Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 3, Issue 2, 2025

LEVERAGING MACHINE LEARNING FOR ENHANCED DETECTION OF BIPOLAR DISORDER: A NEW FRONTIER IN MENTAL HEALTH DIAGNOSTICS

Farhat Jabeen Badar[*1] **Muhammad Zaman**[2] **Muhammad Shahid Mehmood**[3]

[*1,2,3]*Department of Computer Science & IT, Superior University, 10 KM Lahore- Sargodha Rd, Sargodha, Punjab 40100, Pakistan*

[*1]Naseeriamehria@gmail.com, [2]mzamancui@gmail.com, [3]shahidkhan.ak16@gmail.com

DOI: https://doi.org/10.5281/zenodo.14949798

Keywords
Machine Learning, Bipolar Disorder, Mental Health Diagnostics, AI Detection, Predictive Analytics.

Article History
Received on 20 January 2025
Accepted on 20 February 2025
Published on 28 February 2025

Copyright @Author
Corresponding Author: *

Abstract

Mental health disorders have become a critical global concern, affecting individuals across all demographics. Early diagnosis and accurate detection of these conditions are essential for effective intervention, as delayed identification can result in severe consequences, including self-harm and fatal outcomes. This study introduces a novel approach for analyzing facial expressions using the AffectNet and 2013 Facial Emotion Recognition (FER) datasets. Unlike conventional diagnostic methods, this research develops an advanced system that compiles an extensive mental health dataset and predicts mental health conditions based on facial emotional cues.

A distinctive hybrid framework is presented, incorporating the state-of-the-art YOLOv10 object detection algorithm to recognize and classify visual indicators linked to specific mental disorders. To enhance predictive performance, the system employs an ensemble learning model that combines Convolutional Neural Networks (CNNs) with Visual Transformers (ViT), achieving an accuracy of approximately 81% in detecting mental health conditions such as depression and anxiety.

To ensure interpretability and reliability, this study integrates Gradient-weighted Class Activation Mapping (Grad-CAM) and saliency maps to identify significant facial regions influencing model predictions. These insights enhance transparency, allowing healthcare professionals to understand the reasoning behind the system's outputs, thereby fostering confidence in the decision-making process.

A flowchart outlines the mental disorder detection pipeline, beginning with data collection and preprocessing, which then branches into training and testing datasets. Both datasets are processed through a YOLOv10 model, followed by a logic module that links to a dedicated mental disorder dataset. This dataset feeds into an ensemble model consisting of MDNet, ViT, and ResNet50. The final output is refined through Explainable AI (XAI) techniques, ensuring a transparent and interpretable prediction process.

INTRODUCTION

Mental health disorders significantly impact cognitive function, emotional stability, and behavior. These conditions, including anxiety, depression, bipolar disorder, schizophrenia, and eating disorders, pose a growing public health challenge worldwide. Reports indicate a substantial increase in mental health issues, with cases of depression and anxiety rising sharply over the past decade. Early and precise diagnosis is essential for effective intervention, as mental disorders often cause considerable distress and impair daily functioning.

Behavioral indicators such as facial expressions, eye gaze, and head movements provide valuable insights into an individual's mental state. Advances in deep learning (DL) and computer vision have opened new possibilities for automating mental health assessments. CNNs have demonstrated exceptional capabilities in analyzing large datasets, making them effective for applications such as medical diagnostics and object detection. However, deep learning models are often criticized for their lack of interpretability, which raises concerns about their practical use in healthcare.

This research proposes an explainable artificial intelligence (XAI) approach for detecting mental disorders through facial analysis. The developed system not only predicts mental health conditions but also provides clear justifications for its decisions, enabling healthcare professionals to make well-informed diagnoses. The framework employs a hybrid ensemble deep learning model combining CNNs and Vision Transformers (ViT) while utilizing Grad-CAM and saliency maps for interpretability. The system is trained on a specialized dataset and evaluated using accuracy, precision, recall, F1-score, and ablation studies to validate its effectiveness.

Key Contributions:

1. A pioneering approach to facial expression analysis using the AffectNet and FER 2013 datasets, marking the first instance of integrating multiple datasets for mental health prediction.
2. Development of an innovative mental disorder detection pipeline emphasizing explainability and interpretability.
3. Creation of a dedicated mental health dataset based on facial emotional markers.
4. Introduction of a hybrid deep learning model incorporating object detection and ensemble learning techniques to enhance mental disorder predictions.

Paper Structure:

- **Section I:** Introduction to mental disorders and advancements in detection technology.
- **Section II:** Review of related work and contributions of this study.
- **Section III:** Description of datasets, methodology, feature extraction techniques, and learning modules.
- **Section IV:** Experimental setup, evaluation metrics, and performance analysis.
- **Section V:** Discussion of findings, conclusions, and future research directions.
- **Section VI:** Summary of insights and recommendations for further research.

Related Work:

Several studies have explored AI-based approaches for detecting mental disorders using facial emotion recognition. Research by various scholars has utilized machine learning models, including Decision Trees, Random Forests, and Artificial Neural Networks, to analyze facial expressions captured through webcams or images. While these studies demonstrated promising results, their effectiveness was often limited by dataset size and diversity.

Recent studies have leveraged deep learning models such as CNNs to classify emotions into distinct categories, predicting mental states based on facial features. Some researchers have introduced hybrid models combining feature extraction techniques like Local Binary Patterns (LBP) with classifiers such as Support Vector Machines (SVM) to enhance accuracy. Others have integrated multimodal approaches, fusing facial, speech, and neurophysiological data to improve detection performance.

Advanced architectures such as VGG, ResNet, and Inception have been employed to differentiate between depressed and non-depressed individuals, demonstrating the efficacy of CNN-based methods. Additionally, researchers have applied Facial Action Coding Systems (FACS) and region-based CNNs to focus on critical facial areas, refining emotion detection accuracy. Several studies have also explored real-time video analysis for mental health monitoring, highlighting the potential of deep learning in clinical applications.

Despite these advancements, the lack of interpretability remains a major challenge in AI-driven mental health diagnostics. This study addresses this gap by integrating explainability techniques, ensuring that model decisions are transparent and understandable to healthcare professionals.

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

Materials and Methods

Facial expressions serve as vital indicators for diagnosing mental disorders, forming the foundation of this study. The methodology employed is outlined in Figure 1. The research involved acquiring facial expression datasets, preprocessing them, partitioning data into training and testing sets, training models, and evaluating their efficacy. The state-of-the-art YOLOv10 model was utilized for initial training and later adapted for mental disorder detection by analyzing the top two predicted classes. This process led to the formation of a dataset categorizing mental disorders into four distinct groups: anxiety disorder, depressive disorder, no disorder, and other disorders. To enhance classification accuracy, an ensemble model was formulated by integrating three learning architectures: a custom CNN model termed MDNet, a Vision Transformer (ViT), and a pre-trained ResNet50. Each of these models was trained separately on the mental disorder dataset, and their performance was assessed against the proposed ensemble framework. Additionally, ablation studies were conducted to determine the most and least significant components contributing to the ensemble's performance.

To improve the interpretability of model decisions, Grad-CAM and saliency maps were employed to highlight facial regions influencing predictions. Specifically, the Vanilla Saliency method was used to generate saliency maps, as it effectively emphasizes image-specific features associated with different mental disorder classes. This method aligns well with mental disorder analysis, given the variability of these conditions across individuals. The TensorFlow debugging toolkit for Keras models was implemented for this visualization technique. The subsequent sections provide a detailed breakdown of the analysis methodology.

A. Development of the Study's Database

This research is based on two key datasets, selected for their suitability in constructing a robust and accurate mental disorder detection model. The datasets utilized are AffectNet and FER 2013, each offering unique benefits to the study.

AffectNet comprises a vast collection of facial expression images sourced from diverse real-world scenarios. It includes over one million images labeled with seven primary emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset encompasses various lighting conditions, facial orientations, and demographic backgrounds, making it an ideal choice for training deep learning models to recognize emotions in real-world settings. Each image in AffectNet features a single face with variations in resolution, dimensions, and aspect ratios. For consistency in training, all images were resized to 64×64 pixels.

CLASS	TRAIN	TEST
Angry	4800	950
Disgust	4100	1000
Fear	3990	1020
Positive	3430	930
Sad	3340	1250
Total	19660	5150

B. YOLOv10 Model

YOLOv10 is a state-of-the-art real-time object detection model introduced by researchers from Tsinghua University. It builds upon previous YOLO versions by addressing limitations in post-processing and model architecture. A significant innovation in YOLOv10 is the elimination of Non-Maximum Suppression (NMS) through consistent dual assignments, enabling end-to-end deployment with reduced inference latency. The model also incorporates a holistic efficiency-accuracy driven design strategy, optimizing various components to reduce computational overhead while enhancing performance. Extensive experiments have demonstrated that YOLOv10 achieves superior accuracy and efficiency across multiple model scales. For instance, YOLOv10-S is 1.8 times faster than RT-DETR-R18 with comparable Average Precision (AP) on the COCO dataset, while YOLOv10-B offers a 46% reduction in latency and 25% fewer parameters compared to YOLOv9-C, maintaining similar performance levels.

1) YOLOv10 Architecture

YOLOv10 (You Only Look Once version 10) is an advanced real-time object detection model that improves upon previous YOLO versions by enhancing efficiency, accuracy, and speed. The key innovations in YOLOv10 focus on eliminating redundant post-processing, optimizing model architecture, and improving feature extraction.

Key Components of YOLOv10 Architecture

1. **Backbone Network (Feature Extraction)**

o The backbone is responsible for extracting key features from input images.

o YOLOv10 employs a highly optimized **lightweight CNN-based backbone** that balances computational efficiency and accuracy.

o It utilizes **efficient convolutional layers**, **ResNet-like blocks**, and **attention mechanisms** to enhance feature representation.

2. **Neck (Feature Enhancement & Aggregation)**

o The neck structure refines and enhances features extracted by the backbone.

o It uses **Path Aggregation Network** (PAN) and **Feature Pyramid Network** (FPN) structures to improve multi-scale detection.

o These mechanisms help detect objects at different scales by fusing low-level and high-level features.

3.Head (Prediction & Output)

o The prediction head in YOLOv10 generates bounding boxes, class scores, and objectness scores.

o Unlike previous YOLO versions, YOLOv10 eliminates **Non-Maximum Suppression (NMS)** by introducing a **consistent dual assignment strategy**.

o This allows the model to directly predict object locations without additional post-processing, improving inference speed and efficiency.

4.Anchor-Free Object Detection

o YOLOv10 moves towards an **anchor-free detection approach**, reducing computational overhead.

o This improves model performance on small and overlapping objects.

5.Optimized Loss Function

o The loss function in YOLOv10 is improved to enhance localization and classification accuracy.

o It incorporates **IoU-based loss** and **adaptive weighting strategies** to balance different aspects of training.

Advantages of YOLOv10

- **Faster Inference**: Eliminates NMS, reducing post-processing time.

- **Higher Accuracy**: Improved backbone and feature aggregation boost precision.

- **Lightweight Design**: Optimized architecture reduces parameters and memory usage.

- **Scalable**: Supports different model sizes for various applications, from edge devices to large-scale deployments.

YOLOv10 is particularly useful in real-time applications like autonomous driving, security surveillance, and medical imaging due to its high efficiency and accuracy.

C. Predictive Logic and Mental Disorder Dataset

After completing the training of the YOLOv10 model on the study's dataset, the next critical step involved utilizing the model for predictive analysis. In this phase, the trained model was evaluated on the AffectNet dataset using a structured methodology to determine the presence or absence of mental disorders in individual images. Instead of relying solely on the highest prediction probability, we considered the top two predictions when the highest probability was below one.

Following a predefined classification framework, images were categorized based on their dominant emotional expressions. If the top two predictions included **fear and disgust**, the image was identified as indicative of **anxiety disorder**, as referenced in . Likewise, if **sadness and anger** were the predominant predictions, the image was classified as representing **depressive disorder**, as supported by Images lacking positive emotions were grouped under **other disorders**, whereas those displaying positive emotions were categorized as **"no disorder."**

Through this approach, a refined mental disorder dataset was compiled, consisting of **3,601 images**, ensuring a structured and accurate classification process for further analysis.

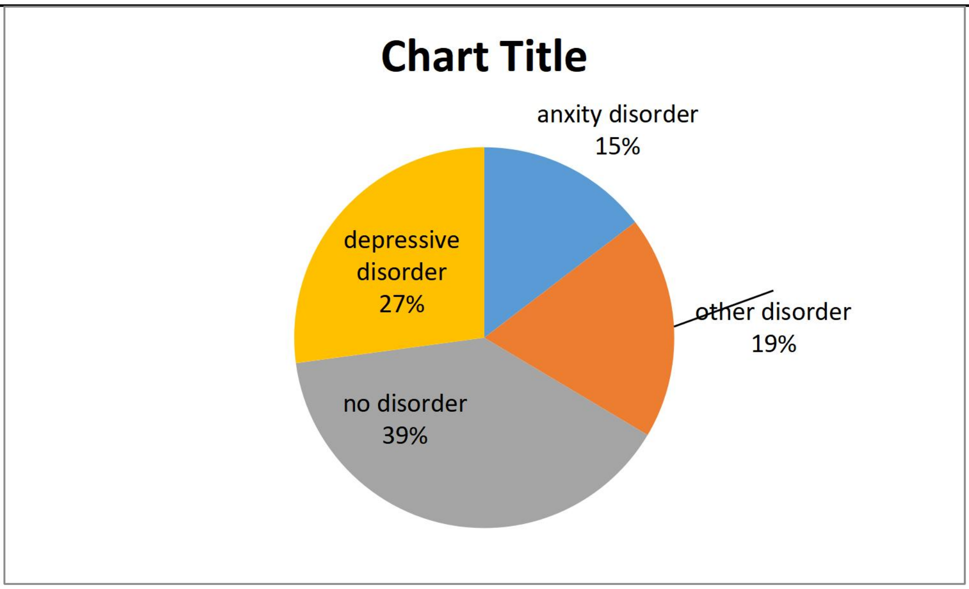

Image distribution in the mental disorder dataset

D. Ensemble Feature Extractors

To enhance the accuracy and precision of mental disorder detection, this study integrates features from three advanced deep learning models: MDNet, Vision Transformer (ViT), and a pre-trained ResNet50. These models are well-established for their proficiency in image classification. Each model underwent rigorous training on the mental disorder dataset, and their outputs were combined using a stacking-based fusion strategy to develop a robust ensemble classifier. By leveraging the distinct strengths of each model, this ensemble approach provides a more comprehensive framework for detecting mental disorders.

MDNet: A Custom CNN Model

The first component of the ensemble is MDNet, a customized convolutional neural network (CNN) specifically designed for feature extraction. CNNs play a crucial role in deep learning applications, particularly in image processing, as they consist of multiple layers that extract hierarchical features from input images. The CNN architecture includes convolutional layers that apply small filters (kernels) to detect fundamental patterns such as edges and textures, followed by pooling layers that reduce spatial dimensions while retaining critical information. Max pooling and average pooling are commonly employed techniques in this process. The final fully connected layer consolidates the extracted features to generate predictions.

For this study, MDNet was developed from scratch and optimized for enhanced performance. The optimal architecture was determined through experimentation, comprising four convolutional layers, four max-pooling layers, five dropout layers for regularization, and a fully connected layer. The total number of trainable parameters in MDNet is 261,188, ensuring a balance between computational efficiency and model effectiveness.

ResNet50: A Pre-Trained Model for Enhanced Feature Extraction

In addition to MDNet, the study incorporates ResNet50, a well-established deep CNN model pre-trained on the ImageNet dataset. ResNet50 features a 50-layer architecture and approximately 24 million trainable parameters. It addresses the vanishing gradient issue that often arises in deep CNNs by employing residual connections. These residual layers enable direct information flow across layers, facilitating more effective training of deep networks. The inclusion of ResNet50 allows the ensemble model to capture intricate patterns in facial expressions, improving classification accuracy.

Vision Transformer (ViT): A Transformer-Based Approach

The third model in the ensemble is ViT, a transformer-based architecture that represents a significant advancement in computer vision. Unlike CNNs, which rely on fixed-size receptive fields, ViT processes images by segmenting them into patches and utilizing a multi-layer transformer encoder with a self-attention mechanism. This approach enables the model to learn spatial relationships between different regions of an image, effectively capturing both local and global features.

The ViT model employed in this study comprises 64 layers and 21.6 million trainable parameters. Its key architectural components include:

- **Patch Embedding:** The input image is divided into fixed-size patches (e.g., 16×16 pixels), which are then transformed into fixed-dimension vectors, converting the image into a sequence of patches.

- **Positional Encoding:** Since transformers do not inherently maintain spatial order, positional encodings are incorporated to retain spatial relationships.

- **Transformer Encoder:** This core component consists of multiple layers of multi-head self-attention and feed-forward neural networks, enabling the model to understand long-range dependencies and capture global context.

- **Classification Head:** The final output of the transformer encoder is aggregated and processed through a fully connected layer to generate predictions.

ViTs have demonstrated strong performance in image analysis, particularly for tasks requiring a broader contextual understanding. While CNNs emphasize localized feature extraction, ViTs process the entire image holistically, making them highly effective in analyzing subtle variations in facial expressions.

Integration of CNNs and Transformers for Improved Performance

Combining CNNs with ViTs enables the model to leverage the advantages of both architectures. CNNs excel at extracting fine-grained local features, whereas ViTs are proficient in understanding global dependencies within an image. This hybrid approach enhances generalization, accuracy, and robustness across diverse conditions, such as variations in lighting, angles, and occlusions. In the proposed framework, CNN layers initially extract local features, which are subsequently processed by ViT to capture global dependencies, leading to a more comprehensive interpretation of facial expressions.

Recent advancements in deep learning have demonstrated the effectiveness of transformer-based models across various domains. To fine-tune the learning models, including MDNet and ViT, extensive experiments were conducted, and hyperparameter optimization was performed using GridSearchCV . The ensemble models were trained on the mental disorder dataset using fivefold cross-validation, ensuring robust evaluation and performance validation.

Experimental Results

The evaluation of the proposed framework was conducted using the dataset described in Section III-A. To assess the effectiveness of the models, various standard performance metrics were employed, including precision, recall, accuracy, and the F1 score. Both quantitative and qualitative analyses were carried out, incorporating the confusion matrix and Receiver Operating Characteristic (ROC) curves to provide a comprehensive assessment. These metrics are based on four fundamental components:

- **True Positives (TP):** Instances correctly classified as positive.
- **True Negatives (TN):** Instances correctly classified as negative.
- **False Positives (FP):** Negative instances incorrectly classified as positive (Type I error).
- **False Negatives (FN):** Positive instances incorrectly classified as negative (Type II error).

The performance metrics are defined as follows:

- **Precision:** This metric determines the proportion of correctly identified positive cases out of all predicted positives and is given by:
Precision=TPTP+FP = {TP}{TP + FP}Precision=TP+FPTP

- **Recall (Sensitivity or True Positive Rate):** It measures the proportion of actual positive cases that were correctly classified by the model, calculated as:
Recall=TPTP+FN = {TP}{TP + FNl=TP+FNTP
- **Accuracy:** This metric represents the overall proportion of correctly classified instances in the dataset, formulated as:
Accuracy=TP+TNTP+TN+FP+FN = {TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN
While accuracy is a key indicator of model performance, additional metrics are necessary when dealing with imbalanced datasets to ensure a fair evaluation.

- **F1 Score:** The F1 score is the harmonic mean of precision and recall, making it particularly useful when managing class imbalances or cases where false positives and false negatives hold varying levels of significance. It ranges from 0 to 1, where 1 signifies optimal performance. The F1 score is computed as:
F1-score=2×Precision×RecallPrecision+Recall{F1-score} = 2 {Precision} \times {Recall}{Precision} + {Recall}}F1-score=2×Precision+RecallPrecision×Recall
- **Standard Deviation:** This metric provides insights into the variation in predictions, where a lower standard deviation indicates consistent model performance, while a higher value may point to potential generalization or stability issues.
- **Confusion Matrix:** This visualization technique helps in analyzing the model's classification performance across individual classes, highlighting potential misclassifications and areas for improvement.
- **ROC Curve:** The ROC curve graphically represents the trade-offs between the true positive rate and the false positive rate, allowing for a comparative evaluation of classification models. It plays a crucial role in assessing the model's ability to differentiate between various classes.

The proposed model demonstrated efficient processing, with a total execution time of 1 minute and 13 seconds. On average, each image was processed in 0.011 seconds, considering both the YOLOv10 model and the ensemble classifier. All experiments were conducted on a high-performance workstation equipped with a 12th Gen Intel® Core™ i7-12700 processor (20 cores), 64 GB of RAM, 1.5 TB of storage, and an NVIDIA GeForce RTX 3060 GPU.

To enhance the robustness of the model and mitigate the risks of overfitting and selection bias, a five-fold cross-validation strategy was employed. This validation technique improves the generalizability of the model to new datasets. During each iteration, the dataset was split into an 80% training set and a 20% testing set, ensuring a balanced evaluation.

The experimental findings are outlined below. Initially, the performance of the YOLOv10 model in classifying emotional categories was examined. The precision, recall, and F1 scores for each category are reported in **Table 3**, Furthermore, **Table 4** presents a comparative analysis of YOLOv10 against other state-of-the-art models trained on the AffectNet dataset.

TABLE 3 YOLOv10 Performance Summary

CLASS	PRECESION	RECALL	F1-SCORE	STANDARD DEVIATION
Angry	0.63	0.68	0.65	1.56
Disgust	0.98	0.96	0.97	0.17
Fear	0.62	0.60	0.61	1.09
Positive	0.76	0.71	0.73	0.77
Sad	0.65	0.68	0.66	1.32
Total	0.72	0.73	0.72	1.29

TABLE 4 Comparison of YOLOv10 Against Other Techniques on the Dataset

Method	Accuracy (%)
gACNN[50]	58.78
IPA2LT[51]	55.71
RAN[52]	52.97
SCN[53]	60.23
DACL[54]	65.20
CNN[55]	56.54
POSTER[56]	67.31
KTN[57]	63.97
TransFER[58]	66.23
YOLOv10	88.00

In the second phase of the experiments, the performance of MDNet, pre-trained ResNet50, and ViT models was evaluated independently when trained on the mental disorder dataset. The average values of key evaluation metrics for each model are summarized in **Table 5**. These visualizations offer deeper insights into the classification effectiveness of each model across different categories.

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

Discussion

Mental health disorders represent a critical public health issue globally, affecting diverse populations and manifesting in various conditions such as depression and anxiety. These disorders are often influenced by daily life stressors, making early detection and accurate diagnosis essential for timely intervention. Delayed recognition can lead to severe consequences, including self-harm, suicidal tendencies, and loss of life. Recent advancements in machine learning and computer vision offer promising solutions to enhance diagnostic accuracy. This study aims to develop a comprehensive analytical framework for detecting mental health disorders while prioritizing transparency and interpretability. By leveraging facial expression analysis, the study employs AffectNet and FER 2013 datasets for model training and validation. The proposed system contributes to the development of a refined mental health disorder dataset and predicts mental health conditions based on facial emotional cues.

A hybrid learning architecture was implemented, integrating pre-trained models, Vision Transformer (ViT), and a shallow Convolutional Neural Network (CNN) into an ensemble classifier. The combined performance of these models yielded an overall accuracy of 78%. However, an ablation study revealed that the exclusion of ViT resulted in the highest accuracy of 81%, surpassing other configurations. These findings highlight the most effective components within the ensemble framework. With advancements in computational capabilities, a balance between model complexity and accuracy is achievable. While incorporating multiple models increases computational demands, experimental results indicate that enhanced hardware significantly improves training efficiency and predictive accuracy. The system's end-to-end testing time was recorded at approximately 60 ± 0.13 seconds. Transparency and interpretability remain fundamental aspects of this research, with techniques such as Grad-CAM and saliency maps utilized to visualize influential regions in input images. These visualization tools provide healthcare professionals with valuable insights into the decision-making process, thereby improving trust and diagnostic reliability.

Performance evaluation demonstrated that the YOLOv10 model excelled in detecting the "Disgust" category, achieving an F1-score of 97%, recall of 96%, and precision of 98%. The model also effectively classified emotions such as "Angry," "Fear," and "Sad," with F1-scores consistently exceeding 60%. Robustness analysis through Receiver Operating Characteristic (ROC) curves further validated the system's effectiveness, with the "Disgust" category attaining the highest Area Under the Curve (AUC) score of 98%, while "Fear" had the lowest at 75%. Additionally, a confusion matrix analysis illustrated the model's classification capabilities, with notable accuracy rates across different emotional states.

Further evaluation of the ROC curves indicated that the pre-trained ResNet50 model outperformed others in classifying emotional categories, as evidenced by its curve positioning in the upper-left quadrant. In contrast, the ViT model exhibited lower classification performance, particularly for Anxiety, Depressive Disorder, and No Disorder categories, with AUC values ranging from 0.61 to 0.76. The MDNet model also had an AUC of 0.65 for Other Disorder classifications. Comparative analysis of feature extractors demonstrated that ResNet50 was the most effective, achieving an accuracy of 72%, surpassing MobileNetv2 and Inceptionv3, which attained 62%, while Xception reached 72%. Due to its strong feature extraction capabilities and efficiency, ResNet50 was selected for further experimentation.

Ensemble and ablation study results, as summarized in Table 6, indicate that excluding ViT from the ensemble led to the highest accuracy (88%), while the full ensemble model, incorporating all components, achieved 78%. The lowest-performing configuration occurred when the pre-trained ResNet50 model was excluded, underscoring its critical role in the classification process. These findings affirm that the integration of ResNet50 with MDNet significantly enhances model performance, whereas ViT contributes minimally..

This research integrates the YOLOv10 object detection framework alongside AffectNet and FER 2013 datasets, advancing traditional methods of mental illness diagnosis. The development of a specialized mental disorder dataset, coupled with a hybrid learning approach that combines CNN and ViT models, resulted in a highly effective ensemble classifier. The proposed methodology demonstrated notable improvements, with the final model achieving 81% accuracy, validating its effectiveness. Prior studies have primarily focused on specific disorders, such as depression, ADHD, ASD, and dyslexia. While previous research has utilized FER 2013 for mental health assessment, its limitations—including class imbalance and small sample size—necessitated the integration of AffectNet to enhance dataset quality and model robustness. Additionally, the incorporation of interpretability methods enhances transparency, providing healthcare professionals with actionable insights.

Limitations and Future Directions

Despite promising results, this study presents certain limitations. The datasets used exhibit class imbalance, which may impact overall performance. Future research should explore larger and more balanced datasets to improve classification accuracy. Additionally, the current system relies solely on facial cues for mental health assessment. Incorporating multimodal data sources, such as speech and contextual information, could enhance predictive capabilities. Furthermore, this study primarily focuses on anxiety and depressive disorders, whereas other conditions may also be identifiable through facial expressions. Expanding the dataset to include a broader range of mental health conditions and advanced versions of modals like yolo, could improve the system's applicability.

Conclusion

This study employed advanced artificial intelligence techniques to develop a transparent and interpretable system for detecting mental disorders. The primary objective was to facilitate early diagnosis and intervention through the integration of computer vision and deep learning methodologies. The YOLOv10 model was utilized to analyze facial expressions associated with various emotions, leading to the creation of a specialized mental disorder dataset. Several CNN-based architectures, including MDNet, ResNet50, and ViT, were trained and evaluated, with ensemble and ablation studies revealing that the combination of pre-trained ResNet50 and MDNet achieved the highest accuracy (88%).

Furthermore, the application of explainable AI (XAI) techniques, such as Grad-CAM and saliency maps, provided deeper insights into the model's decision-making process, enhancing transparency and reliability. Future research should explore transitioning from static image-based analysis to dynamic video-based assessments, allowing for continuous monitoring of mental states over time. Additionally, developing real-time monitoring systems capable of providing instant feedback and distress alerts would be a valuable advancement. Future studies should also investigate transformer-based architectures, alternative feature extraction methods, and the integration of supplementary data sources, such as clinical records and longitudinal emotional trends. These enhancements could significantly improve the accuracy and applicability of AI-driven mental health diagnostics, paving the way for more effective and comprehensive healthcare solutions.

REFERENCES

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

A. Samareh, Y. Jin, Z. Wang, X. Chang and S. Huang, "Detect depression from communication: How computer vision signal processing and sentiment analysis join forces", IISE Trans. Healthcare Syst. Eng., vol. 8, no. 3, pp. 196-208, Jul. 2018.

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

Brian, B., Goruntla, N., Bommireddy, B. R., Mopuri, B. M., Easwaran, V., Mantargi, M. J. S., ... & Ayogu, E. E. (2025). Knowledge, Attitude, and Practice Towards Responsible Self-Medication Among Pharmacy Students: A Web-Based Cross-Sectional Survey in Uganda. Drug, Healthcare and Patient Safety, 7-23.

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

D. F. Santomauro, A. M. M. Herrera, J. Shadid, P. Zheng, C. Ashbaugh, D. M. Pigott, et al., "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic", Lancet, vol. 398, no. 10312, pp. 1700-1712, 2020.

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

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

E. Sadek, N. A. Seada and S. Ghoniemy, "Computer vision techniques for autism symptoms detection and recognition: A survey", Int. J. Intell. Comput. Inf. Sci., vol. 20, no. 2, pp. 89-111, Dec. 2020.

G. Gilanie, M. ul Hassan, M. Asghar, A. M. Qamar, H. Ullah, R. U. Khan, et al., "An automated and real-time approach of depression detection from facial micro-expressions", Comput. Mater. Continua, vol. 73, no. 2, pp. 2513-2528, 2022.

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

H. Hadjar, J. Lange, B. Vu, F. Engel, G. Mayer, P. Mc Kevitt, et al., "Video-based automated emotional monitoring in mental health care supported by a generic patient data management system", Proc. 2nd Symp. Psychol.-Based Technol., Sep. 2020.

J. Aina, "Mental disorder detection system through emotion recognition", Oct. 2023.

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

J. Singh and G. Goyal, "Decoding depressive disorder using computer vision", Multimedia Tools

Appl., vol. 80, no. 6, pp. 8189-8212, Mar. 2021.

L. Duszynski-Goodman and L. Henderson, Mental Health Statistics and Facts in 2024, Jul. 2024, [online] Available: https://www.forbes.com/health/mind/mental-health-statistics/.

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

M. Munsif, M. Ullah, B. Ahmad, M. Sajjad and F. A. Cheikh, "Monitoring neurological disorder patients via deep learning based facial expressions analysis", Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov., vol. 652, pp. 412-423, 2022.

M. Nilsson Benfatto, G. Öqvist Seimyr, J. Ygge, T. Pansell, A. Rydberg and C. Jacobson, "Screening for dyslexia using eye tracking during reading", PLoS ONE, vol. 11, no. 12, Dec. 2016.

M. Tadalagi and A. M. Joshi, "AutoDep: Automatic depression detection using facial expressions based on linear binary pattern descriptor", Med. Biol. Eng. Comput., vol. 59, no. 6, pp. 1339-1354, Jun. 2021.

M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, et al., "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge", Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge, pp. 3-10, Oct. 2013.

M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, et al., "AVEC 2014: 3D dimensional affect and depression recognition challenge", Proc. 4th Int. Workshop Audio/Visual Emotion Challenge, pp. 3-10, Nov. 2014.

P. M. Swamy, P. J. Kurapothula, S. V. Murthy, S. Harini, R. RaviKumar and K. Kashyap, "Voice assistant and facial analysis based approach to screen test clinical depression", Proc. 1st Int. Conf. Adv. Inf. Technol. (ICAIT), pp. 39-44, Jul. 2019.

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

S. Harati, A. Crowell, Y. Huang, H. Mayberg and S. Nemati, "Classifying depression severity in recovery from major depressive disorder via dynamic facial features", IEEE J. Biomed. Health Informat., vol. 24, no. 3, pp. 815-824, Mar. 2020.

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

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

Vigneshwaran, E., Goruntla, N., Bommireddy, B. R., Mantargi, M. J. S., Mopuri, B., Thammisetty, D. P., ... & Bukke, S. P. N. (2023). Prevalence and predictors of cervical cancer screening among HIV-positive women in rural western Uganda: insights from the health-belief model. BMC cancer, 23(1), 1216.

X. Kong, Y. Yao, C. Wang, Y. Wang, J. Teng and X. Qi, "Automatic identification of depression using facial images with deep convolutional neural network", Med. Sci. Monitor, vol. 28, Jun. 2022.

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

Y.-S. Lee and W.-H. Park, "Diagnosis of depressive disorder model on facial expression based on fast R-CNN", Diagnostics, vol. 12, no. 2, pp. 317, Jan. 2022.

Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, et al., "Deep convolution network based emotion analysis towards mental health care", Neurocomputing, vol. 388, pp. 212-227, May 2020.

Zuo, W., Dhal, K., Keow, A., Chakravarthy, A., & Chen, Z. (2020). Model-based control of a robotic fish to enable 3d maneuvering through a moving orifice. IEEE Robotics and Automation Letters, 5(3), 4719-4726.