IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 June 2024, accepted 26 June 2024, date of publication 1 July 2024, date of current version 10 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3421376

RESEARCH ARTICLE

A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition

JOSEPH AINA[1], (Student Member, IEEE), OLUWATUNMISE AKINNIYI[1], MD. MAHMUDUR RAHMAN[2], (Member, IEEE), VALERIE ODERO-MARAH[3], AND FAHMI KHALIFA[1,4], (Senior Member, IEEE)

[1]Electrical and Computer Engineering Department, School of Engineering, Morgan State University, Baltimore, MD 21251, USA
[2]Department of Computer Science, School of Computer, Mathematical and Natural Sciences, Morgan State University, Baltimore, MD 21251, USA
[3]Center for Urban Health Disparities Research and Innovation, Department of Biology, Morgan State University, Baltimore, MD 21251, USA
[4]Electronics and Communications Engineering Department, Mansoura University, Mansoura 35516, Egypt

Corresponding author: Fahmi Khalifa (fahmikhalifa@mans.edu.eg)

This work was supported in part by the National Science Foundation (NSF) Grant #2131307: "Computer and Information Science and Engineering Minority-Serving Institutions (CISE-MSI), Demonstration Projects (DP), Information and Intelligent Systems (IIS), information Integration and Informatics (III) Deep Learning-Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support"; and in part by the National Institutes of Health (NIH)/National Institute of Minority Health and Health Disparities (NIMHD) under Agreement 1OT2OD032581 and Agreement U54MD013376.

ABSTRACT Mental illness has grown to become a prevalent and global health concern that affects individuals across various demographics. Timely detection and accurate diagnosis of mental disorders are crucial for effective treatment and support as late diagnosis could result in suicidal, harmful behaviors and ultimately death. To this end, the present study introduces a novel pipeline for the analysis of facial expressions, leveraging both the AffectNet and 2013 Facial Emotion Recognition (FER) datasets. Consequently, this research goes beyond traditional diagnostic methods by contributing a system capable of generating a comprehensive mental disorder dataset and concurrently predicting mental disorders based on facial emotional cues. Particularly, we introduce a hybrid architecture for mental disorder detection leveraging the state-of-the-art object detection algorithm, YOLOv8 to detect and classify visual cues associated with specific mental disorders. To achieve accurate predictions, an integrated learning architecture based on the fusion of Convolution Neural Networks (CNNs) and Visual Transformer (ViT) models is developed to form an ensemble classifier that predicts the presence of mental illness (e.g., depression, anxiety, and other mental disorder). The overall accuracy is improved to about 81% using the proposed ensemble technique. To ensure transparency and interpretability, we integrate techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and saliency maps to highlight the regions in the input image that significantly contribute to the model's predictions thus providing healthcare professionals with a clear understanding of the features influencing the system's decisions thereby enhancing trust and more informed diagnostic process.

INDEX TERMS Object detection, YOLOv8, feature fusion, saliency maps.

I. INTRODUCTION

In our contemporary world, we grapple with the profound challenge of mental disorders—a silent, yet pervasive affliction that can significantly impact an individual's cognition, emotional regulation, and behavior. It remains a significant global health challenge, affecting individuals across all age groups and socioeconomic backgrounds [1]. There are many different types of mental disorders, such as anxiety disorder, depression, bipolar disorder, eating disorders, Schizophrenia etc. Globally, the number of persons with depressive symptoms increased from 193 to 246 million

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

(almost a 28% increase), while anxiety showed a 25% rise (from 298 to 374 million [2], [3]. Mental disorders are therefore a pressing concern that requires effective early detection and diagnosis, which is crucial for providing timely interventions and thus improving outcomes. Mental disorder can be identified by a clinically significant defect that is occasionally accompanied with distress or impairment in critical areas of functioning [2]. Numerous literature studies have shed light on the significance of unconsciously transmitted behavioral symptoms, such as head pose, eye-gaze direction, and facial expressions, as potential indicators for predicting the presence of diverse mental health disorders. These subtle manifestations that individuals express unknowingly carry valuable information regarding an individual's mental well-being [4], [5], [6].

Advancements in computer vision and Deep Learning (DL) techniques have shown great potential in various domains, especially healthcare developing automated tools that can assist in the detection and diagnosis of medical conditions [7], [8], [9], [10]. Particularly, DL models utilizing Convolutional Neural Networks (CNNs) possess the ability to analyze large datasets and extract meaningful patterns from visual information to assist with remarkable performance in object recognition and disease diagnosis and classification [11]. Therefore, leveraging cutting-edge technologies in the context of mental health to harness these unconscious cues can have a transformative impact on the early detection and targeted intervention of mental illnesses. Although DL-based tools have demonstrated impressive performance in many fields, their decision-making processes are often not transparent; thus, raising concerns about reliability, interpretation, and potential bias. Therefore, in the field of mental health, it is important to develop systems that can provide human-interpretable explanations for their predictions. This would allow health professionals such as psychiatrists and psychologists to understand the rationale behind the model's conclusions and make informed decisions.

This study aims to address the need for an explainable technique for mental disorder detection. Namely, to develop a usable prediction tool for recognizing associated mental disorder's visual signs while providing explicit explanations for its predictions. Specifically, we introduce a hybrid, objective mental disorder detection system that can predict the presence of mental illness by observing human visual cues. A tool that can ultimately assist clinicians in making informed decisions and tailoring subsequent personalized treatment, thereby enhancing the overall well-being and quality of life for individuals experiencing mental health challenges. To achieve our goal we built, an innovative ensemble deep model that integrates CNN-based architectures and Vision Transformers (ViT). Moreover, to enhance the interpretability of the system and facilitate a deeper understanding of the underlying mental illness patterns, we incorporated visual explanations such as Gradient-weighted Class Activation Mapping (Grad-CAM) and saliency maps. The proposed architecture is trained and evaluated on a collected mental disorder dataset. Insight experimental evaluations of the proposed ensemble architecture's detection capabilities are conducted using a spectrum of well-known metrics (e.g. accuracy, precision, recall, and F1-score) and various ablation studies. In essence, the key contributions of this study are as follows:

- Introducing a novel pipeline for the analysis of facial expressions, leveraging both the AffectNet and FER 2013 datasets. To the best of our knowledge, it is the first research to consider the ensemble of more than one dataset for mental disorder prediction.
- Developing a robust analysis pipeline for mental illness detection supported by a level of explainability and interpretability.
- The system architecture is unique as it contains layers that analyze facial emotions and generate a mental disorder dataset from the result of the analysis.
- A comprehensive hybrid architecture that integrates object detection with an ensemble of various learning methods (i.e. shallow, deep neural networks, and ViT) to predict mental disorder.

The paper is structured into six sections, starting with the introduction of mental disorders and technological advancement for the detection and diagnosis in section I, the relevant review of the recent and related literature work followed by the outlined contributions of this work, is given in Section II. Full descriptions of the data, the methodology, and the details of the learnable modules and features extraction strategies are completely given in Section III. Employed evaluation criteria, conducted experiments, and obtained results are given in Section IV. Results discussion, associated conclusions, and future work suggestions are outlined in Sections V and Sections VI, respectively.

II. RELATED WORK

Various literature work have been proposed for the Artificial Intelligence (AI) application in mental illness detection using facial emotion recognition. For example, Vasantha et al. [12] applied several algorithms such as Decision Tree (DT), Random Forest (RF), and Artificial Neural Network (ANN) to detect mental health based on various facial emotions gotten either through live webcam or by uploading facial images. The algorithms were trained and validated on 2,872 webcam images using four emotions: neutral, sad, angry, and happy. The RF algorithm provided the highest accuracy. However, more datasets should be considered to achieve a more generalizable model. Naveen et al. [13] suggested early detection and assistance for challenges in identifying people's mental state using a two-stage—facial emotion recognition model. The latter, is a CNN-based classifier trained using the FER-2013 [14] to predict emotions in seven categories: *Anger*, *Disgust*, *Sadness*, *Neutrality*, *Happiness*, and *Joy*. When an image's mood has predicted by the model is negative, the system suggests coping mechanisms. He et al. [15] proposed a feature descriptor named Median Robust Local Binary Patterns from Three Orthogonal Planes

(MRLBP-TOP), which is able to capture both the microstructure and macrostructure of facial appearance and dynamics. The study aimed at estimating the Beck Depression Inventory-II (BDI-II) score, a clinical depression-specific score. The model was trained on the AVEC2013 [16] and AVEC2014 [17] datasets and the results showed a Root Mean Squared Error (RMSE) and Mean Average Error (MAE) of 9.20 and 7.55, respectively. A study by Fei et al. [18] presented a deep CNN-based algorithm to support mental state detection and diagnosis. The model combined deep features from AlexNet and Linear Discriminant Analysis (LDA) classifier, which was validated against five databases: JAFFE, KDEF, CK+, FER2013 and AffectNet. The findings showed that the suggested model affordability as a friendly cognitive tool to identify, track, and diagnose mental health. In another study, Harati et al. [19] explored the analysis of facial expressions using videos of severely depressed patients before and after an experimental depression treatment (Deep Brain Stimulation (DBS). An elastic net ordinal regression model was developed to classify three levels of severity in patients' depression recovery. The study's findings revealed that unsupervised characteristics collected from these video records, when used in an ordinal regression predictor may distinguish between various degrees of depression severity. Another study by Hadjar et al. [20] employed CNN transfer learning models for real-time video analysis to track emotions so as to support the treatment of patients with primary mental disorders. The results showed that the tiny face detector and Multi-task Cascaded Convolutional Networks (MTCNN) models gave a lower accuracy than the Single Shot Detector (SSD) Mobilenet V1 network, and the accuracy of detecting emotions based on facial expressions declined when the light quality in the experiment site increases. Munsif et al. [21] developed a lightweight DL-based facial expression framework to classify patients with Neurological Disorders (ND). To train the model, images were obtained from a public dataset. Emotions were categorized into *Normal*, *Happy*, *Sad* and *Anger*. The results showed that the model had an accuracy of 93%. Swamy et al. [22] developed a platform that helped to screen depression using image processing and voice assistant. The system combined features extracted from facial analysis with features extracted from questionnaire responses to predict the severity of depression. An 8-layer CNN model was trained on 178 *Angry*, 211 *Happy*, and 208 *Sad* emotion images. The result showed 88% sensitivity and 88% specificity as validated by a mental health professional. In another study, Tadalagi and Joshi [23] addressed illumination variance in machine learning detection of facial emotion to predict depression by using Local Binary Pattern (LBP) as feature extractor and Support Vector Machine (SVM) for depression level detection. The model was trained on the CK dataset as well as other images collected from the internet. The experimental results showed an accuracy of 86%. Hussein et al. [5] used human facial features to develop a model to detect and classify mental health states using a set combination of facial emotions. Haar feature-based cascade was used to extract features from the FER+ dataset then the VGG model was used to predict emotion. In cases of abnormal emotion, the model predicts the presence of depression, anxiety, or other disorder. The result revealed a 95% accuracy.

In addition, other literature research integrated facial and auditory data for mental disorder predictions. Particularly, Mantri et al. [8] proposed a system to automatically diagnose depression using facial characteristics, speech properties, and brain Electrogastrogram (EGG). A decision fusion technique was also implemented to integrate the features in real time. The prediction model classifies the subject based on depression severity. Data from 50 subjects were used to train the model and the result showed that their framework improved depression prediction performance with 80%, 78% and 72% accuracies on the EGG, speech, and facial data, respectively. Singh and Goyal [24] attempted to decode depressive disorder using computer vision. A questionnaire [33] on Attention Deficit Hyperactivity Disorder (ADHD) was administered to a total of 401 volunteers. The responses were studied by clinical psychologists and seventy-two patients were identified as potential depressive disorder patients. A total of 38 out of these patients were invited for an interview. The interview resulted in 18 cases of depressive disorder and 20 cases of non-depressive disorder. A CNN model was trained using the data gathered from these respondents to determine whether a person was depressed or not. The experimental result showed a precision score of 74% in identifying depressive patients. The study therefore concluded that patients with depressive disorders display different facial expressions than others without the condition. Another study by Kong et al. [25] developed a method to recognize depression based on facial images and deep CNN. The images consisted of 1,020 depressed patients and 1,100 healthy participants. Five CNN-based models were trained namely Fully Connected CNN (FCNN), Visual Geometry Group 11 (VGG11), VGG19, deep Residual Network 50 (ResNet50), and Inception version 3. The results demonstrate that depression may be quickly, precisely, and automatically identified using deep CNNs. Gilanie et al. [26] proposed a system to detect depression based on facial expression analysis. Facial Action Coding System (FACS) was used to extract microexpressions as Action Units (AUs), which were correlated with the depressed, disgusted, and contemptuous traits for depression identification. A CNN-based model was further developed to automatically classify depressed patients from images and videos in real-time. The model was trained on 3–5-minute videos collected from 180 patients obtained from Bahawal Victoria Hospital, Bahawalpur in Pakistan. The results revealed a 99.9% validation accuracy. However, more datasets need to be considered to achieve a generalizable model. Another study by Lee and Park [27] proposed a Region-based CNN (R-CNN) model to assist in the diagnosis of depressive disorder. Images were collected from

TABLE 1. Summary of reviewed literature on deep learning approaches to facial emotion detection and mental illness prediction.

Ref	Method	Dataset	Accuracy
[8]	Decision Fusion	Facial Characteristics, EGG, Speech Data	92%
[4]	Haar feature-based cascade, VGG	FER+	95% accuracy
[12]	RF, DT, ANN	2872 webcam images	100%
[13]	CNN	FER 2013	—
[15]	MRLBP-TOP	AVEC2013 (AVEC2014)	RMSE: 9.20 (MAE: 7.55)
[18]	AlexNet, LDA	JAFFE, KDEF, CK+, FER 2013, AffectNet	–
[19]	Elastic Net ordinal regression	Patient's video dataset	—
[20]	SSD MobileNet, Tiny Face Detector, MTCNN	Author's curated video dataset	—
[21]	CNN	KDEF	93% accuracy
[22]	CNN	178 angry, 211 happy, and 208 sad emotion images	88% sensitivity and specificity
[23]	LBP, SVM	CK and Internet images	86% Accuracy
[24]	CNN	ADHD Dataset	Precision: 74
[25]	FCNN, VGG11, VGG19, ResNet50, Inceptionv3	Facial Images of depressed vs non-depressed patients	98.23%, 94.40%, 97.35%, 94.99%, 97.10%
[26]	FACS, CNN	180 depression patients	99.90%
[27]	R-CNN	Facial Expression image data from the internet and direct collection	–
[28]	OpenPose	–	–
[29]	multi-modality fusion model	Audio, Video and Text dataset	RMSE: 5.12& MAE: 4.12
[30]	SVM	97 high-risk patients and 88 low-risk patients	95.60%
[31]	CNN	Face tracking data	96%
[32]	CNN	FER 2013 dataset, RAVDESS, TESS, SAVEE, and CREMA-D	91%, 82%

EGG: ElectroEncephaloGram; **BDI-II**: Beck Depression Inventory-II; **SER**: Speech Emotion Recognition; **FER**: Facial Emotion Recognition; **ASD**: Autism Spectrum Disorder; **ADHD**: Attention deficit hyperactivity disorder; **RF**: Random Forest; **DT**: Decision tree; **MRLBP-TOP**: Median Robust Local Binary Patterns from Three Orthogonal Planes; **LDA**: Linear Discriminant Analysis; **SSD**: Single Shot Detector; **MTCNN**: Multi-task Cascaded Convolutional Networks; **LBP**: Local Binary Patterns; **FCNN**: Fully Connected CNN; **FACS**: Facial Action Coding System; **RCNN**: Region-based CNN; **AVEC**: Audio/Visual Emotion Challenge; **JAFFE**: Japanese Female Facial Expression; **KDEF**: Karolinska Directed Emotional Faces;**CK**: Cohn Kanade; **RAVDESS**: Ryerson Audio/Visual Database of Emotional Speech and Song; **TESS**: Toronto Emotional Speech Set;**SAVEE**: Surrey Audio/Visual Expressed Emotion; **CREMA-D**: Crowd Sourced Emotional Multimodal Actors Dataset; **RMSE**: Root Mean Square Error; **MAE**: Mean Absolute Error

the internet as well as directly from users by capturing their pictures. The facial regions of interest were the eyes and lips. The result revealed that the accuracy of emotion detection can be improved by using the comparative analysis method based on user data to identify emotions through facial expressions. Furthermore, another study by Sadek et al. [28] conducted a comprehensive survey of Autism Spectrum Disorder (ASD), its types, symptoms, prevalence, and developments in its diagnosis. The study also proposed a computer vision-based framework to detect observable symptoms of ASD which involved four main stages known as dataset preprocessing, feature extractions, behavior description and behavior classification. The OpenPose [34] human body pose estimation and skeleton representation ANN model was proposed to extract human body features. With gender variations taken into account, Samareh et al. [29] used the audio, video, and text modalities to find the biomarkers that are predictive of depression. From the study, it was discovered that gender differences in voice and facial emotions are crucial in identifying depression. Furthermore, Benfatto et al. [30] developed a feature extraction and SVM-based model to classify children with dyslexia. The experiment was based on eye tracking data from 185 subjects and the model achieved an accuracy of 95.6%. Another study by Jaiswal et al. [31] proposed a computer-vision-based system to automatically predict ADHD and ASD. The study adopted the dynamic DL method [35] and used face tracking data to determine head movements and facial animation unit characteristics from color and depth (RGBD) photos captured with a Kinect 2.0 sensor camera. The study also created an RGBD database which was used to validate the model on 55 people who had previously been diagnosed with ADHD or ASD as well as subjects from a healthy control group. The results showed that the approach achieved a 96% classification accuracy for the controls vs. condition (ADHD/ASD) group and 94% for the Comorbid (ADHD+ASD vs. ASD alone group). Shanthi et al. [32] developed an integrated approach to access depression scale using four modules namely Scales Questionnaire, Doctor Chat, Speech Emotion Recognition (SER), and Facial Emotion Recognition (FER). The FER2013 dataset is utilized for the FER module; and RAVDESS, TESS, SAVEE, and CREMA-D are all used together for speech-based recognition. The approach resulted in an accuracy of 91% for FER and 82% accuracy for SER. In total, Table 1 shows the summary of reviewed literature with the research objective focused on mental

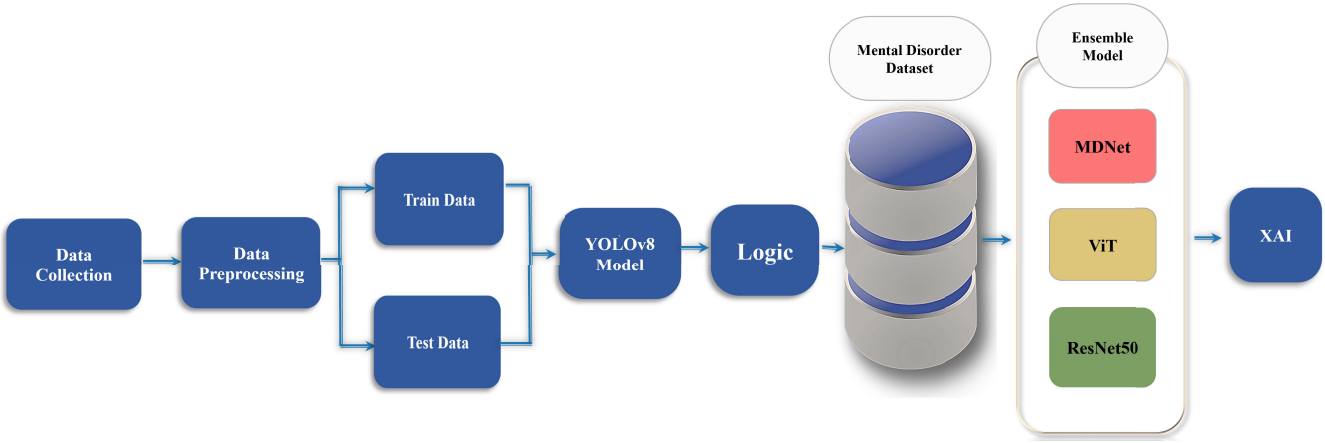


FIGURE 1. Structure of the analysis pipeline for human mental disorders detection through facial emotion recognition.

illnesses/disorders or brain disorders [28], [30], [31] using facial characteristics [4], [8], [12], [21], [22], [23], [25], [26], [27], video/speech properties [8], [15], [20], [22], [29], EGG signals [8].

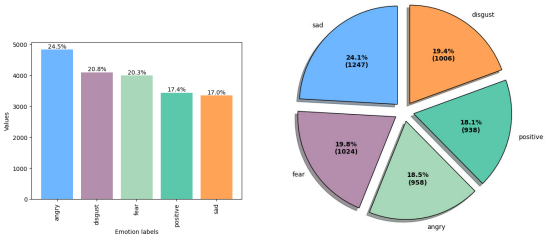
III. MATERIALS AND METHODS

Facial expression is an important feature that could be used to detect mental disorders [36], which is the main objective of the proposed work. The structure adopted to achieve this objective is as depicted in Figure 1. The experiment involved collecting datasets related to human facial expressions, data preprocessing, data split into train and test sets, model training, and evaluation. The State-of-the-art YOLOv8 model was trained on the dataset and extended to identify the presence of mental disorders based on several combinations of the top-2 predictions. The product of this generated a mental disorder dataset with classes of mental disorders namely, anxiety disorder, depressive disorder, no disorder, and other disorders. An ensemble model is proposed which is a concatenated fusion of three learning models: an in-house CNN called MDNet, ViT, and pre-trained ResNet50. The models were trained individually on the mental disorder dataset and the performance was evaluated and compared with the result of the proposed ensemble model. Ablation techniques were also considered to determine the important and less important components of the ensemble model. For explainability, Grad-CAM [37] and saliency maps [38] were adopted to identify facial features that led to the model's decision. The method used to generate the saliency maps is the Vanilla Saliency method [38]. The tensorflow toolkit for debugging tensorflow keras models was used. This method of saliency map was adopted basically due to the image-specific saliency map feature that highlights a given image with respect to the given class. The latter aligns best with the mental disorder domain as disorders are individual-based. The specific process of analysis is expatiated in the succeeding sections.

A. STUDY'S DATABASE GENERATION

The foundation of this research lies in the utilization of two primary datasets, each carefully selected to enable the development of an accurate and robust mental disorder detection system. The datasets employed are the Affect-Net [39] and FER 2013 datasets, each contributing unique attributes and value to our research endeavor. The AffectNet is a large-scale dataset of facial expressions collected from the wild, containing over one million images annotated with seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). The images are diverse in terms of lighting conditions, facial poses, and ethnicities, making them suitable for training deep-learning models to recognize emotions in real-world scenarios. Each image in AffectNet depicts a single face with varying resolutions, sizes, and aspect ratios [39]. The images are resized to 64×64 pixels for training of the proposed model. FER2013 is a widely used benchmark dataset for facial expression recognition, containing 30,000 grayscale images of faces cropped to include only the facial region: 48×48 pixel images. The AffectNet and FER 2013 datasets contain images of human beings with different facial expressions, categorized into distinct emotions; namely, *Neutral*, *Happy*, *Sad*, *Surprise*, *Fear*, *Disgust*, *Angry*, and *Contempt* [40].

The FER 2013 dataset formed the base of the first stage of this research, the YOLOv8 model was trained on this dataset which was then validated on the AffectNet dataset to generate the mental disorder dataset. Preprocessing the data began with closely matching each image to its corresponding label. Next, the images were normalized by dividing them by 255.0, which scaled the pixel values between "0" and "1". The labels related to positive emotions such as *Neutral*, *Happy*, and *Surprise* were sampled and fused into one emotion category named *positive*. Thus, the dataset used for the first stage of model training is presented in Figure 2 and the count of images in each category is presented in Table 2.



(a) (b)

FIGURE 2. Image distribution in the study's database for the (a) train and (b) test sets.

TABLE 2. Count of images per class in the study's database.

Class	Train	Test
Angry	4830	958
Disgust	4097	1006
Fear	3995	1024
Positive	3436	938
Sad	3350	1247
Total	**19708**	**5173**

B. YOLOV8 MODEL

The You Only Look Once - version 8 (or the YOLOv8) model is renowned for its state-of-the-art capabilities in object detection, image classification, and instance segmentation. Its name, "YOLO," succinctly encapsulates its essence, emphasizing the model's efficiency and accuracy. At the time of conducting this research, YOLOv8 emerges as the most recent and promising model in the realm of computer vision. One of YOLOv8's distinguishing features is its ability to maintain a lightweight design while delivering high accuracy. Two distinct components that form the architecture of the model are the backbone and the head component. The backbone forms the model's core architecture, which is an evolution of the CSPDarknet53 architecture, comprising 53 convolutional layers. This sophisticated backbone is instrumental in capturing intricate details from input images [41]. The head component is responsible for generating the model's output, encompassing bounding boxes, class predictions, and other essential information. It encompasses a series of convolutional layers and fully connected layers, which refine and interpret the features extracted by the backbone [42]. In the first phase of the training process, we employed the state-of-the-art YOLOv8 object classification model to establish a foundation for our mental disorder detection system. The training process involves utilizing the study's dataset, which comprises a diverse range of facial images depicting various emotions. One innovative approach applied during this phase involves considering several combinations of the top-2 predictions generated by YOLOv8 on the test set. This approach allows us to create a new database categorized into four distinct mental disorder categories: anxiety, depression, other disorders, and no disorder. This categorization is pivotal for training and evaluating our mental disorder detection system effectively.

1) YOLOV8 ARCHITECTURE

YOLO is a well-known object detection framework designed for real-time processing, and YOLOv8 is the latest iteration, incorporating several advancements to enhance its performance. YOLOv8 retains the core principle of YOLO, where object detection is treated as a single regression problem, directly predicting bounding boxes and class probabilities from full images in one evaluation. This results in high-speed processing while maintaining accuracy. YOLOv8 introduces several architectural improvements:

- **Backbone network** is based on CSPDarknet53, a more efficient and powerful variant of DarkNet. CSPDarknet53 includes Cross Stage Partial connections that enhance gradient flow and reduce computation.
- **Neck network** uses a Path Aggregation Network (PANet) to improve the information flow between different levels of feature maps. PANet enhances localization accuracy and helps in the detection of small objects.
- **Head network** employs anchor-free detection, which simplifies the detection process by predicting key points directly, leading to faster and more accurate predictions.
- **Additional features** such as auto-learning bounding box anchors, mosaic data augmentation, and label smoothing features are integrated with YOLOv8 to further boost detection performance.

These enhancements in YOLOv8 make it particularly effective for our task of facial expression analysis, as it allows for real-time, accurate detection of facial regions and landmarks. The trained YOLOv8 model proposed in this study consists of a total of 183 layers and approximately 36 million trainable parameters.

C. PREDICTIVE LOGIC AND MENTAL DISORDER DATASET

Upon completing the training of the YOLOv8 model on the study's dataset, we transitioned to the critical phase of utilizing this fine-tuned model for predictive tasks. In this phase, the model was applied to the test set of the AffectNet dataset, and a unique approach was adopted to determine the presence or absence of mental disorders in individual images. The probabilities of the top two predictions were considered if the highest probability was less than one. If the top two predictions include fear and disgust emotion, the image is categorized as anxiety disorder as published by [43]. *Sad* and *Angry* suggests depressive disorder as published by [44]. Image predictions that do not contain positive emotions were characterized as other emotions while predictions that contain positive emotions are categorized as no disorder. Thus, the distribution of images in the generated mental disorder dataset is shown in Figure 3. The dataset consisted of a total of 3,601 images. A summary of the logic is shown in Figure 4.

D. ENSEMBLE FEATURE EXTRACTORS

To enhance accuracy and precision in detecting mental disorders, the proposed analysis workflow integrates

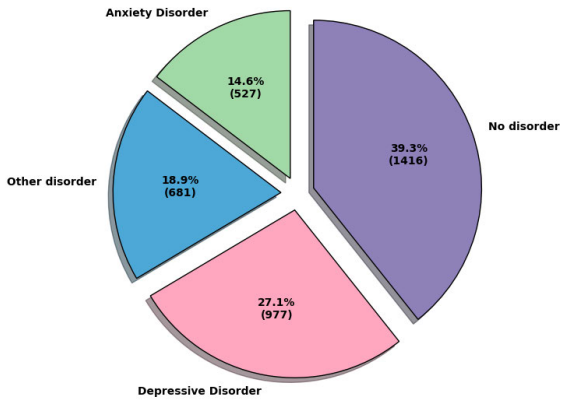


FIGURE 3. Image distribution in the mental disorder dataset.

information-rich features from three learning models, namely, MDNet, ViT, and pre-trained ResNet50. These algorithms are renowned for their proficiency in image classification tasks. In this phase, we meticulously trained each of these individual models on the mental disorder dataset. Furthermore, in order to generate an ensemble classifier model, we employed a fusion strategy in which the layers of these models are joined using the stacking method. This ensemble classifier model capitalizes on the diverse strengths of each individual model, resulting in a comprehensive and robust mental disorder detection framework.

The MDNet model is the first learning module for feature extraction that is based on an in-house CNN architecture. Generally, CNN models represent an advancement in the field of deep learning, developed for image analysis and processing. Its architecture is made up of layers each playing specific roles that enable it to understand and extract intricate features from the images. The layers include convolution layers which employ small filters otherwise known as kernels that scan across the image to extract fundamental features such as edges, textures, and patterns creating feature maps that highlight these specific features. The pooling layers play the critical role of reducing the spatial dimensions of the generated feature maps while retaining the most important information. Common pooling techniques include max pooling and average pooling. The final layer of CNN models is the fully connected layer where the features learned from the previous layers are combined to make predictions. In this study, we developed a CNN model from scratch and named it MDNet. We conducted several experiments to optimize the MDNet architecture hyperparameters and we found that the final structure composed of four convolutional layers, four max-pooling layers, five dropout layers for regularization, and a fully connected layer led to the best performance. The total number of trainable parameters in this model is 261,188.

In addition to the from-scratch-trained CNN (MDNet), we harness the power of well-trained CNN architecture, namely the **pretrained ResNet50** on the ImageNET database. The ResNet50 model is an improvement on the traditional CNN model with 50 layers and approximately 24 million trainable parameters, it has addressed the vanishing gradient problem often encountered by CNN models. The vanishing gradient problem occurs when adding more layers to a CNN model does not result in a better performance of the model. The ResNet50 model circumvents the vanishing gradient problem using the residual layers. The residual layers allow for the direct flow of information from one layer to another, bypassing some layers. This enables the network to learn residual functions, which are effectively the differences between the desired output and the current output at a given layer. By learning these residuals, ResNet-50 can train very deep networks without suffering from vanishing gradients [45]. By incorporating a pre-trained ResNet-50 into our ensemble model, we harness its ability to learn intricate features and patterns from facial expression images.

Finally, the ViT is the third module used in our ensemble architecture. The development of the ViT model is another significant advancement in the field of computer vision. By introducing fixed-sized patch embeddings, the model addresses the drawbacks of CNN models. The latter are passed through a multi-layer transformer encoder and further using a self-attention mechanism. This mechanism enables the ViT model to learn intricate contextual relationships between patches. Unlike CNNs, which operate on fixed-size receptive fields, the self-attention mechanism in transformers allows the ViT to dynamically weigh the importance of each patch in relation to others. This dynamic interplay between patches fosters a more nuanced understanding of the visual context, making it easier for the ViT model to learn intricate features and relationships within the image. Our proposed ViT model consists of a total of 64 layers and the total number of trainable parameters in this model is 21.6 million. By incorporating ViT alongside other models, we harness its unique ability to learn and capture both local and global image features and thus improve the quality of prediction. Key components of the ViT architecture include:

- **Patch Embedding:** The input image is divided into fixed-size patches (e.g., 16×16 pixels). Each patch is then linearly embedded into a vector of a fixed dimension. This process transforms the image into a sequence of embedded patches.
- **Positional Encoding:** Since transformers do not inherently understand the order of the input tokens, positional encodings are added to the patch embeddings to retain spatial information. These encodings help the model understand the relative positions of patches.
- **Transformer Encoder:** The core of the ViT is the transformer encoder, which is made up of feed-forward neural networks and several layers of multi-head self-attention. The self-attention method enables the model to capture long-range interdependence and the global context by allowing it to focus on diverse regions of the image.

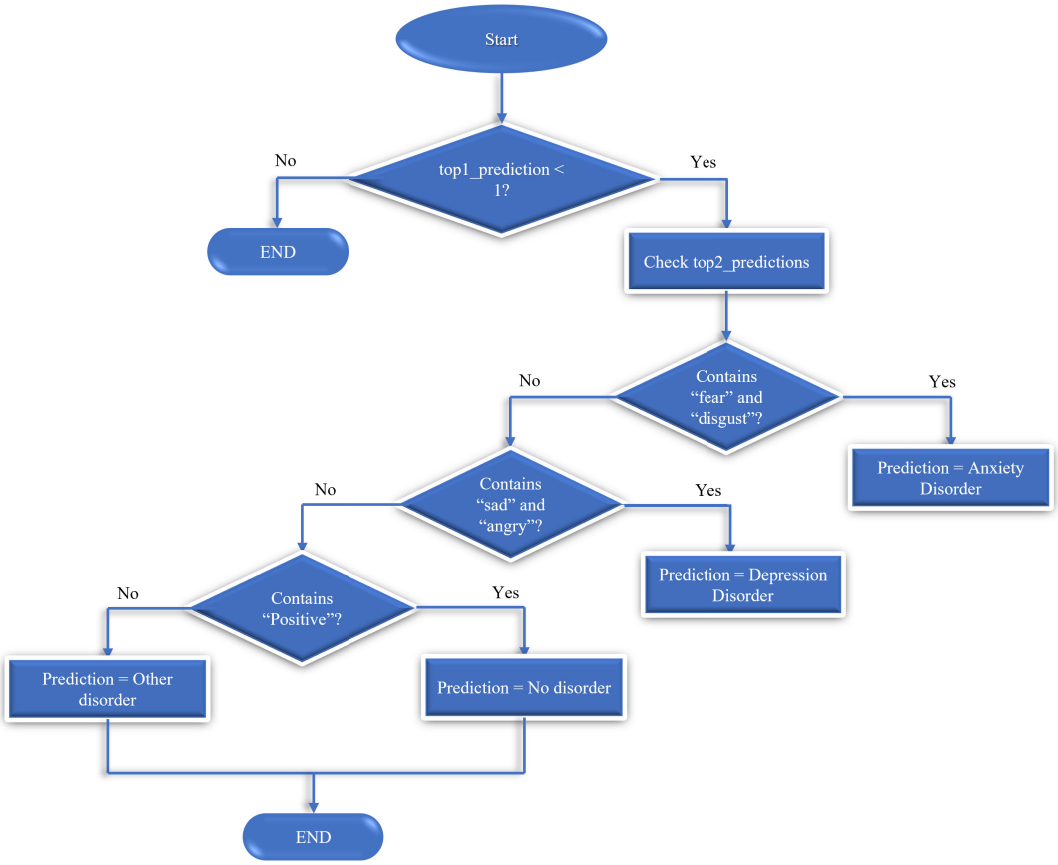


FIGURE 4. A flow chart of the predictive logic adopted in this study.

- **Classification Head:** After processing through the transformer encoder, the output tokens are aggregated, and a classification head (often a fully connected layer) is applied to predict the final output, such as the class of the facial expression.

ViTs have emerged as a powerful architecture for image processing tasks due to their ability to capture long-range dependencies in images more effectively than traditional CNNs. Unlike CNNs, which primarily focus on local features through convolutions, ViTs process the entire image at once, enabling them to understand the global context. This global context is particularly beneficial in facial expression analysis, where subtle variations in different parts of the face contribute to the overall expression. The combination of ViT and CNN aims to leverage the strengths of both architectures to enhance performance in facial expression recognition tasks. CNNs are highly effective at extracting local features but may struggle with capturing global context, especially in complex scenes. ViTs excel at understanding global relationships within the image. By combining CNNs and ViTs, we can thus take advantage of both the global context knowledge of ViTs and the local feature extraction of CNNs, leading to improved accuracy, better generalization across different datasets, and increased robustness to variations in lighting, angles, and occlusions. In our pipeline, we initially use CNN layers to extract rich local features from facial images, which are then fed into the ViT component to capture global context and dependencies, ensuring a comprehensive understanding of facial expressions. This hybrid approach is intended to exploit the best of both worlds, thereby improving the accuracy, generalization, and robustness of facial expression analysis.

Many research work has been introduced to develop transformer-based models in various domains, e.g., [46], [47], and [48]. In choosing the layers of the learning models, such as the MDNet and ViT, some experiments were performed to determine the optimal layers and parameters. Particularly, GridSearchCV [49] was used to determine the optimal layers and parameters. Fivefold cross-validation technique was applied in training the ensemble models on the mental disorder dataset.

IV. EXPERIMENTAL RESULTS

The analysis was conducted on the above-mentioned dataset in Section III-A. To evaluate the models used in the study, common metrics were employed, such as precision and recall scores, accuracy scores, and F1 scores. Quantitative and qualitative representations were also considered for evaluation, such as confusion matrix and the Receiver Operating Characteristics (ROC) curves. Typically, four components are used to determine most of the metrics: **True**

Positives (TP): the correctly identified positive instances; the **True Negatives (TN)**: the correctly identified negative instances; the **False Positives (FP)**: incorrectly classified positive instances (Type I error); and the **False Negatives (FN)**: incorrectly classified negative instances (Type II error). Those metrics are defined as follows:

- **Precision** measures the ratio of correctly predicted positive observations to the total predicted positive observations. It is calculated using the formula in Eq. (1)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

- **Recall** also known as sensitivity or true positive rate, quantifies the proportion of actual positive instances correctly predicted by the classifier. It is calculated using the formula in Eq. (2)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

- **Accuracy** score serves as a fundamental metric, revealing the proportion of correctly classified instances within the dataset. It calculates the proportion of instances that are correctly classified among the total predictions made by the model. Mathematically, it is expressed in Eq. (3)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{3}$$

  A higher accuracy score indicates an enhanced ability of the model to make accurate predictions. Nevertheless, a judicious evaluation necessitates consideration of additional metrics, especially when dealing with imbalanced datasets.
- **F1-score** combines precision and recall into a single metric, providing a balance between them. It's the harmonic mean of precision and recall and is particularly useful when there's an uneven class distribution or when the cost of false positives and false negatives is different. The F1-score ranges between 0 and 1, where 1 indicates perfect precision and recall, and 0 indicates poor performance. The F1-score is calculated with the formula in Eq. (4)

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

- **Standard deviation** is an important evaluation metric in machine learning as it measures the amount of variation or dispersion of a set of predictions. In the context of model performance, a low standard deviation indicates that the model's predictions are consistent and reliable across different datasets or iterations. Conversely, a high standard deviation suggests that the model's performance is highly variable, which may indicate issues with generalization or stability.
- **Confusion Matrix** is a tabular representation that affords a detailed examination of the performance of the models across individual classes within the dataset. It aids in pinpointing the classes where our model tends to misclassify instances. It provides insights into the extent of confusion and helps identify specific areas for improvement.
- **ROC Curve** stands as a powerful tool for the evaluation and comparison of classification models. It provides a graphical representation of a classifier's performance across various trade-offs between true positives and false positives. This curve is particularly valuable for assessing how well the classifier distinguishes between different classes.

These metrics and visualizations provide a multifaceted view of how effectively the models can predict the presence of mental disorders. End-to-end execution time for *testing* the proposed system is one minute and 13 secs which is 0.011 seconds per image on average. That includes both YOLOv8 and the ensemble model. All experiments and analyses are conducted using a workstation with the following characteristics: a 12th Gen Intel® Core™ i7-12700, 20 processors, 64.0 GB of memory, 1.5TB disk capacity, and NVIDIA GeForce RTX 3060 with GPU. Five-fold cross-validation is the method used in the model assessment to evaluate performance. Cross-validation, an unbiased estimator, helps to partially reduce overfitting or selection bias and improves the deep architecture transferability to an independent dataset. In cross-validation, training, and testing are conducted sequentially using a ratio of 80% to 20% of the whole dataset. The experimental results of this study are presented and discussed next.

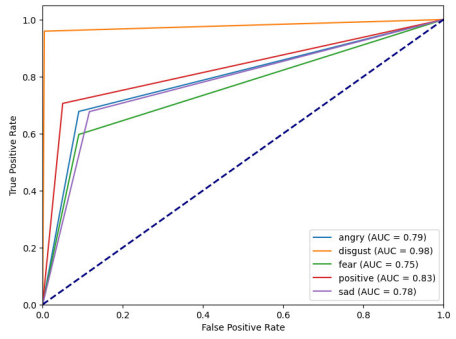
First, the performance of the YOLOv8 model on classifying the categories of emotions is examined and the results for the precision, recall, and F1-scores in each category are presented in Table 3 while the AUC-ROC curve and confusion matrix are showed in Figure 5. Additionally, in Table 4 the performance of the YOLOv8 is compared with existing state-of-the-art models trained on the AffectNet dataset.

TABLE 3. YOLOv8 performance summary.

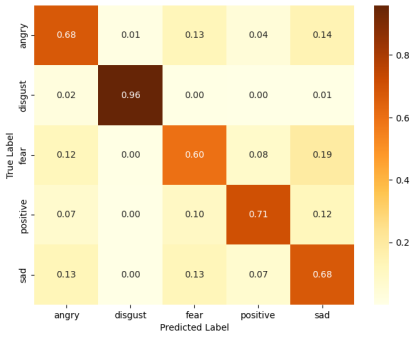
Class	Precision	Recall	F1-score	Standard Deviation
Angry	0.63	0.68	0.65	1.56
Disgust	0.98	0.96	0.97	0.17
Fear	0.62	0.60	0.61	1.09
Positive	0.76	0.71	0.73	0.77
Sad	0.65	0.68	0.66	1.32
Total	**0.72**	**0.73**	**0.72**	**1.29**

In the second set of experiments, the performance of the MDNet, pre-trained ResNet50, and ViT models is evaluated when trained individually on the mental disorder dataset. Table 5 shows the average evaluation scores of each model. Figures 6 drill further to the category-level performance with the Area Under the Curve (AUC) of the ROC (AUC-ROC) and confusion matrix of each model.

To further examine the important components of the ensemble model and the redundant parts, the study considered



(a)



(b)

FIGURE 5. YOLOv8 performance using (a) ROC curves and (b) confusion matrix for each disorder.

TABLE 4. Comparison of YOLOv8 against other techniques on the dataset.

Method	Accuracy (%)
gACNN [50]	58.78
IPA2LT [51]	55.71
RAN [52]	52.97
SCN [53]	60.23
DACL [54]	65.20
CNN [55]	56.54
POSTER [56]	67.31
KTN [57]	63.97
TransFER [58]	66.23
YOLOv8	72.00

TABLE 5. Performance summary of each module.

	MDNet	ResNet50	ViT
Precision	0.61	0.71	0.59
Recall	0.59	0.68	0.55
F1-Score	0.58	0.69	0.56
Accuracy	0.65	0.72	0.61
Standard Deviation	0.68	0.59	0.79

the use of ablation techniques. Ablation is a technique commonly used in machine learning and deep learning to understand the contributions of individual components or models within an ensemble. It involves systematically removing or disabling certain parts of the model to analyze their impact on performance. In this study, an ablation experiment was conducted to determine the contributions of the individual models to the ensemble model's performance. Three different ablation scenarios were explored. First, the contribution of the pre-trained ResNet50 and ViT models was examined while excluding the MDNet model. Secondly, we assessed the contribution of the MDNet and ViT models while leaving out the pre-trained ResNet50 model and finally, the contribution of the MDNet and pre-trained ResNet50 was examined while excluding the ViT model. The performance of the ensemble model is compared as shown in Table 6.

TABLE 6. Summary of the performance of the ensemble vs. ablated models.

	Precision	Recall	F1-Score	Accuracy	Standard Deviation
Ensemble model	0.76	0.75	0.75	0.78	0.53
MDNet ablated model	0.75	0.74	0.74	0.77	0.55
ResNet50 ablated model	0.59	0.55	0.55	0.62	0.62
ViT ablated model	0.81	0.79	0.80	0.81	0.55

Finally, to enhance the transparency and interpretability of the ensemble model's predictions, we considered GRAD-CAM and saliency maps to identify the region within the input images that exert the most significant influence on the model's prediction. The visualizations of the methods are presented in Figure 7.

V. DISCUSSION

Mental health has emerged as a global public health challenge, affecting diverse populations and manifesting in conditions such as depression and anxiety, often stemming from the pressures of everyday life. The timely identification and accurate diagnosis of mental disorders are paramount for effective intervention as delayed recognition may lead to severe consequences, including self-harm, suicidal tendencies, and tragic loss of life. In response to this pressing issue, recent advancements in computer vision and learning techniques offer a promising avenue for innovative solutions. The objective of this study is to develop a robust analysis pipeline for mental illness detection supported by a level of explainability and interpretability. The focus is on identifying and classifying visual cues associated with specific mental disorders, drawing upon the rich datasets of AffectNet and FER 2013 for training and validation. Consequently, this research contributed a system capable of generating a comprehensive mental disorder dataset and, concurrently, predicting mental disorders based on facial emotional cues.

A hybrid learning architecture is crafted that integrates, pre-trained architectures, ViT, and a shallow CNN model into an ensemble classifier. The collaborative efforts of these models demonstrate an overall accuracy of 78%. With an impressive 81% overall accuracy, the ablated ViT model surpasses other models in the subsequent investigation of ablation techniques, which breaks down the entire model under various scenarios.

With advancements in technology, we can trade off more parameters for higher accuracies. Although more models mean more parameters, the experiment from this study shows that improved computational resources enable more efficient learning, leading to better accuracy, as evidenced by the end-to-end testing time of 60 $\pm$ 0.13 seconds. Ensuring transparency and interpretability in the proposed solution is

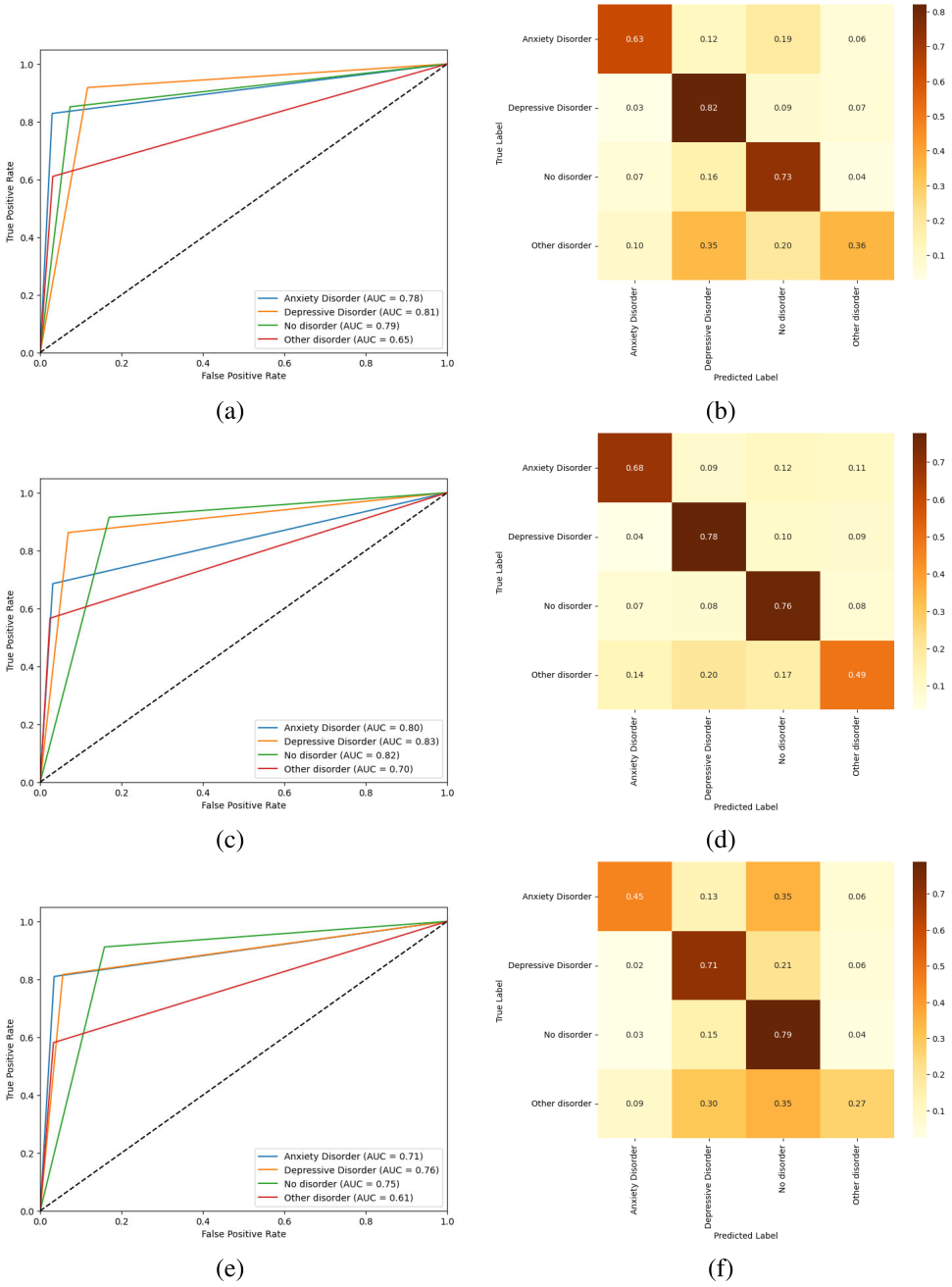


FIGURE 6. ROC curves (left column) and the confusion matrices (right column) of the three individual models MDNet (a,b); ResNet50 (c,d); and ViT (e,f).

fundamental. Therefore, this study incorporates techniques such as Grad-CAM and saliency maps. These techniques shed light on the regions within input images that significantly influence the model's predictions, see Fig. 7. By providing healthcare professionals with clear insights into the features guiding the system's decisions, this approach enhances trust and facilitates a more informed diagnostic process.

As shown in Table 3, the YOLOv8 model demonstrated the highest performance in the *Disgust* class, achieving an F1-score of 97%, a recall rate of 96%, and a precision rate of 98%. In the pursuit of generating a mental disorder dataset for this study, focus is placed on emotions such as *Angry*, *Disgust*, *Fear*, and *Sad*. Encouragingly, the YOLOv8 model consistently achieved F1 scores exceeding 60% across these critical emotion classes. This achievement underscores the model's potential to be a valuable tool in predicting mental disorders based on facial emotional cues.

In addition to quantitative metrics, robustness analysis is also highlighted using ROC curves. The ROC analyses are shown by the interconnected lines in Figures 5 and 6. The

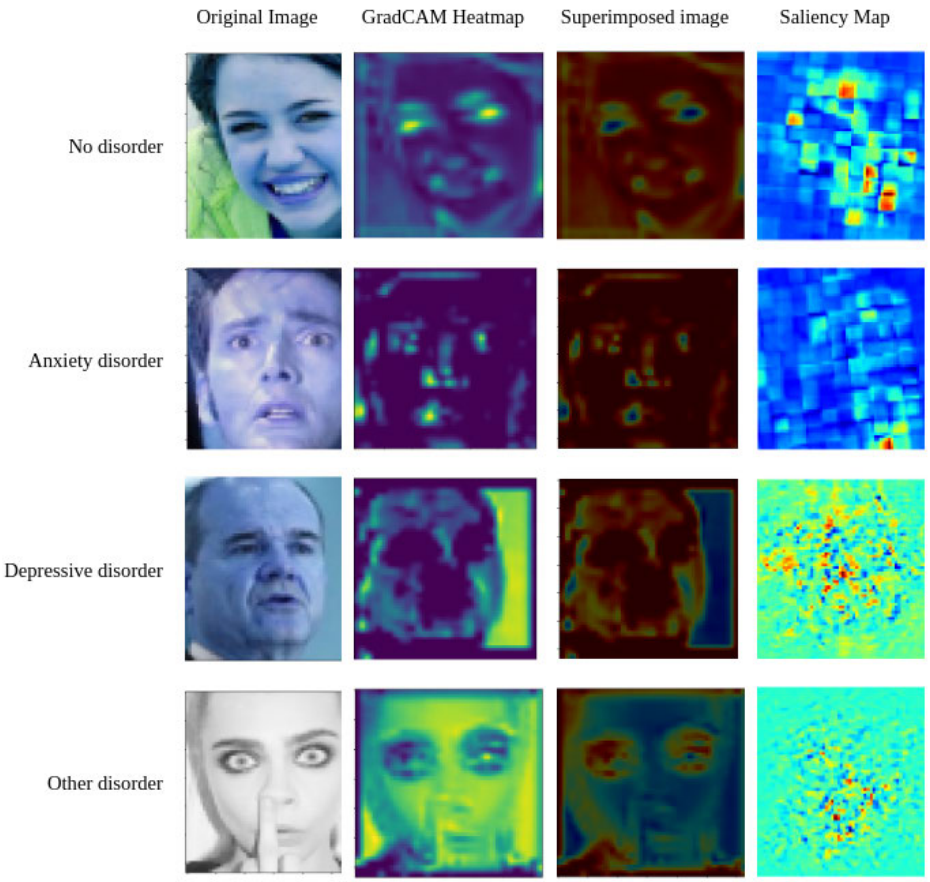


FIGURE 7. XAI Visualization of model's predictions for three samples–*from left-to-right*– Grad-CAM heatmap, superimposed image, and the saliency map [38].

random guess line is portrayed by the broken blue line that runs diagonally across the plot. Curves that fall below this line indicate poor performance of the model in identifying such class. The AUC of such curve quantitatively assesses the model's effectiveness in identifying a specific class with "1" and "0" indicating the best and the worst performance, respectively. Notably, the ROC curve for the *Disgust* class resides closest to the top-left corner (AUC is 98%), while the *Fear* class exhibits the greatest distance from this position (AUC is 75%). In essence, the ROC curve and AUC scores show that the model performed above average in correctly classifying each facial emotion. In Figure 5-b, the confusion matrix further demonstrates the ability of the model to correctly predict 68%, 96%, 60%, 71% and 68% of the *Angry*, *Disgust*, *Fear*, *Positive*, and *Sad* classes respectively.

The pre-trained ResNet50 model performed best in predicting all classes, according to the ROC curves in Figure 6 (a, c, e). This is because its curves are more tilted toward the top left corner of the curve than the ViT ROC curves, which are the furthest from the top left corner among the three models. The AUC values also reveal that the pre-trained ResNet50 model has the best performance on classifying all classes, The ViT model performed worst on the Anxiety disorder, Depressive disorder, and No disorder classes with values of 0.71, 0.76, 0.75, and 0.61 respectively. The MDNet model also had an AUC value of 0.65 for the other disorder predictions. The confusion matrix in Figure 6 (b,d,f) further evaluates the performance of the three models as it shows the proportion of instances each model predicted correctly. The true labels are presented on the y-axis while the predicted labels are presented on the x-axis. The diagonal numbers represent the percentage of images in each class that were correctly predicted. It is observed that the MDNet and ViT models were able to classify over 70% of the Depressive disorder and no disorder classes, over 45% of the anxiety disorder class correctly but had a low performance of classifying less than 35% of the other disorder class correctly. The pre-trained ResNet50 model had the highest performance as it could classify over 50% of the instances correctly.

Moreover, the performance of the ResNet50 backbone as a feature extractor has been investigated. Namely, its performance has been compared with some other well-known off-the-shelf pre-trained models (MobileNetv2, Inceptionv3, and Xception). The result showed that the pre-trained models had close performance (MobileNetv2: 62%, Inceptionv3:

62%,, and Xception: 72%,) and ResNet was chosen due to its famous ability to capture more complex features and well as its ability to mitigate the vanishing gradient problem. The ResNet50 model is also generally considered to be faster and more lightweight than the Xception model. Additionally, Table 5 showed that the best-performing model out of the three models is the pre-trained ResNet50 model with an accuracy score of 72%, average precision, recall, and F1-score of 71%, 68%, and 69%, respectively. The model that performed least was the ViT model with an accuracy of 61%, average precision, recall, and F1-score of 59%, 55% and 56%, respectively.

The ensemble and ablated models' evaluation metrics are summarized in Table 6. The MDNet ablated model results are the performance of the model without the MDNet model. The pre-trained ResNet50 and ViT ablated models' results show the performance of the ensemble model without the pre-trained ResNet50 and ViT models, respectively. Table 6 showed that the ensemble model without the ViT model had the highest accuracy of 81%. The ensemble model of the three models had an accuracy of 78% while the lowest performing scenario was when the pretrained ResNet50 model was excluded from the ensemble model. This underscores the great importance of the ResNet50 model in the ensemble. Though the pretrained ResNet50 model when trained alone had a lower accuracy. However, when fused with the layers of the MDNet, achieved better performance on predicting the presence of mental disorders from images. The performance also reveals the poor performance of the ViT model when trained individually and when fused into an ensemble. To this end, the model adopted in the study is an ensemble of the ResNet50 model and the MDNet.

As visualized in Figure 7, the first column shows samples from the distinct classes: no disorder, anxiety, depressive disorder, and other disorders. Next to these images lies the Grad-CAM heatmap, a visual representation that illuminates the portion of the input image that wielded the most substantial influence over the ensemble model's prediction. These influential areas are indicated by heightened intensity, denoted by vibrant green hues. Upon closer examination of the heatmap, it becomes evident that in instances representing no disorder and anxiety disorder, crucial facial features such as the eyes, segments of the nose, and mouth significantly influenced the model's prediction. Conversely, in the sample representing depressive disorder and other disorders, regions beyond the eyes and nose played a pivotal role in determining the model's performance. The third column of images are the superimposed images, formed by overlaying the heatmap over the original input image. The resulting image accentuates the regions that significantly influenced the model's prediction. These critical areas are distinctly more emphasized compared to surrounding regions, offering a clear visual indication of their importance in the prediction. The last column contains the saliency maps that are also instrumental in revealing the areas within the images that played a pivotal role in influencing the model's decision. They emphasize the most influential regions using warm colors, effectively directing attention to these critical areas. These warm colors serve as a visual indicator, making it easier to discern which specific regions within the images exerted the most substantial influence on the model's decision-making process. This comprehensive visualization technique not only provides insight into the specific features that guided the model's predictions but also enhances the transparency and interpretability of the ensemble model's decision-making process.

By leveraging the YOLOv8 object detection algorithm and utilizing datasets like AffectNet and FER 2013, the study goes beyond traditional diagnostic methods of mental illness. The development of a mental disorder dataset and the implementation of a hybrid learning architecture, combining CNN and ViT models, demonstrate a sophisticated ensemble classifier model. This model exhibits a notable improvement in overall accuracy, reaching around 81%, showcasing the efficacy of the proposed methodology. Prior research endeavors in the area of utilizing facial expressions for the detection of mental disorders have primarily concentrated on specific categories of mental illnesses such as depression [7], [8], [9], [25], [26], [27], ADHD, ASD and Dyslexia [28], [30], [31]. The study in [5] delved into the assessment of anxiety, depression, other disorders, and normal states leveraging the FER 2013 dataset. Given the limitations associated with the small size and class imbalance of the FER 2013 dataset, the present study introduces a novel pipeline for the analysis of facial expressions, leveraging both the AffectNet and FER 2013 datasets. Moreover, the incorporation of interpretability techniques such as Grad-CAM and saliency maps enhances transparency by providing healthcare professionals with insights into the specific regions in input images that contribute significantly to the model's predictions. This transparency not only improves trust in the system but also facilitates a more informed diagnostic process.

While the proposed system demonstrates promising outcomes, several limitations warrant consideration. Firstly, the facial emotion datasets utilized in the study are affected by class imbalance. Consequently, not all images within the dataset could be effectively utilized, significantly impacting the overall performance of the model. To enhance the system's accuracy, it is imperative to explore larger and more balanced datasets.

Although ensemble models result in a highly complex system that requires significant computational power and memory, both during training and inference, technological advancements in the use of GPU-powered machines have made training and inference faster and more efficient. This enables a trade-off between computational complexity and accuracy, allowing for more sophisticated models without prohibitive resource constraints. Despite this, the increased complexity still demands careful management to optimize performance and prevent overfitting.

Secondly, the study exclusively concentrates on facial cues for mental disorder diagnosis. To fortify the system's

robustness, it is crucial to incorporate other elements, such as verbal communication and contextual information. Integrating multimodal data sources can provide a more comprehensive understanding of the individual's mental state, improving the system's predictive capabilities.

Lastly, the research specifically targets anxiety and depressive disorders. However, there are other specific types of mental disorders that are identifiable through facial cues. Expanding the scope of the study to include a broader range of mental health conditions can enhance the applicability and usefulness of the system.

In summary, while the proposed system shows potential, addressing the limitations related to data imbalance, the exclusive reliance on facial cues, and the narrow focus on certain mental disorders is essential for improving its accuracy, robustness, and generalizability. The advancements in GPU technology offer a pathway to manage the computational demands of such a complex ensemble model, making it feasible to develop more accurate and efficient systems.

VI. CONCLUSION AND FUTURE WORK

In this research, we harnessed the cutting-edge capabilities of artificial intelligence tools to develop an advanced, transparent, and interpretable mental disorder detection system. The primary aim was to address the critical need for early detection and timely intervention in the field of mental health. This is achieved through the fusion of potent computer vision and deep learning techniques. The YOLOv8 model was used to identify facial cues attached to various emotions. An innovative logic was applied to the model's predictions to generate the mental disorder dataset. Cutting-edge CNN-based models namely MDNet, ResNet50, and ViT were trained and evaluated on the mental disorder dataset. Various ensemble and ablation techniques were also considered. The experimental results showed that the ensemble of pre-trained ResNet50 and the MDNet model achieved the best predictive performance with an accuracy of 81%. This underscores the importance of the residual layers present in the pre-trained ResNet50 architecture in image classification, particularly mental disorder detection. The integration of XAI techniques, specifically GRAD-CAM and saliency maps, also provided valuable insights into the model's decision rationale. These visualizations highlighted the critical image regions that influenced predictions, enabling healthcare professionals to make informed diagnostic decisions and personalized treatment plans.

To achieve real-time mental disorder detection system, future studies should consider expanding from image-based detection to video analysis to provide more comprehensive insights into an individual's mental health. Developing real-time mental disorder detection systems that continuously analyze a person's behavior and provide immediate feedback or alerts when signs of distress or disorder are detected should also be considered. Future works should also consider comparing transformer-based backbones for feature extraction, fusion classifiers, different saliency maps, etc. Additionally, integrating other components of mental disorder detection, such as clinical reports, emotions over time, and contextual information, will be investigated.

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

REFERENCES

[1] J. Aina, "Mental disorder detection system through emotion recognition," M.S. thesis, Dept. Comput. Sci., Morgan State Univ., Baltimore, MD, USA, Oct. 2023.

[2] L. Duszynski-Goodman and L. Henderson. (2024). *Mental Health Statistics and Facts in 2024*. Accessed: Jul. 21, 2023. [Online]. Available: https://www.forbes.com/health/mind/mental-health-statistics/

[3] D. F. Santomauro, A. M. M. Herrera, J. Shadid, P. Zheng, C. Ashbaugh, D. M. Pigott, C. Abbafati, C. Adolph, J. O. Amlag, and A. Y. Aravkin, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2020.

[4] G. N. Foley and J. P. Gentile, "Nonverbal communication in psychotherapy," *Psychiatry (Edgmont)*, vol. 7, no. 6, p. 38, 2010.

[5] S. A. Hussein, A. E. R. S. Bayoumi, and A. M. Soliman, "Automated detection of human mental disorder," *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, pp. 1–10, Feb. 2023.

[6] Y.-H. Chuang, C.-H. Tan, H.-C. Su, C.-Y. Chien, P.-S. Sung, T.-L. Lee, and R.-L. Yu, "Hypomimia may influence the facial emotion recognition ability in patients with Parkinson's disease," *J. Parkinson's Disease*, vol. 12, no. 1, pp. 185–197, Jan. 2022.

[7] D. Choi, G. Zhang, S. Shin, and J. Jung, "Decision tree algorithm for depression diagnosis from facial images," in *Proc. IEEE 2nd Int. Conf. AI Cybersecurity (ICAIC)*, Feb. 2023, pp. 1–4.

[8] S. T. Mantri, D. D. Patil, P. Agrawal, and V. Wadhai, "Real time multimodal depression analysis," *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, no. 9, pp. 1–7, 2019.

[9] J. H. Majed, S. A. Nasser, A. Alkhayyat, and I. A. Hashim, "Artificial intelligent algorithms based depression detection system," in *Proc. 5th Int. Conf. Eng. Technol. Appl. (IICETA)*, May 2022, pp. 408–413.

[10] B. Alankar, M. S. Ammar, and H. Kaur, "Facial emotion detection using deep learning and Haar cascade face identification algorithm," in *Proc. Advances in Intelligent Computing and Communication: Proceedings of ICAC 2020*. Cham, Switzerland: Springer, 2020, pp. 163–180.

[11] A. Dhillon and G. K. Verma, "Convolutional neural network: A review of models, methodologies and applications to object detection," *Prog. Artif. Intell.*, vol. 9, no. 2, pp. 85–112, Jun. 2020.

[12] S. V. Vasantha and M. D. Ayaz, "Emotion detection using facial image for behavioral analysis," in *Proc. Int. Conf. Futuristic Technol. (INCOFT)*, Nov. 2022, pp. 1–7.

[13] D. Naveen, P. Rachana, S. Swetha, and S. Sarvashni, "Mental health monitor using facial recognition," in *Proc. 2nd Int. Conf. Innov. Technol. (INOCON)*, Mar. 2023, pp. 1–3.

[14] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.

[15] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and Dirichlet process Fisher encoding," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1476–1486, Jun. 2019.

[16] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 3–10.

[17] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, Nov. 2014, pp. 3–10.
[18] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, and H. Zhou, "Deep convolution network based emotion analysis towards mental health care," *Neurocomputing*, vol. 388, pp. 212–227, May 2020.
[19] S. Harati, A. Crowell, Y. Huang, H. Mayberg, and S. Nemati, "Classifying depression severity in recovery from major depressive disorder via dynamic facial features," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 3, pp. 815–824, Mar. 2020.
[20] H. Hadjar, J. Lange, B. Vu, F. Engel, G. Mayer, P. Mc Kevitt, and M. L. Hemmje, "Video-based automated emotional monitoring in mental health care supported by a generic patient data management system," in *Proc. 2nd Symp. Psychol.-Based Technol.*, Naples, Italy, Sep. 2020.
[21] M. Munsif, M. Ullah, B. Ahmad, M. Sajjad, and F. A. Cheikh, "Monitoring neurological disorder patients via deep learning based facial expressions analysis," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, in IFIP Advances in Information and Communication Technology, vol. 652, 2022, pp. 412–423.
[22] P. M. Swamy, P. J. Kurapothula, S. V. Murthy, S. Harini, R. RaviKumar, and K. Kashyap, "Voice assistant and facial analysis based approach to screen test clinical depression," in *Proc. 1st Int. Conf. Adv. Inf. Technol. (ICAIT)*, Jul. 2019, pp. 39–44.
[23] M. Tadalagi and A. M. Joshi, "AutoDep: Automatic depression detection using facial expressions based on linear binary pattern descriptor," *Med. Biol. Eng. Comput.*, vol. 59, no. 6, pp. 1339–1354, Jun. 2021.
[24] J. Singh and G. Goyal, "Decoding depressive disorder using computer vision," *Multimedia Tools Appl.*, vol. 80, no. 6, pp. 8189–8212, Mar. 2021.
[25] X. Kong, Y. Yao, C. Wang, Y. Wang, J. Teng, and X. Qi, "Automatic identification of depression using facial images with deep convolutional neural network," *Med. Sci. Monitor*, vol. 28, Jun. 2022, Art. no. e936409.
[26] G. Gilanie, M. ul Hassan, M. Asghar, A. M. Qamar, H. Ullah, R. U. Khan, N. Aslam, and I. Ullah Khan, "An automated and real-time approach of depression detection from facial micro-expressions," *Comput., Mater. Continua*, vol. 73, no. 2, pp. 2513–2528, 2022.
[27] Y.-S. Lee and W.-H. Park, "Diagnosis of depressive disorder model on facial expression based on fast R-CNN," *Diagnostics*, vol. 12, no. 2, p. 317, Jan. 2022.
[28] E. Sadek, N. A. Seada, and S. Ghoniemy, "Computer vision techniques for autism symptoms detection and recognition: A survey," *Int. J. Intell. Comput. Inf. Sci.*, vol. 20, no. 2, pp. 89–111, Dec. 2020.
[29] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Detect depression from communication: How computer vision, signal processing, and sentiment analysis join forces," *IISE Trans. Healthcare Syst. Eng.*, vol. 8, no. 3, pp. 196–208, Jul. 2018.
[30] M. Nilsson Benfatto, G. Öqvist Seimyr, J. Ygge, T. Pansell, A. Rydberg, and C. Jacobson, "Screening for dyslexia using eye tracking during reading," *PLoS ONE*, vol. 11, no. 12, Dec. 2016, Art. no. e0165508.
[31] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley, "Automatic detection of ADHD and ASD from expressive behaviour in RGBD data," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 762–769.
[32] N. Shanthi, A. A. Stonier, A. Sherine, T. Devaraju, S. Abinash, R. Ajay, V. A. Prasath, and V. Ganji, "An integrated approach for mental health assessment using emotion analysis and scales," *Healthcare Technol. Lett.*, pp. 1–11, Dec. 2022, doi: 10.1049/htl2.12040.
[33] D. B. Yaden, J. C. Eichstaedt, M. L. Kern, L. K. Smith, A. Buffone, D. J. Stillwell, M. Kosinski, L. H. Ungar, M. E. P. Seligman, and H. A. Schwartz, "The language of religious affiliation: Social, emotional, and cognitive differences," *Social Psychol. Personality Sci.*, vol. 9, no. 4, pp. 444–452, May 2018.
[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.
[35] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.
[36] M. L. Birnbaum, A. Abrami, S. Heisig, A. Ali, E. Arenare, C. Agurto, N. Lu, J. M. Kane, and G. Cecchi, "Acoustic and facial features from clinical interviews for machine learning–based psychiatric diagnosis: Algorithm development," *JMIR Mental Health*, vol. 9, no. 1, Jan. 2022, Art. no. e24699.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.
[38] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
[39] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.
[40] W. C. Y. B. Dumitru and Ian Goodfellow. (2013). *Challenges in Representation Learning: Facial Expression Recognition Challenge*. [Online]. Available: https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge
[41] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[42] A. Mehra. (2023). *Understanding YOLOv8 Architecture, Applications and Features.* Accessed: Jun. 5, 2023. [Online]. Available: https://www.labellerr.com/blog/understanding-yolov8-architecture-applications-features
[43] C. Clinic. (9536). *Anxiety Disorders*. Accessed: Jul. 21, 2023. [Online]. Available: https://my.clevelandclinic.org/health/diseases/9536-anxiety-disorders
[44] C. Sawchuk. (2023). *Depression (Major Depressive Disorder) Symptoms and Causes*. Accessed: Jul. 21, 2023. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[46] Y. Lu, J. Fu, X. Li, W. Zhou, S. Liu, X. Zhang, W. Wu, C. Jia, Y. Liu, and Z. Chen, "RTN: Reinforced transformer network for coronary ct angiography vessel-level image quality assessment," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2022, pp. 644–653.
[47] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, and J. Li, "Transformer-based dual relation graph for multi-label image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 163–172.
[48] H.-Y. Zhou, Y. Yu, C. Wang, S. Zhang, Y. Gao, J. Pan, J. Shao, G. Lu, K. Zhang, and W. Li, "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," *Nature Biomed. Eng.*, vol. 7, no. 6, pp. 743–755, Jun. 2023.
[49] S. M. LaValle, M. S. Branicky, and S. R. Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *Int. J. Robot. Res.*, vol. 23, nos. 7–8, pp. 673–692, Aug. 2004.
[50] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[51] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 222–237.
[52] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[53] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6896–6905.
[54] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2401–2410.
[55] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, and H.-C. Hsu, "A study on computer vision for facial emotion recognition," *Sci. Rep.*, vol. 13, no. 1, p. 8425, May 2023.
[56] C. Zheng, M. Mendieta, and C. Chen, "POSTER: A pyramid cross-fusion transformer network for facial expression recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 3146–3155.
[57] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via C-F labels and distillation," *IEEE Trans. Image Process.*, vol. 30, pp. 2016–2028, 2021.
[58] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3581–3590.

JOSEPH AINA (Student Member, IEEE) received the B.Sc. degree in economics from the University of Ilorin, Ilorin, Nigeria, in 2019, and the M.Sc. degree in computer science from Morgan State University, USA, in 2023, where he is currently pursuing the Ph.D. degree in computer and electrical systems engineering. He is a Graduate Research Assistant with the Transportation Center, Morgan State University. He has more than five years of hands-on experience in the fields of data analytics, image processing, machine learning, medical image analysis, and computer vision. He has authored/co-authored about four peer-reviewed publications appearing in high-impact journals and selective peer-reviewed top-rank international conferences.

OLUWATUNMISE AKINNIYI received the B.Sc. degree in computer science from The Federal University of Technology Akure, Nigeria, and the M.Sc. degree in computer science from Morgan State University, USA, in 2023, where she is currently pursuing the Ph.D. degree in Computer and Electrical Systems Engineering. She is a Graduate Research Assistant with the Center for Equitable Artificial Intelligence and Machine Learning Systems, Morgan State University. She has more than five years of hands-on experience in the fields of data analytics, image processing, machine learning, medical image analysis, and computer-aided diagnosis. She has authored/co-authored about five peer-reviewed publications appearing in high-impact journals and selective peer-reviewed top-rank international conferences.

MD. MAHMUDUR RAHMAN (Member, IEEE) received the Ph.D. degree in computer science from Concordia University, Montreal, Canada, in 2008, with an emphasis on medical imaging informatics. He is currently an Associate Professor and an Associate Chair with the Computer Science Department, Morgan State University. He has good expertise in the fields of data science, AI and machine learning, image processing, and computer vision, and their application to classification, annotation, and retrieval of biomedical images from large collections. He has published three books, three book chapters, and around 100 papers in peer-reviewed journals and conference/workshop proceedings. He also received several competitive grants from NSF for both imaging informatics and applied machine learning-based research and CS educational research. His current research interests include deep learning and multimodal language modeling techniques and their application in medical fields, especially for automatic caption and concept generation of medical images, classification, visual question answering (VQA), image-based decision support systems (DSS), and eXplainable AI (XAI).

VALERIE ODERO-MARAH received the Ph.D. degree from The University of Iowa, under the mentorship of Dr. Mary Hendrix, with a focus on breast cancer. She did her postdoctoral training in prostate cancer under the mentorship of Dr. Leland Chung at Emory University. From 2021 to 2024, she was the Executive Director of the Center for Health Disparities Research and Innovation. She is currently a Professor of biology with Morgan State University. At Morgan State University, she is entrusted to build the cancer health disparities research portfolio and contribute also to mentoring junior faculty. Her current research focuses on the role of epithelial-mesenchymal transition (EMT) in prostate and breast health disparities. Internationally, she is the Co-Leader of the Prostate Cancer Working Group, African Caribbean Cancer Consortium (AC3), an organization made up of members within the USA, Africa, and the Caribbean, to study viral, genetic, environmental, and lifestyle factors for cancer risk and outcomes in populations of African descent.

FAHMI KHALIFA (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electronics and electrical communication engineering from Mansoura University, Egypt, in 2003 and 2007, respectively, and the Ph.D. degree in electrical engineering from the Electrical and Computer Engineering Department, University of Louisville, USA, in 2014. He has more than 15 years of hands-on experience in the fields of image processing, machine learning, medical image analysis, computer-aided diagnosis, and digital and analog signal processing. He has authored/co-authored more than 190 peer-reviewed publications appearing in high-impact journals, selective peer-reviewed top-rank international conferences, and leading edited books. He honors and awards include Mansoura University scholarship for distinctive undergraduate students for four consecutive years; Theobald Scholarship Award in 2013 (ECE, UofL); the ECE Outstanding Student award for two times in 2012 and 2014 (ECE, UofL); John M. Houchens award for the outstanding dissertation (UofL); the second-place Post-Doctoral Fellow award in 2014 Research! Louisville, UofL. He was nominated for the ''Faculty Favorite Recognition'' at the Speed School of Engineering, UofL, 2018–2019. He was the recipient of the PowerLIVE Award for Faculty commitment to students and their academic success; and on the final list for the ''Instructional Innovator of the Year,'' Morgan State University, in 2023.