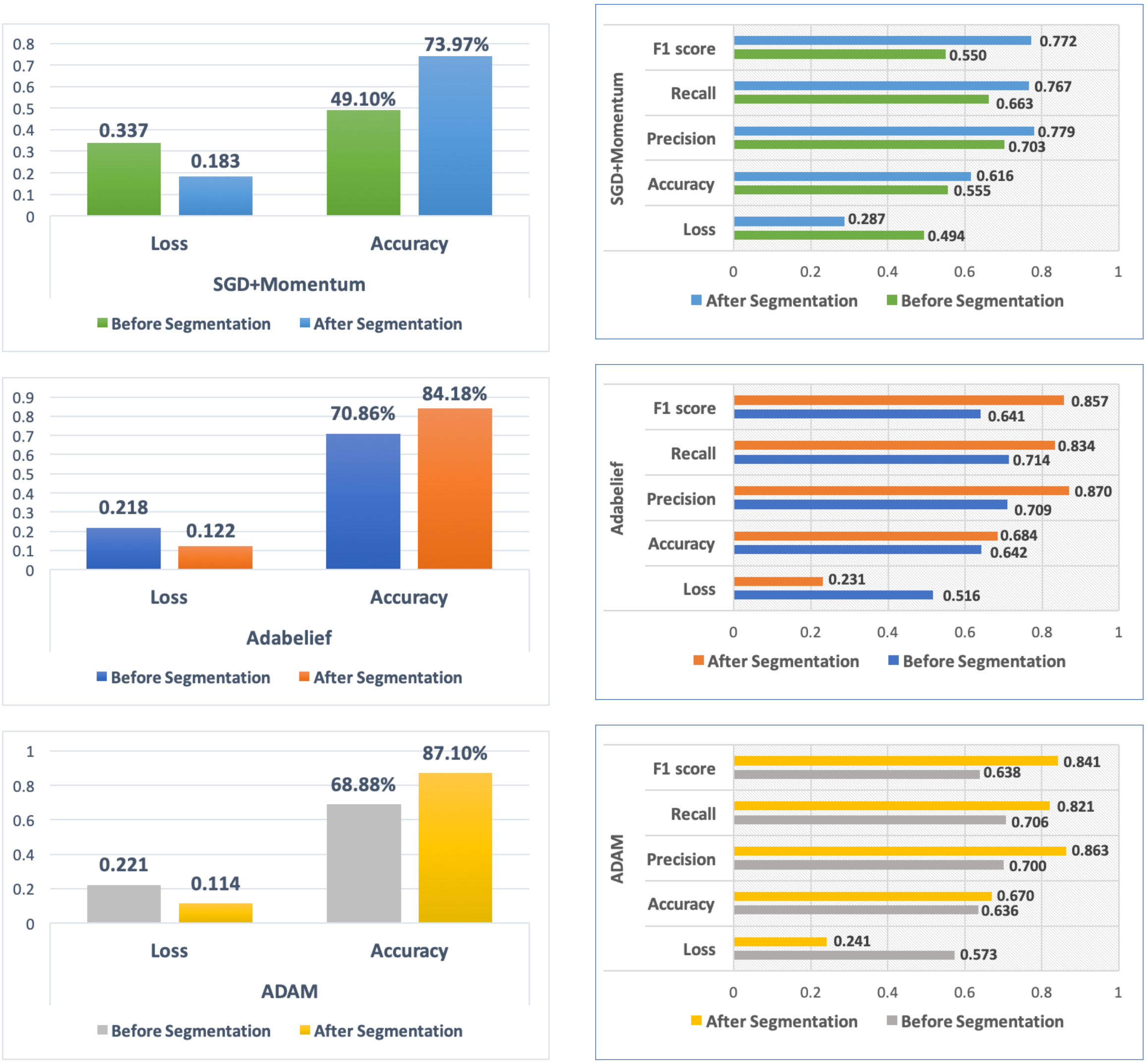
Trait	**Low (n)**	**Avg (n)**	**High (n)**	**Proportions (Max/Min)**	**IR$_{\text{multi}}$**
Extraversion	1058	2287	2420	0.4056 / 0.2531	1.60
Neuroticism	392	3019	2354	0.4571 / 0.1000	4.57
Agreeableness	512	4569	684	0.7122 / 0.1306	5.45
Conscientiousness	233	1803	3729	0.6133 / 0.0594	10.33
Openness to Experience	339	3765	1661	0.6150 / 0.0865	7.11
				Average IR$_{\text{multi}}$	**5.81**

Table 5.2: Number of Class Samples and IR$_{\text{multi}}$ for French After Segmentation

After it is confirmed that segmentation improves the imbalance ratio based on the data presented in Table 5.2, ResNet50 is selected as the base model to evaluate the impact of segmentation on performance. This model is chosen for its balance of computational efficiency and effectiveness, making it suitable for experimentation. The evaluation is conducted on both the original and segmented French dataset, which has more samples and a slightly improved imbalance ratio, using three optimizers: SGD with momentum, AdaBelief, and Adam.

The results in Figure 5.1 clearly show the significant positive impact of segmentation on model performance across various metrics during both the training and testing phases. After segmentation, all models (Adam, AdaBelief, and SGD with momentum) exhibit substantial improvements in accuracy, precision, recall, F1-score, and a reduction in loss. For instance, Adam's loss decreases from 0.221 to 0.114, while its accuracy increases from 68.88% to 87.10%. Similarly, AdaBelief's accuracy improves from 70.86% to 84.18%, and SGD with momentum achieves a remarkable accuracy increase from 49.10% to 73.97%. Precision, recall, and F1-scores consistently improve across all models, highlighting the critical role of segmentation in isolating key features and minimizing noise.

On average, segmentation improves accuracy by approximately 18.8% and reduces loss by 43.5%, confirming its effectiveness in enhancing data quality. These results validate segmentation as an essential preprocessing step, enabling models to generalize more effectively and extract meaningful patterns from the dataset.



(a) Training Results with Different Optimizers.

(b) Test Results with Different Optimizers.

Figure 5.1: Performance of ResNet50 Before and After Segmentation.

5.2 Impact of Focal Loss

In the initial phase of this study, the dataset was labeled for two personality traits: Extraversion and Conscientiousness, and categorized into three classes: low, average, and high. The impact of focal loss is evaluated on this configuration, and the results presented in this section are based on this setup. Through experimentation with various parameter values, the optimal focal loss parameters are determined as $\alpha = 1$ and $\gamma = 2$. These parameters enable the model to prioritize harder examples, improving robustness and mitigating bias toward more frequent labels.

The impact of focal loss is first assessed on the original dataset without segmentation, which has an imbalance ratio of 8.89. Without focal loss, the validation accuracy achieves only 30.94%. However, with the inclusion of focal loss, the validation accuracy significantly improves to 76.79%. This demonstrates that focal loss effectively addresses class imbalance by penalizing easy-to-predict samples and emphasizing harder examples.

For the segmented dataset, where performance is already discussed in Section 5.1, focal loss further enhances model training and generalization. The training curves for all three optimizers (SGD with momentum, AdaBelief, and Adam), shown in Figure 5.2, illustrate faster convergence and improved stability when focal loss is applied compared to training without it.

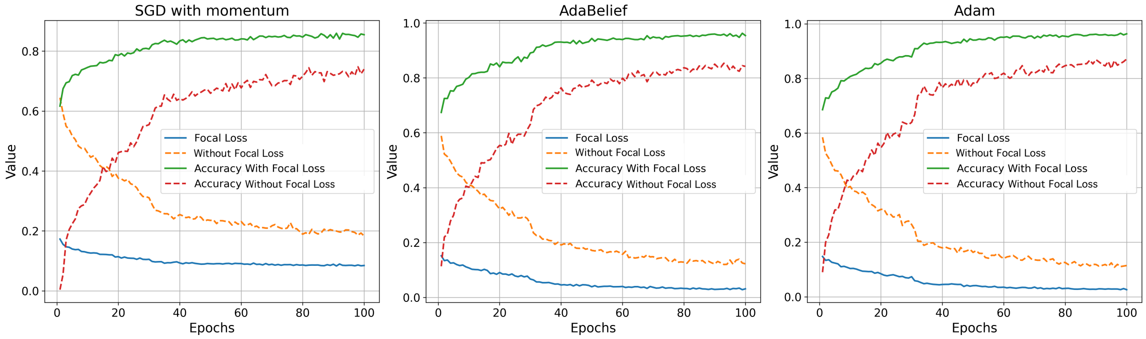
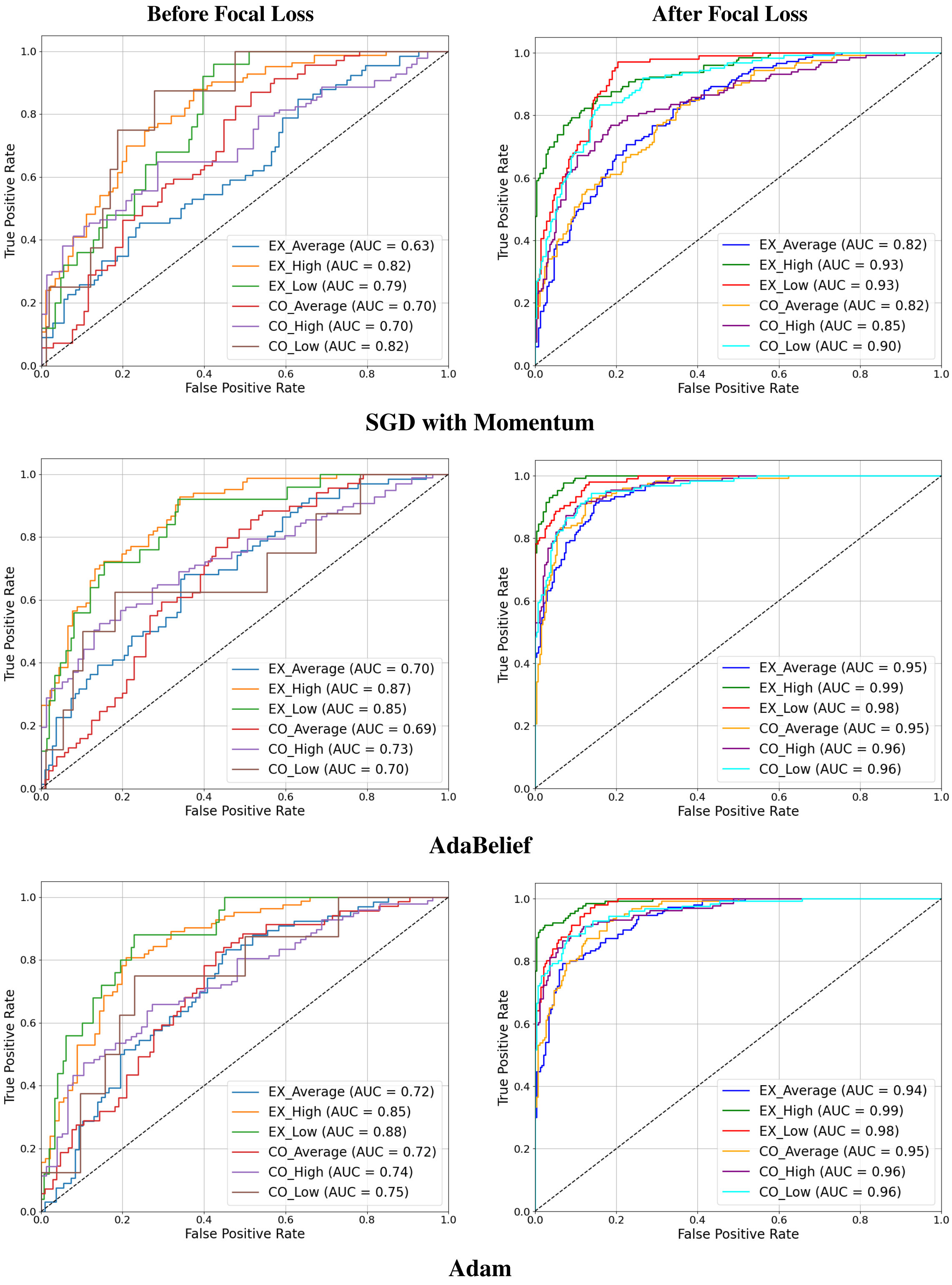


Figure 5.2: Training Curves of ResNet50: Comparison with and without Focal Loss

Figure 5.3 illustrates the performance of each optimizer in ResNet50 on unseen data and highlights the significant impact of focal loss on AUROC scores. Before applying focal loss (left column), the models struggle to differentiate personality traits, resulting in lower AUROC values. After focal loss (right column), a noticeable improvement is observed across all optimizers.



Before Focal Loss

After Focal Loss

SGD with Momentum

AdaBelief

Adam

Figure 5.3: Impact of Focal Loss on AUROC Performance of Optimizers in the Test Phase. (EX: Extraversion, CO: Conscientiousness)

The validation results in Table 5.3 confirm that focal loss consistently improves model performance across all optimizers. Loss decreases significantly, while accuracy shows substantial improvements. For instance, with SGD, AdaBelief, and Adam, accuracy increases by approximately 20% on average, accompanied by a notable reduction in loss.

Table 5.3: Validation Performance of ResNet50 with and without Focal Loss

Optimizer	**Without Focal Loss**		**With Focal Loss**	
	Loss	**Accuracy**	**Loss**	**Accuracy**
SGD	0.256	62.78%	0.100	82.69%
AdaBelief	0.212	71.52%	0.076	90.72%
ADAM	0.197	74.43%	0.066	**91.05%**

In the unseen data performance, as shown in Table 5.4, focal loss continues to demonstrate its effectiveness across all performance metrics, including precision, recall, and F1-score. For example, with the Adam optimizer, accuracy increases from 67.09% (without focal loss) to 90.16% (with focal loss), while the F1-score improves from 0.8419 to 0.8881. Similar trends are observed for the SGD and AdaBelief optimizers, where focal loss consistently outperforms standard cross-entropy loss.

Table 5.4: Impact of Focal Loss on ResNet50 Performance in the Test Phase

Optimizer	**Loss Function**	**Accuracy (%)**	**Loss**	**Precision**	**Recall**	**F1-score**	**AUROC**
SGD	Without Focal	61.65	0.287	0.7796	0.7675	0.7728	0.78
	With Focal	80.22	0.021	0.8367	0.7901	0.8127	0.87
AdaBelief	Without Focal	68.39	0.231	0.8702	0.8341	0.8578	0.82
	With Focal	90.03	0.010	0.8903	0.8837	0.8869	0.96
ADAM	Without Focal	67.09	0.241	0.8637	0.8212	0.8419	0.83
	With Focal	**90.16**	**0.011**	**0.8923**	**0.8844**	**0.8881**	**96.00**

On average, the overall performance of the model improves by approximately 32.1% when focal loss is applied compared to training without it. These findings confirm that focal loss is a critical component for handling imbalanced datasets. By focusing on underrepresented classes and harder examples, focal loss significantly enhances generalization and robustness.

5.3 Performance of Optimizers

The evaluation of the SGD with momentum, Adam, and AdaBelief optimizers on the ResNet50 model using focal loss reveals significant differences in their performance, as summarized in Table 5.4. Among the three, Adam consistently demonstrates the most superior performance across precision, recall, F1-score, and accuracy metrics. With an initial learning rate of 0.0001, Adam achieves the highest validation accuracy, making it the most suitable optimizer for this classification task (Table 5.3).

AdaBelief, which also uses the same initial learning rate, closely follows Adam, delivering comparable results with only minimal differences in performance (Figure 5.2). Even in the segmentation phase discussed in Section 5.1, these two optimizers excelled on the initial dataset, which featured limited samples and a high imbalance ratio (Figure 5.1).

In contrast, SGD with momentum shows noticeably weaker performance. Despite using a MultiStepLR scheduler to adjust the learning rate at critical points (epochs 30 and 80) and starting with a higher initial learning rate of 0.005 combined with a momentum value of 0.9, SGD fails to match the results achieved by Adam and AdaBelief. The optimizer struggles with slower convergence and limited adaptability to the imbalanced nature of the dataset. While SGD makes steady progress over time, its performance plateaus at a much lower level compared to Adam and AdaBelief, as shown in the accuracy and loss trends in Figure 5.4.

The accuracy curves highlight the strengths of Adam and AdaBelief, which maintain high and stable accuracy throughout training and validation. Both optimizers converge quickly during the initial epochs and demonstrate greater consistency in performance, with Adam slightly outperforming AdaBelief in overall precision and recall. On the other hand, SGD exhibits slower convergence and fluctuates more in accuracy, which reduces its reliability for this task.

The loss curves further emphasize the advantages of Adam and AdaBelief. Both achieve rapid loss reductions in the early stages of training and maintain consistently low loss values throughout. AdaBelief slightly outperforms Adam in terms of loss minimization, while SGD falls behind, with consistently higher loss values that indicate its struggle to address the dataset's complexity and imbalance.

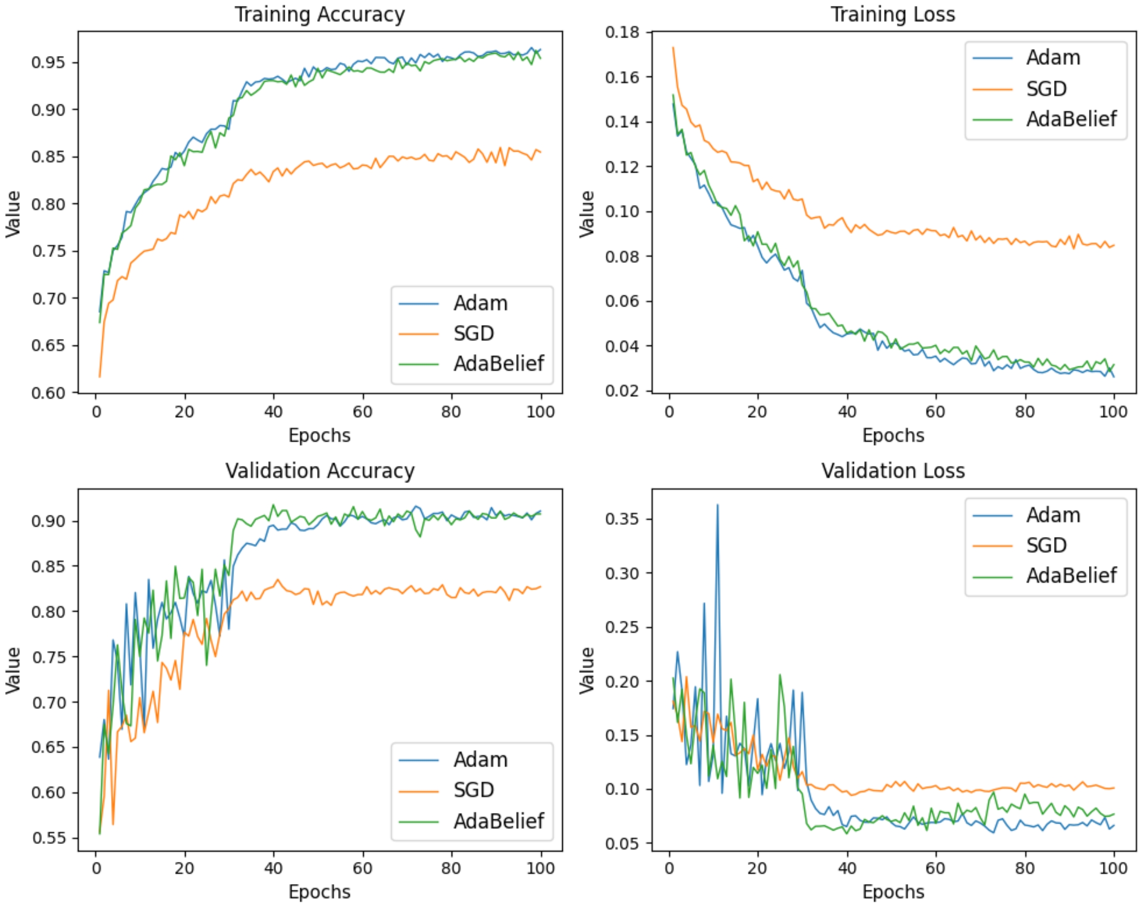


Figure 5.4: Performance of Optimizers: Training and Validation Analysis in ResNet50

Overall, these findings highlight the clear advantages of Adam and AdaBelief as optimizers for handwriting-based personality classification. Among them, Adam is chosen for the continuation of this study due to its superior balance of speed, stability, and precision. While AdaBelief offers a strong alternative with nearly identical performance, Adam's consistent and robust results make it the preferred optimizer. In comparison, SGD with momentum, though computationally efficient, is considered less suitable for this task due to its slower convergence and limited ability to handle imbalanced data effectively.

After these findings, the proposed TrOCR model is trained on the segmented dataset using focal loss and Adam as the best optimizer. On unseen data, the model achieves an accuracy of 90.05%, precision of 89.01%, recall of 88.75%, F1-score of 89.00%, and an AUROC of 97 for the two-class classification task, showing slightly better performance than ResNet50.

5.4 Multi-Label vs. Multi-Class Classification

From this section onward, the research enters its second phase, where the dataset has been expanded to include labels for all five personality traits across the entire dataset. In the multi-class classification approach, each personality trait is predicted as a single multi-class problem using Softmax and cross-entropy loss, with class weighting and focal loss emphasizing minority classes. In contrast, the multi-binary classification approach treats each class (low, average, high) as an independent binary problem, using BCE loss with focal loss to handle imbalances. Results indicate that the multi-binary method captures patterns more effectively, improving performance across all four models.

Based on the results indicated in Table 5.5, in the multi-class classification approach, ResNet50 achieves the highest accuracy of 65.80% and an F1-score of 0.616, showcasing its capability in handling multi-class predictions. However, the overall performance of all models in this approach remains relatively constrained, with TrOCR achieving an accuracy of 61.47% and an F1-score of 0.600, which are slightly lower than ResNet50 but still competitive.

Table 5.5: Comparative Evaluation of Classification Methods on a Validation Dataset

Multi-Class Classification with Cross-Entropy with Softmax					
Models	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ResNet50	0.293	65.80 %	0.636	0.648	0.616
ViT-224	0.499	61.81 %	0.577	0.608	0.576
ViT-384	0.354	63.03 %	0.560	0.620	0.577
TrOCR	0.343	61.47 %	0.566	0.634	0.600
Multi-Label Binary Classification with BCELogitLoss					
Models	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ResNet50	0.136	81.22 %	0.727	0.699	0.712
ViT-224	0.178	77.18 %	0.769	0.762	0.765
ViT-384	0.124	80.89 %	0.771	0.772	0.776
TrOCR	**0.106**	**84.46 %**	**0.808**	**0.807**	**0.810**

In contrast, the multi-label binary classification approach significantly improves the performance metrics across all models, underscoring the advantages of independently optimizing each trait. ResNet50 shows a marked improvement, achieving an accuracy of 81.22% and an F1-score of 0.712, reflecting its enhanced ability to handle imbalanced data when traits are treated as independent binary problems. Similarly, ViT models exhibit notable gains in performance, with ViT-384 attaining an accuracy of 80.89% and an F1-score of 0.776. TrOCR outperforms all other models in the multi-label binary classification, with the highest accuracy of 84.46% and an F1-score of 0.810.

The AUROC scores in Table 5.6 further illustrate the advantages of multi-binary classification in capturing trait-specific distinctions, with consistent improvements across all traits compared to the multi-class approach. The AUROC for the Conscientiousness trait in the proposed TrOCR model increases substantially from 0.5393 in the multi-class approach to 0.8943 in the multi-binary approach. Similarly, notable gains have been observed for Extraversion, Neuroticism, Agreeableness, and Openness to Experience traits. These significant improvements across all traits highlight the effectiveness of the multi-binary classification approach in addressing data imbalance, optimizing each trait independently, and capturing patterns unique to each personality factor, thereby enhancing the overall model performance.

Table 5.6: Comparison of AUROC Scores for Classification Methods Across Models

Model	**Traits**	**Multi-Class AUROC**	**Multi-Binary AUROC**
ResNet50	Extraversion	0.8307	0.8678
	Neuroticism	0.7638	0.8255
	Agreeableness	0.7273	0.8770
	Conscientiousness	0.8412	0.8827
	Openness to Experience	0.8863	0.9291
ViT-224	Extraversion	0.7532	0.8178
	Neuroticism	0.6340	0.7827
	Agreeableness	0.6117	0.8434
	Conscientiousness	0.7694	0.8498
	Openness to Experience	0.8330	0.8971
ViT-384	Extraversion	0.7716	0.8585
	Neuroticism	0.7553	0.8238
	Agreeableness	0.7015	0.8738
	Conscientiousness	0.7813	0.8591
	Openness to Experience	0.8470	0.9192
TrOCR	Extraversion	0.6419	**0.9179**
	Neuroticism	0.6493	**0.8850**
	Agreeableness	0.6642	**0.9138**
	Conscientiousness	0.5393	**0.8943**
	Openness to Experience	0.6719	**0.9334**

5.5 Model Performance

The performance of the proposed TrOCR model is evaluated against three pre-trained deep-learning models: ResNet50, ViT-224, and ViT-384. All models are trained on the same dataset using a consistent classification approach. This ensures that the comparison is reliable and fair, particularly in assessing TrOCR's effectiveness in multi-level classification of personality traits. The evaluation is conducted separately on English and French sub-samples, followed by an analysis of the combined dataset to assess how language influences model performance.

5.5.1 English Language Dataset

The English dataset, comprising 1807 line-segmented handwriting samples, is used to evaluate model performance. During training, TrOCR demonstrates the lowest final training loss and the highest training accuracy, as illustrated in Figure 5.5.

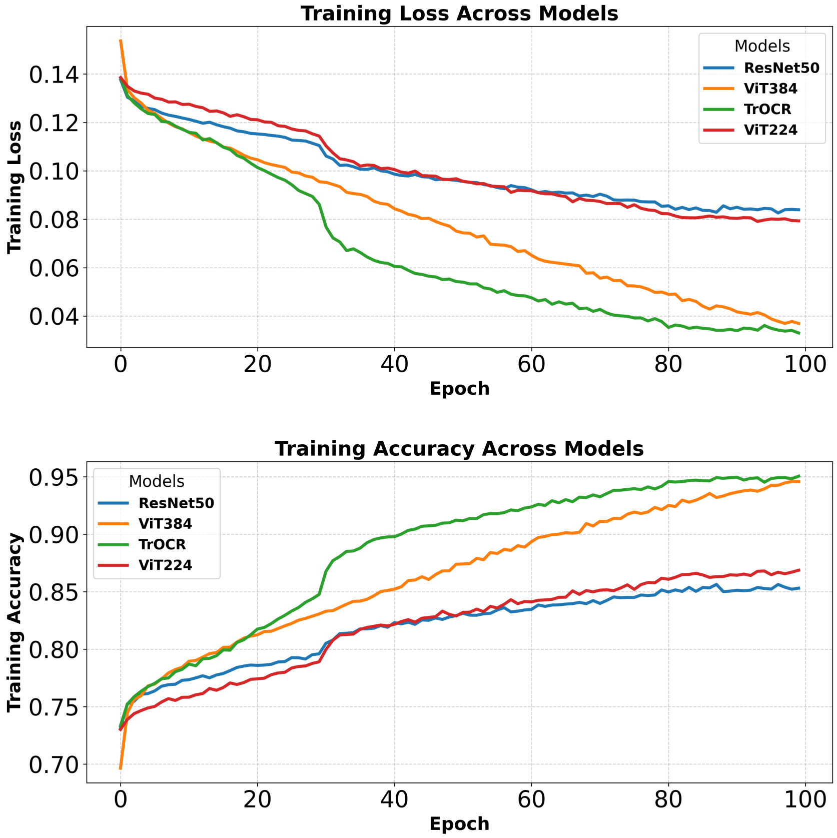


Figure 5.5: Training Results for English Sub-samples

The rapid convergence of TrOCR, with a significant reduction in loss over the epochs, highlights its superior optimization. Similarly, its training accuracy improves steadily to surpass 98%, indicating its ability to effectively capture handwriting patterns. In contrast, ViT-384 achieves comparable but slightly lower performance, while ResNet50 and ViT-224 converge more slowly, with visibly higher loss and lower accuracy values.

The models are validated, and their results are presented in Table 5.7. TrOCR achieves the lowest validation loss (0.107), which indicates its ability to generalize effectively to unseen data. TrOCR also attains the highest validation accuracy (85.22%), precision (82.23%), recall (82.25%), and F1-score (82.27%). These metrics confirm its robustness in learning complex handwriting features for personality classification.

ViT-384 follows TrOCR with a validation accuracy of 82.39% and an F1-score of 78.17%. This model performs well but is less effective than TrOCR in capturing handwriting details. ViT-224 and ResNet50 exhibit lower validation metrics, with ResNet50 achieving the lowest accuracy (78.64%) and F1-score (71.39%).

Table 5.7: Validation Performance Analysis of Models on English Subsamples

Model	**Loss**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ResNet50	0.136	78.64	73.27	69.61	71.39
ViT-224	0.126	81.40	78.53	76.83	77.60
ViT-384	0.127	82.39	78.39	78.63	78.17
TrOCR	**0.107**	**85.22**	**82.23**	**82.25**	**82.27**

The test phase results, shown in Table 5.8, further demonstrate TrOCR's superiority. TrOCR achieves the lowest test loss (0.113) and the highest test accuracy (84.43%) and F1-score (82.37%). Its ability to maintain strong performance in the test phase highlights its reliability.

ViT-384 achieves a test accuracy of 81.24% and an F1-score of 78.68%, performing well but behind TrOCR. ViT224, on the other hand, is the weakest performer among the transformer models. Although it achieves reasonable results, with test accuracy of 81.07%, its metrics are consistently lower than those of ViT384 and TrOCR. This could be attributed to the smaller input resolution (224 x 224), which may limit the model's ability to capture fine-grained handwriting features that are important for accurate classification. ResNet50, while being computationally efficient, has the lowest accuracy at 77.57%. This is likely because it struggles with balancing predictions and sometimes misclassifies majority class samples. However, it performs better than ViT224 in precision, recall, and F1-score. This is because its ability to focus on detailed features helps it classify challenging minority class samples more effectively.

Table 5.8: Test Phase Performance Analysis of Models on English Subsamples

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	0.150	77.57	78.77	78.00	78.35
ViT-224	0.133	81.07	74.85	73.70	74.19
ViT-384	0.131	81.24	79.00	78.41	78.68
TrOCR	**0.113**	**84.43**	**82.58**	**82.17**	**82.37**

The AUROC scores, shown in Figure 5.6, highlight the strengths of TrOCR, which achieves the highest AUROC score of 0.91, demonstrating its effectiveness in distinguishing between personality trait classes. ViT-384 and ResNet50 both achieve AUROC scores of 0.88, while ViT-224 records the lowest score of 0.84.

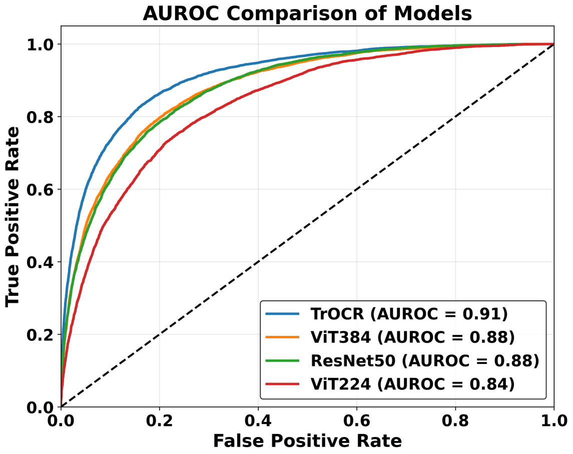


Figure 5.6: AUROC English Sub-samples Test results

5.5.2 French Language Dataset

The French dataset, consisting of 5,765 subsamples, is used to further evaluate model performance. As shown in Figure 5.7, TrOCR achieves the best training results, with the lowest training loss and highest accuracy, demonstrating its adaptability to French handwriting patterns. While

ViT-384 follows a similar trend, it stabilizes with a higher loss and lower accuracy compared to TrOCR. ResNet50, on the other hand, converges more slowly and performs less effectively overall.

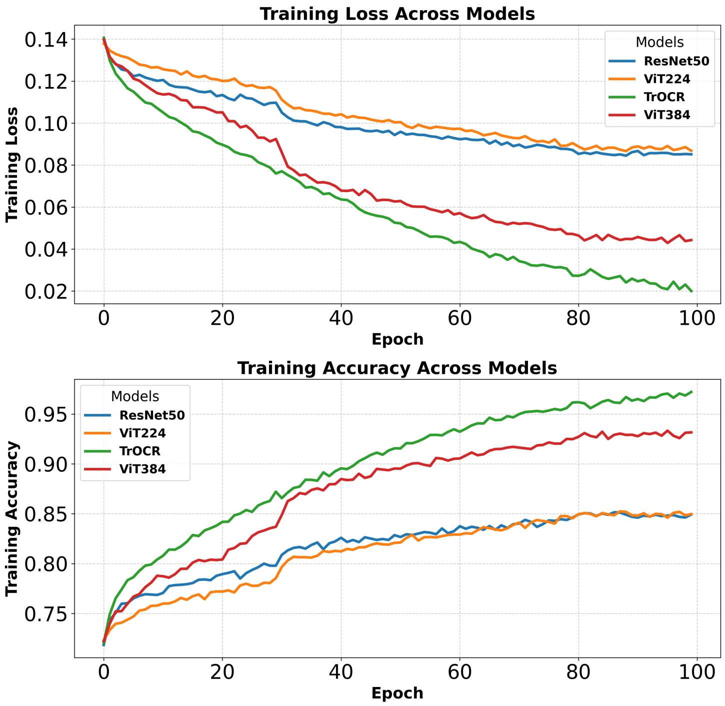


Figure 5.7: Training Results for French Sub-samples

The validation results, summarized in Table 5.9, confirm TrOCR's robustness. TrOCR achieves the lowest validation loss (0.106) and the highest validation accuracy (84.46%), precision (80.83%), recall (80.74%), and F1-score (81.04%). These metrics reflect its ability to generalize effectively to unseen French handwriting data.

ViT-384 achieves a validation accuracy of 80.89% and an F1-score of 77.63%, performing well but below TrOCR. ResNet50 and ViT-224 show lower performance, with ViT-224 having the highest validation loss (0.178).

Table 5.9: Validation Performance Analysis of Models on French Subsamples

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	0.136	81.22	72.74	69.93	71.25
ViT-224	0.178	77.18	76.93	76.21	76.55
ViT-384	0.124	80.89	77.12	77.25	77.63
TrOCR	**0.106**	**84.46**	**80.83**	**80.74**	**81.04**

The test phase results in Table 5.10 confirm TrOCR's superior performance, with the lowest loss (0.106), highest accuracy (84.26%), and best F1-score (83.27%), demonstrating its consistency across all metrics.

ResNet50 outperforms ViT-224, achieving higher accuracy (80.52%), a better F1-score (77.87%), and a lower loss (0.121). This suggests ResNet50's localized feature extraction is more effective for this dataset, whereas ViT-224 struggles with lower resolution data, reflected in its lower accuracy (77.71%) and F1-score (74.25%).

ViT-384 performs closer to ResNet50 with an accuracy of 80.07% and an F1-score of 77.23%, but it still falls short of TrOCR, which leads across all performance metrics.

Table 5.10: Test Phase Performance Analysis of Models on French Subsamples

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	0.121	80.52	78.12	77.56	77.87
ViT-224	0.171	77.71	75.02	73.74	74.25
ViT-384	0.138	80.07	77.78	76.84	77.23
TrOCR	**0.106**	**84.26**	**82.26**	**84.28**	**83.27**

The AUROC scores in Figure 5.8 show TrOCR's superior performance with the highest AUROC (0.91), followed by ResNet50 (0.88). ResNet50 outperforms both ViT-384 (0.87) and ViT-224 (0.84), highlighting its stronger ability to differentiate between personality traits in this dataset.

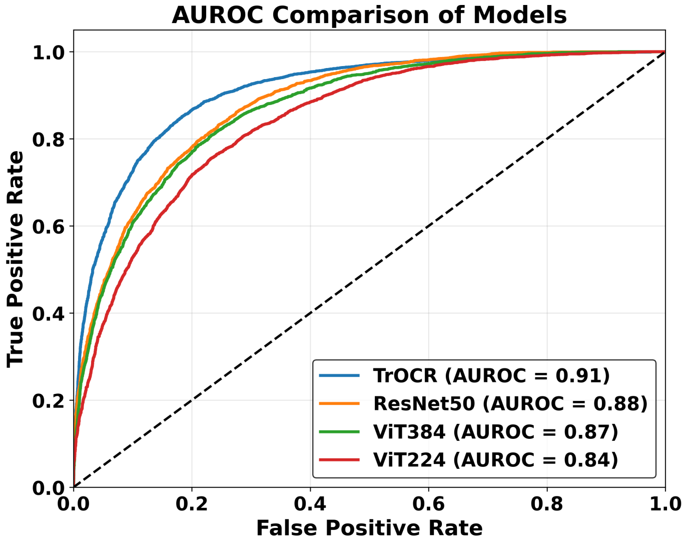


Figure 5.8: AUROC French Sub-samples Test results

Based on the results from both the English and French datasets, we observe that the models performed similarly, with very close outcomes. This is mainly due to the segmentation technique described in Section 5.1, which successfully made the imbalance ratio almost the same for both datasets: 5.81 for French and 6.02 for English. This adjustment made the datasets more comparable despite the French dataset being much larger. The results show that having a larger dataset, such as the French dataset, does not always lead to better performance. Instead, the imbalance within classes has a stronger effect on performance than the total number of samples. This shows that dataset size alone is not the main factor influencing model performance.

In fact, the English dataset, despite its smaller size, slightly performed better than the French dataset in some metrics. This further highlights the importance of other factors, such as good data preparation, balancing class distributions, and addressing the complexity of handwriting data. Additionally, methods like data augmentation and the use of focal loss during training helped the models focus on challenging samples, reducing the impact of dataset size differences. These findings show the importance of high-quality data, balanced classes, and proper data processing over simply increasing the number of samples when aiming to build effective models.

5.5.3 Combined Dataset

In this section, the influence of language on model performance is assessed by combining the English and French datasets, resulting in a total of 7,572 subsamples. The training results on the combined dataset, shown in Figure 5.9, confirm that the performance of all models improves with an average increase of approximately 7%. The proposed model, TrOCR, achieves the lowest training loss and the highest accuracy, demonstrating its strong ability to generalize effectively across the combined dataset. Its consistent superior performance is highlighted by its adaptability to diverse handwriting patterns.

Validating the models, we obtained the results presented in Table 5.11, which show that TrOCR achieves the lowest loss (0.090) and the highest accuracy (90.10%). It also records the best precision (86.78%), recall (86.74%), and F1-score (86.92%), confirming its strong ability to generalize across diverse handwriting samples and perform consistently in multi-level personality classification.

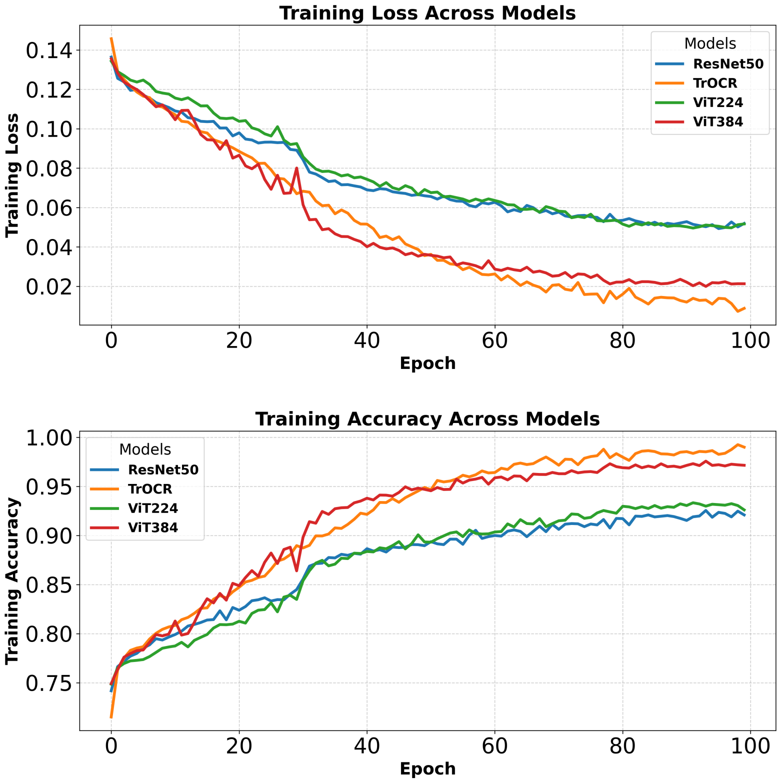


Figure 5.9: Training Results for Total Sub-samples

ViT-384 follows with a validation accuracy of 86.96% and an F1-score of 85.04%, demonstrating solid performance but still behind TrOCR. ViT-224, with an accuracy of 86.32% and an F1-score of 82.95%, performs slightly lower than ViT-384 but remains competitive. ResNet50, while showing improvement compared to its performance on the separate English and French datasets, achieves the lowest validation accuracy (82.79%) and F1-score (79.02%). This suggests that although the larger combined dataset enhances its performance, but it is now less effective than the transformer-based models in handling handwriting variations.

Table 5.11: Validation Performance Analysis of Models on Total Subsamples

Model	**Loss**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ResNet50	0.099	82.79	80.82	77.30	79.02
ViT-224	0.106	86.32	82.84	82.75	82.95
ViT-384	0.101	86.96	85.72	84.36	85.04
TrOCR	**0.090**	**90.10**	**86.78**	**86.74**	**86.92**

The test phase results, presented in Table 5.12, confirm TrOCR as the best-performing model, achieving the lowest loss (0.086) and the highest accuracy (89.01%) on unseen data. It also records the highest precision (87.39%), recall (87.25%), and F1-score (87.32%), demonstrating its ability to effectively handle diverse handwriting styles and generalize well across different samples.

The ViT models follow after TrOCR, with ViT-384 performing better than ViT-224. This suggests that higher input resolution contributes to improved feature extraction, leading to better classification results.

In contrast, ResNet50 shows the weakest performance, recording the lowest accuracy and F1-score. While CNNs like ResNet50 can still capture handwriting features, they struggle more than transformers when dealing with variations in handwriting styles, making them less effective for this classification task.

Table 5.12: Test Phase Performance Analysis of Models on Total Subsamples

Model	**Loss**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ResNet50	0.266	71.18	67.32	66.55	66.87
ViT-224	0.111	84.35	82.56	81.91	82.21
ViT-384	0.091	87.68	86.23	85.94	86.08
TrOCR	**0.086**	**89.01**	**87.39**	**87.25**	**87.32**

The AUROC scores for the combined dataset, shown in Figure 5.10, highlight TrOCR's ability to handle multilingual handwriting data effectively. TrOCR achieves the highest AUROC, while ViT-384 shows improved performance. ResNet50 continues to underperform due to its limited generalization capabilities.

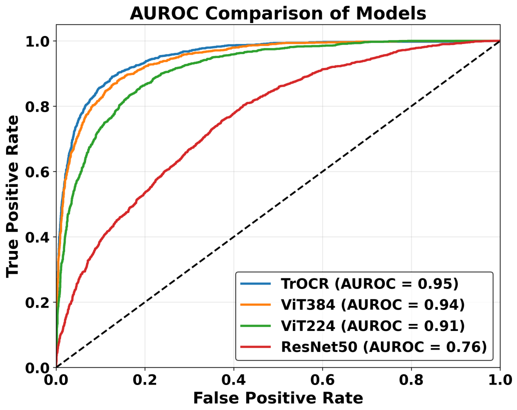


Figure 5.10: AUROC Total Sub-samples Test results

After combining both datasets, we can conclude that the performance of all transformer-based models improved when tested on unseen data, demonstrating their ability to generalize better with a larger and more diverse dataset. TrOCR remained the best-performing model, achieving the highest accuracy, precision, recall, and F1-score, further confirming its effectiveness in handling handwriting variations across different languages. The improved performance of ViT-224 and ViT-384 suggests that increasing the dataset size and incorporating handwriting samples from multiple languages helped these models learn more robust features.

However, ResNet50 did not show the same level of improvement and recorded the lowest accuracy and F1-score. This could be due to its sensitivity to language differences, as CNNs primarily rely on local feature extraction, which may not be as effective when dealing with handwriting variations in different languages. This suggests that language bias may have affected ResNet50's ability to generalize to unseen handwriting samples.

Overall, the findings show that TrOCR outperforms other models in handwriting-based personality classification, adapting better to unseen data and multiple languages.

Comprehensive TrOCR Analysis

Focusing on the proposed TrOCR model as a superior approach for automatic feature extraction in this study, its performance is analyzed in detail. The confusion matrices in Figure 5.11 demonstrate that TrOCR performs effectively for traits like Agreeableness and Openness to Experience, achieving high true positives and low misclassifications. However, traits like Extraversion remain more challenging, with higher false negatives highlighting some difficulty in capturing distinct handwriting features for this trait.

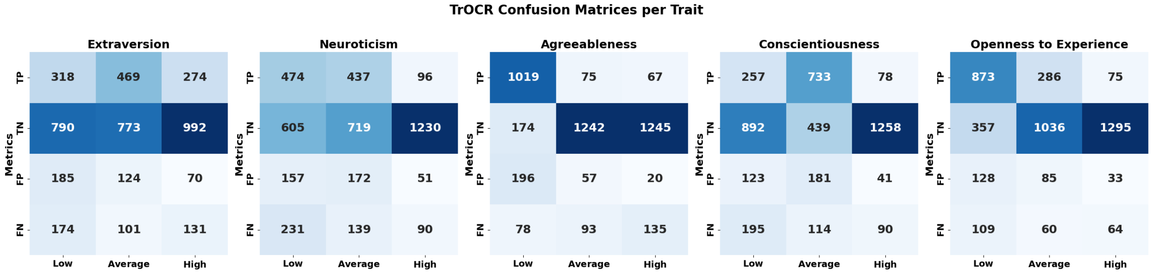


Figure 5.11: Confusion Matrices for TrOCR by Trait on Test Sub-samples

Figure 5.12 shows that TrOCR performs very well across all traits. Agreeableness and Openness to Experience, with an AUROC of 0.96, show clear and nearly perfect differentiation, highlighting the model's ability to pick up distinct handwriting patterns. Neuroticism and Conscientiousness follow closely with an AUROC of 0.95, demonstrating strong classification capabilities. Extraversion, while slightly lower, still performs reliably, showing only minor difficulties in separating features for this trait.

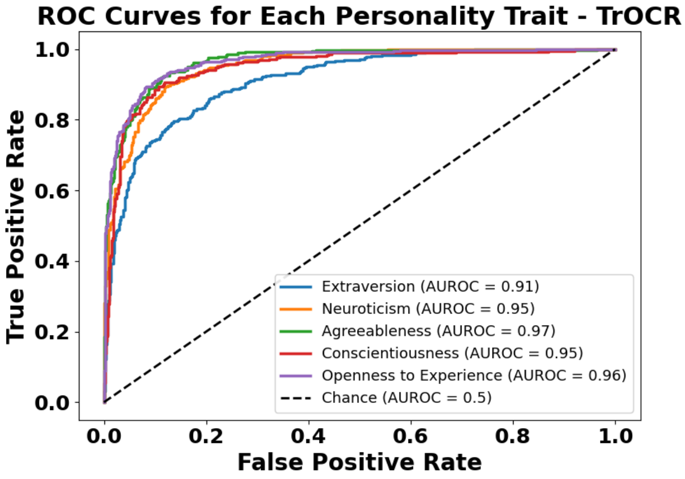


Figure 5.12: AUROC Performance of TrOCR Across Traits

Table 5.13 further supports the AUROC results by providing additional evaluation metrics, demonstrating that the proposed approach effectively classified Agreeableness (F1 Score:89.66%) and Openness to Experience (F1 Score: 88%), indicating these traits have the most distinguishable handwriting patterns. Neuroticism also achieved strong recognition, with an F1 Score of 83%, highlighting the model's reliability in identifying this trait. Similarly, Conscientiousness performed well, attaining an F1 Score of 82%. However, Extraversion remains the most challenging trait to classify, with the lowest F1 Score (75.66%) and Recall (73.99%), suggesting that its handwriting features are less distinct, making accurate classification more difficult.

Table 5.13: Performance Metrics of TrOCR for Personality Trait Classification

Metric	**Extraversion**	**Conscientiousness**	**Neuroticism**	**Agreeableness**	**Openness to Experience**
Precision	77.66	84.66	86.33	91.33	88.00
Recall	73.99	84.00	87.00	87.33	89.33
F1 Score	75.66	82.00	83.00	89.66	88.00
Accuracy	84.35	89.69	90.49	91.42	90.79

The AUROC scores and evaluation metrics for the low, average, and high levels of each personality trait, presented in Figure 5.13 and Table 5.14, provide detailed insights into the model's performance. For **Extraversion**, the ”average” level achieves the highest AUROC of 0.94, followed by the ”high” level at 0.92 and the ”low” level at 0.87, indicating some overlap in handwriting features at lower levels.

For **Neuroticism**, the ”high” level performs the best with an AUROC of 0.98, while the ”low” and ”average” levels both score 0.92, reflecting consistent classification.

Agreeableness shows the strongest overall performance, with the ”high” level achieving an AUROC of 0.99, followed by the ”low” level at 0.91 and the ”average” level at 0.90.

For **Conscientiousness**, the ”high” level scores 0.96, while the ”low” and ”average” levels achieve 0.90 and 0.88, showing slightly weaker separability.

Openness to Experience performs consistently, with ”average” and ”high” levels both scoring 0.95, and the ”low” level achieving 0.93, demonstrating reliable classification across all levels.

Table 5.14: TrOCR Metrics for Personality Traits Across Levels (Low, Average, High)

Trait	**Class**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Extraversion	Low	0.75	0.64	0.69	0.83
	Average	0.81	0.75	0.78	0.86
	High	0.77	0.83	0.80	0.84
Neuroticism	Low	0.81	0.83	0.82	0.87
	Average	0.84	0.82	0.83	0.88
	High	0.94	0.96	0.84	0.96
Agreeableness	Low	0.87	0.97	0.92	0.89
	Average	0.88	0.75	0.85	0.87
	High	0.99	0.90	0.92	0.99
Conscientiousness	Low	0.87	0.84	0.88	0.87
	Average	0.74	0.74	0.68	0.85
	High	0.93	0.94	0.90	0.96
Openness to Experience	Low	0.82	0.86	0.77	0.88
	Average	0.92	0.94	0.92	0.94
	High	0.90	0.88	0.95	0.90

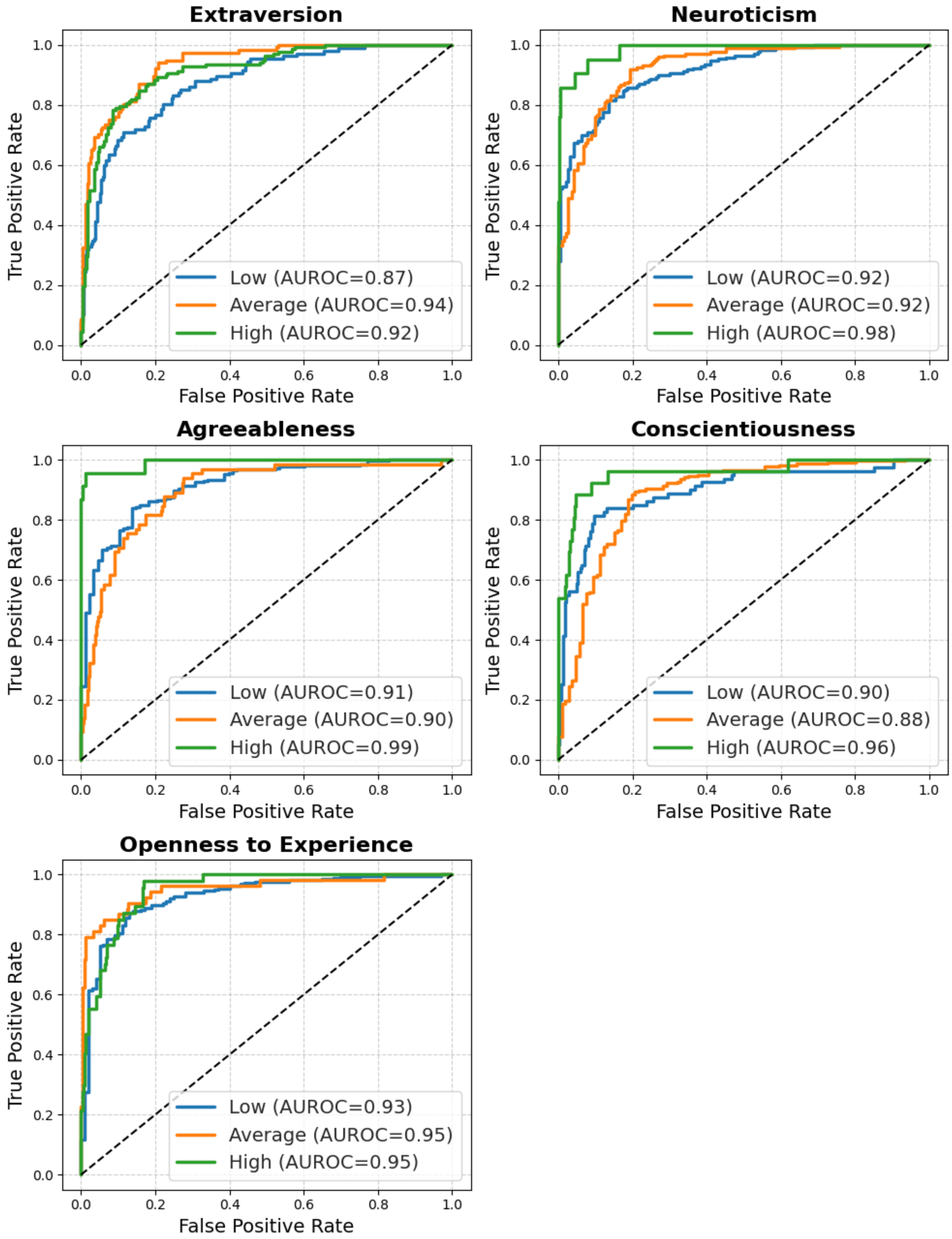


Figure 5.13: AUROC Curves of TrOCR for Personality Trait Levels (Low, Average, High)

5.6 A Comparative Analysis with Related Studies

This section compares our methodology with the most recent studies that share similarities in dataset characteristics, BFFM-based labeling, and the use of deep learning for automatic feature extraction. While these studies follow a similar approach, they exhibit notable limitations in feature learning, classification depth, and generalization.

Rodriguez et al. [47] in Table 5.15 employed a U-Net+CNN architecture for handwriting-based personality trait classification. However, their approach struggled to extract meaningful personality characteristics, leading to low F1 scores. Additionally, U-Net preprocessing may have caused information loss, negatively impacting classification performance.

Table 5.15: F1-Score (%) Comparison of Personality Trait Classification Models

Study	**Method**	**Extraversion**	**Conscientiousness**	**Neuroticism**	**Agreeableness**	**Openness to Experience**
Rodriguez et al. (2023) [47]	U-Net+CNN	61.00	57.00	40.00	59.00	33.00
Proposed Model	**TrOCR**	**75.66**	**82.00**	**83.00**	**89.66**	**88.00**

Table 5.16 shows the study by Safar et al. [51] that used VGG16 as a feature extractor and applied traditional machine learning classifiers. They used SMOTE to address class imbalance, but synthetic samples may have introduced biases that do not reflect genuine handwriting variability.

Although their ensemble learning techniques, Majority Voting and Stacking, led to improved performance, their method only outperformed our approach in Agreeableness. Our proposed TrOCR model outperformed their best results in all other traits, highlighting the advantage of end-to-end deep learning over traditional feature-based methods.

Table 5.16: Comparison of Test Accuracy (%) for ML Classifiers with SMOTE

Study	**Method**	**Extraversion**	**Conscientiousness**	**Neuroticism**	**Agreeableness**	**Openness to Experience**
Safar et al. (2024) [51]	KNN	61.24	80.53	78.35	88.92	82.56
	Random Forest	67.47	81.26	82.08	93.67	82.56
	Bagging Classifier	56.74	77.37	81.09	88.40	78.14
	Extra Trees	64.70	80.77	82.58	92.97	84.54
	Logistic Regression	68.16	77.61	80.09	86.11	84.32
	Majority Voting	67.47	80.04	83.08	93.67	88.52
	Stacking	68.51	81.02	83.58	**95.60**	90.28
Proposed Model	ResNet50	69.98	58.10	65.01	81.03	81.77
	ViT224	79.93	86.37	81.95	85.54	87.94
	ViT384	84.16	89.59	86.10	88.49	90.06
	TrOCR	**84.35**	**89.69**	**90.49**	91.42	**90.79**

Ahmed et al. [33] in Table 5.17 evaluated various deep learning models; however, their performance remained below our proposed model. Furthermore, their classification was only for low and high classes rather than for three classes, which oversimplifies the personality assessment and limits the ability of the model.

Yan et al. [29] evaluated multiple deep-learning models. While ConvNextTiny achieved the highest accuracy (86.84%), it still performed lower than our proposed approach. Moreover, their study was limited to only two personality traits (Conscientiousness and Extraversion), restricting a comprehensive assessment of all Big Five traits.

Table 5.17: Test Accuracy (%) Comparison of Deep Learning Models

Studies	**Methods**	**Test Accuracy**
Ahmed et al. (2024) [33]	VGG16	73.8
	CNN	75.5
	DenseNet201	72.3
	ResNet50	70.5
	InceptionV3	71.0
Yan et al. (2024) [29]	Convnexttiny	86.84
	Densenet121	80.53
	Inceptionv3	77.11
	VGG16	76.05
	Mobilenetv2	76.05
	Nasnetmobile	78.95
	ResNet50v2	77.63
	Xception	78.95
Proposed model	ResNet50	71.18
	ViT-224	84.35
	ViT-384	87.68
	TrOCR	**89.01**

Chapter 6

Conclusion and Future Work

6.1 Summary of Findings

This study explored the application of deep learning models for handwriting-based personality trait classification, focusing on the Big Five personality traits. The primary objectives included evaluating the impact of segmentation on class imbalance, assessing the effectiveness of focal loss, comparing the performance of three optimization techniques, and analyzing the benefits of multi-label binary classification over multi-class classification. The Transformer-based OCR model (TrOCR) was introduced as a new approach to handwriting analysis. Summary of key findings from the study include:

- Larger input sizes improve Transformer performance but increase computational cost. ViT384 performed better than ViT224, showing that higher resolution helps, but it also makes training more expensive.
- Segmenting handwriting into smaller samples improved class balance and model performance. The dataset expansion particularly benefited underrepresented personality traits, leading to better accuracy and more stable training.
- Focal loss significantly improved performance, even before segmentation. It helped models learn from difficult samples, making them more effective on imbalanced data.
- Adam and AdaBelief were the best optimizers for handwriting analysis. They consistently outperformed SGD, showing faster convergence and better generalization on imbalanced

datasets.

- Multi-binary classification was more effective than multi-class classification. Treating each trait as an independent task improved accuracy and feature learning, especially in imbalanced data.
- A larger dataset does not always mean better performance. The smaller English dataset slightly outperformed the larger French dataset, highlighting that class balance is more important than dataset size.
- ResNet50 performed well in single-language training but struggled with mixed-language data. While it sometimes outperformed Vision Transformers on separate English and French datasets, its performance dropped when the datasets were combined, showing that CNNs may not generalize well across languages.
- Transformers adapted better to multilingual handwriting than CNNs. When training on both languages together, Transformer models improved their performance, showing that they are better at handling handwriting variations.
- Our proposed TrOCR achieved the best performance across all models and evaluation metrics. It outperformed ResNet50, ViT-224, and ViT-384, achieving the highest accuracy, F1-score, and AUROC, proving its strength in handwriting-based personality classification.

6.2 Conclusions

This study introduced a new approach by adapting the pre-trained TrOCR model for automatic handwriting feature extraction and multi-label classification of personality traits based on the BFFM. The results demonstrated that TrOCR consistently outperformed ResNet50 as a CNN architecture and Vision Transformers across all datasets, achieving the highest accuracy and F1-scores. On the English dataset, TrOCR reached 84.43% accuracy with an F1-score of 82.37%. For the French dataset, it achieved 84.26% accuracy and an F1-score of 83.27%. When both datasets were combined, TrOCR delivered its best performance with 89.01% accuracy and an F1-score of 87.32%, showing its ability to generalize across different handwriting styles.

Beyond model selection, this study emphasized the role of data preprocessing and loss functions in addressing class imbalance. Segmentation effectively improved class distribution, allowing the models to learn more efficiently, while focal loss helped focus on harder-to-classify samples, leading to better recognition of underrepresented personality traits. The comparison between multi-class and multi-label binary classification further confirmed that handling each personality trait as an independent classification task improves performance, as it enables the model to capture more precise handwriting patterns.

Despite these advancements, some personality traits, such as Extraversion, remain more difficult to classify due to handwriting similarities across its levels.

6.3 Suggestions for Future Research

There are several ways to improve handwriting-based personality classification in future research:

- **Expanding the Dataset** – Collecting more diverse handwriting samples across languages, age groups, and writing styles to improve model generalization.
- **Improving Model Interpretability** – Using attention maps in TrOCR to better understand which handwriting features influence predictions.
- **Enhancing Trait-Specific Performance** – Addressing challenges in classifying Extraversion by incorporating additional handwriting features such as pen pressure and stroke dynamics.
- **Exploring Better Architectures** – Testing newer vision-language models and self-supervised learning techniques to improve feature extraction.
- **Analyzing Bias** – Conducting a comprehensive bias analysis on the dataset and evaluating the influence of data processing tasks on model bias and performance.

Chapter 7

Publications

This chapter presents the publications that resulted from the research conducted for this thesis.

(1) Adeli Shamsabad, M., & Suen, C. Y. (2024). *Deep Multi-label Classification of Personality with Handwriting Analysis*. In *Artificial Neural Networks in Pattern Recognition* (pp. 218–230). Springer Nature, Cham, Switzerland.
DOI: 10.1007/978-3-031-71602-7_19

(2) Adeli Shamsabad, M., & Suen, C. Y. (2025). *Automated Handwriting Pattern Recognition for Multi-Level Personality Classification Using Transformer OCR (TrOCR)*. In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM)* (pp. 141–150). SciTePress, Setúbal, Portugal.
DOI: 10.5220/0013318800003905

References

[1] Mihai Gavrilescu and Nicolae Vizireanu. “Predicting the Big Five personality traits from handwriting”. In: *EURASIP Journal on Image and Video Processing* 2018.1 (2018), p. 57. DOI: 10.1186/s13640-018-0297-3.

[2] Atta Rahman and Zahid Halim. “Predicting the big five personality traits from hand-written text features through semi-supervised learning”. In: *Multimedia Tools and Applications* 81 (Sept. 2022), pp. 1–17. DOI: 10.1007/s11042-022-13114-5.

[3] Fahimeh Alaei and Alireza Alaei. “Handwriting Analysis: Applications in Person Identification and Forensic”. In: *Breakthroughs in Digital Biometrics and Forensics*. Springer International Publishing, 2022, pp. 147–165. ISBN: 978-3-031-10706-1. DOI: 10.1007/978-3-031-10706-1_7.

[4] J. D. McDonald. “Measuring Personality Constructs: The Advantages and Disadvantages of Self-Reports, Informant Reports and Behavioural Assessments”. In: *Enquire* 1.1 (2008), pp. 75–94. URL: https://www.nottingham.ac.uk/sociology/documents/enquire/volume-1-issue-1-dodorico-mcdonald.pdf.

[5] G. Sheikholeslami, S. N. Srihari, and V. Govindaraju. “Computer Aided Graphology”. In: *Proceedings of the Center of Excellence for Document Analysis and Recognition*. Amherst, New York, USA: State University of New York at Buffalo, 1994.

[6] Helmut Ploog. *Handwriting Psychology: Personality Reflected in Handwriting*. iUniverse, 2013. ISBN: 978-1475970234. URL: https://www.iuniverse.com/en/bookstore/bookdetails/430619-Handwriting-Psychology.

[7] Handwriting Analysis. *The Fascinating History of Handwriting Analysis*. Accessed: 2025-01-06. 2023. URL: https://handwriting.feedbucket.com/articles/2023/02/15/the-fascinating-history-of-handwriting-analysis/.

[8] Neo Science Hub. *Beyond the Pen: Evolution & Significance of Graphology*. Accessed: 2025-01-06. 2024. URL: https://neosciencehub.com/beyond-the-pen-evolution-significance-of-graphology/.

[9] Handwriting Analysis. *The Future of Handwriting Analysis: Trends to Watch*. Accessed: 2025-01-06. 2024. URL: https://handwriting.feedbucket.com/articles/2024/06/26/the-future-of-handwriting-analysis-trends-to-watch/.

[10] Roberta Satow and Jacqueline Rector. "Using gestalt graphology to identify entrepreneurial leadership". In: *Perceptual and Motor Skills* 81.1 (1995), pp. 263–270. DOI: 10.2466/pms.1995.81.1.263.

[11] Milton N. Bunker. *Handwriting Analysis: The Science of Determining Personality by Graphoanalysis*. Chicago: Nelson-Hall, 1967. URL: https://archive.org/details/handwritinganaly00bunk.

[12] Klara Goldzieher Roman. *Handwriting: A Key to Personality*. New York: Pantheon Books, 1952. URL: https://archive.org/details/handwritingkeyto0000roma.

[13] Pierre Etienne Cronje. "The Viability of Graphology in Psycho-Educational Assessment". Supervisor: Prof. H. E. Roets. PhD thesis. Pretoria, South Africa: University of South Africa, 2009. URL: http://hdl.handle.net/10500/3134.

[14] Pervez Ahmed and Hassan Mathkour. "On the Development of an Automated Graphology System". In: *Proceedings of the 2008 International Conference on Artificial Intelligence (IC-AI)*. 2008, pp. 897–901.

[15] Charlotte P. Leibel. *Change Your Handwriting, Change Your Life*. New York: Stein and Day, 1972. ISBN: 9780812814163.

[16] R. N. King and D. J. Koehler. “Illusory correlations in graphological inference”. In: *Journal of Experimental Psychology: Applied* 6.4 (2000), pp. 336–348. DOI: 10.1037/1076-898X.6.4.336.

[17] Adrian Furnham, Tomas Chamorro-Premuzic, and Ines Callahan. “Does graphology predict personality and intelligence?” In: *Individual Differences Research* 1.2 (2003), pp. 78–94.

[18] Benjamin Thiry and Odile Rohmer. “Exploring the Validity of Graphology with the Rorschach Test”. In: *European Psychologist* 30.1 (2007), pp. 26–34. DOI: 10.1027/1192-5604.30.1.26.

[19] Claudio Dazzi and Luigi Pedrabissi. “Graphology and Personality: An Empirical Study on Validity of Handwriting Analysis”. In: *Psychological Reports* 105.3_suppl (2009), pp. 1255–1268. DOI: 10.2466/PR0.105.F.1255-1268.

[20] Barbara Gawda. “Lack of evidence for the assessment of personality traits using handwriting analysis”. In: *Polish Psychological Bulletin* 45.1 (2014), pp. 73–79. DOI: 10.2478/ppb-2014-0011. URL: https://doi.org/10.2478/ppb-2014-0011.

[21] Prajakta Harne, Munish Kumar Mishra, and Gurvinder Singh Sodhi. “Analysis of Handwriting Characteristics Based on Diverse Ethnic Distribution”. In: *International Journal of Advanced Trends in Computer Applications (IJATCA)* 5.2 (2018), pp. 1–6. ISSN: 2395-3519. URL: http://www.ijatca.com.

[22] Afnan Garoot. “Handwriting Analysis and Personality: A Computerized Study on the Validity of Graphology”. PhD thesis. Concordia University, 2021. URL: https://spectrum.library.concordia.ca/id/eprint/989091/.

[23] Robert R. McCrae and Paul T. Costa. “A contemplated revision of the NEO Five-Factor Inventory”. In: *Personality and Individual Differences* 36.3 (2004), pp. 587–596. ISSN: 0191-8869. DOI: https://doi.org/10.1016/S0191-8869(03)00118-1.

[24] Lewis R. Goldberg. “An Alternative ”Description of Personality”: The Big-Five Factor Structure”. In: *Journal of Personality and Social Psychology* 59.6 (1990), pp. 1216–1229. DOI: 10.1037/0022-3514.59.6.1216.

[25] Joni Salminen et al. “Enriching social media personas with personality traits: A deep learning approach using the big five classes”. In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 101–120. DOI: 10.1007/978-3-030-50334-5_7.

[26] Majid Ramezani, Mohammad-Reza Feizi-Derakhshi, and Mohammad- Ali Balafar. “Text-based automatic personality prediction using KGrAt-Net: a knowledge graph attention network classifier”. In: *Scientific Reports* 12.1 (2022), p. 21453. DOI: 10.1038/s41598-022-25955-z.

[27] Elma Kerz et al. “Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features”. In: *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics, 2022, pp. 182–194. DOI: 10.18653/v1/2022.wassa-1.17.

[28] Joshua Johnson Sirasapalli and Ramakrishna Murty Malla. “A deep learning approach to text-based personality prediction using multiple data sources mapping”. In: *Neural Computing and Applications* 35.28 (2023), pp. 20619–20630. DOI: 10.1007/s00521-023-08846-w.

[29] Yang Yan et al. “Predicting the Big Five Personality Traits in Chinese Counselling Dialogues Using Large Language Models”. In: *arXiv preprint arXiv:2406.17287* (2024). URL: https://arxiv.org/abs/2406.17287.

[30] Wun Yung Shaney Sze, Maryglen Pearl Herrero, and Roger Garriga. “Personality Trait Inference Via Mobile Phone Sensors: A Machine Learning Approach”. In: *arXiv preprint arXiv:2401.10305* (2024).

[31] Heinrich Peters, Moran Cerf, and Sandra C. Matz. “Large Language Models Can Infer Personality from Free-Form User Interactions”. In: *arXiv preprint arXiv:2405.13052* (2024). DOI: 10.31219/osf.io/apc5g.

[32] Ayoub Ouarka et al. “A deep multimodal fusion method for personality traits prediction”. In: *Multimedia Tools and Applications* (2024), pp. 1–23. DOI: 10.1007/s11042-024-20356-y.

[33] Ahmed Mohamed Ahmed Sayed et al. “Analyzing Handwriting to Infer Personality Traits: A Deep Learning Framework”. In: *2024 Intelligent Methods, Systems, and Applications (IMSA)*. 2024, pp. 58–63. DOI: 10.1109/IMSA61967.2024.10652866.

[34] Sudan Neupane et al. “GraphoMatch: Forensic handwriting analysis using machine learning”. In: *International Journal of Science and Research Archive* 11.02 (2024), pp. 1526–1537. DOI: https://doi.org/10.30574/ijsra.2024.11.2.0643.

[35] Navya K S Shree and Dr. Siddaraju. “Analysis of Personality Based on Handwriting Using Deep Learning”. In: *International Journal of Creative Research Thoughts (IJCRT)* 10.12 (2022), b676–b683. ISSN: 2320-2882. URL: http://www.ijcrt.org/papers/IJCRT2212189.pdf.

[36] Abdellatif Gahmousse et al. “Handwriting based Personality Identification using Textural Features”. In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. 2020, pp. 1–6. DOI: 10.1109/ICDABI51230.2020.9325664.

[37] Peter Drotár et al. “Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson’s disease”. In: *Artificial Intelligence in Medicine* 67.2 (2016), pp. 39–46.

[38] Salankara Mukherjee, Ishita De Ghosh, and Debatree Mukherjee. “Big Five Personality Prediction from Handwritten Character Features and Word ‘of’ Using Multi-label Classification”. In: *Proceedings of the Seventh International Conference on Mathematics and Computing*. Springer Singapore, 2022, pp. 275–299. ISBN: 978-981-16-6890-6. DOI: https://doi.org/10.1007/10.1007/978-981-16-6890-6_21.

[39] Gayathry H. Nair, V. Rekha, and M. Soumya Krishnan. “Handwriting Analysis Using Deep Learning Approach for the Detection of Personality Traits”. In: *Ubiquitous Intelligent Systems*. Springer Nature Singapore Pte Ltd., 2022, pp. 531–539. ISBN: 978-981-16-3675-2. DOI: https://doi.org/10.1007/978-981-16-3675-2_40.

[40] Momin Zaki Mohiuddin et al. “Comparative Analysis of LBP, HOG, and SIFT Techniques for Handwritten Signature Recognition Performance”. In: *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)* 10.11 (2023), pp. 26–29. URL:

https://www.technoarete.org/common_abstract/pdf/IJERCSE/v10/i11/Ext_26713.pdf.

[41] Abdellatif Gahmousse, Rabeb Yousfi, and Chawki Djeddi. “Handwriting Based Personality Traits Identification Using Adaptive Boosting and Textural Features”. In: *MedPRAI 2021, Communications in Computer and Information Science*. Vol. 1543. Springer, 2022, pp. 216–227. DOI: 10.1007/978-3-031-04112-9_16.

[42] Derry Alamsyah et al. “Handwriting Analysis for Personality Trait Features Identification using CNN”. In: *2022 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE. 2022, pp. 232–238. DOI: 10.1109/ICoDSA55874.2022.9862910.

[43] Noor Fazilla Abd Yusof et al. “Extracting Graphological Features for Identifying Personality Traits using Agglomerative Hierarchical Clustering Algorithm”. In: *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. IEEE. 2022, pp. 1–6.

[44] Lakshmi Durga and Deepu R. “A self-adaptive cognitive deep learning framework for classifying graphology features to Big Five personality traits”. In: *International Journal of Advanced Technology and Engineering Exploration* 9.93 (2022), pp. 1151–1167. DOI: 10.19101/IJATEE.2021.875577.

[45] Afnan H. Garoot and Ching Y. Suen. “Measuring the Big Five Factors from Handwriting Using Ensemble Learning Model AvgMlSC”. In: Lecture Notes in Computer Science 13424 (2022), pp. 159–173. DOI: 10.1007/978-3-031-19745-1_12.

[46] Samsuryadi et al. “A Framework for Determining the Big Five Personality Traits Using Machine Learning Classification through Graphology”. In: *Journal of Electrical and Computer Engineering* 2023 (2023), pp. 1–15. DOI: 10.1155/2023/1249004.

[47] Diego A. Peralta-Rodríguez et al. “Automated Handwriting Analysis for Personality Traits Recognition Using Image Preprocessing Techniques”. In: *Research in Computing Science* 152.12 (2023), pp. 93–106. ISSN: 1870-4069. DOI: 10.1007/978-3-031-21648-0_36.

[48] Yashomati R. Dhumal et al. “Automatic Handwriting Analysis and Personality Trait Detection using Multi-Task Learning Technique”. In: *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2023, pp. 348–353. URL: https://ieeexplore.ieee.org/document/10014583.

[49] Bipin Nair et al. “Handwriting Analysis for Classification of Human Personality”. In: *2024 Second International Conference on Advances in Information Technology (ICAIT-2024)*. IEEE. 2024, pp. 1–6. DOI: 10.1109/ICAIT61638.2024.10690448.

[50] H.K. Chethan et al. “Personality Prediction Using Handwriting Analysis”. In: *International Research Journal of Engineering and Technology (IRJET)* 11.4 (2024), pp. 1536–1539. URL: https://www.irjet.net/archives/V11/i4/IRJET-V11I4253.pdf.

[51] Maedeh Safar. “Integrating Handwriting Analysis and Machine Learning for Enhanced Personality Trait Prediction”. Masters thesis. Concordia University, 2024. URL: https://spectrum.library.concordia.ca/id/eprint/994164/.

[52] A. Puttaswamy and R. Thillaiarasu. “Fine DenseNet based human personality recognition using english hand writing of non-native speakers”. In: *Biomedical Signal Processing and Control* 99 (2025), p. 106910. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2024.106910.

[53] Rok Blagus and Lara Lusa. “Evaluation of SMOTE for High-Dimensional Class-Imbalanced Microarray Data”. In: *2012 11th International Conference on Machine Learning and Applications*. Vol. 2. 2012, pp. 89–94. DOI: 10.1109/ICMLA.2012.183.

[54] Justin M. Johnson and Taghi M. Khoshgoftaar. “Survey on deep learning with class imbalance”. In: *Journal of Big Data* 6.1 (2019), p. 27. DOI: 10.1186/s40537-019-0192-5.

[55] Haibo He and Edwardo A. Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[56] Rui Zhu, Yiwen Guo, and Jing-Hao Xue. “Adjusting the imbalance ratio by the dimensionality of imbalanced data”. In: *Pattern Recognition Letters* 133 (2020), pp. 217–223.

ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2020.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0167865520300829.

[57] C. Tomasi and R. Manduchi. "Bilateral filtering for gray and color images". In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. 1998, pp. 839–846. DOI: 10.1109/ICCV.1998.710815.

[58] Connor Shorten and Taghi Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6 (July 2019). DOI: 10.1186/s40537-019-0197-0.

[59] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.

[60] Marzieh Adeli Shamsabad and Ching Yee Suen. "Deep Multi-label Classification of Personality with Handwriting Analysis". In: *Artificial Neural Networks in Pattern Recognition (ANNPR 2024)*. Ed. by Ching Yee Suen et al. Vol. 15154. Lecture Notes in Computer Science. Cham: Springer, 2024, pp. 217–230. DOI: 10.1007/978-3-031-71602-7_19.

[61] Ruoxi Qin et al. "Weighted Focal Loss: An Effective Loss Function to Overcome Unbalance Problem of Chest X-ray14". In: *IOP Conference Series: Materials Science and Engineering* 428 (Oct. 2018), p. 012022. DOI: 10.1088/1757-899X/428/1/012022.

[62] Taissir Fekih Romdhane et al. "Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss". In: *Computers in Biology and Medicine* 123 (2020), p. 103866.

[63] Ning Qian. "On the momentum term in gradient descent learning algorithms". In: *Neural networks : the official journal of the International Neural Network Society* 12 (Feb. 1999), pp. 145–151. DOI: 10.1016/S0893-6080(98)00116-6.

[64] Juntang Zhuang et al. *AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients*. Oct. 2020.

[65] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations* (Dec. 2014).

[66] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[67] Jiaqi Luo, Yuan Yuan, and Shixin Xu. *Improving GBDT Performance on Imbalanced Datasets: An Empirical Study of Class-Balanced Loss Functions*. July 2024. DOI: `10.48550/arXiv.2407.14381`.

[68] Zahra Ebrahimi Vargoorani and Ching Yee Suen. "License Plate Detection and Character Recognition Using Deep Learning and Font Evaluation". In: *Artificial Neural Networks in Pattern Recognition (ANNPR 2024)*. Ed. by Ching Y. Suen et al. Vol. 15154. Lecture Notes in Computer Science. Cham: Springer, 2024, pp. 231–242. DOI: `10.1007/978-3-031-71602-7_20`.

[69] Maithra Raghu et al. "Do vision transformers see like convolutional neural networks?" In: *Advances in neural information processing systems* 34 (2021), pp. 12116–12128.

[70] Michael Koepf, Florian Kleber, and Robert Sablatnig. "Writer Identification and Writer Retrieval Using Vision Transformer for Forensic Documents". In: *Document Analysis Systems*. Springer International Publishing, May 2022, pp. 352–366. ISBN: 978-3-031-06554-5. DOI: `10.1007/978-3-031-06555-2_24`.

[71] Xiaohua Zhai et al. "Scaling Vision Transformers". In: *arXiv preprint arXiv:2106.04560* (2021).

[72] Alexey Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[73] Israel Campiotti and Roberto Lotufo. "Optical character recognition with transformers and CTC". In: Nov. 2022, pp. 1–4. DOI: `10.1145/3558100.3563845`.

[74] Minghao Li et al. *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. 2021. arXiv: `2109.10282 [cs.CL]`.

[75] Ankan Bhunia et al. *Handwriting Transformers*. Apr. 2021. DOI: `10.48550/arXiv.2104.03964`.

[76] Marzieh Adeli Shamsabad and Ching Suen. “Automated Handwriting Pattern Recognition for Multi-Level Personality Classification Using Transformer OCR (TrOCR)”. In: *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods - ICPRAM*. INSTICC. SciTePress, 2025, pp. 141–150. ISBN: 978-989-758-730-6. DOI: `10.5220/0013318800003905`.

[77] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information Processing & Management* 45.4 (2009), pp. 427–437. ISSN: 0306-4573. DOI: `https://doi.org/10.1016/j.ipm.2009.03.002`.