Personality Recognition in Learning Environments: A Multimodality Machine Learning Framework

Yalan Dai
Information Science Division, Graduate School of Science and Technology, Sophia University, Tokyo, Japan
hitomeryu@gmail.com (corresponding author)

Amena Mahmoud
Department of Computer Science, Faculty of Computers and Information, Kafrelsheikh University, Governorate, Egypt
amena_mahmoud@fci.kfs.edu.eg

Eiko Takaoka
Department of Information and Communication Sciences, Faculty of Science and Technology, Sophia University, Tokyo, Japan
m-g-eiko@sophia.ac.jp

Received: 29 January 2026 | Revised: 2 March 2026 and 14 March 2026 | Accepted: 18 March 2026

 DOI: https://doi.org/10.48084/etasr.17778

ABSTRACT

Learning Analytics (LA) and Educational Data Mining (EDM) have emerged as powerful tools for understanding student behavior and improving educational outcomes. This research presents an approach to personality recognition in educational settings by leveraging multimodal data mining techniques across diverse educational datasets. It analyzes student interactions, learning patterns, and behavioral indicators collected from university educational platforms, including learning management systems. Our methodology combines traditional Machine Learning (ML) algorithms with Deep Learning (DL) approaches to process and analyze multiple data streams simultaneously. We employed a comprehensive dataset of 76 students across two different courses to incorporate students' features. Results demonstrate significant improvements in personality recognition accuracy compared to single-model approaches. The proposed system achieved an overall accuracy of 96% in identifying personality traits, with particularly strong performance in recognizing Extraversion and Conscientiousness. Furthermore, this study revealed notable correlations between certain personality traits and specific learning behaviors, providing valuable insights for personalized learning interventions. This research contributes to the field by introducing a scalable framework for personality recognition in educational contexts, offering practical applications for adaptive learning systems and personalized educational support. The findings have important implications for developing targeted interventions and customized learning experiences based on individual personality profiles.

Keywords-personalized learning; Machine Learning (ML); Deep Learning (DL); personality recognition

I. INTRODUCTION

Personality recognition is an important tool supporting personalization in various application domains such as recommendation systems, behavior modeling for human-computer interaction, or investment decisions. In the field of education, recognizing students' personalities should not only drive learning process personalization but also student well-being services and organizational risk detection. Although many studies address personality recognition, most of them deal with the conventional Big Five model [1]. Additionally, their application to educational backgrounds is limited to higher education levels, due to generally unavailable data. Unofficial sources may be a good alternative when dealing with younger generations. The present study starts from the assumption that the effectiveness of recognizing the personalities of digital natives through neural networks depends on the detailed preprocessing of the data input. Therefore, we propose and evaluate Educational Data Mining (EDM) to support personality recognition. For this purpose, a

multimodal framework merging the personality of students perceived by teachers and students themselves was designed to measure both before and after laboratory work [2].

Due to the development of new technologies and digital learning environments, a large amount of educational data is now more readily accessible [3]. Researchers and practitioners may simultaneously gather and examine data to find patterns using powerful data mining techniques. These patterns can be used to predict students with learning difficulties, find ways to improve the learning process [4], and make appropriate decisions that benefit learners. Therefore, two communities focused on exploring data collected from educational environments to enhance learning processes and outcomes—EDM and Learning Analytics (LA)—are widely developing and evolving. EDM focuses on discovering patterns in educational data, whereas LA aims to identify students' learning states by applying information collected from data such as student interactions. Both share the common goal of using information technology to make education more efficient and to timely identify potential problems in students [4].

Personalized education has always been an ideal goal of educational science and practice. Educators and policymakers hope for personalization as a panacea for achievement gaps, lack of student motivation, and more effective teaching. Broadly speaking, personalized education refers to adapting teaching to specific learners, in contrast to "traditional" teaching aimed at an entire group of learners. By changing teaching methods, content, or teaching rates based on certain characteristics of learners, it is suggested that individual shortcomings of learners can be addressed, and their resources can be utilized [5].

The effectiveness of personalized education has been demonstrated in some literature. A key argument for the effectiveness of personalization can be derived from empirical evidence, namely that the learning gains from one-on-one tutoring are two standard deviations higher than traditional classroom teaching [6]. Extending the benefits of one-on-one tutoring to larger learning groups remains one of the driving forces behind personalization research.

To achieve personalized education, it is necessary for teachers to understand the characteristics of each student and teach according to their learning styles. In the numerous studies in the fields of LA and EDM in computer-assisted education, researchers have different research purposes, such as predicting students' learning styles [7], classroom situations [8], etc. In these predictive studies, student modeling or feature engineering is first required.

Personality recognition techniques usually use text or image data as their input. Additionally, text data from drawing prompts and prediction values of emotions and age are also used. Furthermore, data from users when they play and those who design the games are combined. Several studies on personality recognition also use cross-modality data. Authors in [9] used a multimodal approach to extract high-level features by collapsing data from several output measurements into a single input feature set. In their survey, authors in [10] introduced the educational data used in modeling studies to predict students' academic performance, which is referred to as students' attributes in this article. The statistics in this article indicate that 20 studies have used students' personalities as students' attributes to predict academic performance. Other data used for prediction include attendance, scores, social activities, academic activities, and demographics.

In the studies that used personality as modeling data, personality information was obtained through the Big Five personality prediction questionnaire. The Big Five, which reflects stable patterns of individual differences [11], has emerged as a standard method of modeling personality and has become the most widespread and generally accepted model. The Big Five model describes personality across five traits: Openness to Experience, which captures intellectual curiosity, imagination, and receptiveness to novel ideas; Conscientiousness, reflecting self-discipline, organization, and goal-directed behavior; Extraversion, characterized by sociability, assertiveness, and positive emotionality; Agreeableness, associated with interpersonal warmth, altruism, and cooperativeness; and Neuroticism, which pertains to emotional instability, anxiety, and susceptibility to negative affect. The Big Five has been validated across diverse cultures, languages, and age groups, supporting its universality and robustness [12, 13]. As a result, the model has become foundational in personality psychology and has been extensively applied in fields such as organizational behavior, clinical assessment, and especially educational research.

Within higher education, the Big Five personality traits have proven particularly valuable in explaining individual differences in academic performance, learning approaches, and student adjustment. Among university students, Conscientiousness has consistently shown the strongest positive correlation with academic success, as it reflects traits such as diligence, persistence, and organization—qualities critical for managing the demands of autonomous learning environments [14]. Openness to Experience is also positively associated with deeper cognitive engagement and a preference for analytical and reflective learning strategies, which are beneficial in higher-order thinking and complex academic tasks. Conversely, Neuroticism is often linked to lower academic performance due to its association with anxiety, emotional instability, and difficulty coping with academic stress. Research also suggests that integrating personality assessment into student support services may enhance educational outcomes by allowing institutions to offer more targeted interventions, such as study skills training or stress management resources tailored to students' personality profiles [7]. Thus, the Big Five framework offers a robust theoretical and empirical basis for understanding and improving university students' academic experiences.

Beyond the widely accepted Big Five framework, additional personality constructs have been proposed to capture specific cognitive and motivational tendencies that are not fully encompassed by the five-factor model. One such construct is dichotomous thinking, defined as a cognitive style characterized by a preference for black-and-white categorizations and an intolerance of ambiguity. Individuals with a high tendency toward dichotomous thinking often

perceive experiences, judgments, and interpersonal situations in extreme, mutually exclusive terms, which can contribute to cognitive distortions and emotional instability [15]. The Dichotomous Thinking Inventory (DTI) was developed to measure this construct and has demonstrated robust psychometric properties.

Another construct gaining attention in personality research is psychological entitlement, which reflects a pervasive sense of deservingness and expectation of special treatment, irrespective of actual performance or merit. This trait, typically assessed by the Psychological Entitlement Scale (PES), has been empirically associated with narcissistic tendencies, interpersonal conflict, and unethical decision-making [16]. Incorporating such constructs alongside the Big Five offers a more comprehensive understanding of individual differences, particularly in domains related to cognition, affect regulation, and social behavior.

There are very limited studies on DTI and PES related to education. We consider that dichotomous thinking will affect students' understanding of concepts in the learning process [17], whereas psychological entitlement will affect their attitude towards the teacher-student relationship. In order to apply Automated Personality Recognition (APR) technology in the field of learning analysis, we also decided to explore the feasibility of APR for these two traits in the student group.

The dominant personality traits of each student can be deduced using the NEO Five-Factor Inventory (NEO-FFI) questionnaire [18], whose questions are easy to answer for students. However, collecting data through questionnaires has issues such as inefficiency and subjective or random responses from participants. Automated personality detection technologies, which involve automatic data collection and analysis, can mitigate these issues. Automated personality trait recognition aims to use computer science techniques to model the problem of personality trait recognition in cognitive science. It is one of the most important research topics in the field of personality computing [19]. Based on the type of input data, automated personality trait recognition can be divided into single-mode and multi-mode. Specifically, it includes single-mode audio or visual personality trait recognition and multimodal personality trait recognition, which integrates multimodal behavioral data such as audio, visual, and textual information.

Despite the growing body of work on APR, existing studies predominantly rely on social media content, self-authored text, or sensor-based data that are impractical for routine classroom deployment. Moreover, little is known about how different readily available educational data modalities vary in their effectiveness for recognizing specific personality traits. This study addresses this gap by systematically evaluating handwriting, short speech recordings, and learning management system data—collected without specialized sensors—to identify which modalities and learning models are most suitable for personality recognition in real educational environments.

The proposed contributions of the present research are:

- Collecting data from 76 university students over three semesters in different courses, including personality questionnaires, various types of handwriting, audio, and Moodle system data.
- Providing a novel multimodal framework which refers to (1) multiple input modalities (handwriting, audio, and online activity data), (2) multiple classification algorithms, including Multilayer Perceptron (MLP), Random Forest (RF), Convolutional Neural Network (CNN), and Attention-based CNN (Attention-CNN), and (3) multiple target personality traits. The framework does not propose a new algorithm but provides a systematic evaluation pipeline for analyzing how different modality–model combinations affect personality recognition outcomes.
- Evaluating the effectiveness of different personality traits to identify the most suitable personality classification methods and data for each trait.

The proposed contributions can be described in these three research questions:

1. Focuses on the role of data modality: Which classroom-available modalities—handwriting, short audio segments, or learning management system behavioral logs—are most effective for recognizing different personality traits?
2. Examines the impact of modeling approaches: How do traditional Machine Learning (ML) models compare with lightweight Deep Learning (DL) architectures under small-sample conditions typical of real educational environments?
3. Extends beyond the Big Five traits and asks whether multimodal, sensor-free data can feasibly support the recognition of underexplored constructs such as dichotomous thinking and psychological entitlement?

Together, these questions define the scope of our analysis and provide a structured basis for evaluating the feasibility and limitations of multimodal personality recognition in practical classroom settings. This study was conducted with the approval of the Sophia University Ethics Committee for Research on Human Subjects.

II. METHODOLOGY

This section presents data acquisition, data preprocessing, data analysis, and proposed ML and DL models. The data acquisition section mainly introduces detailed information of each course and each type of data collected over three semesters. It is noteworthy that this study avoided the use of expensive sensor equipment and the collection of student privacy information to ensure the applicability and scalability of the research. The data preprocessing section describes the methods used to extract feature vectors for each type of data and how the personality questionnaires were processed into prediction labels. The data analysis section introduces the various classification methods selected for this study, including basic ML and neural networks. The purpose of our comparative experiments is to identify the most suitable data types and

prediction methods for APR based on the above data. Figure 1 illustrates the flowchart of the proposed research.

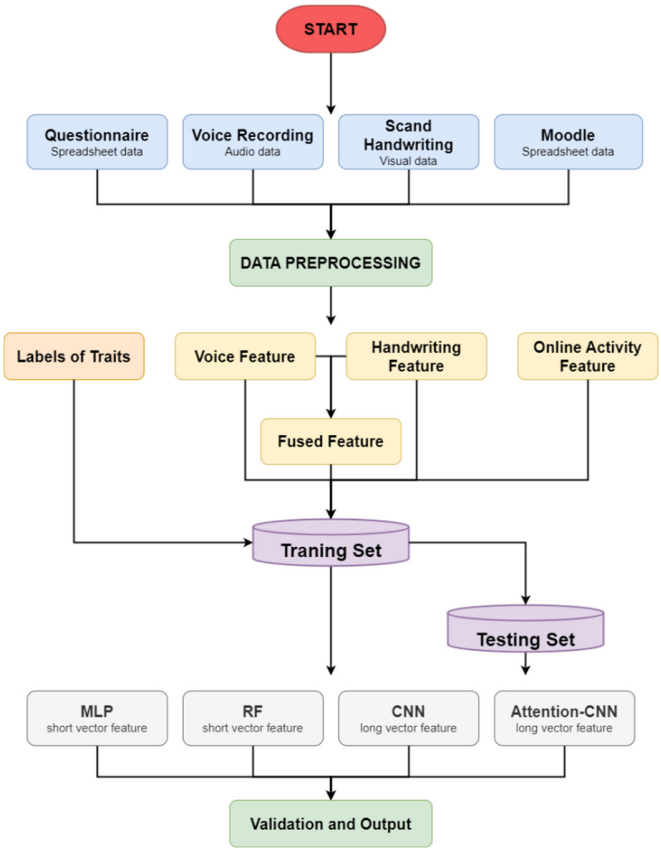

Fig. 1. Proposed research flowchart.

A. *Data Acquisition*

In order to ensure that this LA method can be used in as many schools as possible, we only considered data that can be collected without any sensor or expensive device. Our data were collected from two courses, one is Information and Human Interaction, which is a discussion-based course, and the other is Database, which follows a traditional teaching-based approach.

It should be noted that the collected Moodle data are different as the course changes. Students in the Information and Human Interaction course can come from any major, whereas students in the Database course are exclusively from the Science or Engineering major. The collected data also differ due to the varying teaching methods. However, we classify the data from both classes into four types based on data type: questionnaire data, Moodle data, handwritten data, and audio data. We collected three types of data for training and one for labeling. The descriptions of the data are as follows.

1) *Dataset 1: Personality Questionnaire*

At the beginning of the semester, we provided the Google Forms URL for the personality test questionnaire, which students could fill in after class. The questionnaire has three parts: the Big Five personality traits, the DTI [15] and the PES [16]. Both the Information and Human Interaction course and Database course students were required to fill in the same questionnaire. The questionnaire consists of 151 questions related to student's feelings. Each question has five response options, ranging from "completely disagree" to "completely agree," with gradations in between.

2) *Dataset 2: Moodle Activity*

At the end of the semester, we collected learning data from the Moodle system. As mentioned before, the Moodle activity data that we could collect varies between the two courses. In the Information and Human Interaction course, the data were as follows: binary data on whether a student was the group leader, how many times a student asked questions for the presentation, reaction score, attendance, and final score.

In the Database course, we obtained the following data: attendance, scores from midterm and final closed-book exams (on a 200-point scale), final course assessment results (graded A–D), and weekly online post-class assignment submission details (submission time, submission frequency, total time spent). The attendance data, counts of questions, and reaction score were collected manually, whereas the other data were downloaded from the Moodle system directly.

3) *Dataset 3: Audio Data*

In the Information and Human Interaction course, there were group presentations and question-and-answer sessions. Specifically, each group has a 20 to 30 min presentation about information technology. Students' audio data and presentation content were recorded and collected as experimental data. Group presentations and question-and-answer sessions were held for all students, but presentations during class time were limited to two groups per day, with four to five people per group. Recordings were only made for those who agreed to participate in the study.

In the Database course, due to its nature as a traditional teaching-based course, we incorporated a homework answering session led by the teacher before each class. During this session, three students were selected based on their student IDs to answer some of the previous week's homework questions face-to-face. The answering time, including stating their own name and student ID, was approximately 30 s, with some short-answer questions lasting over a minute. We recorded all of these interactions.

4) *Dataset 4: Handwritten Data*

In the Database course, we scanned papers from midterm examinations and final examinations to collect the characteristics of students' handwriting as experimental data. The entire test paper was scanned, and portions of handwriting were used as the data source. To distinguish the difference between content-independent and content-dependent effects in handwriting features, we chose to use all students' signatures as the content-independent part. For students in the Database course, their fixed fill-in-the-blank answers were selected from the final exam as the content-dependent part. The extracted portions ensured that the content was the same so that the differences in the input data were due to individual differences in handwriting.

We extracted four different types of content from each student's test paper, which were: English, numbers, Japanese kanji, and Japanese katakana. The contents were: CRUD, 1458, 概念 (concept), データ (data). For students in the Information and Human Interaction course, since there was no exam, we asked these students to write four different types of characters, namely English, numbers, Japanese kanji, and Japanese kana as the content-dependent handwriting data source. The contents wer: A B C D, numbers 0–9, and the Japanese writing of Information and Human Interaction (including kanji and kana).

B. Data Preprocessing

1) Questionnaire Preprocessing

Questionnaire data contain unique student IDs and 151 item responses. Processing involves: (1) converting responses to numerical values from 1 to 5 (1 = "completely disagree", 5 = "completely agree"), (2) reversing specific questions per NEO-FFI instructions [20] where 1 replaces 5 and 2 replaces 4, (3) calculating Big Five personality scores (0–5 range) using NEO-FFI mapping tables.

2) Handwriting Image Preprocessing

Scanned signature images were cut to a square format with uniform sizing, grayscale conversion, binarization, and normalization [21]. Images are pixelized into equal-length vectors, then Histogram of Oriented Gradients (HOG) feature extraction is applied for handwriting feature extraction [22].

3) Audio Preprocessing

Audio recordings are processed through: (1) resizing to 16 kHz mono (multi-channel files averaged to a single channel), (2) sampling rate standardization and normalization, (3) windowing to 1 s segments, and (4) Mel-Frequency Cepstral Coefficients (MFCCs) extraction using the librosa library [23]. MFCCs simulate human auditory perception, are converted to logarithmic scale for better frequency representation, and create low-dimensional spectral envelope representations effective for classification tasks [24].

4) Moodle Data Preprocessing

For the Information and Human Interaction course, the data were normalized, and final scores were removed to avoid redundancy. For the Database course, student homework behavior was quantified through two approaches: (1) submission timing analysis using four daily periods (0–6 h early morning, 6–12 h morning, 12–18 h afternoon, and 18–24 h evening), and (2) submission frequency calculated as the total number of submissions divided by the number of weeks to determine the average submissions per homework.

C. Feature Fusion

Since audio and visual signals provide different but related information about the target classes, integrating them at the feature level allows the model to capture cross-modal relationships that may enhance classification performance.

HOG and MFCC features are integrated using an early fusion strategy [25]. Feature vectors are normalized for scale compatibility, then concatenated into composite vectors for joint representation learning, enabling classifiers to capture synergistic cross-modal patterns.

D. Data Analysis

There are 37, 36, and 47 students in the three classes, respectively. After removing the students who did not sign the consent form and the students whose data were incomplete, we obtained a total of 76 students with complete datasets of four data types.

After feature extraction, the dimension of the audio data vector is 416, the dimension of the handwriting data vector is 400, and the dimension of the feature vector after early fusion is 816.

Due to limitations in the number of students participating in the study, we decided to consider the personality recognition task as a classification task. As each personality trait is independent of others, we consider performing a binary classification task on each personality trait, i.e., dividing individuals into positive or negative categories based on their scores on the Big Five personality traits. This classification makes regression or fine-grained multiclass classification statistically unstable. Median-based thresholding is a commonly employed strategy in personality computing research when data availability is constrained. Nevertheless, this approach inevitably reduces trait resolution, and the results should be interpreted as distinguishing relatively high vs. low tendencies rather than precise trait levels.

The mean value of 2.5 was chosen as the threshold for classification. The number of individuals classified in each personality trait is shown in the Figure 2. We found that this threshold may lead to issues of sample imbalance. Consequently, the utilization of the median value of 3 as the threshold for classification was more balanced and was selected as the threshold. Once the threshold was decided, scores on each trait lower than the threshold were classified as negative, and scores higher than the threshold were classified as positive. Following this classification, the students were divided into 2 to the 5th power (i.e., 32) types. We aim to investigate the effects of different data on each personality trait. Therefore, instead of modeling it as a 32-classification problem, we opted to handle binary classification problems separately for each of the five personality traits. The distribution of positive and negative classes for each of the five personality traits is shown in Figure 2.

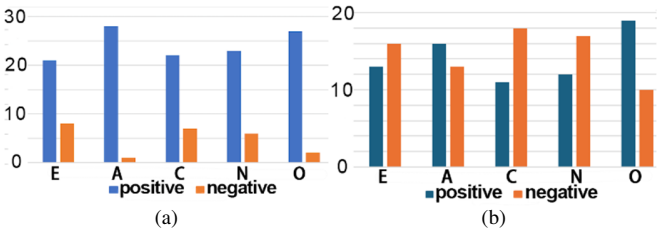

Fig. 2. Distribution of Big Five personality labels in the Database course: (a) with threshold 2.5, (b) with threshold 3.0.

E. Model Comparison

This study investigates ML/DL model performance for personality trait recognition from multimodal learning environment data. Models include traditional ML: MLP [26] and RF [27], and advanced DL: CNN [28] and Attention-CNN [29].

1) Experimental Design

Personality recognition experiments were conducted across seven traits, including the Big Five personality traits, DTI, and PES. The study utilizes multiple data modalities, including Moodle activity data, audio recordings, handwriting images, and multimodal feature fusion.

For Moodle-based data, feature vectors are relatively low-dimensional and were processed using ML algorithms (MLP and RF), which are well suited for tabular data. These experiments were implemented using WEKA software [30] with default parameters and evaluated using 10-fold cross-validation.

For audio and handwriting data, as well as their fused representations, higher-dimensional feature vectors were obtained, making them more suitable for DL models. CNN-based architectures were adopted due to their effectiveness in extracting hierarchical representations from structured input data.

2) Model Architectures

a) Convolutional Neural Network

A standard CNN architecture was employed for feature extraction and classification (Figure 3(a)):

- Feature extraction: Two convolutional blocks (3 × 3 convolution, same padding, stride = 1, ReLU activation, 2 × 2 max-pooling with stride = 2).
- Channels: First layer 16 filters (single-channel input), second layer 32 channels.
- Classifier: Two-layer fully connected (flattened 32-channel maps → 128-dimensional space with ReLU → sigmoid output for binary classification).
- Design: Parameter-efficient through local connectivity (3 × 3 kernels) and convolutional weight sharing.

b) Attention-Based Convolutional Neural Network

An attention-enhanced CNN architecture was employed to improve feature representation (Figure 3(b)):

- Feature extraction: Two convolutional blocks (3 × 3 convolution, same padding, stride = 1, ReLU activation, 2 × 2 max-pooling with stride = 2).
- Attention module: Spatial attention mechanism using parallel average pooling and max pooling along the channel dimension, followed by a 7 × 7 convolution with sigmoid activation to generate attention weights.
- Classifier: Two-layer fully connected (flattened 32-channel maps → 128-dimensional space with ReLU → sigmoid output for binary classification).
- Design: Modulates the original features through element-wise multiplication, effectively highlighting spatially salient regions while suppressing irrelevant areas.

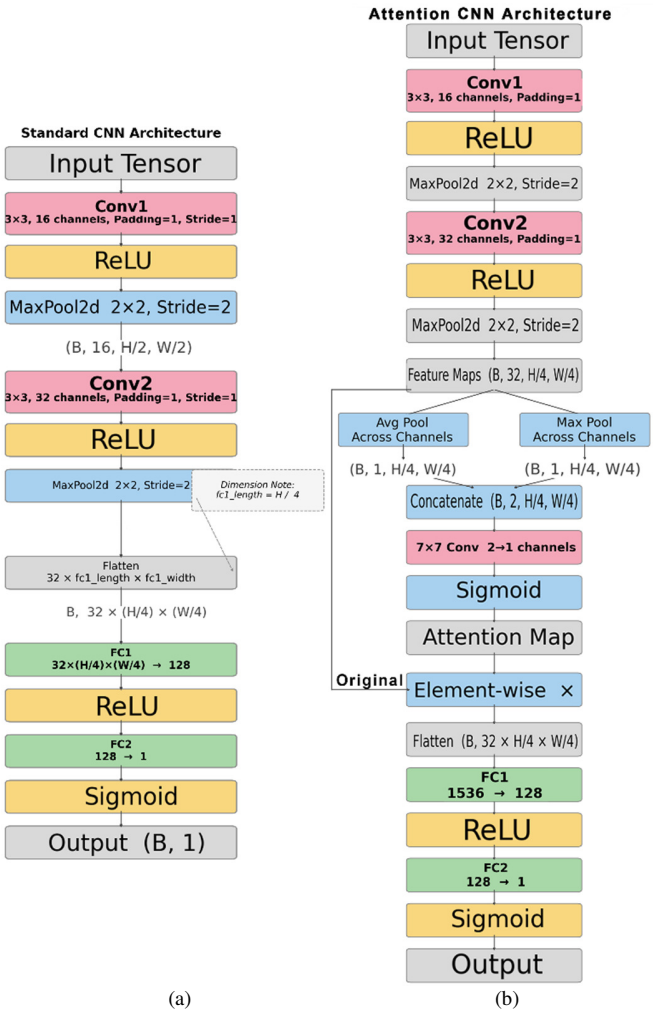

Fig. 3. Structure of the proposed models: (a) CNN, (b) Attention-CNN.

3) Training and Implementation Details

All neural network models were trained using the Adam optimizer, a widely adopted variant of stochastic gradient descent that adaptively adjusts learning rates based on first- and second-moment estimates of gradients, offering faster convergence and improved stability in non-convex optimization landscapes [31]. The initial learning rate was set to 0.001, a commonly used value that balances the speed of convergence with the risk of overshooting optimal minima. The models were trained for 500 epochs, providing sufficient iterations to allow the networks to converge, especially considering the relatively small dataset. For the loss function, Binary Cross-Entropy Loss (BCELoss) was employed, as the classification tasks involved binary labels and the network's output layer was configured with a sigmoid activation function to produce probabilistic outputs in the range [0,1]. This combination of hyperparameters was selected based on

preliminary experimentation and prior empirical evidence, offering a stable configuration for small-sample binary classification tasks involving multimodal features.

III. RESULTS AND DISCUSSION

The experimental results are presented in four sections. The first section presents the APR experimental results of basic ML algorithms on three types of data, along with one fused dataset. The second section presents the APR experimental results of the same algorithm (MLP) on handwritten data with different character types, as well as the results obtained by fusing these handwritten features with the audio features of the same individuals, representing content-dependent handwritten data. The third section presents the APR experimental results of audio features and content-independent handwritten features under different neural networks. The fourth section compares and summarizes all the experimental results.

A. Comparing Four Types of Data on Multilayer Perceptron and Random Forest

We used data collected over three semesters from a total of 76 students for comparative experiments. To evaluate the classification performance across the four data types, we employed the RF and MLP algorithms, which are suitable for handling short feature vectors. As shown in Table I, whether using long feature vectors from audio data or short feature vectors from Moodle data, the classification performance of the RF model is not ideal, with accuracy values generally around 50%, and the best results not exceeding 75%. In contrast, MLP performs better, as shown in Table II, particularly on handwriting and fusion data, where it achieves up to 96% accuracy.

For audio data (Tables I and II), the performance of MLP is comparable to that of the RF model. Moodle data, which have limited feature dimensionality, perform poorly under both algorithms, indicating that it is difficult to automatically recognize students' personalities using only simple and limited behavioral data. Although such features do not capture the full complexity of students' online learning behaviors, their inclusion allows us to evaluate the feasibility of personality recognition using only readily accessible Moodle information.

TABLE I. PERFORMANCE OF THE RF METHOD ON FOUR TYPES OF DATA

Traits	P(Aud)	R(Aud)	F1(Aud)	A(Aud)	P(H)	R(H)	F1(H)	A(H)	P(F)	R(F)	F1(F)	A(F)	P(M)	R(M)	F1(M)	A(M)
Extraversion	0.63	0.63	0.63	0.63	0.54	0.55	0.54	0.54	0.67	0.67	0.67	0.67	0.52	0.53	0.48	0.52
Agreeableness	0.56	0.61	0.57	0.50	0.49	0.56	0.51	0.46	0.59	0.63	0.59	0.53	0.64	0.63	0.63	0.63
Conscientiousness	0.55	0.57	0.55	0.53	0.61	0.63	0.60	0.57	0.50	0.52	0.51	0.49	0.59	0.7	0.64	0.7
Neuroticism	0.47	0.47	0.47	0.47	0.44	0.44	0.44	0.44	0.41	0.41	0.41	0.41	0.66	0.66	0.66	0.66
Openness	0.52	0.55	0.53	0.5	0.55	0.57	0.55	0.54	0.53	0.57	0.52	0.50	0.53	0.52	0.5	0.52
DTI	0.48	0.48	0.48	0.48	0.39	0.39	0.39	0.39	0.48	0.48	0.48	0.48	0.54	0.55	0.54	0.55
PES	0.58	0.57	0.57	0.57	0.46	0.48	0.47	0.45	0.55	0.55	0.55	0.54	0.57	0.55	0.56	0.55

Aud: audio data; H: handwriting data; F: fusion data; M: Moodle data; P: precision; R: recall; F1: F1-score; A: accuracy.

TABLE II. PERFORMANCE OF THE MLP METHOD ON FOUR TYPES OF DATA

Traits	P(Aud)	R(Aud)	F1(Aud)	A(Aud)	P(H)	R(H)	F1(H)	A(H)	P(F)	R(F)	F1(F)	A(F)	P(M)	R(M)	F1(M)	A(M)
Extraversion	0.66	0.66	0.66	0.66	0.98	0.96	0.96	0.96	0.96	0.96	0.95	0.94	0.76	0.73	0.72	0.71
Agreeableness	0.61	0.61	0.61	0.56	0.8	1	0.88	0.82	0.92	0.94	0.92	0.88	0.65	0.65	0.64	0.63
Conscientiousness	0.43	0.45	0.43	0.41	0.87	0.76	0.8	0.92	0.9	0.82	0.85	0.95	0.65	0.65	0.65	0.61
Neuroticism	0.42	0.42	0.42	0.42	0.72	0.83	0.76	0.77	0.79	0.83	0.77	0.76	0.43	0.43	0.43	0.43
Openness	0.54	0.54	0.54	0.52	0.68	0.85	0.73	0.68	0.71	0.84	0.72	0.65	0.67	0.65	0.65	0.65
DTI	0.56	0.55	0.55	0.55	0.58	0.9	0.66	0.6	0.75	0.45	0.52	0.66	0.66	0.65	0.65	0.65
PES	0.56	0.54	0.54	0.54	0.66	0.89	0.74	0.66	0.57	1	0.7	0.57	0.63	0.62	0.61	0.6

Aud: audio data; H: handwriting data; F: fusion data; M: Moodle data; P: precision; R: recall; F1: F1-score; A: accuracy.

B. Comparing Three Types of Handwriting Data

To evaluate the personality recognition performance of various types of content-dependent handwriting features, we conducted the following comparative experiments. Using only one classification method (MLP), we experimented with three types of handwriting (type 1 for English, type 2 for numbers, and type 3 for Japanese) from the same group of students (25 students from the 2024 Spring semester in the Information and Human Interaction course). After comparing the three handwriting types, we also conducted another experiment by fusing the three types of handwriting features with audio feature vectors and comparing the performance using the same classification algorithm. The results are presented in Tables III and IV:

- The classification accuracy for most personality traits remains stable for different handwriting types but a significant variation is observed for Conscientiousness, where performance decreases when using type 3 (Japanese kanji).
- When handwriting features are fused with audio features, the accuracy may remain unchanged or vary by about 10% (Table IV).

- Although the changes in accuracy are observed after feature fusion, the overall performance of the fused features depends on the base handwriting type. The trait Conscientiousness continues to exhibit the same variation trend (Table IV).
- Among all results, the highest accuracy is achieved for Extraversion when type 3 handwriting is fused with audio features, whereas the lowest performance is observed for Neuroticism under the same fusion condition (Table IV).

TABLE III. THREE TYPES OF HANDWRITING CHARACTERS CLASSIFIED USING MLP

Traits	P(1)	R(1)	F1(1)	A(1)	P(2)	R(2)	F1(2)	A(2)	P(3)	R(3)	F1(3)	A(3)
Extraversion	0.65	0.65	0.65	0.86	0.66	0.65	0.64	0.9	0.6	0.5	0.53	0.81
Agreeableness	0.81	0.95	0.84	0.78	0.81	1	0.88	0.81	0.81	1	0.88	0.81
Conscientiousness	0.81	0.9	0.84	0.81	0.7	0.8	0.73	0.73	0.41	0.6	0.48	0.51
Neuroticism	0.53	0.8	0.61	0.53	0.5	0.9	0.63	0.51	0.5	0.8	0.6	0.6
Openness	0.6	0.76	0.64	0.63	0.58	0.61	0.56	0.58	0.68	0.63	0.6	0.58
DTI	0.68	0.8	0.71	0.68	0.51	0.6	0.54	0.51	0.61	0.7	0.64	0.65
PES	0.53	0.65	0.56	0.58	0.58	0.73	0.59	0.51	0.6	0.8	0.63	0.61

1: English characters; 2: Japanese kanji characters; 3: numerical characters; P: precision; R: recall; F1: F1-score; A: accuracy.

TABLE IV. THREE TYPES OF HANDWRITING FEATURES FUSED WITH AUDIO FEATURES AND CLASSIFIED USING MLP

Traits	P(1)	R(1)	F1(1)	A(1)	P(2)	R(2)	F1(2)	A(2)	P(3)	R(3)	F1(3)	A(3)
Extraversion	0.75	0.7	0.71	0.9	0.66	0.65	0.64	0.9	0.8	0.7	0.73	0.93
Agreeableness	0.81	1	0.88	0.81	0.81	0.91	0.82	0.75	0.7	0.9	0.73	0.7
Conscientiousness	0.8	0.85	0.8	0.76	0.6	0.65	0.6	0.68	0.5	0.65	0.54	0.53
Neuroticism	0.53	0.9	0.65	0.61	0.53	1	0.68	0.55	0.36	0.65	0.43	0.41
Openness	0.85	0.86	0.84	0.83	0.65	0.58	0.58	0.76	0.56	0.8	0.63	0.56
DTI	0.7	0.66	0.67	0.76	0.61	0.7	0.64	0.61	0.65	0.8	0.69	0.65
PES	0.68	0.71	0.67	0.73	0.46	0.5	0.43	0.48	0.58	0.8	0.64	0.58

1: English characters; 2: Japanese kanji characters; 3: numerical characters; P: precision; R: recall; F1: F1-score; A: accuracy.

C. *Comparing Neural Networks and All Data Types Based on Three Semesters*

Next, we compare different algorithms for each type of long vector feature data. The experiments used data from all 76 students over three semesters. The handwriting part uses content-independent data, and due to the use of deep neural networks, the comparative data types include only audio, handwriting, and early fusion features, excluding Moodle data with overly short vector lengths.

Figure 4 presents the personality detection experiments using CNN as the classification algorithm on audio, handwriting, and fused features. It can be observed that CNN performs moderately on audio (voice) data (around 0.7–0.8), performs well on handwriting data (greater than 0.8), but performs poorly on the traits of Openness, DTI, and PES (less than 0.6). In the experiments with fused features, CNN performs well on the Big Five traits but poorly on DTI and PES.

To evaluate the classification performance of each data type and classification algorithm, confusion matrices were computed for each personality trait. Each matrix summarizes the counts of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP), allowing for a nuanced analysis of model behavior.

Figures 5–7 show the confusion matrices for the CNN method on the three types of data. When the colors in the top-left (true negative) and bottom-right (true positive) corners of the confusion matrix are significantly darker than those in the bottom-left and top-right corners, the results are considered excellent.

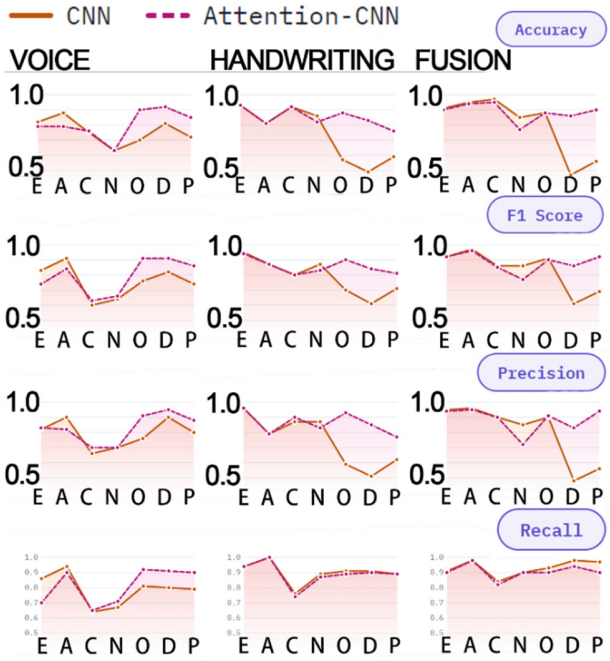

Fig. 4. Classification performance comparison between CNN and Attention-CNN.

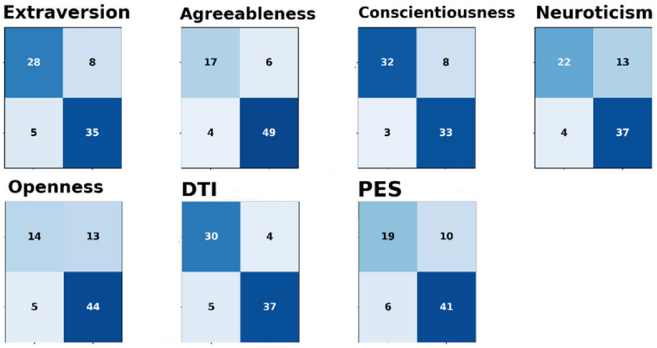

Fig. 5. Confusion matrix of the CNN method for audio data.

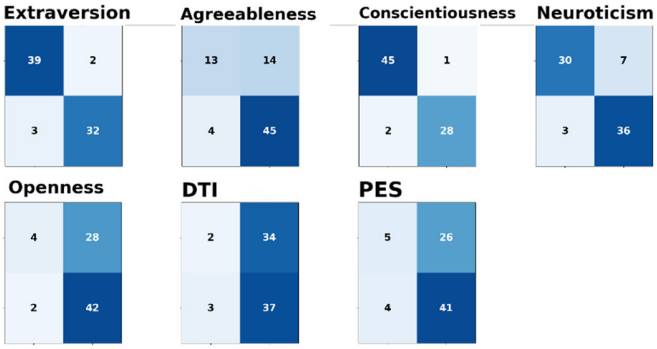

Fig. 6. Confusion matrix of the CNN method for handwriting data.

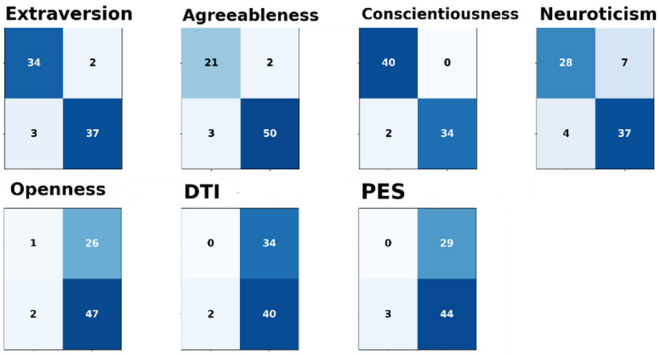

Fig. 7. Confusion matrix of the CNN method for fusion data.

In Figure 4, the results demonstrate predictive capability across personality traits, with particularly high true positive rates (sensitivity) observed in Extraversion, Agreeableness, Neuroticism, Openness, and PES, where the model correctly identified all or nearly all positive samples (e.g., TP = 28 with FN = 0 in Extraversion; TP = 40 with FN = 0 in Agreeableness). However, the relatively higher false positive rates in Neuroticism and Openness suggest a bias toward the positive class, potentially at the cost of precision. DTI presents a slightly different pattern, with a higher number of false negatives (FN = 10), indicating lower sensitivity and suggesting room for improvement in capturing positive instances.

Figure 4 also presents the personality recognition experiments using Attention-CNN as the classification algorithm on audio, handwriting, and fused features. It can be observed that, with the addition of the attention mechanism, most experimental results show slight improvements, especially for DTI and PES, where CNN performed poorly, with more significant improvements. However, some results decreased slightly, such as the Extraversion and Agreeableness traits on audio data. In addition, Figures 8–10 show the confusion matrices for the Attention-CNN method on the three types of data.

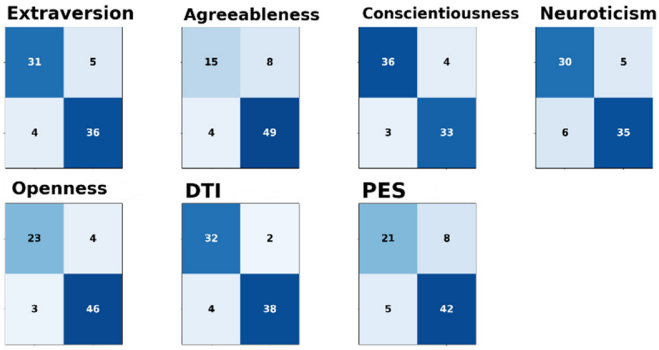

Fig. 8. Confusion matrix of the Attention-CNN method for audio data.

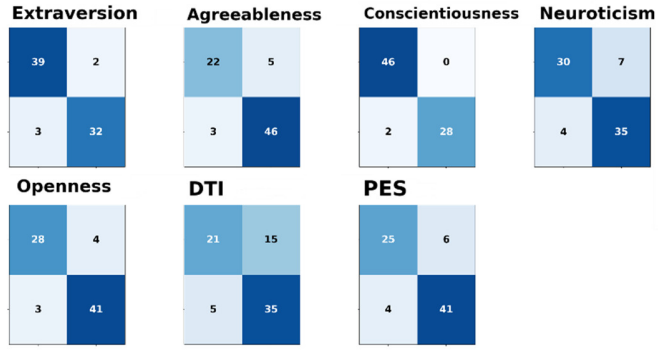

Fig. 9. Confusion matrix of the Attention-CNN method for handwriting data.

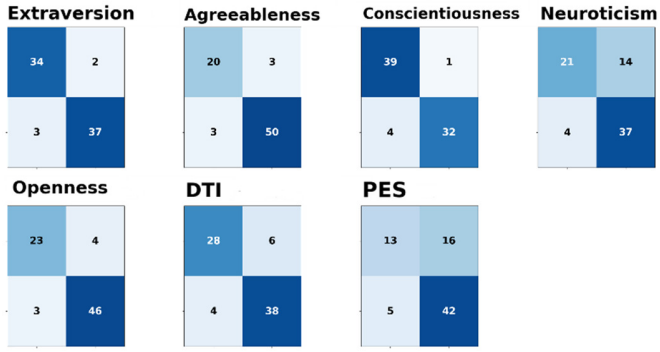

Fig. 10. Confusion matrix of the Attention-CNN method for fusion data.

Given the relatively limited sample size (N = 76), it is essential to assess the statistical robustness of the experimental outcomes using Confidence Intervals (CIs), particularly in the context of neural network-based classification, which can be sensitive to data variability. To evaluate the reliability of the observed performance metrics, 95% CIs were calculated for

both accuracy and F1-scores across all seven binary classification tasks using 10-fold cross-validation in Figure 11. The results reveal that, for several tasks, the width of the CIs was satisfactorily narrow. The tasks where the CI range was less than 0.2 were listed as: traits Extraversion, Agreeableness, Conscientiousness, Openness, and PES classified by CNN using handwriting data; traits Agreeableness and Conscientiousness classified by CNN using fusion data; trait Agreeableness classified by Attention-CNN using audio data; traits Extraversion, Agreeableness, and Conscientiousness classified by Attention-CNN using handwriting data; and traits Agreeableness, Conscientiousness, and Neuroticism classified by Attention-CNN using fusion data. These results indicate relatively stable and consistent model performance. These intervals suggest that the model's generalization ability on these tasks is statistically significant and not likely due to random variation. However, in other tasks, the CIs ranged from 0.2 to 0.45, signaling higher variability and lower certainty in the performance estimates. This level of uncertainty implies that, although the average performance may appear favorable, it may not be reliably replicated across different data splits, and thus the findings should be interpreted with caution.

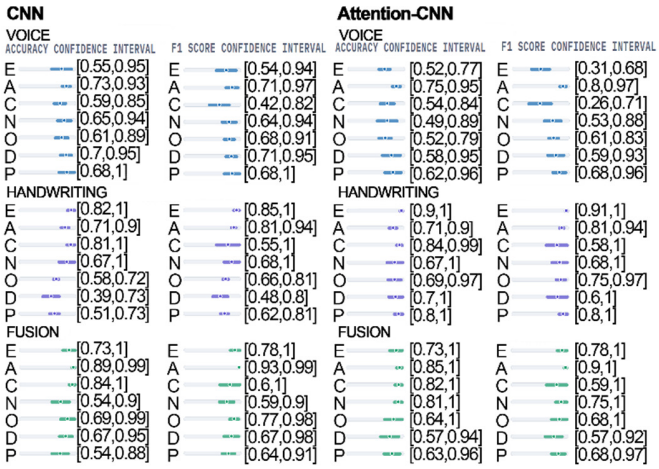

Fig. 11. 95% CIs for CNN and Attention-CNN.

To further investigate the comparative performance of different models and input modalities under limited samples, Cohen's d was used as a measure of effect size to quantify the magnitude of performance differences between different experimental conditions. Specifically, pairwise comparisons were performed between different classification algorithms of the same data type (e.g., CNN vs. Attention-CNN on handwriting features) and different data types using the same algorithm (e.g., handwriting vs. audio inputs under CNN). Cohen's d values provide a scale-independent measure of difference, with thresholds typically interpreted as small ($d \approx 0.2$), medium ($d \approx 0.5$), and large ($d \geq 0.8$) effects [32]. Large effect sizes ($d > 0.8$) indicate significant and practically meaningful differences in performance. For example, when comparing audio and handwriting feature inputs for the trait Extraversion under CNN, Cohen's d exceeded 0.8, indicating that handwriting feature representations significantly outperformed audio features when using CNN. As shown in Figure 12, when using the CNN classifier, handwriting is better than audio in Extraversion; audio is better than handwriting in DTI and PES; fused data are better than handwriting data in Agreeableness, Openness, and DTI; and fused data are better than audio data in Agreeableness and Conscientiousness. When using the Attention-CNN classifier, handwriting data are better than audio data in Extraversion, Conscientiousness, and Openness; fused data are better than handwriting data in Agreeableness; and fused data are better than audio data in Extraversion, Conscientiousness, and Neuroticism.

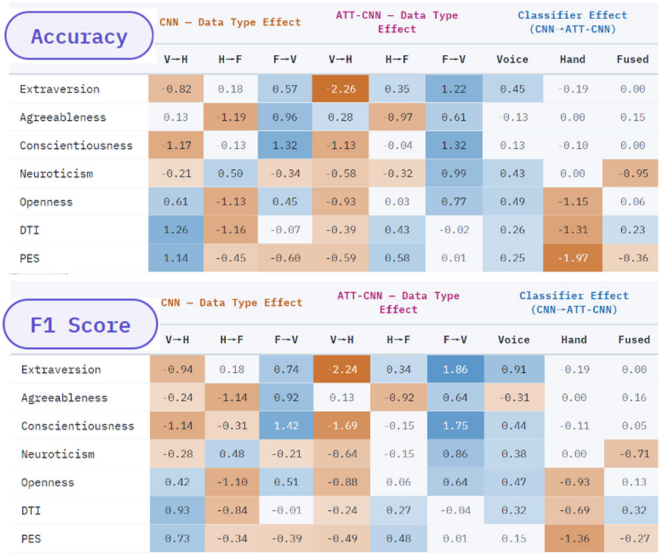

Fig. 12. Cohen's d effect size heatmap. Effect size interpretation: $|d| < 0.2$ (negligible), 0.2–0.5 (small), 0.5–0.8 (medium), and > 0.8 (large). Blue indicates a positive effect, and orange indicates a negative effect.

These findings complement the previous CI analyses, providing a deeper understanding of how performance differences are not only statistically significant but also practically meaningful, reinforcing the importance of model and modality selection in APR tasks based on educational data.

Moreover, a closer examination of the distribution of Cohen's d values across the comparison experiments revealed a notable pattern. When comparing every pair of data types under a fixed model, the number of cases in which Cohen's d exceeded the threshold for a large effect ($d > 0.8$) was substantially higher than the number of large effect sizes observed in comparisons between different models on the same data type. This finding suggests that the choice of input modality—whether handwriting features (HOG), audio features (MFCC), or their fusion—has a more pronounced impact on classification performance than the choice of classifier architecture alone (CNN and Attention-CNN). The implication is that enhancing the informational richness and complementarity of features can yield greater improvements in predictive accuracy than merely switching between learning algorithms. This reinforces the importance of multimodal integration in classification tasks and highlights the sensitivity of model performance to data representation quality. These results also lend empirical support to the hypothesis that

feature-level enhancements often have a greater influence than model-level tuning, particularly in settings constrained by small sample sizes.

D. Discussion of the Best Data and Method for Each Personality Trait

This section discusses which method, and which type of data produce the best results for each trait in the experiments. It also shows both the accuracy and loss curves of those best results.

1) Extraversion

Handwriting and fusion data achieved > 0.9 accuracy (Table II, Figure 4). Fusion performed lower than handwriting alone, with MLP achieving the highest result of 0.96 (Table II). Figures 5–7 and Figures 8–10 show high true positive and true negative rates across all data types and algorithms, indicating that nearly all Extraversion samples were classified correctly regardless of method.

Additionally, Figure 13 shows the following trends:

- Training accuracy: Rapid increase to ~0.96 plateau within initial phase, then stabilizes asymptotically.
- Validation accuracy: Significant growth within the first 100 epochs, followed by a slight decline and stabilization around 0.9.
- Interpretation: Strong training performance but weaker generalization, indicating potential overfitting.

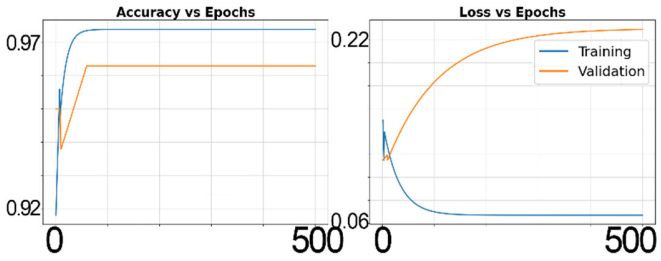

Fig. 13. Best Extraversion accuracy and loss of the MLP method on handwritten data.

In the loss curves, the training loss gradually decreases as training progresses, stabilizing around 0.1 after 100 epochs. However, the validation loss shows a different trend: it continues to rise with increasing epochs without displaying clear convergence. This result further supports the observation from the accuracy graph, indicating that the model performs well on the training set but struggles to generalize to the test set, suggesting overfitting.

Combining the observations from the accuracy and loss curves, it can be concluded that the model demonstrates strong learning capabilities on the training set, but its performance on the validation set does not improve synchronously, with the loss continuously rising, indicating potential overfitting. Therefore, we will not take the MLP method on the handwritten data as the best result for trait Extraversion. We examined other experiments of trait Extraversion and found that most of the accuracy and loss curves have similar trends. The accuracy and loss curves in the Attention-CNN experiment of audio data (Figure 4), which has the only converged loss curve, are shown in Figure 14. The loss of the validation and train set tends to approach 0.3. We take this result as the best result of trait Extraversion.

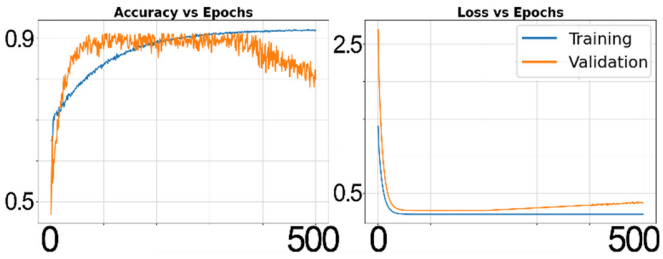

Fig. 14. Best Extraversion accuracy and loss of the Attention-CNN method on audio data.

2) Agreeableness

Only the results obtained using fused data and the CNN and Attention-CNN models achieved above 0.9, with the highest being 0.95 for CNN (Figure 4). The second matrix in Figures 5–7 and Figures 8–10 shows the performance of the Agreeableness trait predicted by all data types and algorithms. All these matrices show high true positive rates, and moderate true negative and false negative rates. Nearly all positive Agreeableness samples can be classified correctly, but some negative samples cannot be identified. Among these moderate TN and FN results, the results of fusion data (Figure 7 and Figure 10) are still satisfactory. Figure 15 shows that the model achieves strong performance on the training set, with accuracy reaching 0.99 and loss stabilizing at 0.1 by epoch 10. This indicates efficient learning on the training data. The validation accuracy stabilizes at 0.96 after epoch 100, and the validation loss also stabilizes near 0.1. The accuracy curve shows strong generalization ability, indicating that the model can adapt well to unseen data; the loss curve shows that the optimization effect is significant, and the model achieves low loss on both the training and validation sets, indicating that it can effectively capture the key features in the data and has strong data adaptability: the validation loss is close to the training loss, further supporting good performance on the validation set.

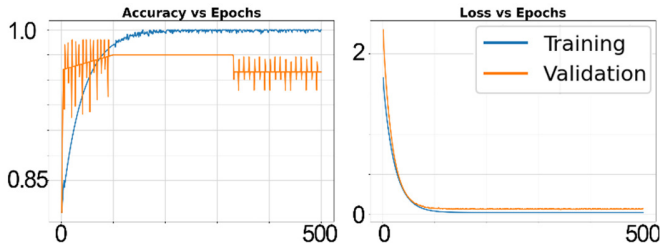

Fig. 15. Best Agreeableness accuracy and loss of the Attention-CNN method on fusion data.

3) Conscientiousness

All results obtained by the three classifiers using handwriting and fusion data achieve above 0.9 accuracy. The

performance of fusion data is higher than handwriting data, and the highest accuracy is 0.97 while using the CNN classifier on fusion data (Figure 4). Figure 16 shows a similar trend to Figure 15, but the performance of the validation set curve is slightly worse than that of Figure 15.

Considering the confusion matrix, the third matrix in Figures 5–7 and Figures 8–10 shows that most samples are TN and TP, whereas stronger performance is observed in identifying negative samples than positive samples.

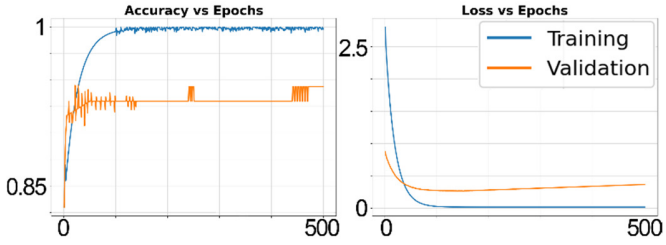

Fig. 16. Best Conscientiousness accuracy and loss of the CNN method on fusion data.

4) Neuroticism

Only the results obtained using handwriting data with the CNN and Attention-CNN models, and those obtained using fusion data with CNN, achieved above 0.8. The fusion effect is decreasing, and the highest result is 0.86 for handwriting data by CNN (Figure 4). The training accuracy rate rises slowly before 200 epochs and then stabilizes at 0.94, indicating that the model learns slowly on the training data but eventually reaches a high fitting level. The validation accuracy rate fluctuates between 0.84 and 0.86 throughout the training process and is always lower than the training set (Figure 17). This indicates that the validation set performance fails to improve synchronously with the training set, and the oscillation is obvious, which means limited generalization ability.

Considering the confusion matrix, the fourth matrix in Figures 5–7 and Figures 8–10 demonstrates balanced discriminative power between classes, with both true positives and true negatives substantially outperforming false predictions. However, a noticeable asymmetry exists in error types: FP and FN exhibit visible prevalence, suggesting class-specific prediction bias.

Despite variations in TP/TN dominance across experiments, the consistent gap between FP/FN and true predictions highlights persistent error patterns.

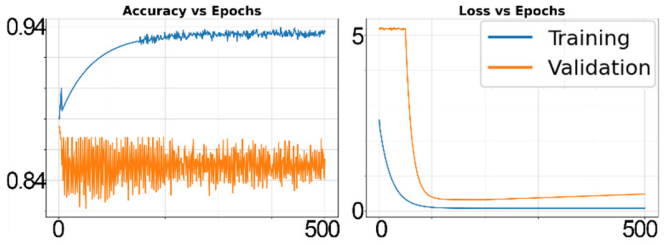

Fig. 17. Best Neuroticism accuracy and loss of the CNN method on handwriting data.

The training loss reaches a minimum value of 0.1 at around 60 epochs and then stabilizes, indicating that the model has achieved good optimization results in a relatively short time. The validation loss also reaches a minimum value of 0.1 at around 60 epochs but is slightly higher than the training loss and then remains stable. This shows that the optimization effect is significant and the generalization gap is small. We examined other experiments on the trait Neuroticism and found that all accuracy and loss curves have similar trends, which means it is hard to recognize the Neuroticism trait from our data and method.

5) Openness

Only the results obtained using audio data and the Attention-CNN model achieved above 0.9 (Figure 4). The accuracy and loss curves (Figure 18) are similar to Figure 16. The accuracy of the training set continues to rise whereas the accuracy of the validation set stagnates, and the validation loss continues to rise slowly after 50 epochs, both indicating that the model overfits the training data. The performance of the validation set curve does not improve synchronously with the training set after 50 epochs, and the validation loss increases slightly, indicating that the model's adaptability to unseen data is not ideal. We examined the curves of other data types/methods and did not find better results. Even though the results were poor, we still chose Attention-CNN and audio data as the optimal combination for the trait Openness.

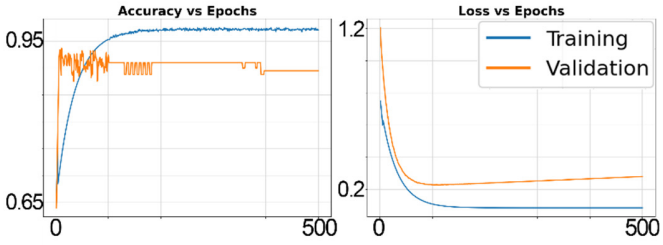

Fig. 18. Best Openness accuracy and loss of the Attention-CNN method on audio data.

Considering the confusion matrix, the fifth matrix in Figures 5–7 and Figures 8–10 shows high true positive rates. In Figures 5–7, low true negative rates and moderate FP/FN rates can be observed. In Figures 8–10, TN rate becomes high, and FP/FN rates are near to 0, which means when Attention-CNN is chosen as the classification method, it achieves near-optimal classification performance, with true positives and true negatives approaching theoretical maxima, whereas false predictions diminish to negligible levels.

6) Dichotomous Thinking Inventory

Only the results obtained using audio data with the Attention-CNN model achieved 0.9, with the highest value reaching 0.92 (Figure 4). A decreasing trend can be observed in fusion data. The accuracy and loss curves (Figure 19) are similar to Figure 18.

Considering the confusion matrix, the sixth matrix in Figures 5–7 and Figures 8–10, it can be observed that the TP rate is high whereas FP/FN is also high. Only in Figure 8

(audio data with attention-CNN) are TN and TP near to the maximum, and FN and FP are close to zero.

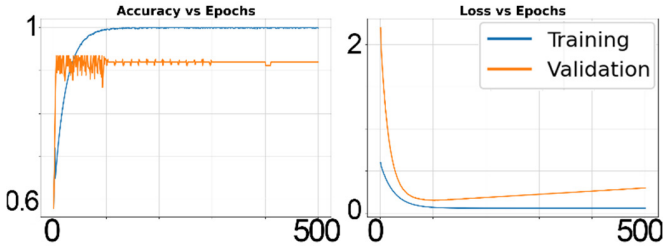

Fig. 19. Best DTI accuracy and loss of the Attention-CNN method on audio data.

7) Psychological Entitlement Scale

Both the result and the curve are almost the same as trait DTI, with the highest being 0.85 for audio data by Attention-CNN (Figure 4). In addition, the accuracy curve of the validation set oscillates significantly after 200 epochs, which indicates that overfitting also occurs on PES and is more serious than on DTI (Figure 20).

Considering the confusion matrix, the seventh matrix in Figures 5–7 and Figures 8–10, it can be observed that the TP rate is high whereas FP is also high. Only in Figure 9 (handwriting data with attention-CNN) are TN and TP near to the maximum, and FN and FP are close to zero.

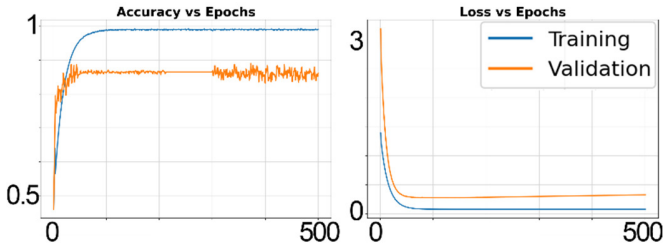

Fig. 20. Best PES accuracy and loss of the Attention-CNN method on audio data.

E. Theoretical Interpretation of Results

For all overfitting curves, we implemented multiple strategies to prevent overfitting and ensure generalizability. First, we reduced model complexity by employing a shallow neural network with L2 weight regularization and dropout (rate = 0.3), and we further adopted early stopping. Hyperparameters were tuned via 10-fold cross-validation to avoid optimistic performance estimates. While larger datasets are typically preferred, our systematic mitigations demonstrate viable paths for small-scale studies.

Although our analytical results demonstrate consistent modality-dependent variations, a theoretical interpretation is necessary to contextualize these findings. Accordingly, we now draw on established research in personality psychology and behavioral expression to explain why different modalities capture particular traits more effectively.

The effectiveness of handwriting in recognizing Conscientiousness may stem from the trait's association with planning, consistency, and motor control, which are reflected in stable handwriting patterns. In contrast, audio-based features appear more informative for traits related to expression and cognitive rigidity, such as Extraversion and DTI, as these traits may manifest more strongly in speech dynamics.

Unlike the Big Five, which have well-documented behavioral and expressive correlates that can manifest in handwriting regularity, vocal prosody, or learning activity patterns, both DTI and PES represent internal cognitive styles and attitudinal beliefs. DTI involves rigid evaluative processing and intolerance of ambiguity, which are unlikely to appear strongly in short speech segments or handwriting samples. Similarly, psychological entitlement tends to emerge in extended interpersonal interactions, relational dynamics, or decision-making contexts rather than in the brief, task-focused behaviors captured in this study. The modalities used here offer limited access to such deeper cognitive or social signals, which explains the weaker predictive outcomes. As such, the lower performance for DTI and PES reflects the inherent behavioral invisibility of these constructs within sensor-free educational data.

Finally, in Table V, we summarize all the best results of each personality trait, and in Figure 21, we provide an overview of the accuracy of all methods and all data types. As Figure 21 shows, the results on the Big Five are generally better than those on DTI and PES, and Attention-CNN is generally better than other methods. Handwritten data can produce better results for each method and trait, but not always the best results. We do not always choose the method and data type with the highest accuracy. The loss curve of the selected method and data is more reliable than that corresponding to the highest accuracy. For example, although the handwritten data under the MLP method has the best accuracy result, its validation loss curve does not converge in Figure 13, so we discard such a high-accuracy solution.

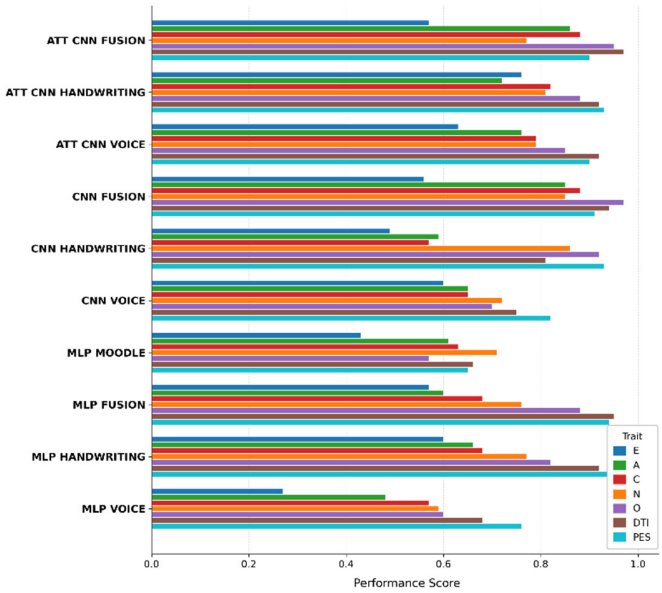

Fig. 21. Accuracy comparison of different methods and data types.

TABLE V. SUMMARY OF BEST AND SELECTED RESULTS FOR EACH PERSONALITY TRAIT

Traits	Best accuracy	Best data	Best method	Selected accuracy	Selected data	Selected method	Accuracy after fusion	Significance (CI range)
Extraversion	0.96	Handwritten	MLP	0.79	Audio	Attention	Decrease	0.25
Agreeableness	0.95	Fusion	Attention-CNN	0.95	Fusion	Attention	Increase	0.15
Conscientiousness	0.97	Fusion	CNN	0.97	Fusion	CNN	Increase	0.16
Neuroticism	0.86	Handwritten	CNN	None	None	None	Decrease	None
Openness	0.9	Audio	Attention-CNN	0.9	Audio	Attention	Decrease	0.27
DTI	0.85	Audio	Attention-CNN	0.85	Audio	Attention	Decrease	0.37
PES	0.77	Audio	Attention-CNN	0.77	Audio	Attention	Decrease	0.34

IV. CONCLUSION

This study is based on four different types of data that were collected from Sophia University students over three semesters for Automatic Personality Recognition (APR) without any sensors. It demonstrated the feasibility of using only handwriting, audio, and Moodle data for APR without relying on personality questionnaires or sensitive student personal information. Features extracted from learning management system (Moodle) logs, questionnaires, audio, and handwriting were used to train and evaluate the models. The experiments identified the best-performing data and methods for each personality trait.

This research highlights the potential of personality recognition in educational contexts and underscores the importance of integrating multimodal data to improve model performance. The findings have practical implications for designing adaptive learning systems and personalized interventions tailored to individual personality profiles. The study explored the feasibility of using the above data for APR with Psychological Entitlement Scale (PES) and Dichotomous Thinking Inventory (DTI) as recognition targets, in addition to the classic Big Five personality traits.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY

The dataset used in this study is a private dataset exclusively used by our research group. Because the data are closely linked to student privacy, data collection was permitted only after obtaining approval from the Ethics Committee of Sophia University. All participants provided written informed consent, and the participant information sheet clearly stated that the data would not be disclosed on the internet. The ethics committee strictly monitored and reviewed the data handling procedures.

AI USE AND DECLARATION OF GENERATIVE AI USE

Generative AI (copilot) was used solely for English translation of portions of the manuscript. The authors have reviewed and edited all AI-generated content and take full responsibility for the content of the published article.

REFERENCES

[1] S. Y. Chen and J.-H. Wang, "Individual differences and personalized learning: a review and appraisal," *Universal Access in the Information Society*, vol. 20, no. 4, pp. 833–849, Nov. 2021, https://doi.org/10.1007/s10209-020-00753-4.

[2] S. G. Essa, T. Celik, and N. E. Human-Hendricks, "Personalized Adaptive Learning Technologies Based on Machine Learning Techniques to Identify Learning Styles: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 48392–48409, 2023, https://doi.org/10.1109/ACCESS.2023.3276439.

[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, May 2020, Art. no. e1355, https://doi.org/10.1002/widm.1355.

[4] A. B. Altamimi, "Big Data in Education: Students at Risk as a Case Study," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11705–11714, Oct. 2023, https://doi.org/10.48084/etasr.6190.

[5] D. Dockterman, "Insights from 200+ years of personalized learning," *npj Science of Learning*, vol. 3, no. 1, Sept. 2018, Art. no. 15, https://doi.org/10.1038/s41539-018-0033-x.

[6] B. S. Bloom, "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher*, vol. 13, no. 6, pp. 4–16, June 1984, https://doi.org/10.3102/0013189X013006004.

[7] M. Komarraju, S. J. Karau, R. R. Schmeck, and A. Avdic, "The Big Five personality traits, learning styles, and academic achievement," *Personality and Individual Differences*, vol. 51, no. 4, pp. 472–477, Sept. 2011, https://doi.org/10.1016/j.paid.2011.04.019.

[8] M. Kitahashi and H. Handa, "Estimating Classroom Situations by Using CNN with Environmental Sound Spectrograms," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 22, no. 2, pp. 242–248, Mar. 2018, https://doi.org/10.20965/jaciii.2018.p0242.

[9] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Information Processing & Management*, vol. 58, no. 3, May 2021, Art. no. 102532, https://doi.org/10.1016/j.ipm.2021.102532.

[10] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies*, vol. 28, no. 1, pp. 905–971, Jan. 2023, https://doi.org/10.1007/s10639-022-11152-y.

[11] D. Cervone and L. A. Pervin, *Personality: Theory and Research*. Hoboken, NJ, USA: Wiley, 2022.

[12] R. R. McCrae and P. T. Costa Jr., "Personality trait structure as a human universal," *American Psychologist*, vol. 52, no. 5, pp. 509–516, 1997, https://doi.org/10.1037/0003-066X.52.5.509.

[13] O. P. John and S. Srivastava, "The Big Five Trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of personality: Theory and research*, 2nd ed., L. A. Pervin and O. P. John, Eds. New York, NY, USA: Guilford Press, 1999, pp. 102–138.

[14] A. E. Poropat, "A meta-analysis of the five-factor model of personality and academic performance," *Psychological Bulletin*, vol. 135, no. 2, pp. 322–338, 2009, https://doi.org/10.1037/a0014996.

[15] A. Oshio, "Development and validation of the Dichotomous Thinking Inventory," *Social Behavior and Personality: an international journal*, vol. 37, no. 6, pp. 729–741, July 2009, https://doi.org/10.2224/sbp.2009.37.6.729.

[16] W. K. Campbell, A. M. Bonacci, J. Shelton, J. J. Exline, and B. J. Bushman, "Psychological Entitlement: Interpersonal Consequences and Validation of a Self-Report Measure," *Journal of Personality Assessment*, vol. 83, no. 1, pp. 29–45, Aug. 2004, https://doi.org/10.1207/s15327752jpa8301_04.

[17] M. Scheuch, J. Scheibstock, H. Amon, G. Fuchs, and C. Heidinger, "Learning Evolution – A Longterm Case-Study with a Focus on Variation and Change," in *13th European Science Education Research Association Conference*, Bologna, Italy, 2019, pp. 119–131, https://doi.org/10.1007/978-3-030-74490-8_10.

[18] J. Anisi, M. Majdiyan, M. Joshanloo, and Z. Ghoharikamel, "Validity and reliability of NEO Five-Factor Inventory (NEO-FFI) on university students," *International Journal of Behavioral Sciences*, vol. 5, no. 4, pp. 351–355, Dec. 2011.

[19] A. Vinciarelli and G. Mohammadi, "A Survey of Personality Computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, July 2014, https://doi.org/10.1109/TAFFC.2014.2330816.

[20] P. T. Costa Jr. and R. R. McCrae, "The Revised NEO Personality Inventory (NEO-PI-R)," in *The SAGE handbook of personality theory and assessment, Vol 2: Personality measurement and testing*, G. J. Boyle, G. Matthews, and D. H. Saklofske, Eds. Thousand Oaks, CA, USA: Sage Publications, Inc, 2008, pp. 179–198, https://doi.org/10.4135/9781849200479.n9.

[21] K. Chaudhari and A. Thakkar, "Survey on handwriting-based personality trait identification," *Expert Systems with Applications*, vol. 124, pp. 282–308, June 2019, https://doi.org/10.1016/j.eswa.2019.01.028.

[22] N. H. Abbas, K. N. Yasen, K. H. A. Faraj, F. A. Razak, and F. L. Malallah, "Offline Handwritten Signature Recognition Using Histogram Orientation Gradient and Support Vector Machine," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 8, pp. 2075–2084, Apr. 2018.

[23] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowledge-Based Systems*, vol. 236, Jan. 2022, Art. no. 107715, https://doi.org/10.1016/j.knosys.2021.107715.

[24] R. Hasan, M. Jamil, G. Rabbani, and S. Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients," in *3rd International Conference on Electrical & Computer Engineering*, Dhaka, Bangladesh, 2004, pp. 565–568.

[25] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, Nov. 2010, https://doi.org/10.1007/s00530-010-0182-0.

[26] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, July 2009.

[27] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, Aug. 2018, https://doi.org/10.1109/TIP.2018.2834830.

[28] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, https://doi.org/10.1109/TNNLS.2021.3084827.

[29] X. Yang, "An Overview of the Attention Mechanisms in Computer Vision," *Journal of Physics: Conference Series*, vol. 1693, no. 1, Dec. 2020, Art. no. 012173, https://doi.org/10.1088/1742-6596/1693/1/012173.

[30] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: a machine learning workbench," in *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, Brisbane, Australia, 1994, pp. 357–361, https://doi.org/10.1109/ANZIIS.1994.396988.

[31] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 30, 2017, https://doi.org/10.48550/arXiv.1412.6980.

[32] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY, USA: Routledge, 2013, https://doi.org/10.4324/9780203771587.