Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2024.057047

ARTICLE

A Deep Learning-Based Automated Approach of Schizophrenia Detection from Facial Micro-Expressions

Anum Saher[1], Ghulam Gilanie[1,*], Sana Cheema[1], Akkasha Latif[1], Syeda Naila Batool[1] and Hafeez Ullah[2]

[1]Department of Artificial Intelligence, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan

[2]Biophotonics Imaging Techniques Laboratory, Institute of Physics, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan

*Corresponding Author: Ghulam Gilanie. Email: ghulam.gilanie@iub.edu.pk

Received: 06 August 2024 Accepted: 22 November 2024 Published: 30 December 2024

ABSTRACT

Schizophrenia is a severe mental illness responsible for many of the world's disabilities. It significantly impacts human society; thus, rapid, and efficient identification is required. This research aims to diagnose schizophrenia directly from a high-resolution camera, which can capture the subtle micro facial expressions that are difficult to spot with the help of the naked eye. In a clinical study by a team of experts at Bahawal Victoria Hospital (BVH), Bahawalpur, Pakistan, there were 300 people with schizophrenia and 299 healthy subjects. Videos of these participants have been captured and converted into their frames using the OpenFace tool. Additionally, pose, gaze, Action Units (AUs), and land-marked features have been extracted in the Comma Separated Values (CSV) file. Aligned faces have been used to detect schizophrenia by the proposed and the pre-trained Convolutional Neural Network (CNN) models, i.e., VGG16, Mobile Net, Efficient Net, Google Net, and ResNet50. Moreover, Vision transformer, Swim transformer, big transformer, and vision transformer without attention have also been used to train the models on customized dataset. CSV files have been used to train a model using logistic regression, decision trees, random forest, gradient boosting, and support vector machine classifiers. Moreover, the parameters of the proposed CNN architecture have been optimized using the Particle Swarm Optimization algorithm. The experimental results showed a validation accuracy of 99.6% for the proposed CNN model. The results demonstrated that the reported method is superior to the previous methodologies. The model can be deployed in a real-time environment.

KEYWORDS

Schizophrenia; deep learning; machine learning; facial expressions; transformers; particle swarm optimization (PSO) algorithm

1 Introduction

Mental problems often exhibit abnormal patterns of thought, perception, and behavior. When the illness reaches a severe stage, those with mental illness have difficulty keeping their minds in the real world and cannot function effectively in daily life [1]. Mental disorders include schizophrenia,

bipolar, anxiety, autism spectrum disorder, post-traumatic stress disorder, neurodevelopment disorder, mood disorder, and anti-social personality disorder. The good news is that these conditions have been easily identified and even cured. Out of these mental health problems, schizophrenia ranks high [2]. Schizophrenia is a severe mental illness with an occurrence rate of about 0.48% over a person's lifespan. This disorder affects approximately 24 million people, representing about 0.32% of the global population. Among individuals, this translates to a rate of 0.45%, or 1 in 222 people [3]. Several developed countries and many emerging economies have a particularly significant problem with mental health. According to the World Health Organization (WHO), one in four persons experiences mental illness throughout their lifetimes. Indonesia has the highest rate of people who have schizophrenia globally. In 2023, around 1.5% of the whole population of Pakistan is affected by schizophrenia.

Furthermore, the situation worsens because three-quarters of persons with severe mental problems do not get treatment. Schizophrenia causes psychotic symptoms and is linked to many disabilities. It can affect all parts of a person's life, such as their personal, family, social, educational, and work life. People with schizophrenia often face shame, discrimination, and human rights violations. It may be responsible for insomnia, regular sleeping patterns, food routines, and child-rearing methods. Mainly, 900,000 people every year commit suicide globally, many of them suffering from severe mental problems [4]. Different types of schizophrenic disorders, i.e., psychotic disorder, schizoaffective disorder, schizophreniform disorder, and catatonia disorder shown in Fig. 1 are categorized clinically.

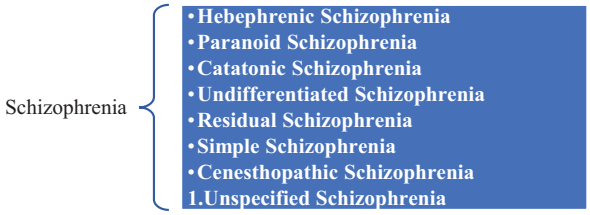


Figure 1: Types of schizophrenia

In recent years, various models have been developed to detect mental illnesses. CNN models are commonly used for recognizing facial expressions. The 2D-CNN model analyzes static images to detect features such as eye movement and lip tightening. However, these static models often struggle to capture the subtle temporal dynamics of micro-expressions, which are crucial for schizophrenia detection. To overcome this limitation, researchers are increasingly employing spatiotemporal models. 3D CNNs are particularly effective because they analyze both the timing and spatial features of facial movements in video sequences, allowing for the detection of subtle facial movements related to schizophrenia that may not be visible in static images. By utilizing brief video recordings, 3D CNNs can learn about facial landmarks and the timing of muscle movements.

We are currently developing a hybrid model that integrates deep learning and machine learning techniques, enhanced by optimization algorithms for improved performance. These methods have not been previously applied to the detection of schizophrenia or other mental disorders. But research on schizophrenia using hybrid model is still limited [5,6].

1.1 Related Work

There are two diagnostic methods to detect schizophrenia i.e., clinical, and automated. The psychiatrist used the clinical method of treatment. It consists of the Clinical Global Impression-Schizophrenia (CGI-SCH) scale [7] and clinical decisions about patients made in light of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V). In recent years, electroencephalograms [8] and eye tracking have also been used to diagnose schizophrenia [9,10]. Current trends in the scientific literature have shown that mental illness can automatically be diagnosed using visual cues. Expressions on the face are classified as either "micro" (lasting only a few hundredths of a second) or "macro" (lasting from one to four seconds) [11]. Various mental health conditions have facial micro-expressions that embody biomarkers. People with severe mental problems often die if not diagnosed and treated [12]. Nearly 800,000 people each year kill themselves just due to schizophrenia. Medicine and psychiatric therapy are the mainstays in treating mental health diseases [13]. Charles Darwin argues that the appropriate coordination of facial muscles can convey an emotion. Paul suggested a Facial Action Coding System (FACS) that assigns numeric values to each of the many possible facial emotions a human can make. The seven basic emotions are just a subset of the 44 Action Units (AUs) in the FACS. Muscle contractions and facial relaxations include tight lips in anger, raised eyebrows, or changing eyelid movements. Additive AUs have no noticeable visual effect; non-additive AUs have a prominent visual effect. Therefore, in the proposed study, we analyze the facial expressions of patients and healthy controls from videos to identify schizophrenia [14].

In this article [15], seven facial expressions focused on schizophrenia have been detected through a mobile app. Twenty-one schizophrenia patients were interviewed using clinical outcome measures. In this research, an International Affective Picture System (IAPS) was used to elicit an emotional response; happiness and sadness were given better frequency than other expressions. The research [16] proposed a deep transfer learning technique of face recognition for computer-aided facial diagnosis. The experiments consisted of computer-aided face diagnosis on single (beta-thalassemia) and multiple diseases (beta-thalassemia, hyperthyroidism, Down syndrome, and leprosy) with a limited dataset. Deep transfer learning from face recognition achieved top-1 accuracy of more than 90%, outperforming existing machine learning approaches and doctors in the studies. This research article [17] measured the intensities of a spontaneous and posed smile focused on the 68 muscle moments of schizophrenia patients. One hundred fifteen schizophrenia patients and 57 average persons participated in this study. This research analyzed the dataset using an open-face toolkit and extracted it from the Convolutional Experts' Constrained Local Model (CE-CLM). The study concluded that upper-face spontaneous smiles are more vital, and the posed smile had higher lower-face intensities. The author of [18] identified schizophrenia from video data and detected frames using Multiple Task Cascaded Neural Networks (MTCNN). The dataset was collected through a Positive and Negative Syndrome Scale (PANSS) scale of 22 patients with schizophrenia. The Speeded Up Robust Features (SURF) algorithm was used for feature extraction and classification by the method of transfer learning using pre-trained deep learning models. The achieved accuracy was 80%. In the research [19], the authors proposed automated schizophrenia detection using facial expressions of 125 schizophrenia patients and 75 average persons through video data. This work was presented using Restnet18 with an accuracy of 89%.

The authors of an article [20], presented FaceDisNet, a computer-aided face diagnosis system that employs a public dataset of photos taken in an unconstrained context and might be used for future comparisons. It employs two feature selection strategies to decrease the huge dimension of features generated by feature fusion. Finally, it creates an ensemble classifier using stacking to do classification. The performance of FaceDisNet demonstrates its capacity to identify both single and

many disorders. Following the ensemble classification and feature selection procedures for binary and multiclass classification categories, FaceDisNet achieved a maximum accuracy of 98.57% and 98%, respectively.

In the research [21], conducted for diagnosis of Parkinson's Disease (PD) using facial expression recognition, the authors collected films of persons with PD and matched them to controls. They employed relative coordinates and positional jitter to extract facial expression features. PD was diagnosed using algorithms from both classic machine learning and advanced deep learning. In the research [20], the authors suggested a model Cross-Modality Graph Convolutional Network (CMGCN) to estimate schizophrenia from Context-Aware Emotion Recognition (CAER) and Audiovisual Emotion Challenge (AVEC) video datasets. The mean average precision of the whole system was 69.52. The CMGCN achieved an accuracy of CAER 87%, and the Root Mean Square Error (RMSE) of AVEC was 8.50. The authors of this research [22] suggested a paradigm for estimating schizophrenia using the facial expression of cannabis patients' data. The dataset comprises images belonging to 2039 schizophrenia patients and 2049 normal. Assessment of the facial expression detection system by Degraded Facial Affect Recognition task (DFAR) obtained results based on angry, fearful, neutral, and happy; the overall confidence interval was 95%.

The authors of another article [23] proposed a paradigm for automatically estimating schizophrenia using facial expressions. The locally developed database was used for experiments. According to the Bayesian classifier and Analysis of Variance (ANOVA), the scaled probability was 0.047, and the variance was 0.05. The authors of this study [24] suggested a method that employs images to diagnose schizophrenia from eye movement. The patient's mental state was accessed using a gradient-boosted decision tree model. Similarly, predictive analysis was performed in R language through ANOVA and chi-squared test. The model was trained using local data from 60 schizophrenia patients and 104 healthy control datasets. The dataset was obtained by using the Psychosocial Risk Scale (PRS). The accuracy of the model was 88%.

In this study, we present a lightweight convolutional neural network (CNN) model for schizophrenia detection, achieving an impressive accuracy of 99.6%. Unlike previous studies that often rely on smaller datasets, our approach utilizes a significantly larger dataset. Moreover, while many prior studies focus on one or two features such as head pose, action units, eye gaze, and facial landmarks but our research takes a more comprehensive approach by integrating multiple features. Additionally, while many previous studies rely on either deep learning or machine learning methodologies, our research employs a hybrid approach that incorporates both techniques. This allows us to not only utilize the strengths of each method but also conduct a comparative analysis of their performance. By addressing these aspects, we aim to provide a more comprehensive understanding of how our proposed model overcomes the limitations of existing methods.

1.2 Motivation

Schizophrenia is a severe mental illness that significantly impacts society, making the need for rapid and efficient identification crucial for early intervention and treatment. This study addresses this need by leveraging deep learning techniques to detect subtle facial micro-expressions associated with schizophrenia, which are often challenging to detect with the naked eye. By combining high-resolution cameras with powerful Convolutional Neural Network (CNN) models, the research presents an innovative strategy to enhance diagnostic accuracy and potentially allow for earlier intervention and treatment in individuals with schizophrenia.

1.3 Contribution of the Study

In recent years, artificial intelligence methods have been employed to analyze facial expressions for diagnosing schizophrenia. Computer vision facilitates the recognition of subtle facial micro-expressions, enabling computer-aided diagnostic systems to screen individuals without hesitation or obstructions. Micro expressions are short and difficult to identify. The major distinction between micro and normal expressions is their duration and appearance. In our research, we used action units to detect micro expressions since they provide important signs that aid in their identification. Furthermore, we recognize that high frame rate video recording improves the detection of micro expressions by capturing the quick facial movements that distinguish these tiny expressions. Both psychiatrists and patients benefit financially and in terms of time from these systems. In the proposed study, the dataset was obtained from Bahawal Victoria Hospital (BVH), Bahawalpur, and Pakistan. Experiments were conducted on a standard scale for schizophrenia diagnosis to infer patients' mental health status using a high-resolution camera that captures expressions that are difficult to detect with the naked eye. We analyzed the action units of facial expressions from both patients and healthy controls in video to identify schizophrenia. The proposed, pre-trained (VGG16, Mobile Net, Efficient Net, Google Net, and ResNet50), and transformers models including Vision Transformer (ViT), Swin Transformers, and Big Transformer (BiT), used on aligned facial images to detect schizophrenia. Additionally, various machine learning classifiers including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) were employed to train the model to achieve better accuracy on features such as action units, gaze, pose, and landmarks. The Particle Swarm Optimization (PSO) algorithm was utilized to optimize extracted features set.

PSO is an effective method for training neural networks to classify images, as it helps find an optimal set of weights that minimizes the loss function and maximizes classification accuracy. As an evolutionary optimization method, PSO employs a swarm in a random population using metaheuristics to search for the best solution, refining it with each iteration. Although PSO demands significant computational power, it can deliver an ideal set of hyperparameters within a designated time frame. Key parameters of PSO include population size, max iterations, cognitive weight, social weight, and inertia weight, each contributing to the optimization process.

The goal of this research is to provide a rapid, efficient, and reliable method for identifying schizophrenia directly from facial expressions, paving the way for early intervention and improved patient outcomes. By leveraging real-time capabilities, this research also aims to enhance accessibility to mental health services, thereby reducing the burden of this debilitating disorder on individuals, families, and communities.

1.4 Novelty of the Study

This is the last level of headings permitted. The proposed model reduces the questioning burden of patients, which psychiatrists use in clinical methods. It saves the time and cost of the patient to identify mental disorders from micro facial expressions, to detect schizophrenia using computer vision methods, to detect it by FACS and AUs, and to diagnose by both deep and machine learning techniques. Analyzing the process of mental disorders is cost-effective, efficient, and easy to use. The key contribution of the research work is as follows:

(a) Development of an efficient model to analyze the AUs through facial expressions of both patients and healthy controls to identify schizophrenia.

(b) Collection of a schizophrenia dataset from a local hospital in Bahawalpur, Pakistan, which is otherwise unavailable publicly.
(c) The model's high performance in analyzing schizophrenia patients with high validation accuracy supports the real-time deployment of the model.

This approach exceeds traditional diagnostic methods by using deep learning, particularly CNN, to analyze facial features and expressions. The integration of PSO enhances the model's efficiency in navigating the search space for optimal parameter settings. The proposed methodology not only achieves exceptional accuracy but also offers real-time applicability, marking a significant advancement in the field of mental health diagnostics.

The remaining parts of the article are structured below, Section 2 describes the proposed methodology of schizophrenia disorder detection, while Section 3 contains the results and discussions. The conclusions and future works are laid out in Section 4.

2 Methodology

In this methodology, the dataset consists of 2 to 3-min videos featuring both schizophrenia patients and healthy controls. These recorded videos were then converted into colour frames. Using the OpenFace tool, aligned and landmarked faces were extracted from the frames. Similarly, the tool generated a Comma Separated Values (CSV) file containing data on pose, gaze, AUs, and landmarks. The 44 AUs were measured in accordance with the FACS. OpenFace is the first open-source tool which is downloaded from the official GitHub repository. OpenFace can detect face landmarks, estimate head poses, recognize facial action units, and estimate eye gazes. The features are extracted from the facial action coding system. Subsequently, the proposed CNN model calculated the features from each frame. This model first normalizes the image, applies convolution and max pooling layers, and then processes the data through flatten, dropout, and fully connected layers. To achieve higher accuracy, the PSO algorithm was employed. The proposed method, which aims to categorize schizophrenia based on facial appearances [25], encompasses several phases, all of which are detailed in Fig. 2.

2.1 Dataset

There are 2 to 3-min videos of schizophrenia patients obtained from the Psychiatry Department of the BVH, Bahawalpur, Pakistan, and some private hospitals are also involved in taking datasets. Each patient has signed a consent form, providing their informed consent for permitting the taking of videos when conducting interviews according to the CGI-SCH [7]. The team of experts, i.e., psychiatrists and psychologists affiliated with BVH, Bahawalpur, and Pakistan also validated the dataset. According to a clinical experiment done by a staff of specialists, there were 300 schizophrenic patients and 299 healthy volunteers (normal), particulars of them are organized in Table 1.

2.2 Preprocessing

The videos were converted into $256 * 256$ colour frames. The length of the video is 2 to 3 min, with thirty frames being taken every second. A total of 3,492,300 frames of schizophrenia and 3,483,000 frames of normal persons were acquired. It was a huge set of data to analyze. Another lighter form of this dataset was also identified by skipping one of the two consecutive frames. Data preprocessed by Dlib or MTCNN (Multi-Task Cascaded Convolutional Network). First we extract facial landmarks to identify key regions related to micro-expressions, then align the faces to a standard position and scale,

normalize pixel intensities to enhance contrast and highlight subtle expressions, and extract features from aligned images to train machine learning models for micro-expression recognition.

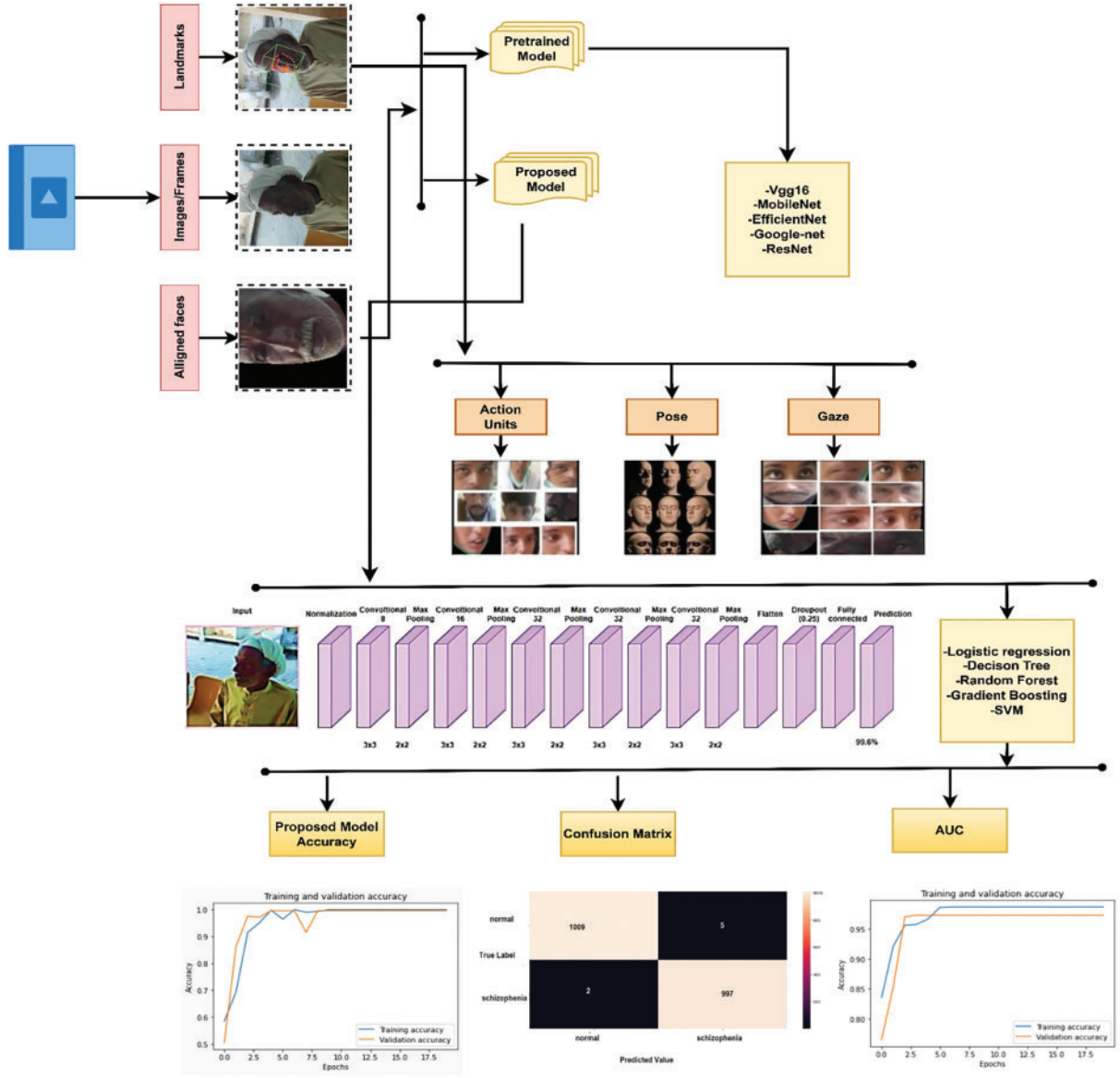


Figure 2: Overview of the proposed methodology

Table 1: Details of patients and healthy controls

Sr#	No. of patients	Gender M/F	Age	Martial status	Occupation
1	30	Male	21–38	Unmarried	Job holder
2	45	Male	31–42	Married	Job holder

(Continued)

Table 1 (continued)

Sr#	No. of patients	Gender M/F	Age	Martial status	Occupation
3	30	Male	40–54	Married	Job holder
4	30	Male	50–65	Married	Job holder
5	34	Female	21–36	Married	Housewife
6	37	Female	31–46	Married	Job holder
7	50	Female	42–55	Married	Housewife
8	40	Female	49–64	Married	Housewife
9	30	Female	49–70	Married	Housewife
10	53	Male	22–40	Unmarried	Job holder
11	42	Female	25–57	Unmarried	Job holder
12	24	Male	32–46	Married	Job holder
13	43	Female	48–60	Married	Housewife
14	51	Male	41–65	Married	Job holder
15	60	Female	57–65	Married	Housewife

2.3 Feature Extraction

After preprocessing the feature extraction technique applied to detect the coordinate of each image, intensity value of each action unit, heads orientation, and the direction of gaze. The OpenFace tool [26] has extracted aligned faces from frames. It also generated pose, gaze, AUs, and landmarking features in CSV format. AUs are based on micro-expressions. The aligned faces are saved as a Joint Photographic Experts Group (JPEG) with dimensions of 256 × 256. Frames were aligned, and then mouth, nose, eyebrows, and cheeks [27] landmarks were created. AUs 1, 2, 4, 5, 6, 7, 9, 12, 15, 16, 20, and 26 are intensely used in schizophrenia. They can appear alone or in combination to express an emotion. Samples of the noisy, aligned, and landmark faces are shown in Fig. 3.

2.4 The Proposed CNN Architecture

CNN is a deep learning model [28]. In the proposed CNN-based architecture, the input layer first applies a convolutional operation with a filter size of 8 and a kernel size of 3. This is followed by a max pooling layer with a pool size of 2. Subsequently, another convolutional layer is applied with a filter size of 16 and a kernel size of 3, followed by a max pooling layer with a pool size of 2. The process continues with a convolutional layer having a filter size of 32 and a kernel size of 3, and then a max pooling layer with a pool size of 2. This pattern repeats with another convolutional layer using a filter size of 32 and a kernel size of 3, followed by a max pooling layer with a pool size of 2. Furthermore, the output shape of another convolutional layer uses a filter size of 16 and a kernel size of 3, followed by a max pooling layer with a pool size of 2. In this model, all convolution and max pooling layers utilize a stride of 1, 'same' padding, and the ReLU activation function. Following these layers, a flattened layer is applied, then a dropout layer with a rate of 0.25. A dense layer follows, and finally, another dense layer with a sigmoid activation function is used. A complete CNN architecture is illustrated in Fig. 4. It is worth stating that the total trainable number of parameters is 545,730. The details of each of the layers configured for the proposed architecture are as follows.

Figure 3: The noisy, aligned, and landmark features

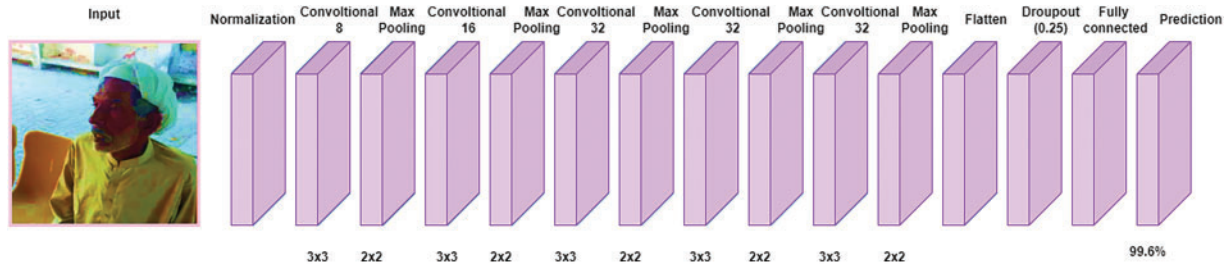


Figure 4: The proposed CNN architecture for schizophrenia detection

2.4.1 Convolutional Layer

Convolutional layers have been specifically and effectively developed to capture localized patterns and hierarchical structures of features within datasets. By utilizing several filters with varying weights, these layers can acquire the ability to identify diverse traits at distinct levels of abstraction. Generally, the initial layers of a CNN are responsible for detecting rudimentary features such as edges and simple patterns. As we progress toward deeper layers, the network becomes capable of discerning more intricate features and objects. The convolutional process is shown in Eq. (1).

$$Y[i,j] = \sum m \sum nX[i+m, j+n] \times K[m,n] + b \tag{1}$$

where $Y\,[I,j]$ = output value in the feature map, $X\,[i+m,\, j+n]$ is the input value, $k\,[m,n]$ is the filter value, and b represented a bias term.

2.4.2 Pooling Layer

The pooling layer, also referred to as a pooling operation, is a crucial element within CNN that serves the primary purpose of reducing the spatial dimensions of feature maps while preserving significant information. Pooling layers are beneficial components in neural networks as they effectively decrease the computational effort required by the entire network and enhance its resilience to minor changes in the input data. The pooling process is shown in Eq. (2).

$$Y[i,j,k] = \llbracket max \rrbracket\_(p=0,\, q=0)\hat{}(f-1)(X[i\times s+p,\, j\times s+q,\, k]) \tag{2}$$

where $Y\,[I,\, j,\, k]$ = output feature map, $\llbracket max \rrbracket\_(p=0, q=0)\hat{}(f-1) = f$ is the size of pooling window, $X\,[i\times s+p, j\times s+q,\, k] = X$ is input feature map, k represents the channel or depth dimension of the feature maps, s is the stride of the pooling operation.

2.4.3 Rectified Linear Unit (RELU) Layer

In the design of a CNN, ReLU activations are commonly employed after convolutional and linear operations within each neuron. This implies that the ReLU function, as shown in Eq. (3), is applied to every element of the outputs from fully connected layers to introduce non-linearity. ReLU activations are frequently used alongside other architectural elements in contemporary CNN architectures to construct highly efficient and successful image classification models.

$$ReLU(x) = max(0,\, x) \tag{3}$$

2.4.4 Fully Connected Layer

Fully connected layers are commonly employed in the final phase of a CNN structure since their purpose is to consolidate and amalgamate the knowledge received from preceding pooling and convolutional layers. This enables a neural network to acquire and comprehend complex traits and patterns that provide value in the result.

2.4.5 Sigmoid Function

The sigmoid function shown in Eq. (4) is implemented in a manner where it is applied individually to each element of the output produced by every neuron within a convolutional layer. Incorporating this transformation introduces non-linearity into the network's computations, enabling it to learn intricate patterns and relationships within the input data. The sigmoid function, used for classification, is an appropriate way to represent data probabilistically. However, it requires significant processing power when applied to large datasets or deep neural networks.

$$\sigma(x) = 1/(1+e\hat{}(-x)) \tag{4}$$

where $\sigma(x)$ represents the output for the input x. e is (approx. 2.71828). x is the input value.

2.5 Parameter Optimization of the CNN Architecture Using PSO Algorithm

PSO uses a swarm to search for the optimal solution in each area. This method involves several key factors. The population size, which represents the number of particles in the search area, is set to 20 in the proposed architecture. The position update, detailed in Eq. (5), shows how each particle moves within the search area. Additionally, the maximum number of iterations for the proposed model

is set to 50. The cognitive weight is a factor that influences a particle's tendency to move toward its personal best position; this is controlled by the velocity update shown in Eq. (6) and is set at 0.5. Similarly, the social weight, set at 0.5, is a factor that influences how strongly a particle moves towards the globally best position found by the swarm. Lastly, the inertia weight controls the extent to which a particle's current velocity influences its next speed update. The simulation parameters of PSO and the hyperparameters of the proposed CNN architecture are shown in Table 2. These factors together help guide the particles towards the best solutions by balancing personal and collective knowledge within the swarm.

$$xi(t+1) = xi(t) + vi(t+1) \tag{5}$$

Table 2: Details of the simulation parameters the PSO and hyperparameters of the proposed CNN

	Parameter	Value
PSO simulation parameters	Population size	20
	Maximum iterations	50
	Cognitive weight	0.5
	Social weight	0.5
	Inertia weight	0.5
CNN simulation parameters	Learning rate	0.001
	Filter size	3×53
	Pooling size	2×2
	Dropout	0.25
	Batch size	64
	Epochs	20

Some of CNN's hyperparameters are the learning rate, the number of filters, the size of the filters, the number of layers, etc. During training, CNN's weights and biases of the convolutional layers, fully connected layers, learning rate, batch size, and activation functions are all adjusted to work best. During training, the CNN changes the values of these parameters to reduce the loss function, which measures the difference between the predicted output and the actual output for a particular input. As part of the optimization process, the gradient of the loss function concerning the values must be calculated.

$$vi(t+1) = w \times vi(t) + c1 \times rand() \times (pbesti - xi(t)) + c2 \times rand() \times (gbest - xi(t)) \tag{6}$$

A swarm of particles changes their relative positions from one cycle to the next, which helps the PSO algorithm search. Each particle moves towards its previous personal best position (pbest) and the swarm's global best position (gbest) to find the best answer. There are different hyper-parameters of the CNN architecture, as a dropout is an overfitting that can be avoided by adjusting this CNN hyper-parameter. The proposed CNN architecture utilizes a fixed dropout value of 0.25. Batch size is a hyperparameter used in the gradient descent algorithm. The batch size value has been set at 64 in the design suggested for CNN, and the learning rate has been set to 0.001 in the proposed model.

3 Results and Discussion

The sample size used in the proposed method consists of 6,975,300 images. We then split the data for training and testing purposes. We considered 70% of the dataset, which amounts to 4,882,710 images, to train the model. Similarly, 15% of the dataset, or 2,092,590 images, was allocated for testing, while the remaining 15%, also 2,092,590 images, was designated for validation. The model is trained with a batch size of 64 and for 20 epochs. Moreover, the reduced version of the dataset consists of 3,487,650 frames, maintaining the same ratios of training (70%), testing (15%), and validation (15%).

As per the results shown in Table 3, when aligned face images are classified using VGG16, the total and trainable parameters are 23,136,577, and the testing accuracy is 94.52% with a batch size of 64 and 20 epochs. In MobileNet, the total number of parameters is 16,107,330, and the trainable parameters are 16,085,442; the testing accuracy is 98.40% with a batch size of 64 and 20 epochs. In EfficientNet, the total and trainable parameters are 61,143,261, and the testing accuracy is 97.10% with a batch size of 64 and 20 epochs. In GoogleNet, the total number of parameters is 40,710,433, and the trainable parameters are 40,676,001; the testing accuracy is 95.46% with a batch size of 64 and 20 epochs. In ResNet50, the total and trainable parameters are 15,927,617, and the testing accuracy is 96.90% with a batch size of 64 and 20 epochs. Similarly, when aligned face images are classified using VGG16, MobileNet, EfficientNet, GoogleNet, and ResNet50, the validation accuracies achieved are 95.67%, 99.0%, 97.63%, 96.57%, and 97.43%, respectively. When four of the latest reported visual models, i.e., ViT, Swin Transformers, BiT, and Vision Transformer without Attention, are used to train the model, they achieve testing accuracies of 56.17%, 71.40%, 84.21%, and 75.16%, respectively. When these four models, viz., ViT, Swin Transformers, BiT, and Vision Transformer without Attention, are used to train the model, they achieve validation accuracies of 51.45%, 70.13%, 81.19%, and 73.19%, respectively. The main reason these models do not achieve higher accuracies is that their weights are not tuned against micro-visual expressions. On the other hand, the proposed model requires only 16 layers and 5 lac parameters. Because of this, it can be called "lightweight," It takes less time to train while still achieving testing accuracy (99.60%), and validation accuracy (99.65%) with 64 batch sizes and 20 epochs.

Table 3: Results of schizophrenia classification through features

Features	Model trained amongst	Parameters million	Testing accuracy	Validation accuracy
Frames having aligned faces	VGG16	23	94.52	95.67
	Mobile Net	16	98.40	99.00
	Efficient Net	61	97.10	97.63
	Google Net	40	95.46	96.57
	ResNet50	15	96.90	97.43
	ViT	–	56.17	51.45
	Swin transformers	–	71.40	70.13
	BiT	–	84.21	81.19
	Vision transformer without attention	–	75.18	73.19
	The proposed CNN-based model	0.5	99.60	99.65

From Table 4, it is evident that combining all characteristics, i.e., pose, gaze, AUs, and landmarks also achieved notable testing accuracies with various classifiers: LR (98.20%), DT (95.60%), RF (98.40%), GB (99.50%), and SVM (90.60%). These are good scores, particularly highlighting the gradient boosting classifier's high accuracy of 99.50%, which indicates a very effective categorization capability. Furthermore, it is essential to mention that several performance evaluation metrics are utilized to obtain these results, including recall, precision, F1-score, and Area under the curve (AUC)-value. These metrics help in providing a comprehensive assessment of the classifiers' performance. Fig. 5 shows the accuracy, loss plots, and confusion matrix of the results.

Table 4: Results of Combined features (6 Pose + 288 Gaze + 35 AUs + 136 Landmark)

Model trained amongst	Accuracy	Recall	Precision	F1-score	AUC-value
LR	98.20%	0.90	0.95	94.64%	0.967
DT	95.60%	0.94	0.96	96.17%	0.985
RF	98.40%	0.98	0.99	98.68%	0.998
GB	99.50%	1.0	1.0	99.68%	1.000
SVM	90.60%	0.90	0.91	91.42%	0.950

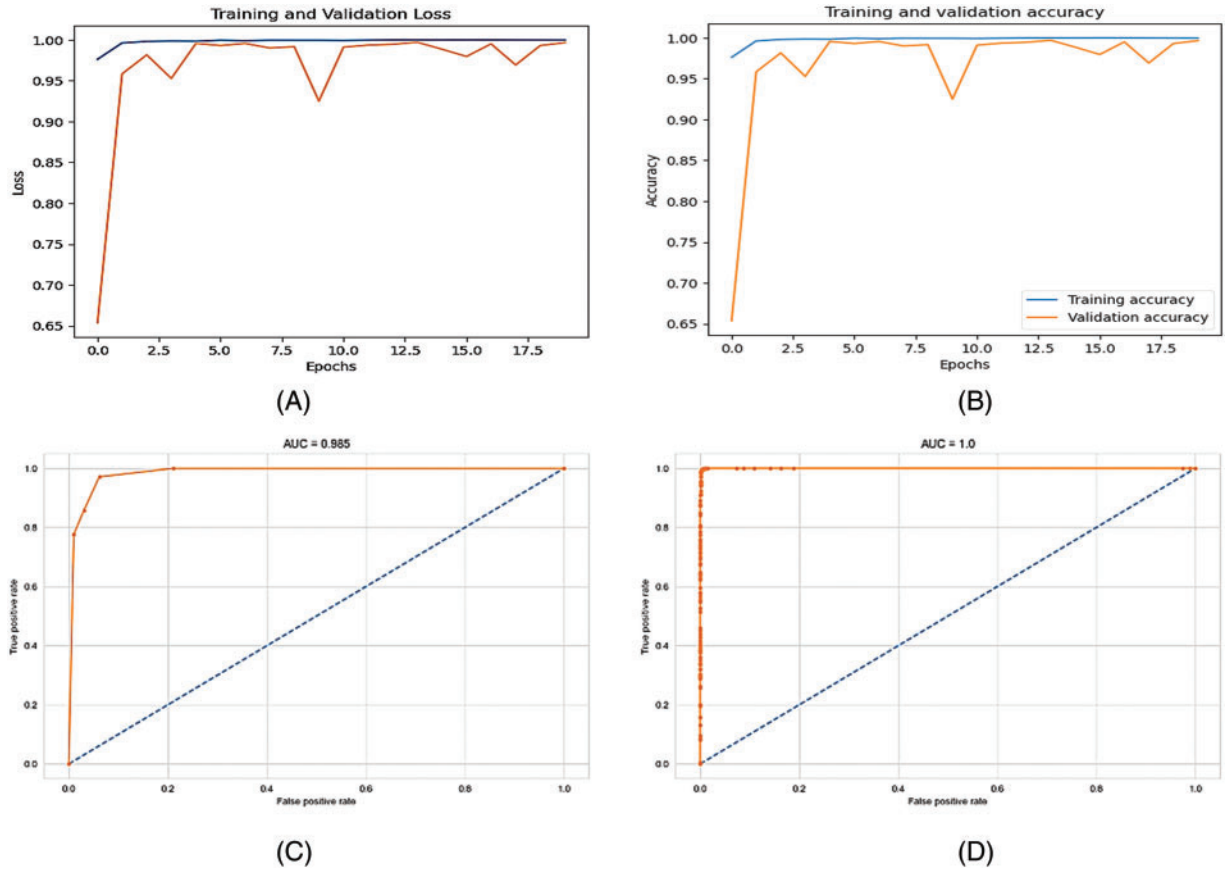
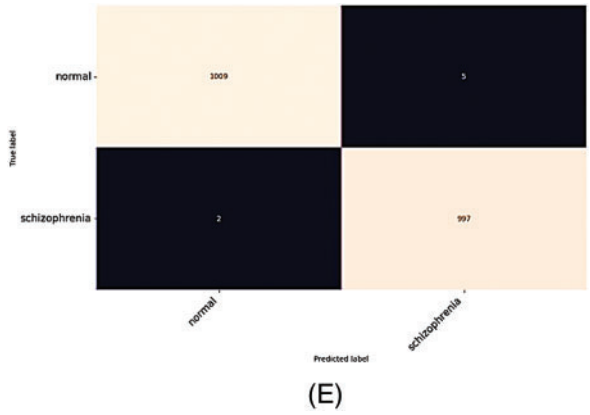


Figure 5: (Continued)



(E)

Figure 5: Accuracy, loss plots, and confusion matrix (A) Accuracy plot of the proposed CNN model on aligned faces, (B) Loss plot of CNN proposed model, (C) AUC of the DT, (D) AUC (Pose + Gaze + AUs + Landmark) of GB, (E) Confusion matrix of the proposed CNN model

Table 5 compares the proposed model's results with recent research on schizophrenia identification using video or image data

Table 5: Comparison of the previous studies

References	Datasets	Clinical techniques	Methodology	Results
[5]	106 schizophrenia and 101 healthy controls	Six specific questions	CNN model	Accuracy 95.18%
[17]	115 schizophrenia and 57 normal persons	Beck Depression Inventory (BDI)	Openface toolkit	Upper-face spontaneous smiles are stronger, and the posed smile had higher lower-face intensities-
		Beck Anxiety Inventory (BAI)	CE-CLM	
[18]	22 videos of a schizophrenic patient	PANSS	MTCNN, SURF, pre-trained deep learning models	Accuracy 80%
[6]	CAER and AVEC video datasets	__	CMGCN	Precision 69.52% Accuracy 87%

(Continued)

Table 5 (continued)

References	Datasets	Clinical techniques	Methodology	Results
[22]	2039 schizophrenia patient images and 2049 healthy images	__	Degraded Facial Affect Recognition Task software version 16.0	Confidence interval 95%
[23]	Chicago Face Database	__	Bayesian classifier and ANOVA scale	Probability 0.047 Variance 0.05
[24]	60 schizophrenia and 104 healthy	PRS	Gradient-boosted decision tree and ANOVA scale	Accuracy 88%, AUC 0.87
[29]	The Chinese Academy of Sciences Micro-Expression dataset (CASME-II), Spontaneous Micro-Expression Corpus-High Speed (SMIC-HS), Spontaneous Actions and Micro-Movements (SAMM)	Clinical Interview	CNN+PSO+SVM	Accuracy 87%
[30]	Micro expression recognition	__	CNN and vision transformer model	Results on the CASME-I, CASME-II, and SAMM benchmark datasets were 95.77%, 98.59%, and 100% accurate, correspondingly.
[31]	Micro expression recognition	___	CNN, Long short-term memory (LSTM), and vision transformer model	Hybrid model perform best. The performance of recognition can be significantly improved via score fusion

(Continued)

Table 5 (continued)

References	Datasets	Clinical techniques	Methodology	Results
The proposed CNN-based model	300 schizophrenic videos and 299 healthy control videos	CGI-SCH	CNN pre-trained models, i.e., VGG16, Mobile Net, Efficient Net, Google Net, and ResNet50 Machine learning models, i.e., logistic regression, decision tree, random forest, gradient boosting, and SVM	Accuracy 99.6%

In this article [5], the author worked on recognizing schizophrenia using facial expressions based on the convolutional neural network. The dataset consists of 106 schizophrenia and 101 healthy controls. Overall accuracy they achieved is 95.18%. In the research [17], the Openface toolkit were utilized for only posed and spontaneous smile of schizophrenic patients of Korean people. The study [18] explores motor function abnormalities in schizophrenia using a computer vision-based approach via smartphones for remote monitoring. Authors used PANSS to obtain the dataset of the head movement of the person with accuracy of 80%. In [6], CMGCN has been proposed. It has been evaluated as per precision (69.52%) and accuracy (87%); accuracy is not up to the mark. In another paper [22], a Degraded Facial Affect Recognition task (DFAR) was used on image data to diagnose schizophrenia. The confidence interval achieved was 95%, but the limited dataset was used.

Another research [23] used the Bayesian classifier networks for schizophrenic emotions. Its probability was 0.047, and its variance was 0.05. This study has a high measure of errors. In the study [24], the gradient-boosted DT was used for encoding the eye movements of schizophrenic patients, with an accuracy of 88%. This study has low accuracy and focuses on eye movement only. In this paper [29] a micro-expression recognition framework combining deep learning models, i.e., AlexNet and VGG16 with PSO for feature selection, achieving an acceptable accuracy of 87.84% on combined datasets, i.e., CASME-II, SMIC-HS, and SAMM. They used SVM for classification.

The conducted study proposed a customized light-weighted CNN model, which consists of 16 layers and 5 million parameters, and archives an overall accuracy of 99.6%.

4 Conclusion and Future Work

Schizophrenia is recognized as a severe condition that can significantly impact health. Analyzing facial action units (AUs) offers considerable insights into diagnosing schizophrenia. Despite numerous efforts to develop models that diagnose schizophrenia based on visual cues, these models have generally underperformed in evaluations. Consequently, we proposed a CNN model complemented by machine learning classifiers, trained on a locally compiled dataset optimized with a particle swarm optimization (PSO) algorithm. Research indicates that our proposed CNN model achieved a validation accuracy of 99.65%, while the features extracted attained a high accuracy of 99.50% using a GB classifier.

Furthermore, our findings demonstrate that this strategy outperforms current best practices. Despite having fewer layers and parameters, our CNN-based model remains reliable, which is particularly significant considering the poor quality of video captured by contemporary cameras. The model is deployable in real-time situations with high-resolution data and achieves excellent validation accuracy. It compares approaches to maximize performance using CNN, machine learning and Transformer models. Its efficacy and versatility are further increased by advanced feature extraction that makes use of OpenFace for pose, gaze, and Action Units and Particle Swarm Optimization for parameter optimization. The model's success is strongly dependent on the quality and diversity of the dataset, but the sample size is limited for Asian people, the findings may not be applicable to larger populations of the world. Taking videos of schizophrenic patients raises ethical concerns about privacy, permission, and potential stigma for people diagnosed with schizophrenia. Real-time deployment may necessitate large amount of hardware resources, which could be a challenge in low-resource environments. Lighting, camera quality, and the participant's condition during data recording can all influence the accuracy of facial expression analysis.

Our long-term goal is to develop a system that utilizes a person's gait, language, and gestures to diagnose a broad spectrum of psychiatric conditions in real time. Specifically, we plan to incorporate an electroencephalogram (EEG) sensor for the early diagnosis of these conditions.

However, the present study has its limitations. First, the model is trained exclusively on datasets collected from local hospitals, limiting its application to recognizing only local visual traits. Second, the model requires high-resolution images to effectively extract facial expressions, thus precluding the use of low-resolution cameras.

The multimodal approach to detect schizophrenia term "gait" describes how a person walks, including their space, rhythm, and stability. Computer vision algorithms are capable of extracting parameters such as walking speed, stride length, and body posture from video images. When it comes to clinical significance, irregularities in gait may be a sign of motor control problems common in people with schizophrenia. Analyzing speech patterns and content, as well as verbal and nonverbal communication, is part of language processing. Techniques for natural language processing (NLP) can evaluate the coherence, fluency, and emotional tone of speech. Analysis is possible for elements like sentence structure and word choice. In a similar vein Machine learning algorithms can be used to assess acoustic characteristics including pitch, tone, and speech tempo in order to spot abnormal patterns in people with schizophrenia. Train models to classify gestures associated with different mental states using labeled data. Features may include gesture frequency, duration, and types of movements. Additionally, the model is not robust across all biomarkers, such as gestures, gait, voice, and handwriting.

Acknowledgement: The researchers would like to thank the BVH, Bahawalpur, Pakistan staff for their cooperation in providing the dataset.

Funding Statement: Self-funded; no external funding is involved in this study.

Author Contributions: Ghulam Gilanie gave the idea of the research topic, selected the title name, and supervised the whole research work; Anum Saher wrote the paper and made a model to perform experiments on the dataset. Similarly, Sana Cheema, Akkasha Latif helped in dataset collection from different private and Government hospitals. Syeda Naila Batool helped in the writing of the literature review. Hafeez Ullah helped in labeling, finalizing & validating the outcomes of the conducted research work. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Ethics Approval: Ethical approval for the study involving human participants was obtained from the Office of Research Innovation and Commercialization (ORIC) at the Islamia University of Bahawalpur, vide letter No. 338. The informed consents were obtained from the participants which from Psychiatry Department of the BVH, Bahawalpur, Pakistan.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] T. Cowan, M. D. Masucci, T. Gupta, C. M. Haase, G. P. Strauss and A. S. Cohen, "Computerized analysis of facial expressions in serious mental illness," *Schizophr. Res.*, vol. 241, pp. 44–51, Mar. 2022. doi: 10.1016/j.schres.2021.12.026.

[2] J. S. Ambikile and M. K. Iseselo, "Challenges to the provision of home care and support for people with severe mental illness: Experiences and perspectives of patients, caregivers, and healthcare providers in Dar es Salaam, Tanzania," *PLoS Glob. Public Health*, vol. 3, no. 1, Jan. 2023, Art. no. e0001518. doi: 10.1371/journal.pgph.0001518.

[3] S. Hasan and M. Adil, "Schizophrenia: A neglected problem in Pakistan," *Lancet*, vol. 394, pp. 115–116, Jul. 2019, Art. no. 10193. doi: 10.1016/S0140-6736(19)30290-9.

[4] C. J. Zhao, X. B. Dang, X. Su, J. Bai, and L. Y. Ma, "Epidemiology of suicide and associated socio-demographic factors in emergency department patients in 7 general hospitals in northwestern China," *Med. Sci. Monit.*, vol. 21, pp. 2743–2749, Sep. 2015. doi: 10.12659/MSM.894819.

[5] X. Zhang *et al.*, "Recognizing schizophrenia using facial expressions based on convolutional neural network," *Brain Behav.*, vol. 13, no. 5, Apr. 2023, Art. no. e3002. doi: 10.1002/brb3.3002.

[6] J. Huang *et al.*, "Automatic recognition of schizophrenia from facial videos using 3D convolutional neural network," *Asian J. Psychiatr.*, vol. 77, Nov. 2022, Art. no. 103263. doi: 10.1016/j.ajp.2022.103263.

[7] S. Z. Domingo, J. Bobes, M. -P. García-Portilla, C. Morralla, and E. -S. S. Group, "Cognitive performance associated to functional outcomes in stable outpatients with schizophrenia," *Schizophr. Res.*, vol. 2, no. 3, pp. 146–158, Sep. 2015. doi: 10.1016/j.scog.2015.03.002.

[8] T. S. Kumar, K. N. Rajesh, S. Maheswari, V. Kanhangad, and U. R. Acharya, "Automated schizophrenia detection using local descriptors with EEG signals," *Eng. Appl. Artif. Intell.*, vol. 117, no. 1, pp. 44–51, Mar. 2023. doi: 10.1016/j.engappai.2023.105602.

[9] G. Gilanie *et al.*, "Coronavirus (COVID-19) detection from chest radiology images using convolutional neural networks," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102490. doi: 10.1016/j.bspc.2021.102490.

[10] D. J. Schretlen, "The use of psychological tests to identify malingered symptoms of mental disorder," *Clin. Psychol. Rev.*, vol. 8, no. 5, pp. 451–476, 1988. doi: 10.1016/0272-7358(88)90074-8.

[11] H. Guerdelli, C. Ferrari, W. Barhoumi, H. Ghazouani, and S. Berretti, "Macro-and micro-expressions facial datasets: A survey," *Sensors*, vol. 22, no. 4, Feb. 2022, Art. no. 1524. doi: 10.3390/s22041524.

[12] M. Peng, Z. Wu, Z. Zhang, and T. Chen, "From macro to micro expression recognition: Deep learning on small datasets using transfer learning," presented at the 13th IEEE Int. Conf. Autom. Face Gesture Recognit., Xi'an, China, 15–19 May, 2018.

[13] V. Esmaeili, M. Mohassel Feghhi, and S. O. Shahdi, "A comprehensive survey on facial micro-expression: Approaches and databases," *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 40089–40134, Nov. 2022. doi: 10.1007/s11042-022-13133-2.

[14] W. Jiang, Y. Wu, F. Qiao, L. Meng, Y. Deng and C. Liu, "Model level ensemble for facial action unit recognition at the 3rd ABAW challenge," presented at the the IEEE/CVF Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 19–20, 2022.

[15] F. L. Siena, M. Vernon, P. Watts, B. Byrom, D. Crundall and P. Breedon, "Proof-of-concept study: A mobile application to derive clinical outcome measures from expression and speech for mental health status evaluation," *J. Med. Syst.*, vol. 44, no. 12, Nov. 2020, Art. no. 209. doi: 10.1007/s10916-020-01671-x.
[16] B. Jin, L. Cruz, and N. Gonçalves, "Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123649–123661, 2020. doi: 10.1109/ACCESS.2020.3005687.
[17] S. Park *et al.*, "Differences in facial expressions between spontaneous and posed smiles: Automated method by action units and three-dimensional facial landmarks," *Sensors*, vol. 20, no. 4, Jun. 2020, Art. no. 1199. doi: 10.3390/s20041199.
[18] A. Abbas *et al.*, "Computer vision-based assessment of motor functioning in schizophrenia: Use of smartphones for remote measurement of schizophrenia symptomatology," *Digit. Biomark*, vol. 5, no. 1, pp. 29–36, Jan. 2021. doi: 10.1159/000512383.
[19] B. J. Lin *et al.*, "Mental status detection for schizophrenia patients via deep visual perception," *IEEE J. Biomed. Health. Inform.*, vol. 26, no. 11, pp. 5704–5715, Nov. 2022. doi: 10.1109/JBHI.2022.3199575.
[20] B. Jin, Y. Qu, L. Zhang, and Z. Gao, "Diagnosing Parkinson disease through facial expression recognition: Video analysis," *J. Med. Internet Res.*, vol. 22, no. 7, Jul. 2020, Art. no. e18697. doi: 10.2196/18697.
[21] O. Attallah, "A deep learning-based diagnostic tool for identifying various diseases via facial images," *Digit. Health*, vol. 8, 2022. doi: 10.1177/20552076221124432.
[22] L. Fusar Poli *et al.*, "The association between cannabis use and facial emotion recognition in schizophrenia, siblings, and healthy controls: Results from the EUGEI study," *Eur. Neuropsychopharmacol.*, vol. 63, pp. 47–59, Aug. 2022. doi: 10.1016/j.euroneuro.2022.08.003.
[23] K. Courtenay, A. H. Wong, R. Patel, and T. A. Girard, "Emotional memory for facial expressions in schizophrenia spectrum disorders: The role of encoding method," *J. Psychiatr. Res.*, vol. 146, pp. 43–49, Dec. 2022. doi: 10.1016/j.jpsychires.2021.12.026.
[24] H. Lyu *et al.*, "Eye movement abnormalities can distinguish first-episode schizophrenia, chronic schizophrenia, and prodromal patients from healthy controls," *Schizophr. Bull. Open*, vol. 4, no. 1, 2023, Art. no. sgac076. doi: 10.1093/schizbullopen/sgac076.
[25] G. Gilanie *et al.*, "An automated and real-time approach of depression detection from facial micro-expressions," *Comput. Mater. Contin.*, vol. 73, no. 2, pp. 2513–2528, Jun. 2022. doi: 10.32604/cmc.2022.028229.
[26] J. Araluce *et al.*, "Gaze focalization system for driving applications using OpenFace 2.0 toolkit with NARMAX algorithm in accidental scenarios," *Sensors*, vol. 21, no. 18, Sep. 2021, Art. no. 6262. doi: 10.3390/s21186262.
[27] C. Correia-Caeiro, A. Burrows, D. A. Wilson, A. Abdelrahman, and T. Miyabe-Nishiwaki, "CalliFACS: The common marmoset facial action coding system," *PLoS One*, vol. 17, no. 5, May 2022, Art. no. e0266442. doi: 10.1371/journal.pone.0266442.
[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. doi: 10.1038/nature14539.
[29] M. Z. Uzun, Y. Çelik, and E. Başaran, "Micro-expression recognition by using CNN features with PSO algorithm and SVM methods," *Int. Inf. Eng. Tech. Assoc.*, vol. 39, no. 5, pp. 1685–1693, 2022. doi: 10.18280/ts.390526.
[30] S. Indolia, S. Nigam, R. Singh, V. K. Singh, and M. K. Singh, "Micro expression recognition using convolution patch in vision transformer," *IEEE Access*, vol. 11, pp. 100495–100507, 2023. doi: 10.1109/ACCESS.2023.3314797.
[31] Y. Zheng and E. Blasch, "Facial micro-expression recognition enhanced by score fusion and a hybrid model from convolutional LSTM and vision transformer," *Sensors*, vol. 23, no. 12, 2023, Art. no. 5650. doi: 10.3390/s23125650.