2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS) | 979-8-3315-4010-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISAS64331.2024.10845404

Classification of Depression, Anxiety and Stress from Handwriting and Drawing with Stacking Models

1st Semra Bayrak
Dept. of Computer Engineering
Karadeniz Technical University
Graduate School of Natural and Applied Science
61080, Trabzon, Türkiye
semraabayrakk@gmail.com

2nd Sedat Golgiyaz
Dept. of Computer Engineering
Bingöl University
Bingöl, Türkiye
sedatg@bingol.edu.tr

3rd Murat Aykut
Dept. of Computer Engineering
Karadeniz Technical University
Trabzon, Türkiye
murat_aykut@ktu.edu.tr

***Abstract*—Recently, emotion analysis from online handwriting and drawings has become an important research topic. This study aims to determine the optimal model combination using dynamic and statistical features obtained from handwriting and drawings for detecting emotional states. The feature vectors are derived from physical (kinematic), statistical and signal processing (spectral, cepstral and frequency domain) analyses. To address the class imbalance problem, ADASYN (Adaptive Synthetic Sampling) and Tomek Links methods were applied together. In this way, the number of samples of the minority class has been increased, while over-sampled classes have been reduced. Dimensionality reduction was realized using PCA, and Gradient Boosting Classifier (GBC) feature selection. After determining the fewest and most effective features, the samples were classified using the two-level stacking ensemble method. eXtreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LightGBM), GBC, Random Forest (RF) methods are selected for base and meta models. Model and hyperparameter optimization was carried out by Optuna frameork. The performance of the models is evaluated through experiments in publicly available EMOTion recognition from HAndWriting and draWing (EMOTHAW) database. The aim was to determine the emotional states of depression, anxiety and stress with the data obtained from 7 different drawing/writing tasks of 129 people. The results obtained from accuracy, f1-score, precision and recall metrics show that the model provides remarkable performance.**

***Index Terms*—Stacking Model, depression detection, online handwriting analysis, emotion recognition.**

I. Introduction

Emotions are one of the most important components of human life and determine how individuals react to certain events and situations. Emotional states such as depression, anxiety, and stress are among the significant health issues in modern society. Recognizing and managing these emotional states play a critical role in the health and well-being of individuals. Advances in the fields of machine learning and deep learning have made it possible to detect emotional states from biometric data, such as handwriting and drawing, with high accuracy. Handwriting analysis has been a method used for many years to understand and predict a person's characteristic traits and emotional state. Professional handwriting analysts, known as graphologists, manually examine an individual's handwriting to classify the writer's personality. However, manual handwriting analysis is time-consuming, costly, and largely dependent on the skills of the graphologists.

In the study by Likforman-Sulem et al. [1], a publicly available database called EMOTHAW (EMOTion recognition from HAndWriting and draWing) was presented for the recognition of emotional states from handwriting and drawing samples. This database includes handwriting samples from 129 participants whose emotional states of anxiety, depression, and stress were assessed using the Depression-Anxiety-Stress Scale (DASS) questionnaire. Time- and ductus-based (movement of the pen) features were calculated from these samples. Feature selection and classification processes were carried out using the Random Forest approach. The results revealed that the recognition performance for anxiety and stress was better than that for depression.

In their study, Chitlangia and Malathi [2] used a dataset consisting of digital handwriting samples from 50 different individuals to automate this process. They extracted features using the Histogram of Oriented Gradient (HOG) technique and used them as input for a Support Vector Machine (SVM) classifier. Two different training/testing splits were performed. In the first, 90% of the data was used for training and 10% for testing, while in the second, Leave-One-Out Cross Validation was applied. In both cases, the classification success rate for the subject's five different personality traits (energetic, extroverted, introverted, disorganized, and optimistic) was reported as 80%. Ayzeren et al. [3] created a new database containing offline and online handwriting and signature biometrics from 134 participants for emotional state detection (happy, sad, stressed). In the study, dynamic features were extracted from raw data and analysed in the frequency domain. In experiments conducted with different classification methods (K-Nearest Neighbor (k-NN), JRip, and Random Forest), notable successes were achieved, especially in stress prediction using

979-8-3315-4010-4/24/$31.00 ©2024 IEEE

handwriting. In the study conducted by Cordasco et al. [4], the handwriting and drawing features of individuals experiencing negative moods (depression, stress, and anxiety) were compared with those of an age- and gender-matched control group. Mixed ANOVA analyses showed significant differences between groups, and these differences were dependent on the relevant exercises and feature categories. The results of the study revealed that time- and frequency-domain features are effective in identifying negative moods. In their study, Nolazco-Flores et al. [5], combined temporal, spectral, and cepstral features of signals captured on a tablet. From the raw data in the EMOTHAW dataset, spectral and cepstral domain features were extracted along with time spent in the air and on paper, task duration, and various other dynamic features. The Fast Correlation-Based Filtering (FCBF) method was used for feature selection, and data augmentation techniques were applied to add synthetic samples to the minority class to ensure data balance. In the study, the Support Vector Machine (SVM) method was used with a Leave-One-Out (LOO) cross-validation strategy. In the study by Rahman and Halim [6], eleven features obtained from handwriting samples were extracted using a graph-based handwriting representation approach. A Semi-supervised Generative Adversarial Network (SGAN) was used to enhance classification accuracy. Experimental results showed that the proposed method was able to recognize personality traits with an accuracy rate of 91.3% by using the handwriting features of 173 participants.

In the study by Nolazco-Flores et al. [7], raw data were transformed into features in the time, kinematic, statistical, spectral, and cepstral domains, and feature selection was performed using PCA and mFCBF methods. Gaussian noise was applied as a data augmentation technique. Classifiers trained and tested using Automated Machine Learning (AutoML) achieved 100% accuracy in detecting two possible mood severity levels. The accuracy rates obtained for detecting three possible mood states were 82.5% for depression, 72.8% for anxiety, and 74.56% for stress.

In the study conducted by Bhattacharya et al. [8], the agglomerative hierarchical clustering technique was utilized. This method groups image pixels through a clustering technique following preprocessing, ensuring that each cluster corresponds to a specific emotion. The model was tested for five emotions (Anger, Sadness, Depression, Happiness, and Excitement) and achieved an accuracy rate exceeding 75%. In the study by Greco et al. [9], a dynamic assessment of handwriting and drawing performance was conducted using handwriting and drawing features to compare healthy participants (n=28) with patients diagnosed with depression (n=27). The obtained data were statistically analysed. The results of the study indicated that all features, except for pressure on paper, successfully distinguished between depressive and non-depressive subjects. In the study by Rahman and Halim [10], signals obtained from handwriting and drawing samples were analysed using temporal, spectral, and MFCC methods. Feature vectors created using a Bidirectional Long Short-Term Memory (BiLSTM) network were classified, and the method was evaluated using multiple publicly available datasets. Experimental results demonstrated that combining features improved recognition accuracy. In the study conducted by Khan et al. [11], an attention-based transformer model was used for features obtained from handwriting and drawing samples. In the proposed method, an accuracy rate of 92.64% was achieved using the EMOTHAW dataset. This study proposes a machine learning model for recognizing emotional states of depression, anxiety, and stress from online handwriting and drawings using the EMOTHAW dataset. While other studies in the literature emphasize the analysis of signal processing, spectral, and cepstral features, this study forms a broader feature set by also considering kinematic and statistical features alongside signal processing. This distinction is a step toward improving the overall performance of the model by using a feature set that contains more comprehensive information. Dimensionality reduction with Principal Component Analysis (PCA) and feature selection with Gradient Boosting Classifier (GBC) techniques were used together. This approach reduces the dataset size to provide a more compact and meaningful representation while enabling the selection of features that contribute the most to model performance in the classification process. In this way, more efficient and effective feature selection has been achieved in high-dimensional and complex datasets. Compared to existing studies in the literature, where methods such as SVM or Random Forest (RF) are commonly used, this study instead preferred stacking ensemble learning approaches optimized with Optuna. The use of this method increased model diversity and improved overall prediction performance by combining different prediction models. In this study, Optuna's Bayesian optimization strategy [12] was preferred for hyperparameter optimization. This strategy offers the advantage of establishing a more flexible and effective hyperparameter range, ensuring cost efficiency in the optimization process. Additionally, unlike techniques commonly used in the literature to address data imbalance, ADASYN and Tomek Links methods were applied sequentially in this study to more effectively address data imbalance. ADASYN increases the number of samples in the minority class, while Tomek Links removes noisy data at the boundary of majority and minority classes, making the dataset more balanced. This sequential method combination aims to create a more balanced dataset and enhance the model's learning performance for the minority class compared to singular methods in other studies.

II. Experimental Design and Dataset

The EMOTHAW dataset used in this study was created by Likforman-Sulem et al. [1] for emotion recognition from handwriting and drawings. This dataset includes 129 participants, aged between 21 and 32 (mean age 24.8, SD = 2.4). The participants consist of 71 female and 58 male students studying at the Second University of Naples. After completing the DASS questionnaire, participants carried out seven designated handwriting and drawing tasks (Table 1 and Figure 1). The DASS is a 42-item self-report questionnaire designed to measure three main negative mood states—depression, anxiety,

and stress—through three separate scales, each containing 14 questions. Data were recorded while participants performed tasks using a digitizing tablet. The dataset includes seven handwriting and drawing tasks, among which are assessments such as clock drawing, the Mini-Mental State Examination [13], and the house-tree-person test, along with four other simple tasks [14].

TABLE I
TASKS PERFORMED FOR EACH PARTICIPANT

1. Copy of two pentagon drawings
2. Copy of a house drawing
3. Writing four Italian words in capital letters
4. Drawing loops with the left hand
5. Drawing loops with the right hand
6. Writing an Italian sentence in cursive
7. Clock drawing

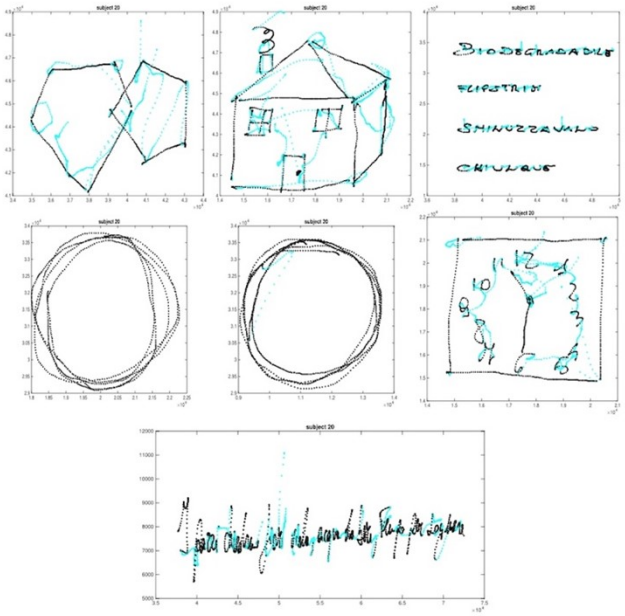

Fig. 1. Examples of Performed Writing and Drawing.

The writing and drawing samples collected from all tasks include pentagon and house drawings, handwriting, loops (left and right hand), clock drawing, and cursive handwriting. Pen-down and pen-up data points are represented in black and blue, respectively. Presented by Likforman-Sulem et al. [1].

A. Sensor Data

The sensor-based data collection process was carried out using the INTUOS WACOM series 4 digitizing tablet and the Intuos Inkpen device [1]. During participants' writing and drawing activities, various parameters such as x and y coordinates, timestamp, pen position (up/down), azimuth and altitude angles, and applied pressure were recorded. The recordings were stored as ASCII files in Wacom's svc format. The obtained data allow for the analysis of movement dynamics, including speed, acceleration, instantaneous trajectory, and displacement. The system is also capable of recording in-air movements of the pen, in addition to movements on paper; however, points are not recorded if the pen is more than 1 cm away from the surface. These recordings are used to conduct an in-depth analysis of writing dynamics and movement features. Figure 2 provides the feature names and sample data for the dataset.

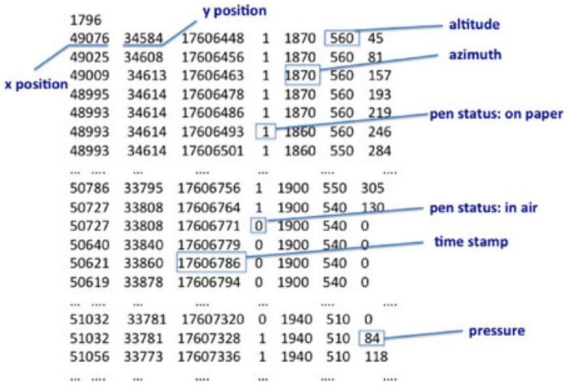

Fig. 2. Summary of an SVC file for the pentagon drawing task.

In Figure 2, the file contains a total of 1,796 points, and seven different parameters are recorded for each point: x and y positions, timestamp, pen status, azimuth angle, altitude, and applied pressure [1].

III. METHOD

A. Feature Extraction and Selection

From the raw data obtained from the tablet and pen, temporal, statistical, kinematic, spectral, and cepstral features were extracted to analyse different aspects of the signal. The dimensionality of the data was reduced using the PCA method, and potential outliers were eliminated. Subsequently, feature selection was performed using a Gradient Boosting Classifier (GBC) model

1) Feature Extraction: Temporal features relate to the movements of the pen in the air and on paper. In this analysis, temporal features were calculated as follows: the total time the pen remained in the air (total_air_time), the total time it moved on paper (total_paper_time), the total duration between the start and end timestamps of the task (total_duration), and the total number of transitions between paper and air (NSt).

A series of statistical methods were applied to gain a deeper understanding of the fundamental trends and distributional characteristics of the data. In this context, measures representing the central tendency of the data, such as arithmetic mean, and measures of distribution and spread, such as standard deviation and median, were calculated alongside the first and third quartiles reflecting the interquartile range of the data. Additionally, skewness was obtained to determine the degree

of symmetry, kurtosis to examine the sharpness of the peaks, and maximum values as statistical features to reflect the most extreme values in the dataset. This comprehensive analysis was conducted to better represent the overall structure of the dataset.

Features were selected to reflect the dynamics of the pen's movement. In this context, the displacement of the pen between successive points was examined, and the mean displacement, standard deviation of displacement, and maximum displacement values were obtained. For velocity, calculated based on displacement and time difference, the mean velocity, standard deviation of velocity, and maximum velocity values were determined. Additionally, acceleration, the derivative of velocity with respect to time, was calculated, and the mean and maximum acceleration values were analysed. For jerk, the derivative of acceleration with respect to time, the mean and standard deviation values were taken as features.

To analyse the frequency-domain features of the signals, frequency-related features were extracted. Raw columns representing X position, Y position, and pen pressure data (columns 0, 1, and 6) were processed. Fast Fourier Transform (FFT) was applied to each signal to transform it into the frequency domain. When calculating the amplitudes of the resulting frequency spectrum, the first element, which is the direct current (DC) component, was removed. Cepstral features were extracted to analyse the variations and periodicities in the underlying frequency structures of the signals more deeply. To extract cepstral features, the FFT of each signal was taken, the logarithm of the spectrum was calculated, and then the inverse Fourier transform was applied to this logarithmic spectrum to obtain the cepstrum. From the resulting cepstrum, statistical features such as mean, standard deviation, and maximum values were extracted.

2) Feature Selection: The EMOTHAW dataset contains a total of 49 features. Through feature extraction techniques, a total of 595 features were obtained. To enhance model performance and reduce computational cost by decreasing the dataset's dimensionality, the Principal Component Analysis (PCA) method was applied. PCA leverages correlations between features in high-dimensional datasets to obtain components that explain the most variance. While these components are fewer than the original dataset's features, they retain the most important structural characteristics of the data. In this study, PCA was configured to select components that explain 97% of the total variance in the dataset. This rate minimizes information loss while eliminating low-variance features that could be considered unnecessary or noise. Thus, the computational load of the model was reduced, while accuracy was maintained.

During the feature selection phase, the features that contributed most to classification performance were identified and selected using GBC. GBC is a powerful ensemble learning method that offers high accuracy rates and was used in this study to enhance model performance. The advantage provided by GBC in feature selection is its ability to identify important and effective features in the data, allowing the model to learn more quickly and efficiently.

The combined use of PCA and GBC was adopted as an effective method to optimize model performance in this study. This process increased the model's accuracy and computational efficiency while reducing the risk of overfitting. While PCA eliminated noise in the dataset, GBC identified the most critical features for classification, making a significant contribution to the model's generalization capacity.

The system structure used in this study is shown in Figure 3 . This diagram illustrates the extraction of various features obtained from sensor data and the process of selecting the best features with the Gradient Boosting model. Features obtained from temporal, kinematic, statistical, spectral, and cepstral domains were ranked using the Gradient Boosting method, and the most important features were selected. This method was used to optimize data dimensionality and maximize classification success to enhance model performance.

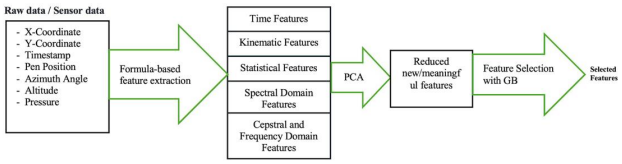

Fig. 3. The schematic structure of the preprocessing process used in the study.

B. Data Preprocessing and Standardization

The data were normalized using the Z-score standardization method to have a mean of 0 and a standard deviation of 1. This standardization process aims to improve model performance by enabling features with different scales to be evaluated on the same scale.

C. Data Augmentation and Balancing

The problem of imbalanced class distribution in the dataset arises due to the uneven distribution of examples in the depression class. In the distribution of DASS scores across the Depression, Anxiety, and Stress scales within the EMOTHAW database, it is observed that 26.4% of participants fall into the positive class for depression. For anxiety and stress, this rate is 42.6%. However, there is a notable data imbalance in the depression class (26.4% positive, 73.6% negative). This situation necessitates addressing the data imbalance issue, particularly in the depression class. To resolve this issue, data balancing was performed using the ADASYN (Adaptive Synthetic Sampling) and Tomek Links methods together. ADASYN analyses the density of samples surrounding the minority class examples and adaptively addresses class imbalance by generating more synthetic samples in areas with low density [15]. This method aims to strengthen the representation of the minority class at class boundaries, allowing the model to learn this class better. After the addition of synthetic samples,

the Tomek Links method was applied to reduce noise in the dataset and clarify class boundaries. Tomek Links identifies pairs of nearest neighbour examples belonging to different classes. These examples represent noisy data at the class boundaries that could potentially cause misclassification [16]. By removing these examples from the dataset, the dataset has become cleaner and more balanced.

D. Classification

In this study, powerful machine learning models, including LightGBM, XGBoost, Random Forest, and Gradient Boosting, were used for the classification of depression, anxiety, and stress. By using the stacking classification method as an ensemble learning approach, robust machine learning models were created. Hyperparameter optimization was conducted to ensure accurate model training and to achieve efficient results.

The Optuna framework was chosen for the optimization method. In this study, a Bayesian optimization method called Tree-structured Parzen Estimator (TPE), which is adopted by Optuna by default, was used. Unlike traditional grid search or random search methods, TPE selects new samples based on the performances of previously observed samples. This approach aims to reach the globally optimal parameter combination with fewer trials. In the hyperparameter search, essential parameters such as the number of estimators, learning rate, number of leaf nodes, and maximum depth were optimized, and the generalization ability of the models was tested with 5-fold cross-validation. Optuna's flexible and fast optimization structure contributed to improving model performance while also reducing computational costs. Finally, a StackingClassifier model, composed of combinations of the models demonstrating the best performance, was established. The block diagram of the proposed model is provided in Figure 4. In the stacking method, independent learners are combined by an ensemble learner [17]. The independent learners are referred to as base learners, while the ensemble learner responsible for the outputs of these learners is called the meta or final learner. The main idea of stacking is to create a new dataset to be applied to the meta learner by training the base learners on the initial datasets. The base learner forms the input features of the labelled new dataset. Although complex stacked ensembles can be obtained by using different learning algorithms, simpler homogeneous ensembles can also be created. However, to reduce the risk of overfitting, the meta learner should be trained on the new dataset created by the base learners; otherwise, using the same dataset for both the base and meta learners may increase this risk. The performance of the models was evaluated using metrics such as F1 Score, Precision, Recall, and Accuracy.

IV. Experiments Conducted

In this study, two-level stacking model combinations were created for the classification tasks of depression, anxiety, and stress. The performance of each model was evaluated using accuracy, precision, F1 score, and recall performance metrics. These metrics revealed not only the overall accuracy

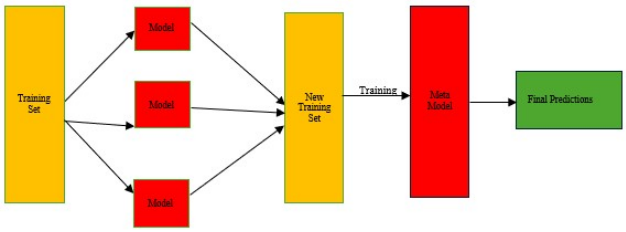

Fig. 4. The basic architecture of the stacking process.

of the models but also how well they predicted the positive classes (precision), the balance of false positives and false negatives (F1 score), and the sensitivity of the model (recall). In the experiments, Gradient Boosting, XGBoost (Extreme Gradient Boosting), and LightGBM (Light Gradient Boosting Machine) were used as base (level 1) models, while XGBoost (Extreme Gradient Boosting) and Random Forest were used as the final (level 2) model. The evaluation process was conducted using 5-fold cross-validation. Table 4 provides the classification performance resulting from the combined use of data from writing and drawing tasks across three different stacking approaches.

TABLE II
Performance Comparison of Stacking Models in Depression, Anxiety, and Stress Categories Using Data Obtained from Writing and Drawing Tasks.

Category	Model	F1	Recall	Precision	Acc
Depression	LGBM + XGB → XGB	60.19	54.29	72.89	81.15
	LGBM + GB → XGB	63.44	62.86	65.64	80.0
	LGBM + XGB → RF	46.96	45.71	51.83	74.23
Anxiety	LGBM + XGB → XGB	75.26	82.73	69.56	76.92
	LGBM + GB → XGB	76.37	82.73	72.53	78.08
	LGBM + XGB → RF	69.54	74.55	66.36	73.46
Stress	LGBM + XGB → XGB	76.31	81.82	72.01	78.46
	LGBM + GB → XGB	72.65	74.55	72.22	76.54
	LGBM + XGB → RF	73.76	76.36	71.94	76.92

Table 2 presents a comprehensive comparison of stacking model combinations in the classification of depression, anxiety, and stress categories based on data obtained from writing and drawing tasks. In depression classification, the LGBM + GB → XGB model achieved the highest F1 score (63.44%) and recall (62.86%), indicating its effectiveness in identifying positive cases, though with slightly lower precision (65.64%). The LGBM + XGB → XGB model, on the other hand, reached the highest precision (72.89%) and accuracy (81.15%) for depression, suggesting a tendency to reduce false positives but with a moderate trade-off in recall. In anxiety classification, the LGBM + GB → XGB model showed the best performance, with the highest F1 score (76.37%), recall (82.73%), and accuracy (78.08%), demonstrating balanced and robust capability in capturing positive cases while minimizing false negatives. This performance is closely followed by the LGBM + XGB → XGB model, which achieved a comparable recall (81.82%) in the stress classification, suggesting an efficient approach for sensitive detection of positive cases in stress-related data.

However, the LGBM + XGB → RF model demonstrated the most balanced performance in stress classification, with an F1 score of 73.76% and high recall (76.36%), along with solid precision (71.94%), making it particularly effective in distinguishing between positive and negative cases for stress. Overall, the results indicate that each model configuration presents unique strengths across different emotional categories, emphasizing the adaptability and effectiveness of stacking models in emotional state detection.

TABLE III
PERFORMANCE COMPARISON OF STACKING MODELS IN DEPRESSION, ANXIETY, AND STRESS CATEGORIES USING DATA OBTAINED FROM DRAWING TASKS.

Category	Model	F1	Recall	Precision	Acc
Depression	LGBM + XGB → XGB	61.88	65.71	69.85	78.08
	LGBM + GB → XGB	62.78	67.14	61.64	78.85
	LGBM + XGB → RF	52.87	52.86	61.73	77.31
Anxiety	LGBM + XGB → XGB	76.64	81.82	73.43	78.85
	LGBM + GB → XGB	76.09	80.91	73.09	78.46
	LGBM + XGB → RF	76.65	81.82	73.60	79.23
Stress	LGBM + XGB → XGB	76.38	85.45	69.71	77.31
	LGBM + GB → XGB	77.52	80.91	74.86	80.0
	LGBM + XGB → RF	77.82	85.45	72.10	79.23

Table 3 provides a detailed comparative analysis of stacking model combinations for the classification of depression, anxiety, and stress categories using data obtained exclusively from drawing tasks. For depression classification, the LGBM + GB → XGB model achieved the highest F1 score (62.78%) and recall (67.14%), indicating its strength in identifying positive cases, albeit with a moderate precision (61.64%) and accuracy (78.85%). The LGBM + XGB → XGB model demonstrated slightly lower performance in terms of F1 score (61.88%) and recall (65.71%), but its higher precision (69.85%) suggests a balance in minimizing false positives. In anxiety classification, the LGBM + XGB → XGB model performed notably well, achieving an F1 score of 76.64% and recall of 81.82%, which indicates a strong sensitivity to positive cases. However, the LGBM + GB → XGB model closely followed with an F1 score of 76.09% and a recall of 80.91%, combined with accuracy (78.46%).This reflects the ability of both models to maintain a balance between capturing positive cases and minimizing false positives in anxiety classification, with minor differences in performance metrics. For stress classification, the LGBM + XGB → RF model excelled, achieving the highest F1 score (77.82%), recall (85.45%), and balanced precision (72.10%) with an accuracy of 79.23%. This model's strong recall underscores its capability to accurately identify stress-related positive cases, while its balanced precision indicates a reduction in false positives. The LGBM + GB → XGB model, with a comparable F1 score (77.52%) and recall (80.91%), displayed a similarly balanced performance but with slightly higher accuracy (80.0%). Overall, these results suggest that different stacking model configurations offer varying levels of performance across depression, anxiety, and stress classifications. The findings underscore the adaptability and utility of stacking models in achieving robust classification accuracy in emotional state detection based on drawing data. Each configuration presents distinct strengths, with the LGBM + XGB → RF model particularly excelling in stress detection, while the LGBM + XGB → XGB and LGBM + GB → XGB models demonstrated balanced performance in anxiety and depression classification.

TABLE IV
PERFORMANCE COMPARISON OF STACKING MODELS IN DEPRESSION, ANXIETY, AND STRESS CATEGORIES USING DATA OBTAINED FROM WRITING TASKS.

Category	Model	F1	Recall	Precision	Acc
Depression	LGBM + XGB → XGB	72.70	75.71	72.44	85.0
	LGBM + GB → XGB	66.75	67.14	67.42	81.92
	LGBM + XGB → RF	64.20	64.29	66.10	81.15
Anxiety	LGBM + XGB → XGB	76.31	81.82	72.11	78.46
	LGBM + GB → XGB	78.07	83.64	73.80	80.0
	LGBM + XGB → RF	73.17	76.36	71.01	76.15
Stress	LGBM + XGB → XGB	77.54	76.36	79.83	81.15
	LGBM + GB → XGB	77.35	80.0	75.33	80.0
	LGBM + XGB → RF	73.45	72.73	75.45	78.08

Table 4 provides a comprehensive analysis of stacking model combinations in the classification of depression, anxiety, and stress categories using data derived from writing tasks. In depression classification, the LGBM + XGB → XGB model achieved the highest F1 score (72.70%) and recall (75.71%), coupled with a balanced precision (72.44%) and accuracy (85.0%), highlighting its strength in detecting positive cases accurately while maintaining low false positives. The LGBM + GB → XGB model, with an F1 score of 66.75% and accuracy of 81.92%, showed moderate recall (67.14%) and lower precision (67.42%), indicating limitations in capturing positive cases for depression. In comparison, the LGBM + XGB → RF model yielded the lowest F1 score (64.20%) and recall (64.29%) in this category, reflecting relatively weaker performance in positive case detection. For anxiety classification, the LGBM + GB → XGB model demonstrated the best overall performance, achieving the highest F1 score (78.07%), recall (83.64%), and precision (73.80%) with an accuracy of 80.0%. This model's high recall indicates its capability in capturing positive instances effectively, which is crucial for anxiety detection. The LGBM + XGB → XGB model followed closely with an F1 score of 76.31% and a recall of 81.82%, though with slightly lower precision (72.11%) and accuracy (78.46%), suggesting a balanced performance but with a minor trade-off in precision. In stress classification, the LGBM + XGB → XGB model once again showed strong performance, achieving the highest precision (79.83%) and an F1 score of 77.54%, with accuracy of 81.15%. This model's precision indicates its effectiveness in minimizing false positives, which is advantageous in stress detection tasks. The LGBM + GB → XGB model demonstrated similar strength, with an F1 score of 77.35%, recall of 80.0%, and precision of 75.33%, emphasizing its balanced and reliable performance. The LGBM + XGB → RF model, with an F1 score of 73.45% and recall of 72.73%, showed somewhat lower accuracy (78.08%), indicating moderate performance

in stress classification. Overall, these findings underscore the effectiveness of stacking models in emotional state detection using writing data, with each model combination displaying unique strengths across different emotional categories. The LGBM + XGB → XGB and LGBM + GB → XGB models are particularly noteworthy for their robust performance in stress and anxiety classifications, respectively.

TABLE V
COMPARISON OF MODEL PERFORMANCES FOR DIFFERENT EMOTIONAL STATES

Emotion	Model	Task	Acc(%)
Depression	LGBM+XGB → XGB	Both	81.15
	LGBM+GB → XGB	Draw	78.85
	LGBM+XGB → XGB	Write	85.0
Anxiety	LGBM+GB → XGB	Both	78.08
	LGBM+XGB → RF	Draw	79.23
	LGBM+GB → XGB	Write	80.0
Stress	LGBM+XGB → XGB	Both	78.46
	LGBM+GB → XGB	Draw	80.0
	LGBM+XGB → XGB	Write	81.15

Table 5 compares the accuracy (%) rates of models used to classify three main emotional states depression, anxiety, and stress across different tasks (writing, drawing, and both combined). For depression, the LGB + XGB → XGB model achieved the highest accuracy (85.0%) with writing tasks, highlighting the effectiveness of writing data for depression detection. Anxiety classification showed balanced performance with the LGB + XGB → RF model, achieving 80.0% accuracy in both writing-only and combined tasks. In stress classification, the LGB + XGB → XGB model reached an accuracy of 81.15% with writing tasks, indicating writing data's robustness in stress detection, though drawing tasks also provided solid accuracy (80.0%). Overall, these results suggest that writing tasks are particularly valuable for detecting depression and stress, while combining writing and drawing provides stable performance across all emotional states, supporting the application of stacking models in mental health diagnostics.

V. CONCLUSIONS AND DISCUSSION

When examining the entirety of the experiments conducted, we observe that the modeling results on data obtained from both writing and drawing tasks demonstrate effective performance in classifying different emotional states (depression, anxiety, stress). In the analyses conducted for the categories of depression, anxiety, and stress, each model combination has its advantages and disadvantages. While different model combinations provided high accuracy and precision for depression, it was observed that different models performed better in the classification of anxiety and stress. This indicates that each emotional state can be distinguished based on different writing and drawing features. Overall, all model combinations demonstrated good performance, but the model combination showing the best performance varied depending on the specific emotional state. This study has established a solid foundation for emotional state identification using behavioral features such as writing and drawing, and it has particularly shown that stacking model combinations adapt well to different states. The results indicate that stacking model combinations are successful in emotional state classification. Specifically, ensemble models have demonstrated good performance in detecting complex emotional states such as anxiety and depression. Table 6 provides a comparative analysis of the accuracy (%) of the proposed method against existing literature on the EMOTHAW dataset for classifying depression, anxiety, and stress based on writing, drawing, or combined tasks.

TABLE VI
COMPARISON OF RESULTS USING THE EMOTHAW DATASET

Emotion	Task	Accuracy (%)				
		Likforman-Sulem et al. [1]	Nolazco-Flores et al. [7]	Rahman & Halim [6]	Khan-Xia et al. [11]	Proposed Method
Depression	Drawing	72.80	75.59	83.28	86.15	78.85
	Writing	67.80	80.31	89.21	91.39	85.0
	Both	71.20	74.01	87.11	92.64	81.15
Anxiety	Drawing	60.50	67.71	76.12	79.51	79.23
	Writing	56.30	68.50	74.54	77.38	80.0
	Both	60.00	72.44	80.03	83.22	78.08
Stress	Drawing	60.10	67.71	75.39	78.76	80.0
	Writing	51.20	67.71	75.17	79.41	81.15
	Both	60.20	70.07	74.38	78.05	78.46

For depression, the proposed method shows competitive accuracy rates: 78.85% for drawing, 85.0% for writing, and 81.15% for both tasks combined. While the accuracy for the proposed method falls slightly below Khan-Xia et al.'s results (86.15% for drawing, 91.39% for writing, and 92.64% for both tasks), it remains higher than the earliest study by Likforman-Sulem et al. and close to the results of Nolazco-Flores et al. The proposed method demonstrates a reliable performance, particularly in writing tasks, even if it does not surpass the highest-performing models. For anxiety, the proposed method achieves 79.23% accuracy for drawing, 80.0% for writing, and 78.08% for both tasks. These values are close to the accuracy achieved by Khan-Xia et al. (79.51% for drawing, 77.38% for writing, and 83.22% for combined tasks) and significantly higher than Likforman-Sulem et al. This indicates that the proposed method performs reliably for anxiety detection and is comparable to some of the recent literature in this category. In stress classification, the proposed method's accuracy is 80.0% for drawing, 81.15% for writing, and 78.46% for both tasks, which is competitive compared to Rahman & Halim and Nolazco-Flores et al. Although it doesn't outperform Khan-Xia et al. for each task, the proposed method maintains consistent and balanced performance across both individual and combined tasks. In summary, while the proposed method does not always achieve the highest accuracy values in each category, it consistently provides competitive results across all tasks and emotional states.

REFERENCES

[1] L. Likforman-Sulem, A. Esposito, M. Faundez-Zanuy, S. Clémençon, and G. Cordasco, "EMOTHAW: A novel database for emotional state recognition from handwriting and drawing," IEEE Transactions on Human-Machine Systems, vol. 47, no. 2, pp. 273–284, 2017.

[2] A. Chitlangia and G. Malathi, "Handwriting analysis based on histogram of oriented gradient for predicting personality traits using SVM," Procedia Computer Science, vol. 165, pp. 384–390, 2019.
[3] Y. B. Ayzeren, M. Erbilek, and E. Çelebi, "Emotional state prediction from online handwriting and signature biometrics," IEEE Access, vol. 7, pp. 164759–164774, 2019.
[4] G. Cordasco, F. Scibelli, M. Faundez-Zanuy, L. Likforman-Sulem, and A. Esposito, "Handwriting and drawing features for detecting negative moods," Quantifying and Processing Biomedical and Behavioral Signals, vol. 27, pp. 73–86, 2019.
[5] J. A. Nolazco-Flores, M. Faundez-Zanuy, O. A. Velázquez-Flores, G. Cordasco, and A. Esposito, "Emotional state recognition performance improvement on a handwriting and drawing task," IEEE Access, vol. 9, pp. 28496–28504, 2021.
[6] A. U. Rahman and Z. Halim, "Predicting the big five personality traits from hand-written text features through semi-supervised learning," Multimedia Tools and Applications, vol. 81, no. 23, pp. 33671–33687, 2022.
[7] J. A. Nolazco-Flores, M. Faundez-Zanuy, O. A. Velázquez-Flores, C. Del-Valle-Soto, G. Cordasco, and A. Esposito, "Mood state detection in handwritten tasks using PCA–mFCBF and automated machine learning," Sensors, vol. 22, no. 4, p. 1686, 2022.
[8] S. Bhattacharya, A. Islam, and S. Shahnawaz, "TEmoDec: emotion detection from handwritten text using agglomerative clustering," in 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), pp. 1–6, IEEE, March 2022.
[9] C. Greco, G. Raimo, T. Amorese, M. Cuciniello, G. Mcconvey, G. Cordasco, and A. Esposito, "Discriminative Power of Handwriting and Drawing Features in Depression," 2023.
[10] A. U. Rahman and Z. Halim, "Identifying dominant emotional state using handwriting and drawing samples by fusing features," Applied Intelligence, vol. 53, no. 3, pp. 2798–2814, 2023.
[11] Z. A. Khan, Y. Xia, K. Aurangzeb, F. Khaliq, M. Alam, J. A. Khan, and M. S. Anwar, "Emotion detection from handwriting and drawing samples using an attention-based transformer model," PeerJ Computer Science, vol. 10, p. e1887, 2024.
[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.
[13] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "Mini-mental state: a practical method for grading the cognitive state of patients for the clinician," Journal of Psychiatric Research, vol. 12, no. 3, pp. 189–198, 1975.
[14] P. Kline, Handbook of Psychological Testing. Routledge, 2013.
[15] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328, IEEE, June 2008.
[16] D. Devi and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance," Pattern Recognition Letters, vol. 93, pp. 3–12, 2017.
[17] P. Smyth and D. Wolpert, "Stacked density estimation," Advances in Neural Information Processing Systems, vol. 10, 1997.