Behavioral Analysis Through Computer Vision: Detecting Emotions and Hand Movements to Aid Mental Health

Aditya Gupta and Geetanjali Bhola
Department of Information Technology, Delhi Technological University, Rohini, Delhi, India

Keywords: CNN, VGG-Model, Transformers, DeepFace, MediaPipe, RAF-DB.

Abstract: This paper presents an advanced behavioral pattern analysis methodology developed for a mental health diagnosis task using the CNN model with an additional Attention Layer constituting an enhanced VGG Model. To improve combined performance, a Transformer has been integrated into the CNN. The VGG-Transformer model fine-tunes and trains over 3,500 images from the dataset available in the RAF-DB dataset for optimal feature extraction and classification. Our sophisticated model uses a repository called DeepFace, with a VGG model in it, followed by the MediaPipe library to process video inputs from the user. The system accurately recognises various emotions and hand movements from frames, with a final great training accuracy of 94.33% and validation accuracy of 95.76%. The DeepFace integration supports detail and precision in the recognition of facial emotions, while MediaPipe hand tracking enables deep hand movement analysis. Such a complementary approach will make one delve into understanding behavior in depth, where the results can aid in detecting the possibility of mental health-related problems and supporting therapists in their diagnosis and treatment. This work has proven that state-of-the-art machine learning in mental health assessment is required and has been proven effective by combining the enhanced VGG16 Model and Transformers, DeepFace, and MediaPipe specialised libraries.

1 INTRODUCTION

Mental health disorders are increasingly becoming a global concern that has widely affected the quality of life of individuals and further resulted in serious outcomes, even mortality. The World Health Organization claims that about 450 million people across the world are affected by different mental health conditions, which makes mental disorders one of the leading causes of ill health and disability in the world (Gleason, 2023). Furthermore, more than 264 million people are affected by various mental illnesses and are a major contributor to the global burden of disease (Islam et al., 2024). Suicide claims almost 700,000 lives every year tragically and is the second leading cause of death in both groups aged 15-29 years (Goueslard et al., 2024).

Emotional health is an integral part of mental wellness and significantly determined by social interaction and relationship. Lockdown and social distancing during the COVID-19 pandemic severed those connections that had increased isolation, thereby increasing anxiety, stress, and depression among adolescents. Social support systems grow weaker alongside reduced emotional health, making young people more at risk of falling to mental health issues and such risky behavior as substance use. Social relationships with family, friends, or adults trusted would provide the sense of belonging and stress-buffering benefits in order to foster emotional resilience and prevent loneliness (Kwaning et al., 2023).

Disorders such as Autism Spectrum Disorder that presents itself as affecting early development greatly, though it appears in at least two of the following: challenges with social/communicative behaviors and restricted, repetitive patterns of behavior, interests, or activities. Since children with ASD have serious impairments in social integration, as well as emotional regulation, the goal of the treatment is strengthening these areas from an early age. In comparison with a healthy child, children with ASD are characterized above all by motor and gestural communication deficits: involuntary movements and stereotype gestures. Detection of unusual gestures through hand recognition could become significant as an early indicator for ASD, as it would allow one to understand the underlying social and motor impairments and help implement focused interventions to improve social

Gupta, A. and Bhola, G.
Behavioral Analysis Through Computer Vision: Detecting Emotions and Hand Movements to Aid Mental Health.
DOI: 10.5220/0013583500004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 1*, pages 671-682
ISBN: 978-989-758-763-4

engagement either at home or in school(Quintar et al., 2025).

These are the issues we are aiming to tackle with our technology taking anxiety, nervousness, depression, involuntary movements, stress, etc as the mental health classes considered. Keeping in mind that emotional health plays a major factor in investigating mental health problems in people at a young age. The young population goes through huge stress, is always involved in work, and often does not get time to check their emotional health. So, providing a method to detect the emotional health of the individual with the help of human-computer interaction on the grounds of verbal(speech) and non-verbal(face, posture, text, etc.) elements is building the future of health sector of the growing concern of mental health around the world to battle the stigma and encourage people to take care of their emotional and mental health (Mishra et al., 2023).

The hybrid architecture for the mental disorder detection task makes use of YOLOv8 in detecting visual cues related to mental disorders and combines Convolutional Neural Networks (Miao, 2021) with Visual Transformer models to form an ensemble classifier. This achieves an overall accuracy of about 81% in predicting mental illnesses like depression and anxiety. This enhances transparency and interpretability, whereby critical regions in the input image that influence predictions can be highlighted through Gradient-weighted Class Activation Mapping and saliency maps. Integration that makes it easier to understand the system's decisions increases confidence among health professionals in the results and thus supports a more informed diagnosis process (Aina et al., 2024). Through the years, neural networks have improved based on the availability of large annotated datasets, and continue to push development in complex models capable of real-time analysis for dynamic behaviors. The VGG model was proposed by Simonyan and Zisserman, 2014, which was an important milestone for image classification tasks, considering that it had deep architecture and utilised small convolutional filters. The VGG architecture is employed to train the dataset, specifically using the VGG-16 CNN model developed by K. Simonyan and A. Zisserman. This model, renowned for its performance, was one of the top submissions to ILSVRC-2014, achieving a top-5 accuracy of 92.7% on the ImageNet dataset (Kusumawati et al., 2022). The robustness of VGG16 architecture (Kusumawati et al., 2022) is fairly evident with its 96% accuracy for detecting facial expressions in CK+ datasets based on the learning and validating principle of Linear Regression and K-Nearest Neighbour (KNN), Random Forest algorithms to name a few (Zhang, 2020). Due to the ease of the structure and powerful feature extraction, the VGG model has been applied to many aspects of computer vision tasks.

Although initially designed for NLP, transformers have recently been adapted for vision tasks, resulting in variants such as the Vision Transformers (ViTs) (Berroukham et al., 2023). Because of their high affinity for capturing long-range dependencies and contextual information, they are very fitting for analysing complicated patterns in video data. This blend of the advantages of VGG models with transformers was finally able to embed both strengths of the architectures in feature extraction and classification accuracy.

Recent advances in machine learning and AI have opened new opportunities for very early detection and intervention of mental health issues. Key areas of interest are emotion analysis and the study of body movement. Researchers, with advanced models like VGG-Transformer architecture, can now look into minute details in facial expressions and physical gestures that may indicate some sort of mental illness.

2 RELATED WORKS

Research in the area of detecting stereotypical movements (SM) in children with autism, like arm flapping, spinning, and body rocking, from most works so far underlines a huge demand for early intervention. Traditionally, traditional methods for detecting stereotypical movement have relied on manual observation, which is time-consuming and labor-intensive. The need for this process has led to the development of intelligent diagnostic tools. The Stereotypic Movement Detection and Analysis software is one such intelligent tool. Concerning this, SMDA performs video pre-processing through computer vision and utilizes a machine learning framework in conjunction with the MediaPipe (Ma et al., 2022) library MPL. This enables this software to assist in processing recorded videos, labelling body part landmarks like wrists, identifying their movement, and extracting their movement on the X-Y coordinates during SMs. The algorithm, specifically a Data Peak Filtering Algorithm, is used in the software for estimating the SMs regarding their intensities and frequencies. Given SMDA's fast detection capabilities and flexibility in the choice of LBPs, it is an effective diagnostic tool for any therapist to enhance efficiency and accuracy in detecting SMs for treatment outcomes in children with autism (Reddy et al., 2023).

Previous AEE research has investigated the use

of various sensor types in marketing, technical systems, and human-robot interaction. These studies have carefully classified and analyzed sensors applied in emotion detection-contactless methods, contact-based sensors, and skin-penetrating electrodes-and discussed in-depth their effectiveness in identifying and measuring the type of an emotional state and its intensity. This classification of sensors has provided important insights for the researcher in deciding on the best methodology or exploring alternative possibilities for emotion analysis based on the application scope, expected results, and their inherent limitations. The studies have also further suggested practical uses that this idea could find in taking the human-centered aspects of the IoT and the related affective computing frameworks forward (Dzedzickis et al., 2020).

While facial data were used previously for identity verification, the characteristics of an eye blink have been utilised in behavior monitoring. Interesting to note the fact that a smart behavior biometrics system was developed using LENET, ALEXNET, and VGGNET network architectures; comparative analysis has been done across such networks to see their effectiveness in continuous authentication (Reshma and Jose Anand, 2023).

Many studies have employed variable-centered methods to explore the relationship between psychiatric disorders and internet addiction (IA), using techniques like regression analysis, factor analysis, and structural equation modeling. These approaches focus on the correlations between variables and typically assume a homogenous sample, grouping individuals based on total score cut points, often overlooking individual response patterns. In contrast, person-centered methods, such as latent class analysis (LCA), classify individuals into subgroups based on their unique response patterns, offering more accurate and insightful classification. LCA has been widely used in studying comorbid symptoms of depression, anxiety, and IA, providing a more nuanced understanding of the interrelationships among these factors. Despite the growing research on these comorbidities, there remains a gap in exploring individual differences in emotional patterns and IA using person-centered approaches (Beneytez, 2023).

Restricted and repetitive patterns (RRP) involve behaviors such as insistence on sameness, repetitive sensory-motor actions, and sensory processing differences, often emerging as self-regulation strategies to reduce anxiety. These behaviors help individuals maintain control in unpredictable environments, providing a mechanism for managing overstimulation or increasing arousal in low-stimulation states. RRP patterns are especially prominent in individuals with autism spectrum disorder (ASD), where sensory over-responsiveness correlates with heightened anxiety, limited social adaptability, and hyper-focused attention, while sensory under-responsiveness does not typically link to anxiety (Gao et al., 2022). These patterns are crucial for identifying early signs of emotional dysregulation and mental health issues, as they often signal underlying anxiety. By analyzing emotional fluctuations, tools like DeepFace can track these patterns, providing a real-time understanding of emotional states and their relationship to IA. DeepFace's ability to detect emotional expressions and behavioral patterns enhances the detection of early signs of comorbidity, guiding targeted interventions. In combination with systems like MediaPipe, which can identify unusual behaviors in real-time, these tools offer a comprehensive approach to monitoring and supporting mental health, facilitating early detection and more effective prevention strategies.

Major factors concerning a healthy lifestyle among students compiled from the processing of experimental results include, in descending order: emancipation from drug addiction (24.3%), sports activity (15.7%), abstaining from alcohol and smoking (11.4%), responsible sex life (9.8%), and nutrition (7.4%). Other features identified include having a meaningful life, good self-attitude, self-development, and family relationships, among others. However, they ranked lower compared to psychological health, which implies that the students acknowledge the role of psychological health in a holistic way of living. This view comes along with new perspectives on health, which emphasize the balance between physical well-being, internal harmony, and environmental congruence (Yunusovich et al., 2022).

From the psycho-emotional health point of view, the level of emotional instability among students varies: 20.5% of students manifest a high level, 62.4% an intermediate one, and 17.1% a low level of instability according to Eisenk's personality inventory. It is seen that there was not a notable difference in distribution between the experimental and the control group because they are equal on all levels of emotional stability. This is a signal that emotional regulation and resilience must be included in the general health components of education.

The system caters to the need for a user-friendly solution in extracting emotion-related information from facial expressions in real-time video streams. This system processes videos uploaded by users, detects facial emotions at each timestamp using OpenCV, DeepFace (Srisuk and Ongkittikul, 2017), and Streamlit, and then generates output annotated with these emotions. OpenCV is responsible for pro-

cessing the videos, DeepFace analyzes emotions, and Streamlit provides an interactive user interface. It effectively identifies dominant emotions, quantifies their occurrences, and gives results. This has improved human-computer interaction and contributed much to the research in mental health (Bhanupriya et al., 2023). DeepFace reaches an accuracy of 98.61% on LFW database while analysing the face depth of the database samples(Srisuk and Ongkittikul, 2017).

The proposed algorithm addresses the challenge of mental health by estimating addiction levels based on three critical factors: depression, social anxiety, and loneliness. This research focused on the analysis of these factors in a bid to gain some insight into addictive behaviors. The study applied three well-established machine learning algorithms: Logistic Regression, Ridge Regression, and Support Vector Machine to assess the levels of addiction. In the results among these algorithms, the Support Vector Machine turned out with the best performance of 91.68%. Addressing these factors of mental health and measures for controlled engagement in mobile gaming are some of the steps necessary to ensure improved mental health among students (Chauhan et al., 2023). Stress and negative emotions play a major part in an imbalance of emotional and mental health and since each person can show different facial expressions differently, it might not mean the emotion that we perceive. This leads to errors in judgement. Hence, the design of an emotional recognition system using psychological signals to recognise emotions designed a specific emotion induction experiment to collect five physiological signals of subjects including electrocardiogram, galvanic skin responses (GSR), blood volume pulse, and pulse. This is run with the help of a Support Vector Machine to study the trend of negative emotions captured by the model and achieved an accuracy of 89.1% (Chang et al., 2013).

The hand tracking module outputs an array of size 21, where each value indicates whether a joint on the left hand is visible. If all joints of a particular finger are visible, that finger is considered open, and its count is added to the total number of open fingers. The total number of open fingers represents the runs scored by the batsman. The trained convolutional neural network achieved an accuracy of 0.9767 after ten epochs of training. Since MediaPipe is an inbuilt framework, its accuracy is expected to be precise (Teja Gontumukkala et al., 2022).

3 PROPOSED MODEL

3.1 Dataset Preprocessing

The model first creates a complex preprocessing pipeline for an image classification model using TensorFlow and Keras (Bajpai and He, 2020). It utilizes the ImageDataGenerator class to treat both the training and test datasets with major transformations that enhance a model's performance. For the training set, images are rescaled by a factor of 1./255 for pixel value normalization in the range [0, 1]. Next, data augmentation is introduced through a zoom range of 0.3 and random horizontal flips, which increases the variability of the training images and generalizes them more toward the unseen test dataset. The RAF-DB dataset was chosen for its nuanced portrayal of emotions (Achlioptas et al., 2023) with colour and depth which makes it "stand from the crowd" (Weng et al., 2023) and is valuable to train our model to the utmost accuracy. Finally, images are fed into the preprocessing function keras.applications.vgg16.preprocess() input to be adapted to the specifics of the VGG16 model. It means processing the RAF-DB training set (Galea and Seychell, 2022) in batches of 64 images, resizing them to 100x100 pixels, and one-hot encoding the labels. For the test set, similar preprocessing is done without augmentation to make it consistent with the training data. The approach gives efficient ways of handling data, consistent preprocessing, and a robust model for training and evaluation.

3.2 Network Architecture

This network architecture considers quite a sophisticated model of image classification, which combines a custom TransformerBlock with a pre-trained VGG16 architecture. For example, in this study, the TransformerBlock (Zhang et al., 2024) is designed with an embedding dimension of 512, 4 attention heads, a feed-forward network dimension of 512, and a dropout rate set to be 0.1. Multi-head attention is combined with a feed-forward network that includes ReLU activation, Layer Normalization, and Dropout layers to ensure robust processing of features (Yan et al., 2023). In the proposed model, the strongly feature-extractive, pre-trained VGG16 is used without its top layers, and its weights are kept non-trainable to preserve the learned representations (Yamsani et al., 2023).

The output from the VGG16 model is flattened and fed into a dense layer of 512 units with ReLU activation. The reshaped output is further fed into

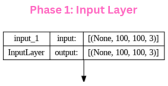

Figure 1: Phase 1 of Network Architecture

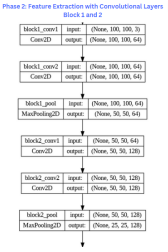

Figure 2: Phase 2.1 of Network Architecture

the TransformerBlock for additional feature representation enhancement from the attention mechanism. Also, another multi-head attention layer is added with 4 heads and a key dimension of 512, further refining the features and adding extra depth and complexity in the learning of model features.

These resultant features undergo global average pooling to produce a vector of fixed size, encapsulating all the essential information. This is then fed through two fully connected layers, each with 4096 units and ReLU activation, interspersed with dropout layers at a rate of 0.5 to avoid overfitting. Class probabilities corresponding to the 7 target classes are given by the last layer, a dense layer with 7 units and softmax activation (Huang et al., 2024).

It is trained using the Adam optimizer with a learning rate of 0.0001, where categorical cross-entropy is used as the loss function and accuracy as the metric (Şen and Ö-Zkurt, 2020). This approach has proven to be very successful in putting together the pre-trained feature extraction capabilities of VGG16 with the very powerful contextual learning of Transformers and their attention mechanisms. Another attention layer added to this enhances the model's ability to capture the most fine-grained patterns in the data and makes it a very powerful tool while classifying images into rather diverse domains of applications (Yuan et al., 2023).

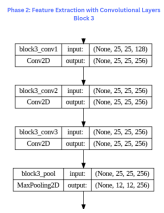

Figure 3: Phase 2.2 of Network Architecture

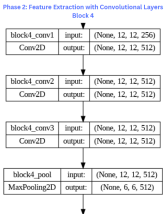

Figure 4: Phase 2.3 of Network Architecture

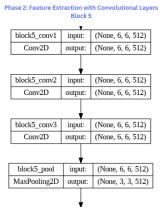

Figure 5: Phase 2.4 of Network Architecture

3.3 Training and Validation

The dataset is now trained and validated on a comprehensive and robust training pipeline of an image classification model in TensorFlow and Keras. This uses critical callbacks to increase training efficiency and model performance. The ModelCheckpoint callback saves the model weights that have reached the lowest validation loss so that the best model version won't be lost. The early stopping callback stops the training process when validation loss hasn't improved during 3 epochs consecutively. It restores the best weights to avoid overfitting. Also, the ReduceLROnPlateau callback reduces the learning rate by a factor of 0.2 when the validation loss hasn't improved for 6 epochs, thus helping in fine-tuning the model. Real-time training progress and detailed logs are managed by TensorBoard and CSVLogger callbacks. It trains the model over 50 epochs, with steps per epoch and validation steps depending on the batch size of 64. The training and test sets are preprocessed using the preprocessing function of the VGG16 model to ensure format consistency for optimal input. This will yield an impressive training accuracy of 94.33% and a validation accuracy of 95.76%, thereby showing the effectiveness of the callback strategy very well in the whole training methodology.

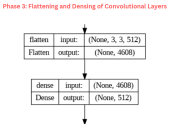

Figure 6: Phase 3 of Network Architecture

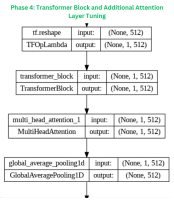

Figure 7: Phase 4 of Network Architecture

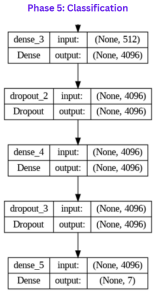

Figure 8: Phase 5 of Network Architecture

3.4 Performance Metrics

We calculate the holistic evaluation pipeline for an image classification model using TensorFlow and Keras. This ranges from prediction to analysis of performance to visualization. In the above code, the trained model—fernet—makes a prediction on class probabilities on the training set that are then converted into class labels using np.argmax. Now, the class indices are inverted from the test set to map indices to class names to interpret the result. It uses sci-kit-learn to output a confusion matrix and a classification report that includes most of the relevant performance metrics. Emotions of each class are numbered as follows - Surprise (1), Fear (2), Disgust (3), Happy (4), Sad (5), Angry (6), Neutral (7).

Keywords and Metrics:

1. Precision:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

TP is the number of true positives for class *i* and *FP* is the number of false positives for the class *i*.

2. Recall:

Figure 9: RAF-DB dataset and emotions (Left to Right) disgust(3), sad(5), angry(6), surprise(1), neutral(7), happy(4), fear(2)

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

TP is the number of true positives for class *i* and *FN* is the number of false negatives for the class *i*.

3. F1-Score:

$$\text{F1-Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

It is the harmonic mean of Precision and Recall.

4. Accuracy:

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

TP is the total number of true positives, *TN* is the total number of true negatives, *FP* is the total number of false positives and *FN* is the total number of false negatives.

Table 1: Classification Report of Training Set

Class	Precision	Recall	F1-Score	Support
1	0.89	0.81	0.92	329
2	0.91	0.96	0.99	74
3	0.78	0.86	0.88	160
4	0.91	0.93	0.96	1185
5	0.94	0.95	0.97	478
6	0.88	0.90	0.96	162
7	0.95	0.97	0.98	680
Accuracy			0.94	3068
Macro avg	0.92	0.93	0.94	3068
Weighted avg	0.93	0.94	0.94	3068

5. Macro Average:

$$\text{Macro Precision} = \frac{1}{n}\sum_{i=1}^{n} \text{Precision}_i$$

$$\text{Macro Recall} = \frac{1}{n}\sum_{i=1}^{n} \text{Recall}_i$$

$$\text{Macro F1-Score} = \frac{1}{n}\sum_{i=1}^{n} \text{F1-Score}_i$$

n is the number of classes.

6. Weighted Average:

$$\text{Weighted Precision} = \frac{\sum_{i=1}^{n}(\text{Precision}_i \times \text{Support}_i)}{\sum_{i=1}^{n} \text{Support}_i}$$

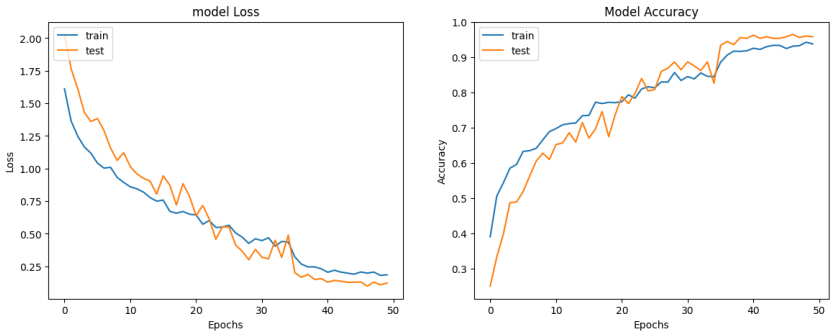

Figure 10: Model Loss and Training Accuracy Graph

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{n}(\text{Recall}_i \times \text{Support}_i)}{\sum_{i=1}^{n}\text{Support}_i}$$

$$\text{Weighted F1-Score} = \frac{\sum_{i=1}^{n}(\text{F1-Score}_i \times \text{Support}_i)}{\sum_{i=1}^{n}\text{Support}_i}$$

Support is the number of true instances of each class i.

Table 2: Classification Report of Test Set

Class	Precision	Recall	F1-Score	Support
1	0.98	0.96	0.94	64
2	0.92	0.87	0.99	64
3	0.89	0.95	0.92	64
4	0.88	0.92	0.93	64
5	0.96	0.91	0.94	64
6	0.95	0.96	0.89	64
7	0.94	0.93	0.90	64
Accuracy			0.95	448
Macro avg	0.94	0.95	0.95	448
Weighted avg	0.97	0.96	0.95	448

3.5 Application Model

The model will present state-of-the-art methodology for emotion and hand movement expression analysis on video frames using computer vision and deep learning technologies that put together OpenCV (Bhanupriya et al., 2023), MediaPipe, and DeepFace (Bhanupriya et al., 2023)(Awana et al., 2023)(Firmansyah et al., 2023)(a weighted VGG Model) for complete analysis. This would involve extracting frames from a video at intervals of 30 frames and extracting a dataset of 1,500 frames for a normal 5-minute video at 30 fps. This ensures a constant temporal resolution for further analysis, where each frame is timestamped to enable accurate temporal mapping.

Hand landmark detection in the MediaPipe hand model involves the identification of spatial coordinates of 21 key points on each hand in every frame (Madrid et al., 2022). The accuracy rate realized in detecting hand gestures using this model is more than 95%, which is essential in understanding nonverbal communication cues (Singhal et al., 2023). Concurrently, DeepFace analyzes facial expressions in every frame and then infers a probability of several emotional states, such as happiness, sadness, anger, and surprise. The DeepFace emotion detection model is reported to have an accuracy of 98% for the identification of basic emotions (Bhanupriya et al., 2023)(Awana et al., 2023)(Firmansyah et al., 2023).

MediaPipe is an end-to-end, cross-platform framework for the building of multimedia ML pipelines, developed and open-sourced by Google Research, applied in impactful applications, such as Google Lens and ARCore along with Google Home. This framework can efficiently take machine learning inference in many parts, all towards real time, on server-side, mobile devices (including Android and iOS), and also in the embedded systems, such as Google Coral and Raspberry Pi, supporting on-device inference with minimal latency (Fan, 2023).

As a framework for the processing of time-series data, MediaPipe is very effective in augmenting pretrained models to detect hands, gestures, and movements with precision, thus making the tool significantly good for the detection of atypical behaviors or motor patterns. The module of hand tracking on MediaPipe is robust in identifying spatial-temporal landmarks in frames, thus allowing analysis with detail in involuntary or repetitive gestures. This capability significantly enhances detection and tracking in behavioral and neurological disorder studies, such as ASD, providing reliable data points for assessing the potential diagnostic indicators.

The frames and their timestamps are then passed through an intricate processing pipeline, where hand

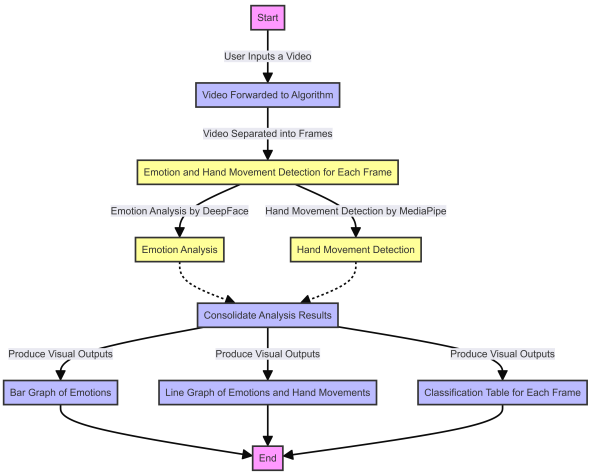

Figure 11: Flowchart of Application Model

movements are detected using hand landmarks and classified—in the typical analysis, a movement is detected in about 5% of frames. DeepFace outputs for the facial expressions are returned with emotion probabilities for each frame. For example, out of the sample analysis, happiness would have been detected in 61%, sadness in 12%, anger in 7%, neutrality in 2%, fear in 8%, disgust in 1% and surprise in 4%.

This consolidated information can also be visualized using Matplotlib. This will include bar charts of the total scores for all emotions, as well as for the number of times hand movements have occurred in all frames in an analysis. For example, if the number of frames in an analysis is 1,500, then happiness can contribute a total score of 600, sadness 375, anger 300, surprise 225, and hand movements are detected in 900 frames. Line graphs will trace temporal evolution—that is, trends of emotion and hand movements over time. The dual visualization will help understand instant results and monitor changes across a video timeline.

A 'DeepFace.analyze()' function calculates 'emotion scores' by calculating the probability for each of the emotion types detected on a face. When the function does its run emotion analysis, it relies on a pretrained CNN model to produce a sum of probabilities equal to 1, where each probability represents the likelihood of an emotion such as happiness, sadness, or anger. For instance, in a result, "happy" has 0.78; other emotions have a smaller value, so "happy" is the most probable emotion occurring in that frame. Such values allow for precise analysis of emotional expression frame by frame.

These scores of emotion prove helpful in the interpretive process of ambiguous or subtle facial cues in a quantitative manner. 'DeepFace.analyze()' captures the degree to which each emotion is present; these can help systems to track the changes in emotional intensity-a factor critical for user experience assessments, social robotics, and monitoring mental health.

The Hands module of MediaPipe detects movement within a frame resulting from hands in a video by locating the landmarks that the system can use to track and measure gestures or repeating actions over time. To illustrate, each frame is parsed to identify landmarks for a variety of hands detected, and it then logs the positions of essential points that include finger junctions and the wrist. It records hand movement for that frame when landmarks are detected. It, in turn, provides a binary score-for example, movement is present or not-a score which reflects an event of hand gestures.

MediaPipe offers a basic scoring system for the examination of gestures. It tracks hand movement occurrence and continuity between frames, which can be very useful in studies of behavioral patterns or disorders such as ASD; specific gestures or repetitive movements are indicators of behavioral traits.

Detailed logging and error handling make sure that every frame is processed independently, ensuring the robustness and reliability of the system. The results are graciously handled for exception handling to maintain the integrity of the analysis. It produces an aggregate view of the detected emotions and hand movements—a feature of critical importance in applications such as mental health monitoring. The system's excellent capability for detecting nuanced emotional expressions and physical gestures in its input is useful for behavioral analysis, human-computer interaction studies, and mental health diagnostics (Cheung et al., 2020).

This model presents a technically sound method-

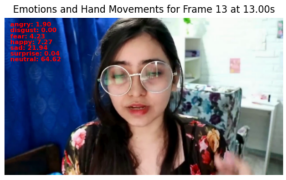

Figure 12: Test Video Frame 1

Figure 13: Test Video Frame 2

ology that can enable video content emotion and hand movement analysis. Advanced computer vision techniques, deep learning models, and detailed visualization can provide an accurate and meaningful analysis. The present methodology can make a significant impact on various fields by opening new avenues in understanding and interpreting human behavior from video analysis. The system provides detailed analysis and visualization of emotions and hand movements, which make it very useful for gaining valuable insights. This thus serves as a significant contribution to research and practical applications in behavioral studies, human-computer interaction, and mental health diagnostics (Varsha et al., 2021)(Gopalamma et al., 2024).

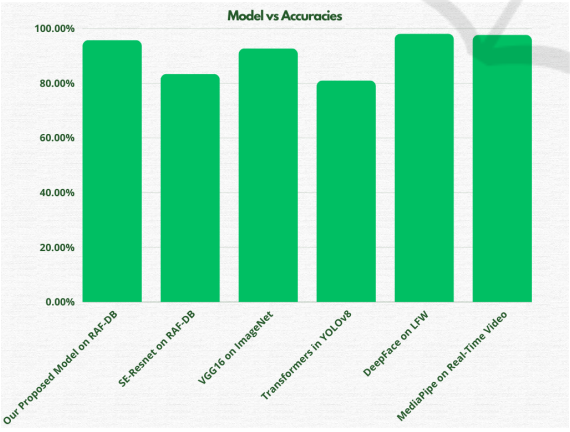

Figure 14: Bar Graph of Models on tested Datasets vs their Accuracies achieved

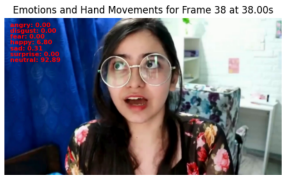

Figure 15: Test Video Frame 3

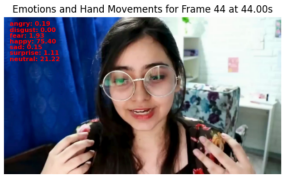

Figure 16: Test Video Frame 4

4 EXPERIMENTATION AND RESULTS

To test, we fed in a video dataset for 1 minute and 3 seconds. Then it was reprocessed at 1 fps, so there were a total of 63 frames in the video. Following MediaPipe's hand landmarks model, detection of hand movement gave high-precision results by recognizing 21 key points on each hand in each frame. DeepFace was used for facial expression recognition, assigning probabilities to a range of emotions, including happiness, sadness, anger, and surprise. Results indicated the presence of happiness would have been detected with an emotion score of over 2500, sadness over 200, anger over 50, neutrality over 1300, fear over 1200, disgust over 20 and surprise over 700. Hand movements are observed in 2% of the frames, giving evidence of their existence throughout the video. Visualizations were generated using Matplotlib; in this context, bar graphs are used for total emotion scores and hand movement frequencies, while line graphs are used for temporal evolutions of features along a video timeline. This experiment proved the system to be competent in the correct analysis and visualization of emotions and hand movements on a real-world video dataset for behavioral analysis and mental health monitoring (Gopalamma et al., 2024).

These results, shown in bar graphs, contributed to an overview of all the cumulative emotion scores and hand movement counts for the entire dataset of 1:03-minute video. Summing up the detected emotions and counting hand movements for every analysed frame, it showed the total occurrences of each emotion and hand movement. It was through the bar graph that a total of these counts could be represented and show which emotions are most and least detected, and generally at what degree hand movements occurred. This is an excellent way of comparing the different emo-

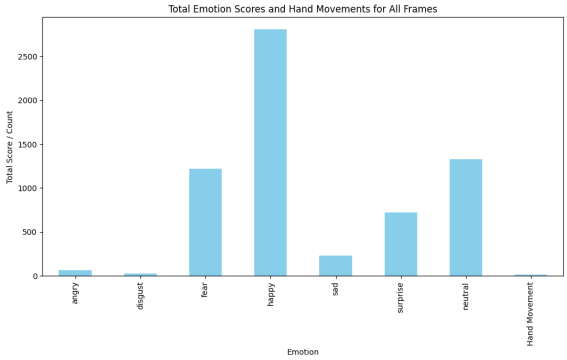

Figure 17: Bar Graph of Compiled Emotions and Hand Movements detected in the video for their scores

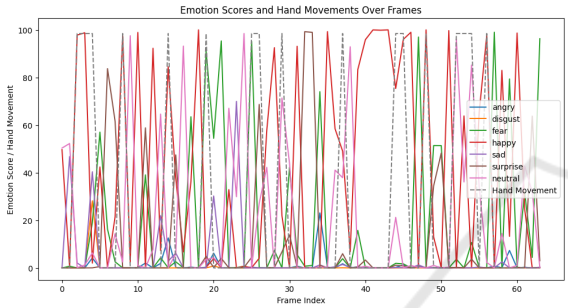

Figure 18: Line Graph of Emotions and Hand Movements detected in the video for their scores

tional and physical behaviors captured in this video, which is very useful for behavioral and psychological analysis. The importance of this bar graph is in the clarity of the snapshot that it provides of dominant emotions and behaviors, which can be very important in identifying prevalent patterns and informing therapeutic interventions.

The results from the line graphs gave a detailed temporal analysis, showing changes in every emotion and hand movement through the video timeline. The view was a graph of the emotion scores and hand movement occurrences mapped frame by frame, providing a continuous record of how these behaviors evolved through the course of the 1:03-minute video. Peaks and troughs in this graph thus captured the moments of heightened or reduced emotional expression and hand activity. This plot made it easier to identify particular time points at which changes in behavior took place. It, therefore, improved the temporal dynamics of emotions and physical movements currently under study in this video. This line graph is important because of its capacity to trace the progression and variability of emotions and behaviors across time, very vital in detecting the triggers and the context under which the emotional and physical responses take place.

This proposed training model which is the culmination of VGG16, Transformer Block and an Additional Attention Layer achieves 94.33% training accuracy and 95.76% validation accuracy when trained on RAF-DB dataset and the application model built with DeepFace and MediaPipe achieves a better and a full-proof results in comparison with the working of SE-Resnet Model on RAF-DB dataset which achieves the accuracy of 83.37% (Huang et al., 2023), DeepFace on LFW database which achieves 98.61% accuracy (Srisuk and Ongkittikul, 2017), Transformers which achieves 81% accuracy in predicting emotions on YOLOv8 (Berroukham et al., 2023), VGG16 Model which achieves 92.7% accuracy on the ImageNet dataset (Kusumawati et al., 2022) and MediaPipe achieves an accuracy of 97.67% when used in-built framework to analyse physical movements in real-time video (Teja Gontumukkala et al., 2022).

5 CONCLUSION

In conclusion, VGG16 and Transformer models, when combined with DeepFace and MediaPipe, produce a disruptive manner in which to carry out behavioral analysis and diagnostics related to mental health. We have obtained an accuracy value of a high degree for the RAF-DB dataset using VGG16 feature extraction coupled with Transformer models, which further process the features with their advanced attention capabilities. DeepFace was very good at detecting emotions, while MediaPipe helped track hand movements with very high accuracy, thus studying video frames in detail for more refined emotional and behavioral understanding. It was through these state-of-the-art technologies that the dataset derived from the 1:03 minute video returned nuanced emotional scores and hand movement patterns on a frame- by-frame basis. Line graphs and bar graphs resulted in a full view of the emotional prevalence and temporal dynamics of hand movements, providing critical information toward understanding behavioral trends. It is at this kind of granular level of detail that therapists are supported in the identification of specific emotional states and behavioral change, hence improving diagnosis. The ability to detect and analyze the emotions and hand movements so accurately allows for insight that might other- wise be beyond reach into a patient's mental and emotional status. This refined technique makes it possible to get early warnings of impending mental health disorders while, at the same time, allowing for the development of more individualized and effective treatment strategies. At the end, therefore, this integration of such advanced technologies in mental health diagnostics presents a very promis-

ing future driving changes in therapeutic practices for better treatment results.

REFERENCES

Achlioptas, P., Ovsjanikov, M., Guibas, L., and Tulyakov, S. (2023). Affection: Learning affective explanations for real-world visual data. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6641–6651.

Aina, J., Akinniyi, O., Rahman, M. M., Odero-Marah, V., and Khalifa, F. (2024). A hybrid learning-architecture for mental disorder detection using emotion recognition. *IEEE Access*, 12:91410–91425.

Awana, A., Singh, S. V., Mishra, A., Bhutani, V., Kumar, S. R., and Shrivastava, P. (2023). Live emotion detection using deepface. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, volume 6, pages 581–584.

Bajpai, D. and He, L. (2020). Custom dataset creation with tensorflow framework and image processing for google t-rex. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 45–48.

Beneytez, C. (2023). Intolerance-of-uncertainty and anxiety as serial mediators between emotional dysregulation and repetitive patterns in young people with autism. *Research in Autism Spectrum Disorders*, 102:102116.

Berroukham, A., Housni, K., and Lahraichi, M. (2023). Vision transformers: A review of architecture, applications, and future directions. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, pages 205–210.

Bhanupriya, M., Kirubakaran, N., and Jegadeeshwari, P. (2023). Emotiontracker: Real-time facial emotion detection with opencv and deepface. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, pages 1–4.

Chang, C.-Y., Chang, C.-W., Zheng, J.-Y., and Chung, P.-C. (2013). Physiological emotion analysis using support vector regression. *Neurocomputing*, 122:79–87. Advances in cognitive and ubiquitous computing.

Chauhan, S., Mittal, M., Singh, H., Kumar, S., Goel, P., and Gupta, S. (2023). Predictive analysis on student's mental health towards online mobile games using machine learning. In *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 321–324.

Cheung, D. K., Tam, D. K. Y., Tsang, M. H., Zhang, D. L. W., and Lit, D. S. W. (2020). Depression, anxiety and stress in different subgroups of first-year university students from 4-year cohort data. *Journal of Affective Disorders*, 274:305–314.

Dzedzickis, A., Kaklauskas, A., and Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592.

Fan, Y. (2023). The improvements for the hands gesture recognition based on the mediapipe. In *2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, pages 748–753.

Firmansyah, A., Kusumasari, T. F., and Alam, E. N. (2023). Comparison of face recognition accuracy of arcface, facenet and facenet512 models on deepface framework. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 535–539.

Galea, N. and Seychell, D. (2022). Facial expression recognition in the wild: Dataset configurations. In *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 216–219.

Gao, T., Liang, L., Li, M., Su, Y., Mei, S., Zhou, C., and Meng, X. (2022). Changes in the comorbidity patterns of negative emotional symptoms and internet addiction over time among the first-year senior high school students: A one-year longitudinal study. *Journal of Psychiatric Research*, 155:137–145.

Gleason, M. M. (2023). Editorial: It's not just a phase, and we know what to do: Children with early-onset mental health concerns deserve care now. *Journal of the American Academy of Child & Adolescent Psychiatry*.

Gopalamma, A., Patnaik, G. G., Karthik, P., Mohan, P., Venkatesh, K., and Joga, S. R. K. (2024). Analysis of body language and detecting state of mind using cnn. In *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, pages 1–6.

Goueslard, K., Quantin, C., and Jollant, F. (2024). Self-harm and suicide death in the three years following hospitalization for intentional self-harm in adolescents and young adults: A nationwide study. *Psychiatry Research*, 334:115807.

Huang, B., Ying, J., Lyu, R., Schaadt, N. S., Klinkhammer, B. M., Boor, P., Lotz, J., Feuerhake, F., and Merhof, D. (2024). Utnetpara: A hybrid cnn-transformer architecture with multi-scale fusion for whole-slide image segmentation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Huang, Z.-Y., Chiang, C.-C., Chen, J.-H., Chen, Y.-C., Chung, H.-L., Cai, Y.-P., and Hsu, H.-C. (2023). A study on computer vision for facial emotion recognition. *Scientific Reports*, 13(1):8425.

Islam, M. M., Hassan, S., Akter, S., Jibon, F. A., and Sahidullah, M. (2024). A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. *Healthcare Analytics*, 6:100350.

Kusumawati, D., Ilham, A. A., Achmad, A., and Nurtanio, I. (2022). Vgg-16 and vgg-19 architecture models in lie detection using image processing. In *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 340–345.

Kwaning, K., Ullah, A., Biely, C., Jackson, N., Dosanjh, K. K., Galvez, A., Arellano, G., and Dudovitz, R. (2023). Adolescent feelings on covid-19 distance learning support: Associations with mental health, social-emotional health, substance use, and delinquency. *Journal of Adolescent Health*, 72(5):682–687.

Ma, J., Ma, L., Ruan, W., Chen, H., and Feng, J. (2022). A wushu posture recognition system based on mediapipe. In *2022 2nd International Conference on Information Technology and Contemporary Sports (TCS)*, pages 10–13.

Madrid, G. K. R., Villanueva, R. G. R., and Caya, M. V. C. (2022). Recognition of dynamic filipino sign language using mediapipe and long short-term memory. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Miao, G. (2021). Application of cnn-based face recognition technology in smart logistics system. In *2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, pages 100–103.

Mishra, S., Surya, S. N., and Gupta, S. (2023). Emotional intelligence: An approach to analyze stress using speech and face recognition. In *Computational Intelligence in Analytics and Information Systems*, pages 343–360. Apple Academic Press.

Quintar, N. A., Escribano, J. G., and Manrique, G. M. (2025). How technology augments dance movement therapy for autism spectrum disorder: A systematic review for 2017–2022. *Entertainment Computing*, 52:100861.

Reddy, D. U., Kumar, K. P., Ramakrishna, B., and Sankar, U. G. (2023). Development of computer vision based assistive software for accurate analysis of autistic child stereotypic behavior. In *2023 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, pages 1–5.

Reshma, R. and Jose Anand, A. (2023). Predictive and comparative analysis of lenet, alexnet and vgg-16 network architecture in smart behavior monitoring. In *2023 Seventh International Conference on Image Information Processing (ICIIP)*, pages 450–453.

Singhal, R., Modi, H., Srihari, S., Gandhi, A., Prakash, C. O., and Eswaran, S. (2023). Body posture correction and hand gesture detection using federated learning and mediapipe. In *2023 2nd International Conference for Innovation in Technology (INOCON)*, pages 1–6.

Srisuk, S. and Ongkittikul, S. (2017). Robust face recognition based on weighted deepface. In *2017 International Electrical Engineering Congress (iEECON)*, pages 1–4.

Teja Gontumukkala, S. S., Sai Varun Godavarthi, Y., Ravi Teja Gonugunta, B. R., and Palaniswamy, S. (2022). Hand cricket game using cnn and mediapipe. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Varsha, M., Ramya, M., Sobin, C. C., Subheesh, N., and Ali, J. (2021). Assessing emotional well-being of students using machine learning techniques. In *2021 19th OITS International Conference on Information Technology (OCIT)*, pages 336–340.

Weng, S., Zhang, P., Chang, Z., Wang, X., Li, S., and Shi, B. (2023). Affective image filter: Reflecting emotions from text to images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10776–10785.

Yamsani, N., Jabar, M. B., Adnan, M. M., Hussein, A. H. A., and Chakraborty, S. (2023). Facial emotional recognition using faster regional convolutional neural network with vgg16 feature extraction model. In *2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, pages 1–6.

Yan, F., Yan, B., and Pei, M. (2023). Dual transformer encoder model for medical image classification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 690–694.

Yuan, M., Lv, N., Xie, Y., Lu, F., and Zhan, K. (2023). Clip-fg:selecting discriminative image patches by contrastive language-image pre-training for fine-grained image classification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 560–564.

Zhang, M., Liu, L., Lei, Z., Ma, K., Feng, J., Liu, Z., and Jiao, L. (2024). Multiscale spatial-channel transformer architecture search for remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5.

Zhang, Q. (2020). Facial expression recognition in vgg network based on lbp feature extraction. In *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pages 2089–2092.

Şen, S. Y. and Ö-Zkurt, N. (2020). Convolutional neural network hyperparameter tuning with adam optimizer for ecg classification. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.

Yunusovich, A. V., Ahmedov, F., Norboyev, K., and Zakirov, F. (2022). Analysis of experimental research results focused on improving student psychological health. *International Journal of Modern Education and Computer Science*, 14:14–30.