Automated Handwriting Pattern Recognition for Multi-Level Personality Classification Using Transformer OCR (TrOCR)

Marzieh Adeli Shamsabad[a] and Ching Yee Suen[b]
Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, Montreal, Quebec, Canada
{m_adelis, suen}@cenparmi.concordia.ca

Keywords: Handwriting Analysis, Automated Feature Extraction, Imbalanced Dataset, Multi-Level Classification, TrOCR, Big Five Personality Traits.

Abstract: Automated personality trait assessment from handwriting analysis offers applications in psychology, human-computer interaction, and personal profiling. However, accurately classifying different levels of personality traits remains challenging due to class imbalances in real-world datasets. This study addresses the issue by comparing multi-class and multi-label binary classification methods to predict levels of the Big Five personality traits: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness, each categorized as low, average, and high in an imbalanced dataset of 873 French handwriting samples. A new approach is introduced by adapting the TrOCR pre-trained model for feature extraction, modifying its encoder to capture local and global handwriting features relevant to personality classification. This model is compared with three other pre-trained models: ResNet50 and Vision Transformer base 16 with input resolutions of 224 and 384. Results demonstrate that multi-label binary classification, which treats each trait level as an independent binary task, effectively addresses data imbalance, enhancing accuracy and generalization. The proposed TrOCR model achieves the highest performance, with an accuracy of 84.26%, an F1-score of 83.26%, and an AUROC of 91% on the test set. These findings emphasize the effectiveness of the presented framework for automated multi-level personality trait classification from handwriting in imbalanced datasets.

1 INTRODUCTION

Personality represents the unique patterns of thoughts, emotions, and behaviors that define an individual's character and influence their interactions with the world (Costa and McCrae, 1997). Understanding personality traits can provide insights into how individuals respond in various situations, shaping personal, social, and professional outcomes (Roberts and Mroczek, 2008). Traditionally, personality assessment is achieved through self-reported questionnaires, such as the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits model, which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. However, these methods are often time-consuming and can be influenced by the individual's self-awareness and response biases (Alshouha et al., 2024). As a result, alternative methods for personality prediction gained attention like handwriting.

Handwriting analysis, or graphology, suggests that an individual's handwriting reflects their psychological state and personality traits. This idea arises from the observation that handwriting is a psychomotor activity, where the brain guides the hand's movements, resulting in unique patterns in letter formation, slant, pressure, and spacing. Thus, handwriting is shaped by physiological factors, socio-cultural influences, and personal experiences (Rahman and Halim, 2022). Manual handwriting analysis for personality assessment is subjective and requires expertise, leading researchers to explore automated solutions through image processing and machine learning techniques (Gavrilescu and Vizireanu, 2018).

Many studies relied on manual feature extraction and traditional machine learning classifiers. Mukherjee et al. (Mukherjee et al., 2022) approached personality prediction by extracting character-based features such as specific letters (e.g., 'a', 'g', 'n', 't') and the word "of," from handwritten samples and used classifiers like K-nearest neighbor (KNN) and Multi-layer Perceptron (MLP) for prediction. Nair et al. (B J et al., 2024) manually extracted features such as stroke pressure and letter slant, then classified

[a] https://orcid.org/0009-0005-7680-6689
[b] https://orcid.org/0000-0003-1209-7631

Adeli Shamsabad, M. and Suen, C. Y.
Automated Handwriting Pattern Recognition for Multi-Level Personality Classification Using Transformer OCR (TrOCR).
DOI: 10.5220/0013318800003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 141-150
ISBN: 978-989-758-730-6; ISSN: 2184-4313

personality traits using algorithms like Support Vector Machine (SVM), Decision Trees, and Random Forest. Chin et al. (Chin et al., 2021) used a Histogram of Oriented Gradients to extract handwriting features and classified them using multiclass Support Vector Machines, with logistic regression. Another early work by Gavrilescu and Vizireanu (Gavrilescu and Vizireanu, 2018) employed a semi-automatic approach involving handwriting feature extraction followed by a neural network-based model for classifying the Big Five personality traits.

With artificial intelligence and deep learning advancements, automatic feature extraction methods have become more prevalent. Ahmed et al. (Sayed et al., 2024) proposed a deep learning framework using Convolutional Neural Networks (CNNs) such as VGG16, DenseNet201, ResNet, and InceptionV3 to automatically extract handwriting features like letter size, slant, and pressure on the IAM handwriting database. Similarly, Nair et al. (Nair et al., 2021) compared the performance of ResNet50 and CNN for handwriting analysis to predict personality traits, indicating that deep learning models provide a robust solution for capturing handwriting patterns. Additionally, Puttaswamy et al. (Puttaswamy and Thillaiarasu, 2025) employed Fine DenseNet with attention mechanisms to extract intricate handwriting features for personality classification, showcasing the role of attention-based feature extraction in enhancing classification accuracy.

Dhumal et al. (Dhumal et al., 2023) applied Transformer and LSTM networks to automatically extract handwriting features and predict personality traits, exploring the multi-label classification approach. Shree et al. (Shree and Dr.Siddaraju, 2022) used YOLO v5 and ResNet34 for feature extraction and personality classification, demonstrating an effective strategy to improve accuracy and efficiency. Recent advancements in attention mechanisms and vision transformers have further improved the feature extraction capabilities (Koepf et al., 2022), allowing for more analysis of handwriting features and improved personality prediction performance.

Another line of research focused on semi-supervised and hybrid models for personality classification. Rahman and Halim (Rahman and Halim, 2022) employed a Semi-supervised Generative Adversarial Network (SGAN) to classify personality traits based on handwriting samples, utilizing a combination of labeled and unlabeled data to improve classification accuracy. This approach highlighted the efficacy of semi-supervised learning in addressing the challenges posed by limited labeled data.

Previous studies have advanced the field, but many of them focused on independent trait classification or relied on manual feature extraction, which limits their scalability and adaptability. Our research aims to overcome these limitations by developing an end-to-end deep learning framework with fully automated feature extraction. Building on our previous work, where only two personality traits were analyzed, this research expands the dataset to have all five traits outlined by the Big Five Factor Model (BFFM). This broader approach enables multi-level classification of all five traits simultaneously, facilitating a more comprehensive personality analysis.

A French handwriting dataset from the CENPARMI lab is used in this study, expanded from 873 full-page handwriting images to 5,765 line-segmented images to increase dataset size while preserving essential handwriting patterns. The model implicitly learns handwriting patterns through deep learning mechanisms like convolutional filters or attention heads. Based on our previous research, Focal Loss is applied to effectively address class imbalance (Adeli Shamsabad and Suen, 2024), in combination with Softmax for multi-class classification and BCE with logit loss for multi-binary classification.

A new method for handwriting feature extraction is proposed, using the base version of TrOCR, a model with an encoder-decoder structure pre-trained on general text data and fine-tuned on the IAM Handwriting Database with an input size of 384 to adapt to handwriting-specific features. Instead of using the decoder to generate text, TrOCR is modified to output encoder features and then fed for classification. To the best of our knowledge, this represents the first use of TrOCR in a classification framework, specifically for predicting different levels of personality traits from handwriting images, showcasing its potential beyond traditional OCR applications.

To evaluate the proposed approach, three other pre-trained deep learning models are compared: ResNet50, chosen for its efficient CNN architecture and proven performance in prior work (Adeli Shamsabad and Suen, 2024), and Vision Transformer (ViT) base 16 with input sizes of 224 and 384 that are pre-trained on ImageNet by Google that allows for an analysis of performance differences between transformer models tailored for OCR tasks and those designed for general-purpose vision tasks.

This paper is organized as follows: Section 2 details the materials, methods, and evaluation metrics. Section 3 analyzes the results and compares them with existing methods. Section 4 concludes the study and suggests future work.

2 MATERIALS AND METHODS

2.1 Data Collection

For this research, a custom dataset is created to enable automatic handwriting analysis for predicting multi-level personality traits based on the BFFM. Since no public dataset is available to meet these specific requirements, data collection is conducted with ethical approval from Concordia University's Human Research Ethics Committee. The dataset is composed of two main parts: responses to a BFFM personality questionnaire and corresponding handwriting samples. Participants are recruited from Concordia University to ensure a diverse dataset, and to maintain consistency and minimize external influences on handwriting, data collection is carried out in a controlled environment within a dedicated lab at Concordia University's CENPARMI research center.

Participants are asked to complete the BFFM questionnaire, which assesses five personality traits: Extraversion (EX), Neuroticism (NE), sometimes referred to as Emotional Stability in its inverse form, Agreeableness (AG), Conscientiousness (CO), and Openness to Experience (OE) (Costa and McCrae, 1997). Scores for each trait are categorized into three levels: low, average, and high. Each handwriting sample is manually analyzed by a professional graphologist for specific features, such as slant, spacing, and letter formation, which correspond to each of the BFFM traits. Each sample is labeled with all five personality traits, with a unique level assigned to each trait, indicating that every sample reflects different levels for each of the five traits. In total, 1110 handwriting samples are collected, digitized at a high resolution of 600 dots per inch (DPI), and processed to remove noise and irrelevant marks.

The dataset includes handwriting samples in multiple languages, with the majority in French (873 samples), followed by English (181 samples), along with smaller quantities in languages such as Persian and Korean. However, the distribution of personality traits is severely imbalanced, with the medium level being predominant in each trait category. Due to the large number of French samples, this study primarily focuses on analyzing this subset and its personality trait distribution, as follows:

- EX: Low - 125, Average - 333, High - 415;
- NE: Low - 88, Average - 473, High - 312;
- AG: Low - 96, Average - 703, High - 74 ;
- CO: Low - 44, Average - 319, High - 510;
- OE: Low - 38, Average - 558, High - 277.

In our previous study, the dataset included labels for only two personality traits, Extraversion and Conscientiousness (Adeli Shamsabad and Suen, 2024). For this research, the dataset has been expanded to include all five BFFM traits, enabling comprehensive analysis through multi-level classification and allowing simultaneous prediction of all five personality traits.

2.2 Data Preprocessing

The primary goal of this research is to develop an end-to-end automated system for handwriting pattern recognition using deep learning techniques. To enhance the model's learning capacity and ensure effective generalization across various handwriting styles, it is essential to increase the number of handwriting samples and address dataset imbalance (Shorten and Khoshgoftaar, 2019). Additionally, the original TrOCR base model is designed for single-line text segments, optimizing its performance for line-by-line tasks rather than continuous multi-line or paragraph text (Li et al., 2021). To address these limitations, each handwriting sample is segmented line by line using OpenCV, with careful attention to preserving the original text structure. This segmentation approach ensures that the most significant features of each letter shape are retained, as illustrated in Figure 1.

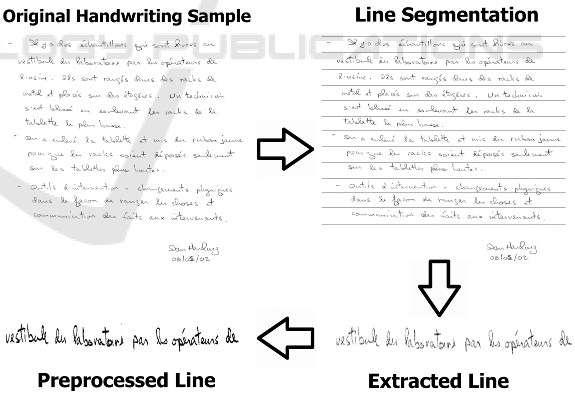

Figure 1: Line Segmentation Process.

Through line-level segmentation, the dataset is expanded to 5,765 handwriting subsamples. This segmentation enhances the distribution of traits, particularly for those traits that initially had fewer samples (Table 1).

The handwriting samples are digitized through scanning, with some images resulting in low quality, which necessitates further preprocessing. Based on our previous findings, Otsu's binarization method combined with bilateral filtering is the most effec-

Table 1: Trait Distribution After Line Segmentation.

Total Number of Sub-Samples: 5765			
Trait	**Low**	**Average**	**High**
EX	1058	2287	2420
NE	392	3019	2354
AG	512	4569	684
CO	233	1803	3729
OE	339	3765	1661

tive preprocessing approach (Adeli Shamsabad and Suen, 2024). Otsu's binarization converts the images into a binary format, enhancing the contrast between text and background, while bilateral filtering reduces noise while preserving important edge details (Figure 1). These methods significantly improve handwriting clarity, thereby facilitating more effective feature detection by the model (Xu et al., 2024).

After processing, the images are converted into a three-channel format to ensure compatibility with neural networks, which typically require RGB input channels. The dataset is then divided into 60% for training, 20% for validation, and 20% for testing. This preprocessing pipeline including line segmentation and image enhancement ensures that the dataset is optimized for training a neural network model, promoting improved generalization capabilities and greater potential for accurate personality trait prediction.

2.3 Model Development

A new approach is introduced to automatically extract relevant features for classifying multi-level personality traits by adapting the TrOCR model for classification tasks. For a baseline comparison, the performance of the proposed model is evaluated against three pre-trained models: ResNet50 and Vision Transformer (ViT) base 16 at two input resolutions (224 × 224 and 384 × 384). Two classification approaches: multi-class and multi-label binary classification, are investigated to identify the method that best addresses data imbalance and supports effective learning for traits with limited samples. In both approaches, focal loss is employed to minimize the impact of well-classified instances, focus the model on challenging samples, and manage the imbalanced distribution of trait levels (Lin et al., 2017).

2.3.1 Classification Layer

Multi-Class Classification: In this approach, each personality trait is treated as a separate task with three mutually exclusive levels: low, average, and high. Five classification heads are used, each dedicated to one trait, producing logits for these three levels. Softmax activation followed by cross-entropy loss is applied to assign probabilities across the levels for each trait (Goodfellow et al., 2016). To address the dataset imbalance, targeted data augmentation techniques including random rotation, affine transformations, perspective distortion, color jitter, and Gaussian blur are selectively applied to minority classes to enrich their representation, thus promoting more balanced learning and improving generalization (Shorten and Khoshgoftaar, 2019).

To further reduce the effects of imbalance, 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 (Luo et al., 2024).

Multi-Label Binary Classification: The model in this approach is configured to use 15 binary classification heads, one for each level across all five traits. Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) combined with sigmoid activation is applied, allowing each level within each trait to be treated as an independent binary classification task (Nam et al., 2014). In this framework, data augmentation is applied uniformly across all classes, allowing the model to develop a balanced understanding of each trait level, regardless of class distribution. This configuration enables the model to learn each trait's characteristics independently, without being influenced by the distribution of other levels, making it particularly effective for traits with imbalanced levels.

These two approaches establish a structure for comparing the efficacy of multi-class and multi-label binary classification in managing imbalanced data. By evaluating both methods, findings are provided into which approach better supports balanced learning and improves classification performance across all personality traits.

2.3.2 CNN Architecture: ResNet50

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 (Vargoorani and Suen, 2024). ResNet50, a deep CNN architecture, is selected in this study for its powerful feature extraction capabilities and its proven effectiveness in handwriting classification tasks shown in our previous research. It demonstrates strong performance in capturing meaningful features from handwriting data, outperforming other models in simpler classification settings (Adeli Shamsabad and Suen, 2024). ResNet50 offers a favorable balance between performance and computational efficiency, making it a practical choice for handwriting feature extraction, especially when compared to more resource-intensive models like transformers (Raghu et al., 2021).

2.3.3 Vision Transformer: ViT Base 16

The Vision Transformer (ViT) is a transformer-based model adapted specifically for computer vision tasks. Unlike CNNs, which rely on convolutional filters to capture local patterns, ViT splits images into patches, treats each patch as a token, and processes these tokens using a transformer encoder. This structure enables ViT to capture global dependencies across the entire image, which can be particularly advantageous for tasks requiring an understanding of spatial relationships, such as handwriting analysis (Koepf et al., 2022).

To explore the effectiveness of ViT at different scales and to compare its performance with TrOCR, two configurations are evaluated in this study: ViT base 16 with input resolutions of 224 x 224 and 384 x 384 (Zhai et al., 2021). The ViT base 16-224 is chosen as a computationally efficient baseline, offering faster processing and a general overview of handwriting features. In contrast, the ViT base 16-384 allows for the capture of finer handwriting details, providing insights into how higher resolutions impact feature details and classification performance (Dosovitskiy, 2020).

These configurations evaluate the applicability of vision transformers for handwriting-based personality trait classification, focusing on computational efficiency and feature extraction compared to TrOCR. Both models use a transformer encoder with self-attention mechanisms to process patch embeddings, identifying spatial and stylistic handwriting patterns.

2.3.4 The Proposed Transformer Model: TrOCR

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 (Campiotti and Lotufo, 2022), 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 (Li et al., 2021). This architecture effectively captures local(e.g., character styles, ligatures) and global (e.g., ink width, slant) dependencies 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 (Bhunia et al., 2021).

In this study, the pre-trained TrOCR model, fine-tuned on the IAM handwriting dataset, is repurposed for personality trait classification from handwriting images. Through this modification, instead of converting handwriting images into text, TrOCR's encoder processes handwriting images by transforming them into a sequence of visual tokens, capturing essential handwriting characteristics such as stroke patterns, shapes, and distinctive features. These visual tokens serve as the basis for identifying patterns associated with different personality traits. The text decoder is replaced by a custom classification head that outputs predictions for each personality trait, allowing it to perform both multi-class and multi-label classification depending on the task requirements (Figure 2).

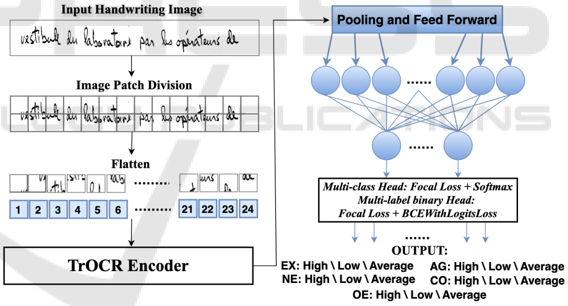

Figure 2: Proposed TrOCR Model for Classification.

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.

2.4 Performance Evaluation Metrics

The effectiveness of each model in classifying personality traits from handwriting images is evaluated using a range of performance metrics. Given the dataset's imbalance and the multi-class, multi-label classification structure, metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUROC) are applied to ensure a comprehensive assessment (He and Garcia, 2009). These metrics are derived from values in the confusion matrix, which reflects model performance by displaying the counts of true positives, true negatives, false positives, and false negatives for each class i:

- True Positives (TP_i): Instances that are Correctly identified as class i.
- True Negatives (TN_i): Instances that are Correctly identified as not class i.
- False Positives (FP_i): Instances that are Incorrectly identified as class i.
- False Negatives (FN_i): Instances of class i Incorrectly identified as another class.

2.4.1 Accuracy

This metric measures the proportion of correct predictions among the total predictions. However, due to the imbalanced dataset, accuracy alone may not reflect the model's true performance across all classes, especially on minority traits (Tanha et al., 2020). Therefore, accuracy is considered alongside other metrics for a more balanced evaluation.

2.4.2 Precision

Precision calculates the ratio of correctly predicted positive instances to the total predicted positives. 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.

2.4.3 Recall

Recall (or sensitivity) is the ratio of correctly predicted positive instances to all actual positives. High recall means the model effectively identifies the target class, minimizing false negatives. This metric is essential for ensuring that all personality traits, especially those with fewer samples, are accurately detected by the model.

2.4.4 F1-Score

The F1-score, calculated as the harmonic mean of precision and recall, provides a single metric that balances both measures. The F1-score is particularly relevant in the case of imbalanced data, as it offers a more comprehensive view of a model's performance when precision and recall are equally significant. For the multi-class and multi-label classification tasks, a weighted average F1-score is calculated across all traits to assess overall performance.

2.4.5 AUROC

AUROC is used to evaluate the model's ability to distinguish between classes, a high AUROC score reflects strong performance, balancing sensitivity (true positive rate) and specificity (true negative rate) indicating better separability. 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.

Each of these metrics is calculated individually for each personality trait, and the average performance across all traits is reported. This approach provides a clear assessment of the proposed method's effectiveness and allows for meaningful comparisons with other deep-learning models. It ensures the evaluation captures the challenges of classifying individual traits while highlighting how well each model handles class imbalance and distinguishes between personality traits.

3 RESULTS AND DISCUSSION

This section presents the result of two classification strategies: multi-class classification using Cross-Entropy with Softmax and multi-label binary classification using BCEWithLogitsLoss, to predict multi-level personality traits from imbalanced handwriting data. The performance of the proposed TrOCR model is analyzed in comparison to three baseline models: ResNet50 and ViT base 16 with input resolutions of 224 and 384. All models were trained for 100 epochs using an NVIDIA A100 Tensor Core GPU and the outcomes of these experiments are detailed in the subsequent subsections.

3.1 Multi-Label vs. Multi-Class Classification

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 2, in the multi-class classification approach, ResNet-50 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 ResNet-50 but still competitive.

In contrast, the multi-label binary classification approach significantly improves the performance metrics across all models, underscoring the advantages of independently optimizing each trait. ResNet-50 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 configuration, achieving the highest accuracy of 84.46% and an F1-score of 0.810.

Table 2: 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**

The AUROC scores in table 3 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 CO 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 are observed for EX, NE, AG, and OE 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 3: Comparison of AUROC Scores for Classification Methods Across Models.

Model	**Traits**	**Multi-Class AUROC**	**Multi-Binary AUROC**
ResNet-50	EX	0.8307	0.8678
	NE	0.7638	0.8255
	AG	0.7273	0.8770
	CO	0.8412	0.8827
	OE	0.8863	0.9291
ViT-224	EX	0.7532	0.8178
	NE	0.6340	0.7827
	AG	0.6117	0.8434
	CO	0.7694	0.8498
	OE	0.8330	0.8971
ViT-384	EX	0.7716	0.8585
	NE	0.7553	0.8238
	AG	0.7015	0.8738
	CO	0.7813	0.8591
	OE	0.8470	0.9192
TrOCR	EX	0.6419	**0.9179**
	NE	0.6493	**0.8850**
	AG	0.6642	**0.9138**
	CO	0.5393	**0.8943**
	OE	0.6719	**0.9334**

3.2 Performance Comparison of TrOCR vs. Other Models

To provide a reliable comparison for evaluating the performance of the proposed TrOCR model, three pre-trained deep learning models, ResNet50, ViT-224, and ViT-384, are trained using the same classification approach and dataset. This approach ensures a reasonable and consistent comparison, as one of the primary objectives of this study is the multi-level classification of personality traits. Since no directly comparable work is found addressing multi-level personality classification from handwriting using a similar methodology, these models are selected to benchmark the effectiveness of TrOCR.

The results reveal distinct differences in the models' abilities to process handwriting data. As shown in the training accuracy plot (Figure 3), TrOCR achieves the highest training accuracy, exceeding 95% in later epochs. ViT-384 demonstrates competitive accuracy but remains below TrOCR. ResNet50 shows slower improvements and surpasses ViT-224, but both models perform lower than TrOCR and ViT-384.

The training loss plot (Figure 3) further emphasizes TrOCR's superior performance. The lowest final loss values are achieved by TrOCR, reflecting its ability to learn handwriting features effectively and minimize errors efficiently. While ViT-384 shows moderate performance with lower loss values than ResNet50 and ViT-224, these baseline models exhibit higher loss values, indicating less effective learning and feature extraction.

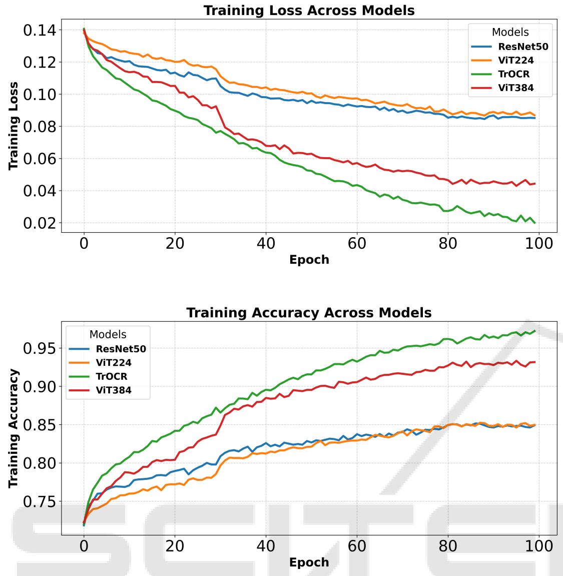

Figure 3: Training Performance Comparison Across Models in Multi-Label Binary Classification.

The AUROC comparison, shown in Figure 4, highlights the performance differences among the models in distinguishing between classes in the multi-label binary classification task. The TrOCR model achieves the highest AUROC of 0.91, demonstrating its strong capability in extracting handwriting features and separating classes effectively. This score reflects the model's ability to handle the complexities of handwriting data with precision.

ResNet50 achieves an AUROC of 0.88, showing competitive performance but still falling short of the TrOCR model. Among the Vision Transformer models, ViT-384 performs slightly better with an AUROC of 0.87, while ViT-224 achieves a lower AUROC of 0.84. This performance suggests that the higher input resolution used by ViT-384 aids in capturing more detailed handwriting features compared to ViT-224.

Looking at the test metrics in Table 4, TrOCR once again stands out as the top performer, with an accuracy of 84.26%, precision of 0.823, recall of 0.842, and F1-score of 0.832. These results show that TrOCR handles unseen data more effectively.

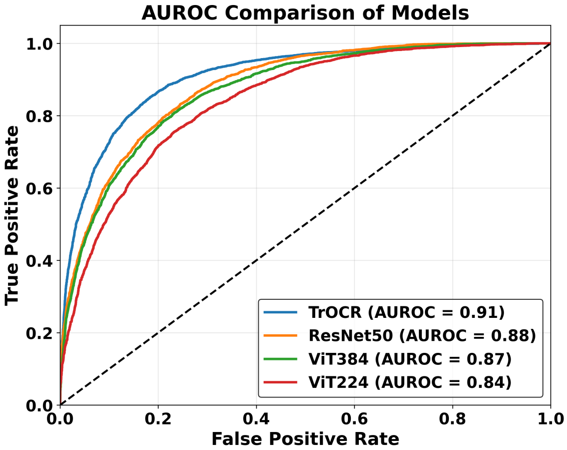

Figure 4: AUROC Score Comparison Across Models in Multi-Label Binary Classification.

ResNet50 follows with an accuracy of 80.52% and an F1-score of 0.778, showing reasonable but lower performance. ViT-224 and ViT-384 perform less effectively, with ViT-224 achieving the lowest test accuracy at 77.71% and an F1-score of 0.742. ViT-384 performs slightly better, with an accuracy of 80.07% and an F1-score of 0.772, close to the results of ResNet50. The consistently high performance of TrOCR across training accuracy (Figure 3), AUROC score (Figure 4), and test metrics (Table 4) indicate that TrOCR captures handwriting patterns more effectively than the other models.

Table 4: Performance of Models on Test Dataset for Multi-Label Binary Classification.

Models	Loss	Accuracy	Precision	Recall	F1-Score
ResNet50	0.121	80.52 %	0.781	0.775	0.778
ViT-224	0.171	77.71 %	0.750	0.737	0.742
ViT-384	0.138	80.07 %	0.777	0.768	0.772
TrOCR	**0.106**	**84.26 %**	**0.823**	**0.842**	**0.832**

The confusion matrices (Figure 5) demonstrate that TrOCR performs effectively in capturing patterns for most personality traits, with notable TP rates such as 466 for AG in the Low class and 389 for OE in the Low class. High TN values are also observed across traits, including 678 for OE in the High class, reflecting robust class separation. However, higher FN counts in traits like CO and NE, particularly in the Low class, indicate challenges in distinguishing these levels due to the smaller number of samples available. These results highlight TrOCR's capability to extract distinctive handwriting patterns and classify personality traits accurately, even when faced with imbalanced data.

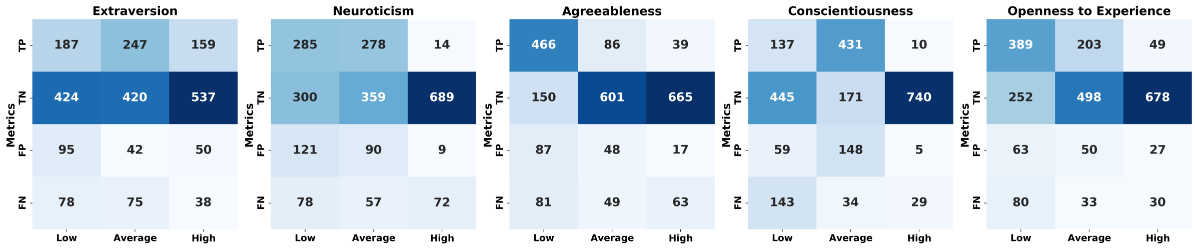

Figure 5: TrOCR Confusion Matrix per Trait.

4 CONCLUSION

The study introduces a new approach by adapting the pre-trained TrOCR model for automatic handwriting pattern recognition and multi-level personality trait classification. The results show that TrOCR consistently outperforms ResNet50, ViT-224, and ViT-384 across all metrics. On the test dataset, TrOCR achieves an AUROC score of 0.91, the highest accuracy of 84.26%, a precision of 0.823, a recall of 0.842, and an F1-score of 0.832, demonstrating its superior capability to learn and generalize handwriting patterns associated with personality traits. The model also records the highest training accuracy and lowest training loss, further emphasizing its effectiveness. Confusion matrix analysis highlights its strong True Positive (TP) and True Negative (TN) rates for traits like Agreeableness and Openness to Experience, though challenges persist for low-class samples in traits such as Conscientiousness and Neuroticism due to limited representation.

The findings underline the effectiveness of the multi-label binary classification approach in addressing class imbalances, which are common in personality trait datasets. By treating each trait independently, this approach enhances the model's ability to learn from underrepresented classes, thereby improving its overall performance.

Future work will focus on expanding the diversity and size of handwriting datasets to improve the model's robustness and generalizability. Combining the strengths of CNNs and transformers through ensemble modeling could also help achieve better accuracy and stability. Since the model learns handwriting patterns through mechanisms like convolutional filters and attention heads, techniques such as Grad-CAM or SHAP could be used to highlight which handwriting features or regions influence predictions the most. Aligning these findings with graphology principles could clarify how handwriting relates to personality traits, making the model more practical and transparent.

Additionally, applying these techniques would build trust in applications like psychological assessments and forensic analysis. Collaborating with experts in psychology and linguistics to tailor the model for real-world use will help validate its effectiveness and refine its design, ultimately making handwriting-based personality analysis more scalable and reliable.

REFERENCES

Adeli Shamsabad, M. and Suen, C. Y. (2024). Deep multi-label classification of personality with handwriting analysis. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 218–230. Springer.

Alshouha, B., Serrano-Guerrero, J., Chiclana, F., Romero, F. P., and Olivas, J. A. (2024). Personality trait detection via transfer learning. *Computers, Materials & Continua*, 78(2).

B J, B., S, K., Suraj, A., K, N., and Venkitesan, P. (2024). Handwriting analysis for classification of human personality. pages 1–6.

Bhunia, A., Khan, S., Cholakkal, H., Anwer, R., Khan, F., and Shah, M. (2021). Handwriting transformers.

Campiotti, I. and Lotufo, R. (2022). Optical character recognition with transformers and ctc. pages 1–4.

Chin, X. Y., Lau, H. Y., Chong, Z. X., Chow, M. P., and Salam, Z. A. A. (2021). Personality prediction using machine learning classifiers. *Journal of Applied Technology and Innovation (e-ISSN: 2600-7304)*, 5(1):1.

Costa, P. T. and McCrae, R. R. (1997). Chapter 11 - longitudinal stability of adult personality. In Hogan, R., Johnson, J., and Briggs, S., editors, *Handbook of Personality Psychology*, pages 269–290. Academic Press, San Diego.

Dhumal, Y. R., Shinde, A., Chaudhari, K., Oza, S., Sapkal, R., and Itkarkar, S. (2023). Automatic handwriting analysis and personality trait detection using multi-task learning technique. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 348–354.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gavrilescu, M. and Vizireanu, N. (2018). Predicting the big five personality traits from handwriting. *EURASIP Journal on Image and Video Processing*, 2018(1):57.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Koepf, M., Kleber, F., and Sablatnig, R. (2022). *Writer Identification and Writer Retrieval Using Vision Transformer for Forensic Documents*, pages 352–366. Springer International Publishing.

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Luo, J., Yuan, Y., and Xu, S. (2024). Improving gbdt performance on imbalanced datasets: An empirical study of class-balanced loss functions.

Mukherjee, S., Ghosh, I., and Mukherjee, D. (2022). *Big Five Personality Prediction from Handwritten Character Features and Word 'of' Using Multi-label Classification*, pages 275–299.

Nair, G., Rekha, V., and Krishnan, M. S. (2021). *Handwriting Analysis Using Deep Learning Approach for the Detection of Personality Traits*, pages 531–539.

Nam, J., Kim, J., Gurevych, I., and Fürnkranz, J. (2014). Large-scale multi-label text classification — revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452, Berlin, Heidelberg. Springer Berlin Heidelberg.

Puttaswamy, B. S. and Thillaiarasu, N. (2025). Fine densenet based human personality recognition using english hand writing of non-native speakers. *Biomedical Signal Processing and Control*, 99:106910.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128.

Rahman, A. and Halim, Z. (2022). Predicting the big five personality traits from hand-written text features through semi-supervised learning. *Multimedia Tools and Applications*, 81:1–17.

Roberts, B. W. and Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science*, 17(1):31–35. PMID: 19756219.

Sayed, A. M. A., Selim, A. W. G., Ashraf, A., and Emam, E. (2024). Analyzing handwriting to infer personality traits: A deep learning framework. In *2024 Intelligent Methods, Systems, and Applications (IMSA)*, pages 58–63.

Shorten, C. and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60.

Shree, N. and Dr.Siddaraju (2022). Analysis of personality based on handwriting using deep learning.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., and Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7(1):70.

Vargoorani, Z. E. and Suen, C. Y. (2024). License plate detection and character recognition using deep learning and font evaluation. In Suen, C. Y., Krzyzak, A., Ravanelli, M., Trentin, E., Subakan, C., and Nobile, N., editors, *Artificial Neural Networks in Pattern Recognition*, pages 231–242. Springer Nature Switzerland.

Xu, Y., Tang, Y., and Suen, C. Y. (2024). Two key factors in handwriting analysis for personality prediction. In *Fifth International Conference on Image, Video Processing, and Artificial Intelligence (IVPAI 2023)*, volume 13074, pages 102–107. SPIE.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2021). Scaling vision transformers. *arXiv preprint arXiv:2106.04560*.