Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:35
https://doi.org/10.28919/cmbn/7916
ISSN: 2052-2541

ANALYZING IMPORTANT STATISTICAL FEATURES FROM FACIAL BEHAVIOR IN HUMAN DEPRESSION USING XGBOOST

BRILYAN NATHANAEL RUMAHORBO[1], KENJOVAN NANGGALA[1,*], GREGORIUS NATANAEL ELWIREHARDJA[2,3], BENS PARDAMEAN[1,2]

[1]Computer Science Department BINUS Graduate Program – Master of Computer Science Program Bina Nusantara University, Jakarta 11480, Indonesia

[2]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[3]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract. Major Depressive Disorder (MDD) has been known as one of the most prevalent mental disorders whose symptoms can be observed from changes in facial behaviors. Previous studies had attempted to build Machine Learning (ML) models to assess depression severity using such features but few have utilized these models to determine key facial behaviors for MDD. In this study, we used video data to assess the severity of MDD and determine important features based on three approaches (XGBoost, Spearman's correlation, and t-test). In addition, there is the Facial Action Coding System (FACS) framework that allows visual data such as changes in facial behavior to be modeled as time series data. The results show that the XGBoost model obtained the best results when trained using features selected through the t-test statistical method with 5.387 MAE, 6.266 RMSE, and 0.042 R^2. The majority of the important features consist of Action Unit (AU) and Features 3D around the regions of the left eye, right cheek, and lip area. However, the majority of the important features discovered from the three approaches, are the first derivatives of the 3D facial landmark coordinates of the cheeks, eyes, and

*Corresponding author
E-mail address: kenjovan.nanggala@binus.ac.id
Received February 18, 2023

lips, especially along the z-axis. However, the variables used in this research are limited to the first derivatives, which meant that usages of wider variations of facial behavior data may further be studied so that Computer-Aided Diagnosis (CAD) systems for mental disorders may be realized in the future.

Keywords: machine learning; depression; XGBoost; major depressive disorder (MDD); facial behavior analysis.
2020 AMS Subject Classification: 68T05, 68T01, 26A24 .

1. INTRODUCTION

Major Depressive Disorder (MDD) is known as one of the most prevalent mental disorders affecting physiological and psychological aspects of humans, including sadness or prolonged sorrows among its symptoms. Untreated MDD cases may induce anxiety, feelings of isolation, and even suicidal thoughts, which may later lead to illegal drug usage or suicide [1]. MDD can be seen based on the symptoms experienced by the patient. One such symptom is the changes in facial behaviors. Research by Bodenschatz et al. concluded that depressed subjects show significantly more sad facial expressions compared to a healthy person [2].

Usually, the evaluation for MDD can be done by self-reporting of the patient's symptoms and filling in the eight-item Patient Health Questionnaire (PHQ-8). However, the PHQ-8 data is subjective and the psychologist's diagnosis results are influenced by their level of expertise [3]. The advent of Artificial Intelligence (AI) technology made it possible for AI models to provide more objective results if it is trained with the appropriate data. AI possesses huge potential in medical fields [4], especially for Computer-Aided Diagnosis (CAD) technology. This technology is intended to assist doctors in making more accurate and precise diagnoses and provide appropriate follow-up actions for patients [3].

In this study, we performed a depression severity assessment using Facial Action Coding System (FACS) which allows visual data such as changes in facial behavior to be modeled as time series data. Song et al. had done feature selection using Correlation-based Feature Selection (CFS) named "Voted-CFS" which details are presented in Related Works [5]. However, the discovered important statistical features from the raw FACS data, as well as their first and second derivatives, were not listed. It means that questions related to feature importance had yet to be answered. In other words, the explainability of their proposed model was not elaborated. Therefore, our study focused on the first derivatives or changes to each feature per frame.

We used XGBoost, a widely used model for Explainable Artificial Intelligence (XAI), to determine important statistical features from these changes. Then, the results were compared with variables that have high Spearman's correlation value to the PHQ-8 scores, as well as variables that have significant differences between depressed and normal subjects obtained from a pooled t-test. Features found to be important from Spearman's correlation and the t-test were also fit into XGBoost to test whether the selected first derivative features allow XGBoost to make better predictions in assessing depression severity. All in all, the main contribution of this study is determining statistical features from the first derivative variables of FACS variables that can be considered important for Machine Learning (ML) models in assessing MDD severity through statistical and XAI methods. To the best of our knowledge, no such research has been conducted. In other words, the findings of this research can serve as the foundation for developing more advanced and accurate ML or deep learning models for video-based MDD severity assessment, which results from this study require enhancements before the implementations of mental disorder CAD systems.

2. Related Works

Akbar et al. used FFNN (Feedforward Neural Network) to predict MDD severity by comparing three algorithms, which are LM (Levenberg Marquardt), BR (Bayesian Regularization) Backpropagation Algorithm, and Scaled Conjugate Gradient Backpropagation Algorithm to recognize MDD from extracted Facial Action Units (FAU) and find a reduced set of FAU features [6]. This study used the DAIC-WOZ dataset and concluded that Particle Swarm Optimization (PSO) was able to improve the performance of the Backpropagation Algorithm used with the best results obtained by Bayesian Regularization with 97.83% accuracy of 20 PSO iterations. Similarly, Mulay et al. made a depression classification system that focused on visuals, including images and videos by training a Convolutional Neural Network (CNN) model using a dataset from Kaggle, achieving 66.45% accuracy with a ratio of 80% training, 10% validation, and 10% test [7]. In a similar study, Jiang et al. also trained a CNN which obtained an average Area Under Curve (AUC) of 0.721 and an average test accuracy of 0.706 in 10 trials [8].

Song et al. conducted research on depression analysis that focuses on statistical features of human behavior, such as gaze directions, facial action units, and so on [5]. This study used

the Support Vector Machine (SVM) algorithm on the statistical features, which was compared to CNN. Compared to other previously proposed vision-based systems, this study obtained an 18.1% improvement in Root Mean Squared Error (RMSE) and 25.7% in Mean Absolute Error (MAE) which concluded that the CNN's performance dramatically improved in assessing depression severity. Rathi et al. proposed a similar approach that aims to assist clinicians in obtaining accurate and objective assessments in detecting MDD using three popular feature selection filters used, such as Fisher Discriminant Ratio (FDR), Mutual Information (MI), Pearson's correlation (PC) [9]. To carry out the classification, using a Decision Tree (DT), Linear Discriminant Analysis (LDA), k-Nearest Neighbor (KNN), and SVM. They then determined the level of MDD using regression techniques, such as Decision Tree (DT), Linear Regressor (LR), Partial Least Square (PLS), and Support Vector (SVR). The results of this research were the combination of FDR and LDA outperformed all classification models. Moreover, the combination of PC and LR surpassed the existing regression model in assessing MDD severity since both of them are based on the linear correlation between the three univariate filter feature selection procedures (FDR, MI, and PC) and the response variable. PC-based feature selection followed by LR produced the best performance in terms of MAE and RMSE.

Ray et al. conducted a study on depression with text, audio, and video analysis [10]. They used Long Short-Term Memory (LSTM) to identify low-level descriptors that focus on pose, gaze, and FAU. The proposed method increased the accuracy in the video aspect, but it was still inferior when compared to the approach using audio and text data. Similarly, Yoon et al. used Gaussian Mixture Model (GMM) clustering and fisher vector for visual data while the classification task was performed by a Support Vector Machine (SVM) and neural networks [11]. Their research used a dataset called D-Vlog which is a collection of vlog videos from YouTube that consists of 961 videos and provided higher depression detection performance than those trained with DAIC-WOZ. In addition, Zhang also conducted research on MDD using feature selection with the purpose of automatically detecting someone's depression based on text, audio, and video. The study used the XGBoost model in video-based depression detection to perform feature selection and detection and concluded that XGBoost feature selection can achieve the best performance for the video modality [12], similar to the research by Eteng that

focuses on audio and video using a random forest classifier, SVM, and also XGBoost. In the research, the best accuracy was obtained by the XGBoost algorithm with an accuracy of 0.82 for 2-bin classification and 0.639 for 3-bin classification [13].

Muzammel et al. reported an audio and video-based clinical examination of depression using LSTM and CNN architectures [14] to handle the early-level and model-level fusion of deep audio information with visual and textual features. The model-level fusion of deep audio and visual characteristics using the LSTM network gave the best performance with an accuracy of 77.16%, a precision of 53% for the sad class, and a precision of 83% for the non-depressed class.

All of this related works conclude that deep learning is indeed quite popular in this task and is capable of producing extraordinary accuracy. However, research with deep learning is still classified as a 'black box' and extra methods are needed to extract feature importance. In addition, the results obtained by researchers for video-based depression classification generally still have relatively low accuracy or high RMSE values which may be due to the presence of noisy variables. Therefore, the research conducted by Song et al. can be a baseline by proving that statistical features performed only with ML are able to provide performance that is not inferior to deep learning.

3. Research Methodology

3.1. XGBoost. XGBoost is currently the most popular algorithm in many applications among other Gradient Boost Methods [15]. XGBoost is an advanced Gradient Boosting Tree-based (GBDT) method that can efficiently deal with large-scale problems with very limited computing resources. Since this method was introduced, XGBoost won various machine learning competitions such as Kaggle and Knowledge Discovery and Data Mining (KDD) Cup [16, 17] and became a powerful and efficient solution for solving various problems classification [17]. XGBoost was developed with 10 times faster optimization than other gradient boosting methods.

XGBoost is a GBDT ensemble approach. The leaf nodes with the scores represent the outcomes in a regression tree, whereas the interior nodes carry the values for the test variables. The

prediction result is the number of scores predicted by the K tree, as shown in the equation:

$$\hat{y}_l = \sum_{k}^{K} f_k(x_i), f_k \in F \tag{1}$$

where $\hat{y}_1$ represents the model prediction value, $f_k(x_i)=\omega_q(x)$ is the space of Classification and Regression Trees (CART), $\omega_q(x)$ is the score of samples x, each tree's structure is represented by q, the number of trees is represented by K, and each f_k equates to an independent tree structure. q as well as leaf weight. XGBoost is a model that has many parameters, meaning that it requires more time to determine the value of each parameter. Unlike GBDT, XGBoost adds a regularization term to the goal function to avoid overfitting. The objective function is illustrated as follows:

$$O = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{k=1}^{t} R(f_k) + C \tag{2}$$

where $L(y_i, F(x_i))$ is the loss function, $R(f_k)$ indicates the regularization term at iteration time k and C is a constant term that can be eliminated providently. The regularization term $R(f_k)$ is demonstrated as:

$$R(f_k) = \propto H + \frac{1}{2}\eta \sum_{j=1}^{H} w_j^2 \tag{3}$$

where $\propto$ indicates the complexity of the leaves, H represents the number of leaves, η denotes the penalty parameter, and w_j^2 denotes the output result of each leaf node. In particular, a leaf denotes a predicted category following the classification rules and a leaf node denotes an indivisible tree node. In addition, rather than employing first-order derivatives like in GBDT, XGBoost utilizes a second-order Taylor set of objectives. If the mean squared error is utilized as the loss function, the objective function is as follows:

$$O = \sum_{i=1}^{n} \left[p_i w_{q(x_i)} + \frac{1}{2}\left(q_i w_{q(x_i)}^2 \right) \right] + \propto H + \frac{1}{2}\eta \sum_{j=1}^{H} w_j^2 \tag{4}$$

$q(x_i)$ denotes the function that assigns data points to the corresponding leaf, p_i and q_i denotes the first and second derivatives of the loss function respectively. The final loss value can be calculated by summing the loss values of the leaf nodes because the sample corresponds to the nodes in the decision tree.

As a result, the objective function is often written as follows:

$$O = \sum_{j=1}^{T} \left[P_j \mathrm{w}_j + \frac{1}{2} \left(Q_j + \eta \right) \mathrm{w}_j^2 \right] + \propto H \tag{5}$$

where $P_j = \sum p_i$, $Q_j = \sum i \in I \; q_i \; j_i \in I_j$, and I denotes all samples in a leaf node j. In other words, the objective function's optimization is changed in the case of selecting the minimum of the quadratic function. Besides, XGBoost also has a better ability to overcome overfitting problems.

3.2. Spearman's correlation coefficient. Spearman's correlation coefficient, also known as Spearman's rank correlation coefficient, is a coefficient implying the degree of association between two variables obtained by ordering or sorting of correlations. Spearman's correlation is a powerful way to measure the monotonic association between variables [18]. In this study, we compared every feature from the first derivative with the PHQ-8 score that we encoded into several categories which are referred to as "encoded PHQ-8 Score".

Spearman's correlation has a relationship with Pearson's correlation coefficient [18]. However, the usage of Spearman's correlation (ρ) is convenient and has many rank-bound ability values. Spearman's correlation can also determine linear or non-linear monotonic data relationships, while Pearson's correlation is only suitable for linear correlation evaluation [19].

Mathematically, Spearman's correlation measures the individual coefficients between two variable columns [18]. Spearman's correlation has a relay range of +1 or -1 which shows the correlation value of the monotonic relationship [20]. The formula for calculating the Spearman's correlation Coefficient can be expressed as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n \left(n^2 - 1 \right)} \tag{6}$$

where d is the difference between the set variables x and y. The variables x and y are two random variables whose number of elements is both n.

3.3. T-test (Pooled t-test). In this study, we used the voted version of the pooled t-test, which is referred to as the "voted t-test" for the rest of this paper. The voted t-test is a test method of parametric statistical tests. The statistical t-test is a test that measures how significant the mean is between 2 different groups. In this method, we test for significant differences between

normal subjects and depressed subjects. The t-test calculates p-values and compares them to the threshold value of 0.05 (α=5%) with certain criteria as follows :

- If the p-value is >= 0.05, then it means that normal subjects do not have a significant influence on depressed subjects.
- If the p-value is <0.05 then it means that the normal subjects variable has a significant influence on the depressed subjects [21, 22]. P-values are calculated from t-values, which are obtained by using the following formula:

$$\text{T-value} = \frac{\text{mean 1} - \text{mean 2}}{\frac{(n1-1)\times \text{var}1^2+(n2-1)\times var2^2}{n1+n2-2} \times \sqrt{\frac{1}{n1}+\frac{1}{n2}}} \tag{7}$$

where mean1 and mean2 are the sample sets' average values, var1 and var2 are the number of records in each sample set, and n1 and n2 are the numbers of records in each sample set. In this study, we used three sample sets (random state = 0,1,2) and because the data are imbalanced between the normal and depressed categories (more records on the normal), we undersampled the data, resulting in 30 samples each for the normal and depressed subjects, respectively. Hence, the method is referred to as the "voted t-test" as the test was conducted on multiple groups sampled from the original population.

3.4. Dataset. The DAIC-WOZ (Distress Analysis Interview Corpus) database is a large database of clinical interviews from the corpus by a virtual interviewer, named Ellie. This research used the DAIC-WOZ database because it has been widely used in previous studies. During the interview, the patient was identified for signs of psychological disorders verbally and non-verbally [23].

This database provides recordings, in the form of audio, video, and psychiatric responses in text form [24]. To evaluate the severity of the patient's depression, interviews were conducted using the PHQ-8. As a standard to differentiate them, depressed patients have a PHQ-8 score >= 10, and non-depressed patients have a PHQ-8 score <10 [25].

The DAIC-WOZ dataset consists of 189 participants, which were divided into 3 sets, namely training set (107 participants, 57%), validation set (35 participants, 19%), and test set (47 participants, 25%). Each interview lasted 7-33 minutes [24]. The dataset contains zipped files which are coded for each patient's number and also 3 CSV files (train, test, and dev) which contain the

patient's number. Each patient file encloses video, audio, and text data. For this research, we focus on the video data which consists of:

- Facial Action Units (FAU)

  The FAU is an important component in analyzing a person's facial expressions [26]. Each Action Unit (AU) has dots that mark parts of the face. The following are the 30 AU points which are divided into two groups, the upper face and lower face AU listed in Table 1 and Table 2, respectively [27]:

TABLE 1. Upper Face Action Unit

AU 01	Inner Brow Raiser	AU 41	Lid Droop
AU 02	Outer Brow Raiser	AU 42	Slit
AU 04	Brow Lowerer	AU 43	Eyes Closed
AU 05	Upper Lid Raiser	AU 44	Squint
AU 06	Cheek Raiser	AU 45	Blink
AU 07	Lid Tightener	AU 46	Wink

TABLE 2. Lower Face Action Unit

AU 09	Nose Wrinkler	AU 18	Lip Puckerer
AU 10	Upper Lip Raiser	AU 20	Lip Stretcher
AU 11	Nasolabial Deepener	AU 22	Lip Funneler
AU 12	Lip Corner Puller	AU 23	Lip Tightener
AU 13	Cheek Puffer	AU 24	Lip Pressor
AU 14	Dimpler	AU 25	Lip Part
AU 15	Lip Corner Depressor	AU 26	Jaw Drop
AU 16	Lower Lip Depressor	AU 27	Mouth Stretch
AU 17	Chin Raiser	AU 28	Lip Suck

- 3D features

  This file consists of 68 3D facial landmarks in millimeters in the world coordinate space, with the axes aligned to the camera and the camera being at (0,0,0) in the (X, Y, Z) axes.

- Gazes

  This file consists of x_0, y_0, z_0, x_1, y_1, z_1, x_h0, y_h0, z_h0, x_h1, y_h1, and z_h1. The output are 4 vectors that were divided into two groups, the first group consists of two vectors that describe the gaze direction of both eyes and the second group also consists of two vectors that describe the gaze in head coordinate space (this means that the direction of the vectors will be indicated based on the gaze, not the head position).

- Pose

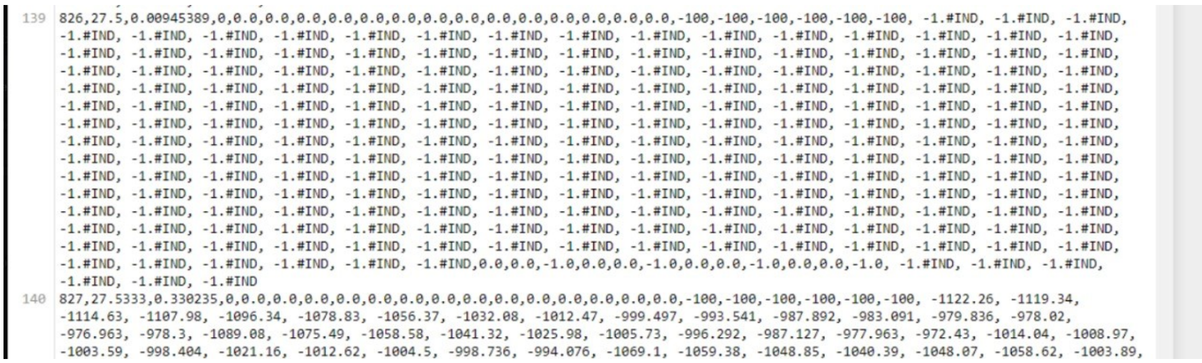
  This file consists of 6 items, which are Tx, Ty, and Tz which represent the position coordinates in millimeters, and Rx, Ry, and Rz which represent the head rotation coordinate in radians and in the Euler angle convention.

3.5. Data Preprocessing. As the output of OpenFace, which is the software utilized in extracting the FACS features, was stored in separate txt files in the dataset, these files were first combined into a data frame for each patient. Next, we processed the timestamp cropping `(start_time, stop_time)` based on the transcript as a cut reference. Then, we exported with the same name so that it overwrites the merged CSV file and we then performed the first derivative step for each patient file, which means we look for the difference for every 2 rows of the patient data (difference from row 0 and 1, 1 and 2, 2 and 3, and so on).

$$f'(x) = \lim_{x \to 0} \frac{f(x+h) - f(x)}{h} \tag{8}$$

On the first derivative step, an anomaly was found in the combined file for subject number 432 where rows 0 to 139 in the file contained data that could not be processed due to a technical error during the recording of the interview so that rows 0 to 139 in the file were removed.

Next, we performed an aggregation based on the first derivative result file for each patient. At this step, we calculated the mean, standard deviation, min, and max values for each feature column for each patient. This aggregation stage produced 971 columns which we then export into CSV form.




```
826,27.5,0.00945389,0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,-100,-100,-100,-100,-100,-100, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,0.0,0.0,-1.0,0.0,0.0,-1.0,0.0,0.0,-1.0,0.0,0.0,-1.0, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND
827,27.5333,0.330235,0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,-100,-100,-100,-100,-100,-100, -1122.26, -1119.34,
-1114.63, -1107.98, -1096.34, -1078.83, -1056.37, -1032.08, -1012.47, -999.497, -993.541, -987.892, -983.091, -979.836, -978.02,
-976.963, -978.3, -1089.08, -1075.49, -1058.58, -1041.32, -1025.98, -1005.73, -996.292, -987.127, -977.963, -972.43, -1014.04, -1008.97,
-1003.59, -998.404, -1021.16, -1012.62, -1004.5, -998.736, -994.076, -1069.1, -1059.38, -1048.85, -1040.39, -1048.07, -1058.62, -1003.89,
```


FIGURE 1. An example of anomaly found in the dataset

3.6. Model Training. The aggregated data were fit into the XGBoost model to predict the PHQ-8 Score using the trained model and to get the feature importance from the model's function. In this stage, we carried out several XGBoost prediction methods, which were divided into three approaches:

(1) XGBoost predictions without feature selection

We used aggregated data in the train and dev folders where the train folders were used as fitting or training data for the XGBoost model and the dev folder were used to validate the model's performance throughout the training process. After that, we tested the aggregated results file in the test folder with the XGBoost model that has been trained.

(2) XGBoost predictions based on the features from the t-test

We divided the data into two groups based on the binary categorization of the PHQ-8, which are normal subjects and depressed subjects. Afterwards, we sampled data on normal subjects with a random state = 0,1,2 and use the voted t-test model to select features with p-values < 0.05.

(3) XGBoost predictions based on features from Spearman's correlation.

In Spearman's correlation, we first encoded the PHQ-8 Score into categorical or ordinal. The range of the PHQ-8 score is between 0 to 24 points with the score between 0 to 4 points set into 0 and representing the normal category, the score between 5 to 9 set into 1 and representing the mild depression category, the score between 10 to 14 are set into 2 and represent the moderate depression category, the score between 15 to 19 are set into 3 and represent the moderately severe depression category, and the score between 20 to

24 are set into 4 and represent the severe depression category [28]. Then, we compared the correlation between all features with the encoded PHQ-8 score and used Spearman's correlation to select features with p-values < 0.05.

4. Results and Discussion

4.1. Feature Selection. In feature selection, we used the first derivative which meant that we investigated whether the speed at which these landmarks or features change positions will have a significant difference between the depressed and normal subjects. Usually, depressed subjects make more sad facial expressions than normal subjects so the speed of changing landmark positions in depressed subjects tends to be lower than in normal subjects [2]. The top 20 features selected based on the p-values were listed in the following explanation, which was obtained from the t-test and Spearman's correlation.

According to Table 3, we determined which features can be correlated with PHQ-8 scores with p-values < 0.05. From the features obtained, it can be concluded that the maximum changes in these features have statistically significant p-values from the t-test on features3D, especially in AU which include AU04, AU12, AU15, AU23, AU28, AU45 which are brow lowerer, lip corner puller, lip corner depressor, lip tightener, lip suck, and blink. Additionally, the Y coordinate also has slightly less number of features than AU, which include Y19, Y20, Y21, Y27, Y39, Y40, and Y41 which is the area of the right face consisting of the eyebrows, eyes, and bridge of the nose. Apart from that, from the features obtained, it can be seen that the AU feature and the Y coordinate have the potential to become important features, especially in certain areas which are the eyes area for the Y coordinate and the lip area for the AU features.

According to Table 4, we also determined which features can be correlated with the p-value of Spearman's correlation. From the features obtained, it can be concluded that changes in the mean of these features possess statistically significant p-values of the Spearman's correlation on AU which include AU04, AU12, AU23, and AU45 which are areas of brow lowerer, lip corner puller, lip tightener, and blink. Additionally, Features3D, especially in the Y coordinate, has slightly less number of features than AU which include Z0, Z1, Z2, Z3, Z4, Z5, Z6, and Z7 which is the right cheek area. Similar to the t-test, it can be seen that the AU features and the

TABLE 3. Top 20 features selected based on the p-value obtained from t-test

No	Features	No	Features
1	max_AU04_c	11	min_AU15_c
2	max_AU12_c	12	min_AU23_c
3	max_AU15_c	13	min_AU28_c
4	max_AU23_c	14	min_AU45_c
5	max_AU28_c	15	stdev_z_0
6	max_AU45_c	16	max_Y19
7	max_Y20	17	max_Y27
8	max_Y21	18	max_Y39
9	min_AU04_c	19	max_Y40
10	min_AU12_c	20	max_Y41

TABLE 4. Top 20 features selected based on the p-value obtained from Spearman's correlation

No	Features	No	Features
1	mean_AU12_c	11	max_AU45_c
2	mean_AU23_c	12	mean_AU12_r
3	mean_y_1	13	mean_Z0
4	mean_y_h1	14	mean_Z1
5	min_AU04_c	15	mean_Z2
6	min_AU12_c	16	mean_Z3
7	min_AU45_c	17	mean_Z4
8	max_AU04_c	18	mean_Z5
9	max_AU12_c	19	mean_Z6
10	max_AU23_c	20	mean_Z7

Z coordinate have the potential to become important features but both of them only focused on the lip and right cheek area.

TABLE 5. The optimal hyperparameter values used in training the models

Hyperparameter	Values
max depth	10
min child weight	10
learning rate	0.001
subsample	0.5
colsample bytree	0.6
reg lambda	0.1
num boost round	5000
gamma	39
alpha	21

4.2. Training Results. In Table 5, the hyperparameter column represents the type of hyperparameter and the values column represents the optimal values of the hyperparameter found through the manual search. These values were used in training all of the XGBoost models

Figure 2 showed the performance of each model during the training session. It can be seen that the XGBoost with the raw data is able to obtain lower validation loss compared to those with features selected by Spearman's correlation and t-test with validation loss of 6.82 compared to 7.21 in Spearman's correlation and 7.24 in the t-test. For the training loss, the t-test filtered model is able to obtain the lowest train loss among the three approaches with 4.20 compared to 3.56 in XGBoost and 4.17 in Spearman's correlation-filtered model. This indicated that XGBoost filtered model has the largest gap between the train and validation losses which means that the model overfits. This phenomenon may have been caused by an insufficient number of train samples, or the training data may not be representative enough to model the general pattern of MDD.

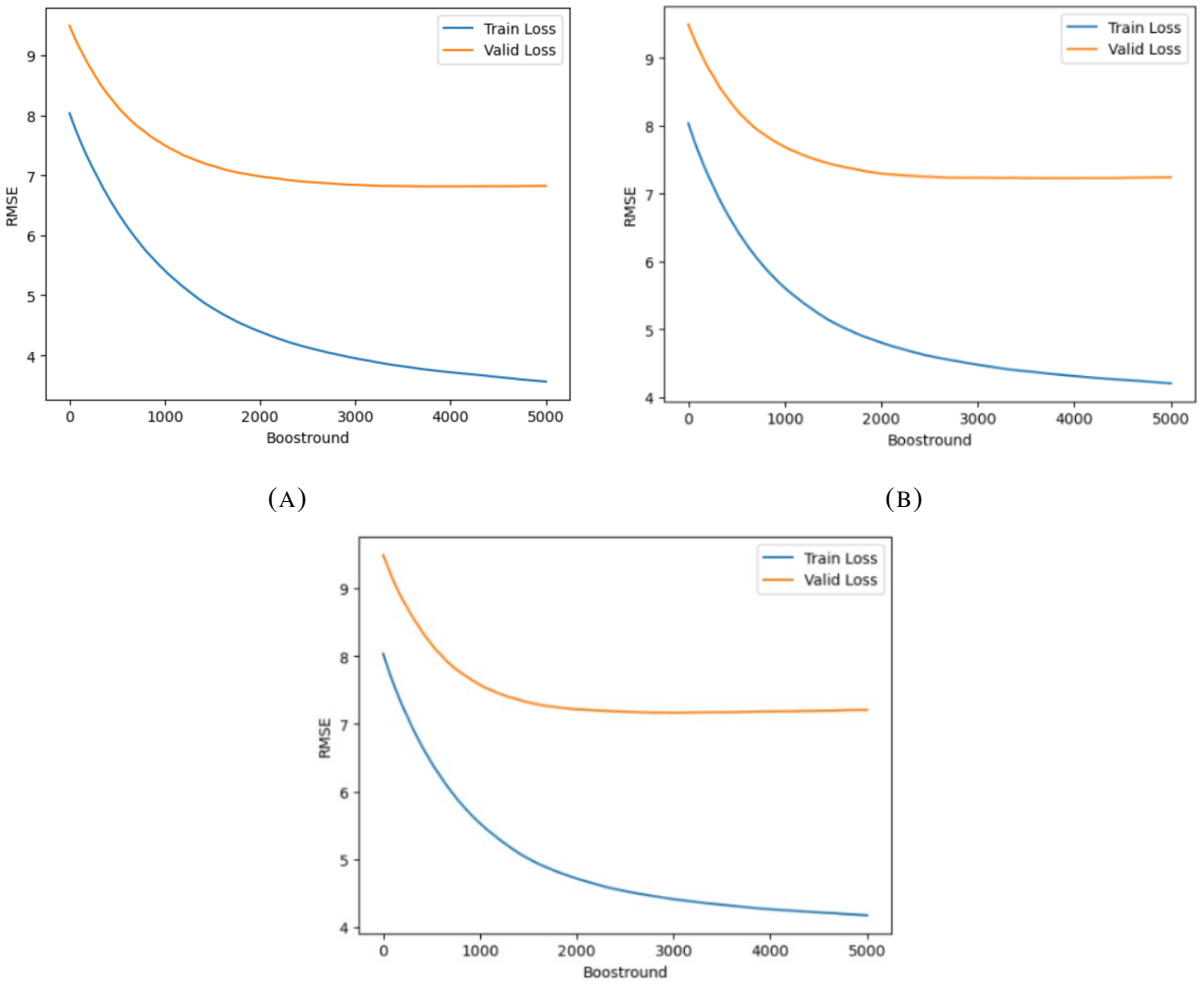


FIGURE 2. Comparison of the training and validation losses of each model trained using XGBoost with: (A) XGBoost training result without feature selection, (B) XGBoost training result with t-test statistical method, (C) XGBoost training result with Spearman's correlation

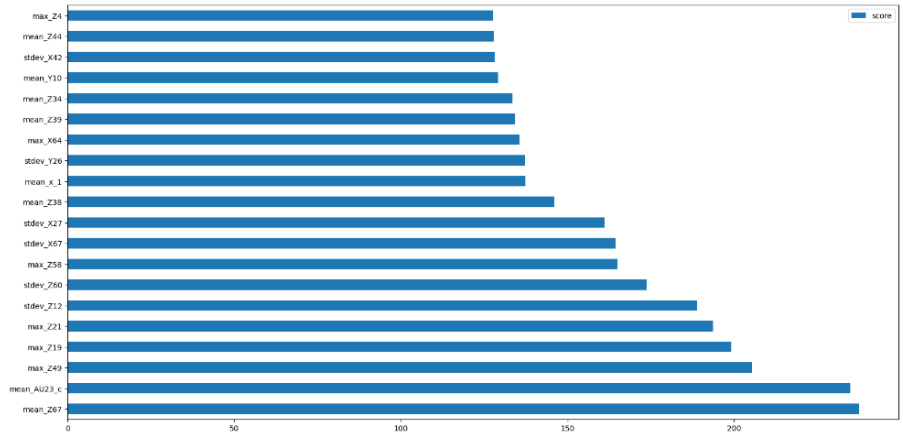
TABLE 6. Evaluation results on the test set for each approach

Parameter	XGBoost	**XGBoost + t-test**	XGBoost + Spearman's correlation
MAE	5.416	**5.387**	5.561
RMSE	6.328	**6.266**	6.402
R^2	0.023	**0.042**	0.0002

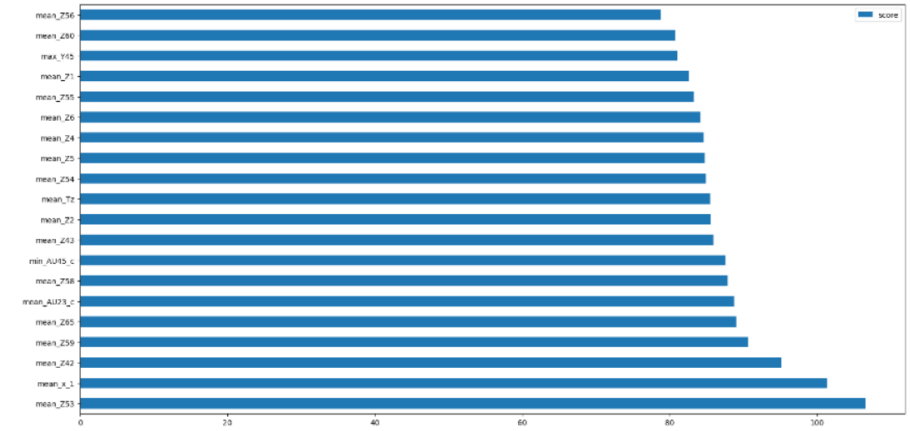
In Table 6, the XGBoost column represents the test results from the XGBoost model using raw first derivative data, and the XGBoost + t-test column represents the test results from the XGBoost training model using the statistical method t-test based on the feature selection shown in Table 3 while the XGBoost + Spearman's correlation column represents the test results from training the XGBoost model using Spearman's correlation based on the feature selection that shown in Table 4.

According to Table 6, it can be concluded that the statistical t-test method is the best method when compared to Spearman's correlation and XGBoost algorithms in selecting the important features. It should be noted that due to data imbalance between depressed and normal subjects, the deployed voted t-test used a poll with three random states or three population pairs between depressed and normal subjects to be used as a feature selection to minimize the effects of sampling bias. Additionally, training the XGBoost model without feature selection is the least recommended to be used among the three. Therefore, it can be inferred that some of the input features may have brought the noise to the data, which was removed through the feature selections.

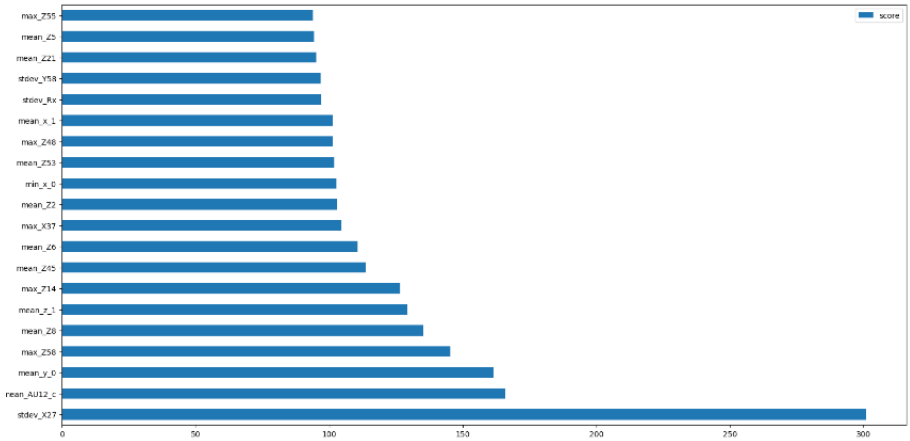
4.3. Feature Importance Results Based on Each Approach. In this section, we have visualized the top 20 important features based on each approach using the results of plotting data in Figure 3 and also displayed the top 20 important features in Table 7. In Table 7 the XGBoost column represents the top 20 important features using the XGBoost model without feature selection, the XGBoost + t-test column represents the top 20 important features using the statistical t-test method with the XGBoost model, and XGBoost + Spearman's correlation column represent the top 20 important features using Spearman's correlation with the XGBoost model.



(A)



(B)



(C)

FIGURE 3. Top 20 gain features important of each model trained using XGBoost with: (A) XGBoost without feature selection, (B) XGBoost with t-test statistical method, (C) XGBoost with Spearman's correlation

TABLE 7. Top 20 important features based on their gains on the trained XGBoost model for each scenario

No	XGBoost	XGBoost + t-test	XGBoost + Spearman's correlation
1	mean_Z67	mean_Z53	mean_Z59
2	mean_AU23_c	mean_x_1	mean_Z64
3	max_Z49	mean_Z42	mean_x_1
4	max_Z19	mean_Z59	mean_Z6
5	max_Z21	mean_Z65	mean_Z54
6	stdev_Z12	mean_AU23_c	mean_Z7
7	stdev_Z60	mean_Z58	mean_Tz
8	max_Z58	min_AU45_c	mean_Z2
9	stdev_X67	mean_Z43	mean_Z52
10	stdev_X27	mean_Z2	mean_Z42
11	mean_Z38	mean_Tz	stdev_Y67
12	mean_x_1	mean_Z54	mean_Z0
13	stdev_Y26	mean_Z5	mean_Z1
14	max_X64	mean_Z4	max_X0
15	mean_Z39	mean_Z6	mean_Z43
16	mean_Z34	mean_Z55	mean_Z57
17	mean_Y10	mean_Z1	mean_z_1
18	stdev_X42	max_Y45	mean_Z17
19	mean_Z44	mean_Z60	mean_Z5
20	max_Z4	mean_Z56	mean_Z65

According to Figure 3 and Table 7, we can conclude that the majority of important features in XGBoost, t-test, and Spearman's correlation are features3D especially in Z coordinate. In XGBoost with raw data, the majority of the features are in the cheek, eyes, ala nasi, and the right side of the lip. In the t-test, the majority of the features are in the left eye, right cheek, and lip area, and in Spearman's correlation, the majority of the features are exactly the same

TABLE 8. Comparison of our best model with previous studies using the same dataset.

Reference	Input Feature	Method	MAE	RMSE
[5]	mean, std, max, and quartile of raw of the first and second derivative.	SVR	4.37	5.84
[9]	faceHOG features	Combination PC and LR	4.64	5.98
[12]	mean, max, min, skewness, kurtosis, standard deviation, median, root mean square level, peak-magnitude to root-mean-square ratio, interquartile range of AU, 3D Landmarks, Head Pose, Eye Gaze, and Geometric Distance	XGBoost	4.97	6.45
Ours	**mean, std, max, min of the first derivative**	Voted t-test with XGBoost model	**5.387**	**6.266**

as the t-test, which are the left eye, right cheek, and lip area. From the three models, it can be concluded that the majority area of the important features is the cheek, eyes, and lip.

4.4. Discussion. Based on our research, the t-test statistical method is the best method compared to the XGBoost and Spearman's correlation models.

According to Table 8, when compared with Song et al.'s research which focused on features such as AU, Gaze, and Head Pose which use the SVR model, their model produces smaller MAE and RMSE values than this study. Research by Song et al. produced MAE and RMSE values of 4.37 and 5.84 which focuses on the AU, Gaze, and Head Pose modalities using SVR. While our proposed method has an MAE of 5.387 and RMSE of 6.266. Such results may have been generated from the fact that they used more features from the raw data and second derivatives, which may include features more relevant than the first derivatives. In addition, Song et al. use the CFS method to reduce overfitting on the classifier or regressor in its analysis. However, Song et al. did not explain what important features could affect the severity of depression,

which is the reason why we experimented with the XGBoost model and statistical methods to determine the important features [5].

Rathi et al.'s research [9] produced better MAE and RMSE values by using LR and PC feature selection because PC and LR are indeed based on linear correlation between features and variable responses. However, this research only focuses on the faceHOG feature set which has high dimensionality to the number of samples in the DAIC-WOZ dataset. If the dimensionality is high, the complexity of the models built for depression detection is also high [9]. Additionally, the more features used, the greater the possibility of noises appearing in the data. Therefore, feature selection is needed to reduce dimensionality and improve the model's performance. Besides, research conducted by Zhang [12] can produce a smaller MAE value but a higher RMSE compared to the voted t-test. However, this study only focused on FAU features, because according to Zhang [12], FAU is the most correlated feature with depression because it can describe various emotions associated with depression while other features cannot provide information as useful as FAU. So, it can be concluded that not all features require first derivatives in their processing stage and not all features must be used in such research. For example, the research conducted by Zhang only focused on raw AU features and Rathi et al. only focused on the FaceHOG feature but can produce better MAE and RMSE results. Therefore, we used the first derivatives that did not focus only on AU or FaceHOG features and proved that there are other features that are also correlated with depression, namely Features3D. This research also proved that the feature selection used in statistical methods can create a model that has a positive R^2 value in this complex data so that this research can open the focus of the research on feature selection algorithms for similar cases in the future.

5. CONCLUSION

In this study, selecting FACS features using t-test and fitting them to XGBoost produced the highest R^2 value compared to other methods. It can be concluded that the R^2 value is relatively small due to limited data exploration. For example, we did not use the FaceHOG feature set like Rathi et al.'s research [9]. Nonetheless, statistical methods can still influence the assessment of a person's level of depression. Besides, we used the first derivative to validate whether changes in the value of the FACS variable are proven to have an association with MDD levels

and resulted that the most impactful feature in detecting MDD severity is Features3D with an average percentage of about 80% of the list of top 20 features for each approach. This research also contributed to producing important features based on the three approaches that were carried out in which these important features would be useful for researchers. By only using features that are relevant to the model, researchers can shorten research time and get better performance in their research. However, in this study, the available dataset was still relatively small and limited to only using the first derivatives. In the future, researchers can conduct MDD detection research using raw data combined with the second derivative method for further research or even using important features that have been obtained using deep learning because although conventional ML is still reliable, deep learning still provides better results [29]. Besides, deep learning has proven that the performance produced in recent years has been outstanding due to the development of technology that is constantly evolving towards better, larger datasets, and deeper network architecture [30].

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] N. Sartorius, Depression and diabetes, Dialog. Clinic. Neurosci. 20 (2018), 47-52. https://doi.org/10.31887/dcns.2018.20.1/nsartorius.

[2] C.M. Bodenschatz, M. Skopinceva, T. Ruß, T. Suslow, Attentional bias and childhood maltreatment in clinical depression - An eye-tracking study, J. Psych. Res. 112 (2019), 83–88. https://doi.org/10.1016/j.jpsychires.2019.02.025.

[3] A. Othmani, A.O. Zeghina, A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept, Healthcare Analytics. 2 (2022) 100090. https://doi.org/10.1016/j.health.2022.100090.

[4] C.A. Lovejoy, Technology and mental health: The role of artificial intelligence, Eur. Psychiatr. 55 (2019), 1-3. https://doi.org/10.1016/j.eurpsy.2018.08.004.

[5] S. Song, L. Shen, M. Valstar, Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, Xi'an, 2018: pp. 158–165. https://doi.org/10.1109/FG.2018.00032.

[6] H. Akbar, S. Dewi, Y.A. Rozali, et al. Exploiting facial action unit in video for recognizing depression using metaheuristic and neural networks, in: 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), IEEE, Jakarta, Indonesia, 2021: pp. 438-443. https://doi.org/10.1109/ICCSAI53272.2021.9609747.

[7] A. Mulay, A. Dhekne, R. Wani, et al. Automatic depression level detection through visual input, in: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), IEEE, London, United Kingdom, 2020: pp. 19-22. https://doi.org/10.1109/WorldS450073.2020.9210301.

[8] Z. Jiang, S. Harati, A. Crowell, et al. Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions, IEEE Trans. Biomed. Eng. 68 (2021), 664-672. https://doi.org/10.1109/tbme.2020.3010472.

[9] S. Rathi, B. Kaur, R.K. Agrawal, Enhanced depression detection from facial cues using univariate feature selection techniques, in: B. Deka, P. Maji, S. Mitra, D.K. Bhattacharyya, P.K. Bora, S.K. Pal (Eds.), Pattern Recognition and Machine Intelligence, Springer International Publishing, Cham, 2019: pp. 22–29. https://doi.org/10.1007/978-3-030-34869-4_3.

[10] A. Ray, S. Kumar, R. Reddy, et al. Multi-level attention network using text, audio and video for depression prediction, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, ACM, Nice France, 2019: pp. 81–88. https://doi.org/10.1145/3347320.3357697.

[11] J. Yoon, C. Kang, S. Kim, J. Han, D-vlog: Multimodal Vlog Dataset for Depression Detection, in: Proceeding of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), 36 (2022), 12226–12234. https://doi.org/10.1609/aaai.v36i11.21483.

[12] W. Zhang, Biomedical engineering application: disease diagnosis and treatment, Thesis, University of Alberta, (2022). https://doi.org/10.7939/R3-1WAQ-PX35.

[13] P. Eteng, Machine learning for mental health screening, Ph.D. thesis, Worcester Polytechnic Institute, (2022).

[14] M. Muzammel, H. Salam, A. Othmani, End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis, Computer Methods Programs Biomed. 211 (2021), 106433. https://doi.org/10.1016/j.cmpb.2021.106433.

[15] T.W. Cenggoro, B. Mahesworo, A. Budiarto, et al. Features importance in classification models for colorectal cancer cases phenotype in Indonesia, Procedia Computer Sci. 157 (2019), 313–320. https://doi.org/10.1016/j.procs.2019.08.172.

[16] B. Mahesworo, T.W. Cenggoro, A. Budiarto, et al. Phosphorylation site prediction using gradient tree boosting, Commun. Math. Biol. Neurosci. 2020 (2020), 48. https://doi.org/10.28919/cmbn/4653.

[17] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016: pp. 785-794. https://doi.org/10.1145/2939672.2939785.

[18] K. Cheng, M.S. Khokhar, M. Ayoub, et al. Nonlinear dimensionality reduction in robot vision for industrial monitoring process via deep three dimensional Spearman correlation analysis (D3D-SCA), Multimedia Tools Appl. 80 (2020), 5997–6017. https://doi.org/10.1007/s11042-020-09859-6.

[19] M.S. Khokhar, K. Cheng, M. Ayoub, et al. Multi-dimension projection for non-linear data via spearman correlation analysis (MD-SCA), in: 2019 8th International Conference on Information and Communication Technologies (ICICT), IEEE, Karachi, Pakistan, 2019: pp. 14-18. https://doi.org/10.1109/ICICT47744.2019.9001973.

[20] B. Liu, X. Tan, Y. Jin, et al. Application of RR-XGBoost combined model in data calibration of micro air quality detector, Sci. Rep. 11 (2021), 15662. https://doi.org/10.1038/s41598-021-95027-1.

[21] T.K. Kim, J.H. Park, More about the basic assumptions of t-test: normality and sample size, Korean J. Anesthesiol. 72 (2019), 331-335. https://doi.org/10.4097/kja.d.18.00292.

[22] R. Kelter, Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research, BMC Med. Res. Methodol. 20 (2020), 88. https://doi.org/10.1186/s12874-020-00968-2.

[23] U. Arioz, U. Smrke, N. Plohl, et al. Scoping review on the multimodal classification of depression and experimental study on existing multimodal models, Diagnostics. 12 (2022), 2683. https://doi.org/10.3390/diagnostics12112683.

[24] A. Saidi, S.B. Othman, S.B. Saoud, Hybrid CNN-SVM classifier for efficient depression detection system, in: 2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET), IEEE, Hammamet, Tunisia, 2020: pp. 229-234. https://doi.org/10.1109/IC_ASET49463.2020.9318302.

[25] A. Mallol-Ragolta, Z. Zhao, L. Stappen, et al. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews, Interspeech. 2019 (2019), 221-225. https://doi.org/10.21437/interspeech.2019-2036.

[26] G. Li, X. Zhu, Y. Zeng, Q. Wang, L. Lin, Semantic Relationships Guided Representation Learning for Facial Action Unit Recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33 (2019), 8594-8601. https://doi.org/10.1609/aaai.v33i01.33018594.

[27] R. Zhi, M. Liu, D. Zhang, A comprehensive survey on automatic facial action unit analysis, Vis. Comput. 36 (2019), 1067–1093. https://doi.org/10.1007/s00371-019-01707-5.

[28] K. Kroenke, T.W. Strine, R.L. Spitzer, et al. The PHQ-8 as a measure of current depression in the general population, J. Affect. Disorders. 114 (2009), 163-173. https://doi.org/10.1016/j.jad.2008.06.026.

[29] A. Mitchell, E. Edbert, G. Elwirehardja, et al. Offline signature verification using transfer learning and data augmentation on imbalanced dataset, ICIC Express Lett. 17 (2023), 359-366. https://doi.org/10.24507/icicel.17.03.359.

[30] M.V. Ramadhan, K. Muchtar, Y. Nurdin, et al. Comparative analysis of deep learning models for detecting face mask, Procedia Computer Sci. 216 (2023), 48-56. https://doi.org/10.1016/j.procs.2022.12.110.