Computers in Biology and Medicine 179 (2024) 108822

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Optimized efficient attention-based network for facial expressions analysis in neurological health care

Muhammad Munsif [a,1], Muhammad Sajjad [b,c,1], Mohib Ullah [d], Adane Nega Tarekegn [c], Faouzi Alaya Cheikh [c], Panagiotis Tsakanikas [e], Khan Muhammad [f,*]

[a] Sejong University, Seoul, 143-747, Republic of Korea
[b] Digital Image Processing Lab, Department of Computer Science, Islamia College, Peshawar, 25000, Pakistan
[c] Department of Computer Science, Norwegian University for Science and Technology, 2815, Gjøvik, Norway
[d] Intelligent Systems and Analytics Research Group (ISA), Department of Computer Science, Norwegian University for Science and Technology, 2815, Gjøvik, Norway
[e] Institute of Communication and Computer Systems, National Technical University of Athens, 15773 Athens, Greece
[f] Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea

ARTICLE INFO

Keywords:
Neurological disorder
Parkinson
Information processing
Mobile health care
Facial expression analysis
Deep learning

ABSTRACT

Facial Expression Analysis (FEA) plays a vital role in diagnosing and treating early-stage neurological disorders (NDs) like Alzheimer's and Parkinson's. Manual FEA is hindered by expertise, time, and training requirements, while automatic methods confront difficulties with real patient data unavailability, high computations, and irrelevant feature extraction. To address these challenges, this paper proposes a novel approach: an efficient, lightweight convolutional block attention module (CBAM) based deep learning network (DLN) to aid doctors in diagnosing ND patients. The method comprises two stages: data collection of real ND patients, and pre-processing, involving face detection and an attention-enhanced DLN for feature extraction and refinement. Extensive experiments with validation on real patient data showcase compelling performance, achieving an accuracy of up to 73.2%. Despite its efficacy, the proposed model is lightweight, occupying only 3MB, making it suitable for deployment on resource-constrained mobile healthcare devices. Moreover, the method exhibits significant advancements over existing FEA approaches, holding tremendous promise in effectively diagnosing and treating ND patients. By accurately recognizing emotions and extracting relevant features, this approach empowers medical professionals in early ND detection and management, overcoming the challenges of manual analysis and heavy models. In conclusion, this research presents a significant leap in FEA, promising to enhance ND diagnosis and care.The code and data used in this work are available at: https://github.com/munsif200/Neurological-Health-Care.

1. Introduction

Recognizing and articulating emotions is critical in enabling efficient social interaction and bolstering cognitive development [1,2]. The efficacy of an emotion recognition system predominantly hinges on its capacity to accurately discern and interpret human sentiments, subsequently facilitating the inference of an individual's emotional state [3]. FEA has found significant application in the therapeutic rehabilitation of patients grappling with NDs, for instance as Parkinson's [4], Stroke [5], and Alzheimer's [6]. As a tool for human emotions recognition, FEA is an integral element in the therapeutic regimen for these conditions. This methodology is lauded for its efficacy in treating patients with neurological disorders and is widely employed within rehabilitative care [7]. Patients with neurological disorders often exhibit cognitive impairments, resulting in hindered emotional development and frequent expressions of distressing emotions such as sadness and anger [8]. Extant studies suggest that efficient emotion analysis and therapeutic interventions can prove advantageous for patients with neurological disorders, both in terms of estimating the severity level of symptoms and bolstering rehabilitation efforts [9]. It is also evidenced that effective training can enhance emotional recognition skills in autistic patients, thereby improving their cognitive abilities [10]. Furthermore, cognitive behavior therapy has demonstrated efficacy

* Corresponding author.
E-mail address: khan.muhammad@ieee.org (K. Muhammad).
[1] These authors contributed equally to this work and are co-first authors.

https://doi.org/10.1016/j.compbiomed.2024.108822
Received 24 January 2024; Received in revised form 25 June 2024; Accepted 25 June 2024
Available online 11 July 2024

in treating depressive disorders. Monitoring the emotional states of patients with neurological disorders throughout their treatment can aid in evaluating their overall health conditions. Similarly, the analysis of clinical and emotional attributes can serve as potential biomarkers to assess treatment efficacy in these patients [11]. Thus, the integration of emotion analysis and intervention into the diagnostic and treatment procedures for patients with neurological disorders can offer invaluable insights and contribute positively to their rehabilitative journey.

FEA is a crucial technology in the field of emotion perception and recognition [12–14]. It offers the opportunity to examine FEA systems and assess their impact on the quality and standards of human–computer interactions in rehabilitation systems and techniques [15,16]. The primary objective of FEA is to predict various emotions of humans from their facial images, infer human emotions [17], and facilitate the system in making responses such as creating a report, storing, or broadcasting the data [18]. There has been a growing interest among researchers in fields such as cognitive psychology, human–robot interactions, health management, and autonomous driving in the application of FEA [19]. FEA can be achieved through various techniques, including analyzing facial regions and extracting FEA-related features. FE features involve the face's shape and appearance-related features [20]. For the shape information, facial landmark points are extracted as feature vectors [21]. For other such as appearance features, a holistic spatial analysis technique is utilized. However, a drawback of these techniques is their reliance on handcrafted features, which can be limited in adapting to changing appearance environments [22]. To address this issue, convolutional neural network (CNN) have been used to obtain facial information. Furthermore, a hierarchical architecture of transformer neural with multi-scale features learning utilized aggregation technique in [23] and achieved convincing results for ordinary FEA. Several studies have focused on the psychological conditions [24] and monitoring of patients, such as Liao et al. [25] proposed a computationally expensive spatiotemporal framework for sequence-level affective level estimation, integrating facial expression features pyramid network and a temporal transformer encoder. This method, evaluated across multiple datasets for tasks like engagement prediction and pain assessment, shows promising results compared to existing algorithms. Ye et al. [26] introduces a cascaded spatiotemporal attention network for dynamic facial expression recognition and uses spatial and temporal attention modules to improve facial expression analysis. Results demonstrate its utility in capturing the nuances of facial expressions. Additionally, in a study conducted in [27] on Alzheimer's patients, the impact of Alzheimer's on facial expressions was investigated. The results showed a deficit in the facial expressions of the affected individuals. Another study analyzed the effects of neurodegenerative disorders on human facial expressions and found that patients with these disorders tend to express negative emotions such as anger and sadness. In a study by Bevilacqua et al. [28], FEs were utilized in identifying neurological disorders. The study proposed a method that can detect NDs by having the patients mimic various facial expressions demonstrated in a photo/video. The system interpreted patient expressions by determining the intensity of the replicated expression, a factor subsequently used to predict the disease's status. Parallel research, as presented in [29], proposed a computer vision-centric framework for FEA of severely dementia-stricken individuals during musical therapy. This framework was capable of categorizing activities and expressions under various conditions, such as speaking, singing, elation, and normalcy. Furthermore, Dapogny et al. [30] proposed an innovative 3D mobile gaming application, JEMlmE, designed to enhance the expressive abilities of children diagnosed with autism spectrum disorders. The proposal involved training a machine learning (ML) model based on the children's range of expressions (such as sadness, happiness, anger, neutral, etc.) and integrating this model into the JEMlmE application. A rewarding system was established where children accumulated positive points for correct emotional expressions and, conversely, lost points for incorrect ones. Similarly, Jin et al. [31] conducted a comparative analysis of DL and ML-based methodologies for diagnosing Parkinson's disease through an in-depth study of FEA. They collected videos of healthy as well as patient individuals diagnosed with Parkinson's disease, all of which featured smiling faces. They utilized Face++ API and traditional ML techniques such as Decision Trees (DT), Logistic Regression (LR), and Random Forests (RF) with the addition of DL-based sequence learning techniques (Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)) for different purposes, including feature extraction and subsequent classification [32,33].

While these studies demonstrate significant advancements in using FEA for monitoring psychological conditions and diagnosing neurological disorders, however, existing frameworks are not suitable for real-time applications due to their large number of parameters and high computational demands. Techniques such as Zhang et al. [34]; Tao and Duan [35]; Huang and Tsai [36]; Liao et al. [25]; and Ye et al. [26] show promising results but are computationally expensive. Similarly, methods developed for specific conditions like Alzheimer's, NDs, dementia, autism, and Parkinson's disease involve complex and handcrafted feature extraction algorithms that require substantial computational resources. This complexity makes it challenging to deploy these models in real-time settings, highlighting the need for more efficient algorithms that can achieve comparable accuracy with lower computational requirements. Furthermore, recognizing facial expressions with weak emotion intensities or indistinct visual appearances adds another layer of complexity to existing methods. Analyzing the facial expressions of patients with neurological disorders becomes even more challenging due to variations in the physiological effects of diseases, which directly impact the patients' emotional states. To cope with these gaps, we propose an efficient and robust FEA framework for patients with neurological disorders based on a CBAM-embedded DL network. The proposed framework leverages facial RGB images to capture detailed facial information and handle indistinct visual appearances. We design a custom CNN specifically tailored to extract deep features from pre-processed facial images. Additionally, incorporated an efficient CBAM-based feature refinement block to enhance features extracted from different input images automatically. In addition, our framework employs an efficient attention fusion method to enhance the spatial correlations between different facial feature maps instead of directly concatenating the extracted features. This approach optimizes the representation of facial expressions. The key contributions of this work are as follows:

1. Existing ND patients' FEA techniques are limited in their ability to extract meaningful and relevant features. This study proposed a lightweight DL-based network using a CBAM to accurately monitor ND patients in the early stages of their disease, specifically Parkinson's and Alzheimer's. It can extract the most relevant features from facial images and can improve the accuracy of ND diagnosis.
2. The model has a compact size, with a mere 3 MB, making it easily deployable on resource-restricted devices such as smartphones and Raspberry Pi, thus allowing for practical use by medical practitioners. The efficiency of DL models is greatly dependent on the number of parameters and the size of the model, and our model has been optimized to strike a balance between performance and size, ensuring its practicality in real-world applications.
3. The performance of the model in real-time by considering actual ND patients' data, collected from various sources is a critical aspect of automatic FEA methods. Our method demonstrated exceptional performance on ordinary as well as with up to 73.2% accuracy on ND's patients' collected data, comprising facial expressions from individuals of diverse gender and age groups.

The structure of this paper is organized as follows: Section 2 introduces the methodology, which encompasses data preprocessing, model architecture, and feature extraction. The process of feature refinement, implemented through spatial and channel attention, is detailed further

in Section 3. Section 4 provides information about the data used to train the model, explains the evaluation criteria, and presents both the ablation study and latency analysis. Lastly, Section 5 concludes the paper, outlining potential avenues for future research.

2. Facial expression analysis method

This Section discussed the proposed FEA method for ND patients. The factorial representations of the overall process are given in Fig. 1, which mainly consists of four subsections: Data acquisition and pre-processing, model architecture, realtime monitoring, and prognosis. Further explanation of each subsection is given below.

2.1. Data-pre-processing

This study utilizes standard as well as patient facial emotion data, the specifics of which are provided in the results section. One of the pivotal stages in creating a DL model is data pre-processing, given its substantial influence on the model's performance during the training phase. The primary purpose of pre-processing within this context is to extract pertinent regions from raw images and eliminate superfluous pixels. This measure is crucial in maintaining the model's precision and efficiency. The pre-processing procedure in our methodology commences with face detection, a task that presents challenges due to variations in angles and lighting conditions. To mitigate these issues, we utilize the Viola–Jones algorithm [37], which is renowned for its precision during the face detection phase. The initial step involves converting RGB images to grayscale before they are input into the Viola–Jones algorithm. This conversion enhances the accuracy of the algorithm and diminishes computational costs. Following this step, the identified face is isolated from the image, a process illustrated in Fig. 2. In order to further minimize computational costs, the isolated images are down-sampled to a resolution of 148 × 148 pixels prior to being input into the proposed training model. The down-sampling process reduces the size of the images, thereby improving training speed and decreasing memory usage.

2.2. Model architecture

This subsection describes the overall architecture of the proposed network. Our proposed network consists of features extraction module followed by two consecutive attention blocks that refine the extracted features. The factorial representation of the overall network is depicted in Fig. 1, step two, and the further details of the network are as follows.

2.3. Feature extraction

Facial expression recognition heavily relies on the involvement of CNN layers, which serve a critical role in extracting high-level features from input images. These layers employ filters that scan across the image, enabling the detection and learning of local patterns within the data. As the data undergoes successive processing through multiple CNN layers, the extracted features gradually become more abstract and representative of the overall content of the image. Consequently, when the features reach the final layer, they encapsulate essential information pertaining to the facial expression as depicted in Fig. 1. The proposed model is composed of six CNN layers, incorporating kernels of varying sizes ranging from 3 × 3 with a stride of one to 4 × 4 with the same stride in the last layer, thus facilitating the comprehensive representation of facial features. To refine the features obtained from the CNN layers, the CBAM is employed as an attention mechanism. CBAM enhances the overall quality of the extracted features by learning to emphasize the most relevant ones while suppressing the less significant information. This is achieved through the utilization of two parallel branches: the spatial attention branch and the channel attention branch, which collaboratively determine the importance of each feature. The spatial attention branch evaluates the significance of each feature location in the feature map by subjecting it to global average pooling followed by a 1 × 1 convolutional layer. Consequently, a weighted feature map is obtained, where the weights reflect the importance assigned to each feature location. In contrast, the channel attention branch focuses on determining the importance of each feature channel. This is accomplished by processing the feature map through two fully connected layers. These layers provide a weighted feature map, with the weights indicating the importance assigned to each feature channel. The attention branches are subsequently combined to yield the final attention map, which is used to weigh the features and refine them accordingly. The concluding stage of the facial expression recognition model involves a fully connected layer. This layer is responsible for mapping the input features to their respective class labels. By taking the high-level features extracted by the CNN layers, the fully connected layer processes them through multiple dense layers, where each layer is interconnected with all neurons in the preceding layer. Ultimately, the final fully connected layer produces predicted class probabilities, facilitating the determination of the facial expression depicted in Fig. 1, step 2. During training, the fully connected layer is optimized to minimize a defined loss function, quantifying the discrepancy between the predicted and ground truth labels.

3. Attention module

A prevailing objective among researchers is to develop deep networks that deliver high performance with minimal parameters [38–40]. The enhancement of network depth [41,42] and width [43,44] have traditionally been the most straightforward strategies adopted in pursuit of this objective. Recently, the notion of attention has emerged as a significant point of focus, with an emphasis on cardinality [45,46] and attention mechanisms. Inspired by the human visual system, it has been applied to DL mechanisms to enhance representation capability and keep focus on the most important information derived from the input data. This work uses CBAM [47] to obtain refined feature maps by amalgamating channel and spatial information. Differing from the approach in [47], this study utilized CBAM subsequent to extracting feature maps from the proposed CNN network. It bolsters the information flow across the network layers, facilitating information highlighting or suppression and yielding superior relevant information to predict the state of the ND patients. The mathematical formulations of the attention module is given in Eqs. (1) and (2). The feature maps denoted as $F \in \mathbb{R}^{C\times H\times W}$. The channel attention derives a 1D channel attention representation map denoted as $M_c \in \mathbb{R}^{C\times 1\times 1}$, as well as a 2D representation map represented as $M_s \in \mathbb{R}^{1\times H\times W}$, as illustrated in Figs. 3, 4, and 5.

$$F_{channel} = M_c(F) \otimes F \tag{1}$$

$$F_{spatial} = M_s(F_{channel}) \otimes F_{channel} \tag{2}$$

F denotes feature maps obtained from CNN. The refined spatial feature maps are denoted by $F_{spatial}$, and the channel attention refined feature maps are denoted by $F_{channel}$. The symbol ⊗ is used to represent the element-wise multiplication operation.

3.1. Channel attention module (CAM)

The CAM revolves around the focus on the critical feature maps within the input data. In this work, we used the formulation of Woo et al. [47] for channel attention (CA), employing both average and max-pooling and reducing the spatial dimension of the input feature maps acquired from the CNN. To shape it a mathematical representation, the average-pooling is represented as F^c_{avg}, along with the max-pooled feature, denoted as F^c_{max}, is subsequently introduced into a Multi-Layer Perceptron (MLP) with a single intermediate layer. The MLP generates

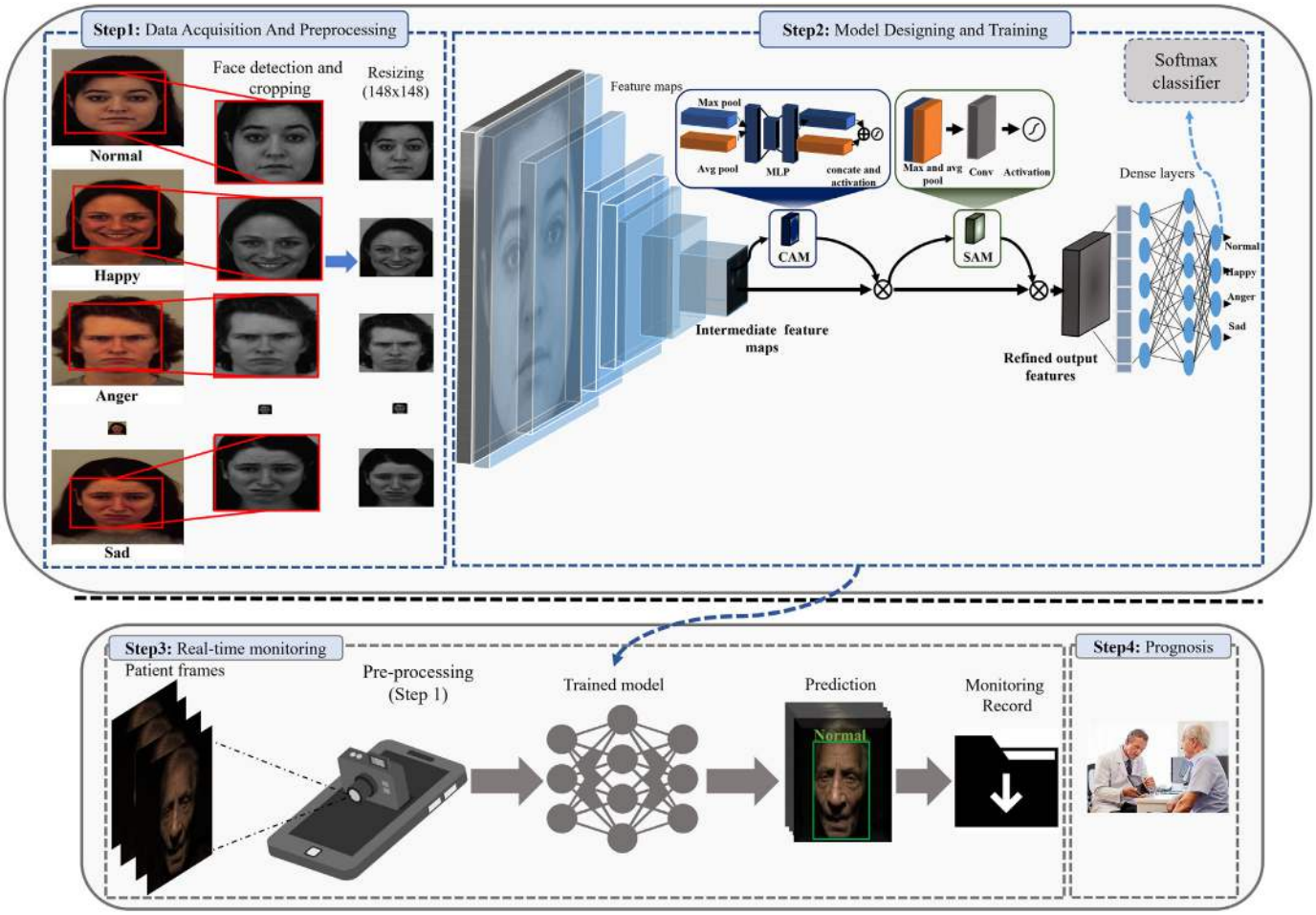

Fig. 1. The proposed framework: Comprises four key stages. Initially, acquired data undergoes a series of pre-processing techniques, including face detection, cropping, and resizing. Subsequently, a CBAM-DL model is designed and trained using the pre-processed data. Lastly, the trained model undergoes evaluation using real data from patients with NDs. The system maintains records of patients, which is utilized by medical professionals to offer tailored prognostic advice to patients.

Fig. 2. Face detection and cropping.

the CA map, denoted as $M_c \in \mathbb{R}^{C\times1\times1}$. Furthermore, this technique can be succinctly represented as follows:

$$M_c(F) = \sigma(MLP(Avg_{pool}(F)) + MLP(Max_{pool}(F))) \quad (3)$$

The feature maps produced by CNN are denoted by F and have dimensions of $\mathbb{R}^{C\times H\times W}$. In addition, average and max pooling operations are represented by Avg_{pool} and Max_{pool}.

$$M_c(F) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \quad (4)$$

The primary activation function in the CAM is the Sigmoid function σ. $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C\times C/r}$ are the weight parameters from the input to the hidden layer and from the hidden layer to the output, respectively, for the MLP. To maintain a small number of parameters in the MLP, the activation size of the hidden layer is $\mathbb{R}^{C/r\times1\times1}$, where r is the reduction ratio.

3.2. Spatial attention module (SAM)

SAM operates in harmony with CA, the core objective being to pinpoint the most distinguishing information within the obtained feature maps. The computation of SA commences with the application of both average and max pooling. The ensuing out maps are then consolidated to produce an efficient feature representation. Subsequently, this represents processing by a convolution layer, yielding the spatial attention map $M_s(F) \in \mathbb{R}^{H\times W}$. The mathematical representation of

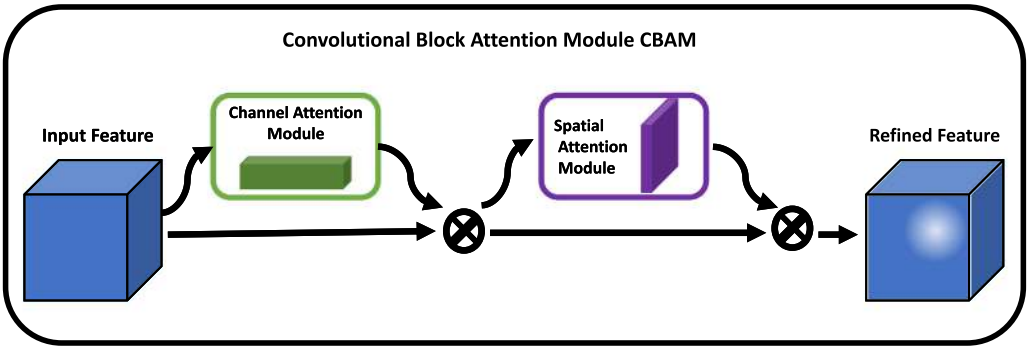

Fig. 3. Feature refinement with convolution block attention.

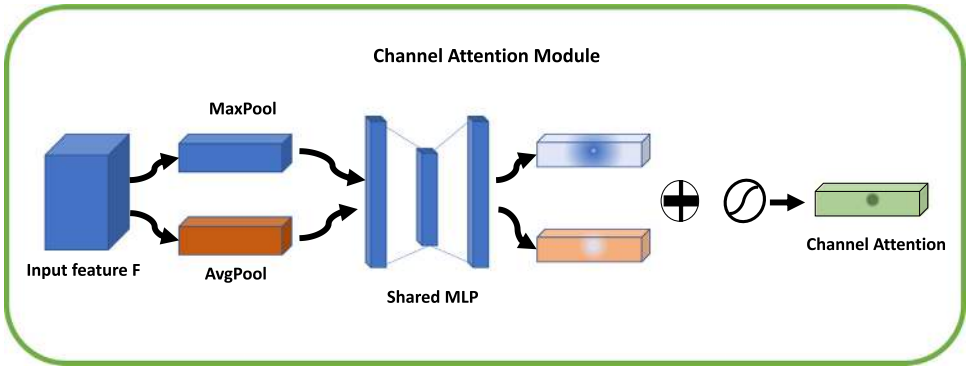

Fig. 4. Feature refinement with channel attention.

spatial attention is outlined below, where F stands for the input feature maps as derived from the CNN:

$$M_s(F) = \sigma(Conv^{9\times9}([Avg_{pool}(F); Max_{pool}(F)])) \tag{5}$$

$$M_s(F) = \sigma(Conv^{7\times7}([F^s_{avg}; F^s_{max}])) \tag{6}$$

Here, $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{max} \in \mathbb{R}^{1\times H\times W}$ correspond to the average and max pooling operations respectively. The Sigmoid function, denoted by σ, is utilized as the primary activation function, while $Conv^{7\times7}$ symbolizes the convolution operation employing a kernel size of 7×7. The final maps of features are then utilized as the slice of information and provided to the final prediction model as shown in Fig. 1. In addition, in the quest to achieve more precise localization of the most informative regions within the final feature maps, employ spatial attention, which serves as a complement to the CA mechanism. This process begins with the both average and max pooling to the feature maps. Subsequently, the output information maps are concatenated to create efficient feature representations. This descriptor is then processed through a convolution layer, resulting in the generation of the SA map $M_s(F) \in \mathbb{R}^{H\times W}$. For mathematical representations $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F^s_{max} \in \mathbb{R}^{1\times H\times W}$ denote the average and max pooled feature maps respectively. The Sigmoid function σ is implemented as the primary activation function, and $Conv^{7\times7}$ indicates the convolution operation with a filter size of 7×7. Finally, the refined feature map is fed into a dense layer, followed by a softmax function, which classifies the feature maps into predefined categories.

3.3. Deployment and real-time monitoring

Our framework can test and integrate into mobile and surveillance systems seamlessly as shown in Fig. 1, step 3, to enhance the treatment and monitoring of patients with neurological disorders. The system perceives input via cameras and averages the emotional state of a patient with each 5 to 10-minute analyzed clip at a time, leveraging the proposed DL model specifically optimized for deployment on resource-constrained devices, it allows for the efficient analysis of subtle facial cues directly on mobile or embedded devices within existing surveillance infrastructure. This real-time capability is critical for capturing micro-expressions and diminished facial movements indicative of various neurological conditions, such as for Parkinson's disease, providing continuous, non-invasive insights into the emotional and cognitive states of patients. The model's lightweight architecture ensures that it operates seamlessly on devices with limited computational power, enabling real-time data collection and analysis without the need for heavy computational resources. This facilitates the generation of automatic, comprehensive reports that synthesize behavioral and emotional changes, for medical intervention and treatment adjustment. For healthcare professionals, this means access to timely, actionable insights delivered directly through an optimized system that integrates smoothly into the daily management of neurological conditions [48]. By enhancing remote monitoring capabilities, our framework can not only reduce the frequency of in-person visits but can also significantly improve the ecological validity of patient assessments, ultimately leading to better patient outcomes and optimized healthcare strategies.

4. Results

This section presents a comprehensive discussion regarding the performance of our proposed network within our defined experimental setting. An in-depth analysis encompasses the experimental setup, employed datasets, comprehensive ablation study, as well as real-time testing. This discussion profoundly comprehends the model's capabilities and performance characteristics.

4.1. Experimental setting and implementation details

Our proposed method is implemented, and experiments were conducted using a Python version 3.7 virtual environment hosted on a

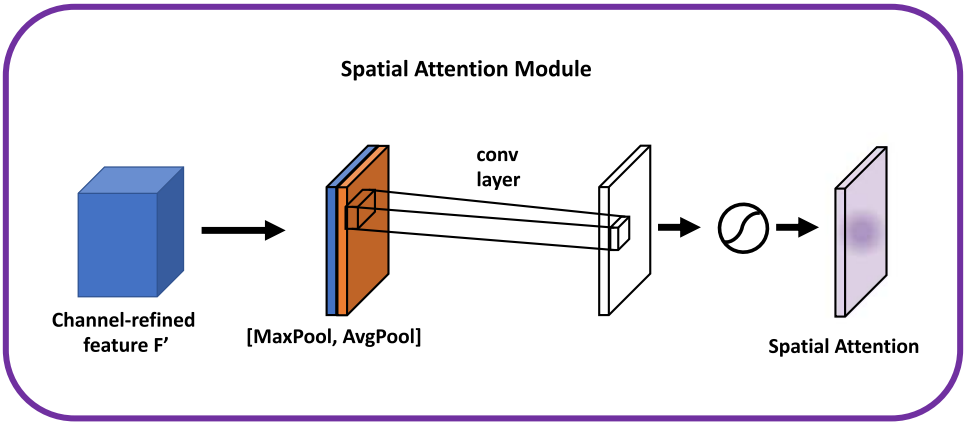

Fig. 5. Feature refinement with spatial attention.

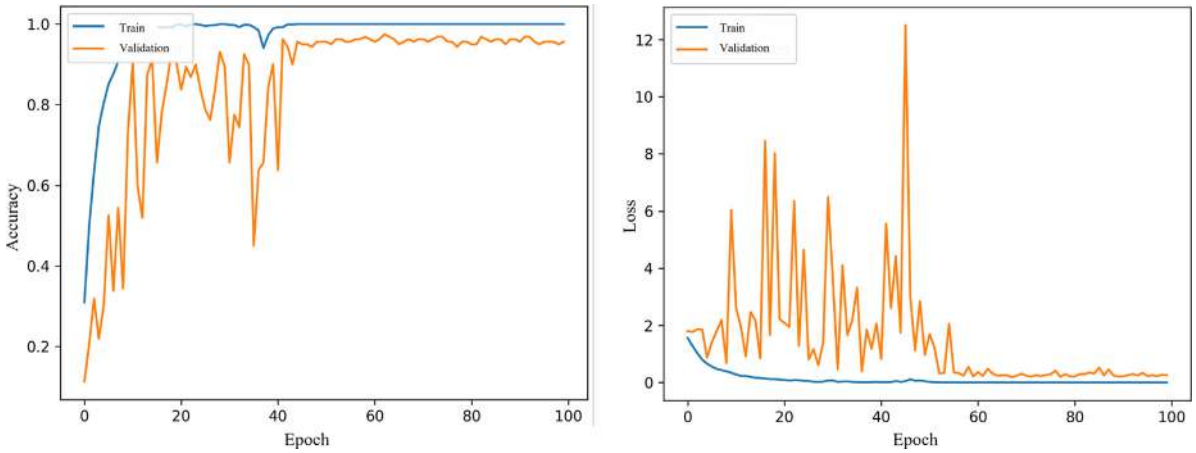

Fig. 6. Accuracy and loss graphs of the proposed framework on patients data.

PC equipped with an Intel Xeon X5560 processor, 8.00 GB of RAM, and an NVIDIA GeForce GTX 1070 GPU. The DL model was architecturally designed, trained, and evaluated, leveraging the TensorFlow-GPU framework (version 2.0.0) with the Keras-GPU front end. Training of the model utilized the categorical cross-entropy loss function and optimized our model's performance using a grid search strategy for hyperparameters selection. This strategy systematically explored combinations of key parameters such as the learning rate, weight decay, and momentum, ultimately yielding effective parameter values. The training of the end-to-end modules utilized the ADAM optimizer with fixed hyperparameters with a learning rate (lr) of 0.00001. The training was performed in mini-batches of 32 over the course of 100 epochs. The entire process depends on the size of the data and hyperparameters.Furthermore, The training performance graph of the model is shown in Fig. 6.

4.2. Data used for training and evaluation

The collection, annotation, and arrangement of datasets, particularly for FEs of ND patients, is a complex and resources demanding task. This requires a substantial number of ND patients or skilled actors capable of exhibiting genuine expressions similar to ND patients. Both options entail significant financial and human resources, including the involvement of researchers, doctors, and patients. To mitigate these challenges, we opted to utilize publicly available datasets for training, such as The Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS) [27] and the KEDF [28].

The preparation of the datasets for the DL model is described in the method section of data pre-processing. By leveraging publicly available datasets, we were able to streamline the data collection process and reduce the resource constraints associated with creating a dataset from scratch. Further, for a fair evaluation of the proposed model, we collected data from YouTube of real ND patients and conducted a comprehensive evaluation of the model after fine-tuning using the patient's data and testing in real time. Samples of each dataset, including the ND patients, are given in Fig. 7, and details of each dataset are provided as follows.

The second dataset that is utilized RAVDESS [49] is a multi-modal dataset developed in collaboration between the psychology department at Ryerson University, Canada, and computer science and information system department at the University of Wisconsin, United States. The aim was to provide high-quality data to researchers with a highly validated audiovisual dataset for automatic analysis of audio, video, or both to facilitate rehabilitation through the analysis of neurological disorders. The RAVDEES database was recorded by 24 professional actors, consisting of 12 male and 12 female actors aged between 21 and 33. The dataset contains 7356 recordings, including audio-only recordings of songs and speech and videos containing speech and a sequence of frames. Our work focuses on the analysis of facial expressions from visuals, so we selected only the visual component of the dataset, which contains videos of various facial expressions, extracted frames from the videos with a frame skip of 20 to avoid redundancy and reduce the training time after extraction of frames pre-processed it according to the requirements of our problem. In comparison, in the KDEFE dataset, the five common expression classes selected included angry, afraid, disgust, happy, and neutral samples from the happy and sad classes, where each class of the data consists of 2000 image samples from happy and sad

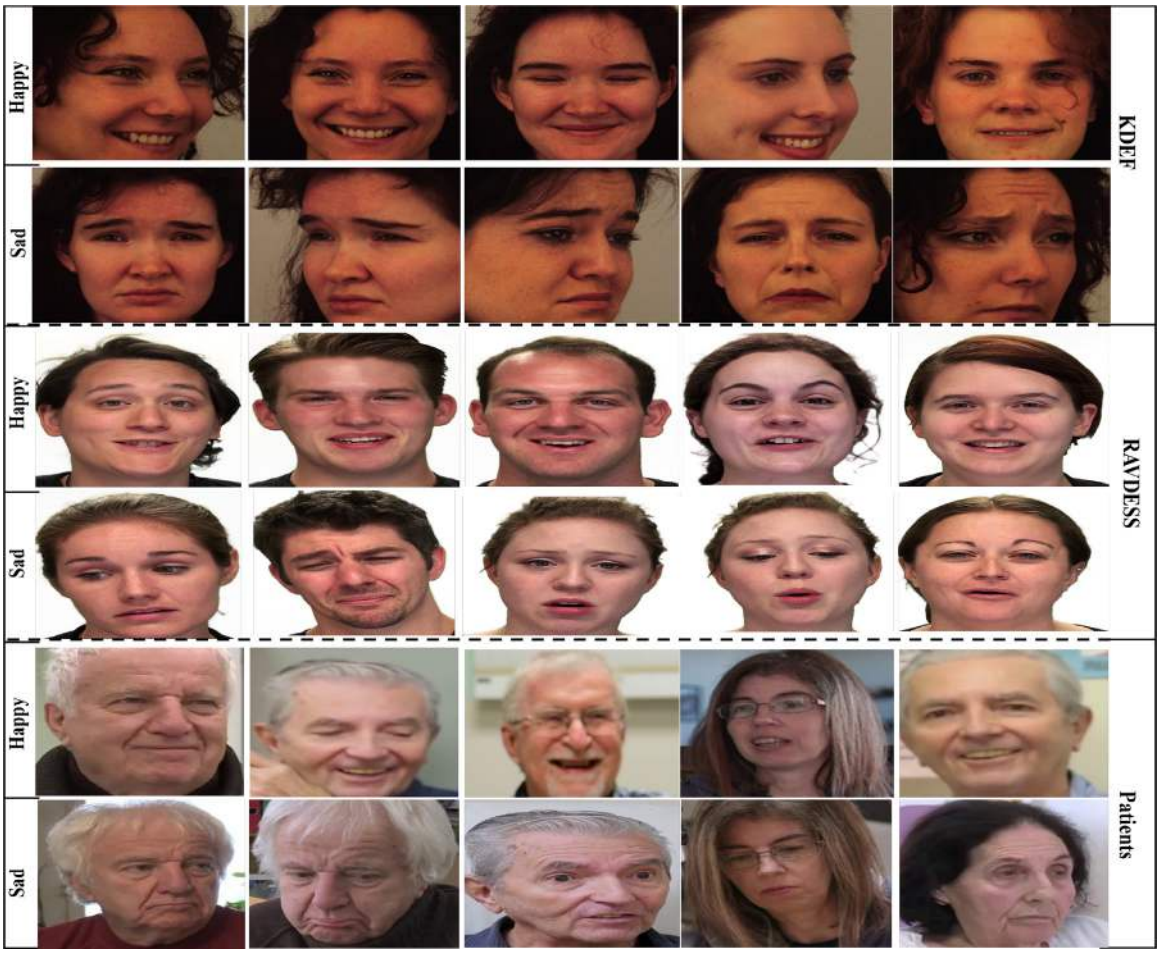

Fig. 7. Samples from the datasets: contains happy and sad facial expressions from KDEF, RAVDESS, and Patient's datasets.

classes are depicted in Fig. 7. To obtain results of real-time testing, we collected full-length videos of NDs patients from well-known channels, such as Michigan Medicine and 60 min Australia BAYSTATEHEATH, on the YouTube platform. After collection, extracted frames from each video are arranged in corresponding folders and then utilized for model evaluation, samples are shown in Fig. 7.

4.3. Evaluation criteria

The performance evaluation of the proposed model encompassed various evaluation metrics, including testing accuracy, precision, recall, and inference time. These metrics were computed based on a confusion matrix denoted as C, a 2-dimensional array. In the confusion matrix, the element $C_i j$ represents the count of instances predicted as $class_i$ and actual belonging to class j. The speed of the model was quantified using the inference time, which indicates the time taken by the model to make predictions on the test data. To assess the model's performance, Eqs. (7)–(10) are used to calculate loss, accuracy, precision, and recall. Accuracy (A) refers to correct predictions out of the total instances. While precision (p) shows the fraction of correctly predicted positive noted as TP instances out of the total instances predicted as positive including false positives (FP). Recall (R) is the fraction of correctly predicted positive instances out of the total positive instances including false negative (FN) samples.

$$L = \frac{1}{N} * \sum_{i=1}^{N} L_i \tag{7}$$

$$P = \frac{TP}{TP + FP} \tag{8}$$

Table 1
Ablation Study investigating variants of the proposed network on patients data.

# of Conv Layers	Attention mechanism	Accuracy (%)	Inference time (s)
1	–	30.0	0.1
2	–	33.0	0.2
3	–	38.0	0.3
4	–	47.0	0.5
5	–	52.0	0.5
6	–	57.0	0.6
6	CAM	67.0	0.6
6	SAM	69.0	0.6
6	CBAM	73.2	0.7

In Eq. (7), N is the number of training examples, and L_i is the loss for each example i. These evaluation metrics provide a comprehensive understanding of the model's performance and enable a thorough comparison with other models.

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$A = \frac{TP + TN}{TP + FN + TN + FP} \tag{10}$$

4.4. Ablation study

The results were obtained from a series of experiments conducted. These experiments aimed to investigate the impact of increasing the number of convolutional layers and altering the attention mechanism from spatial attention to channel attention, as well as their combination. In addition, Table 2 includes a comparative evaluation of various

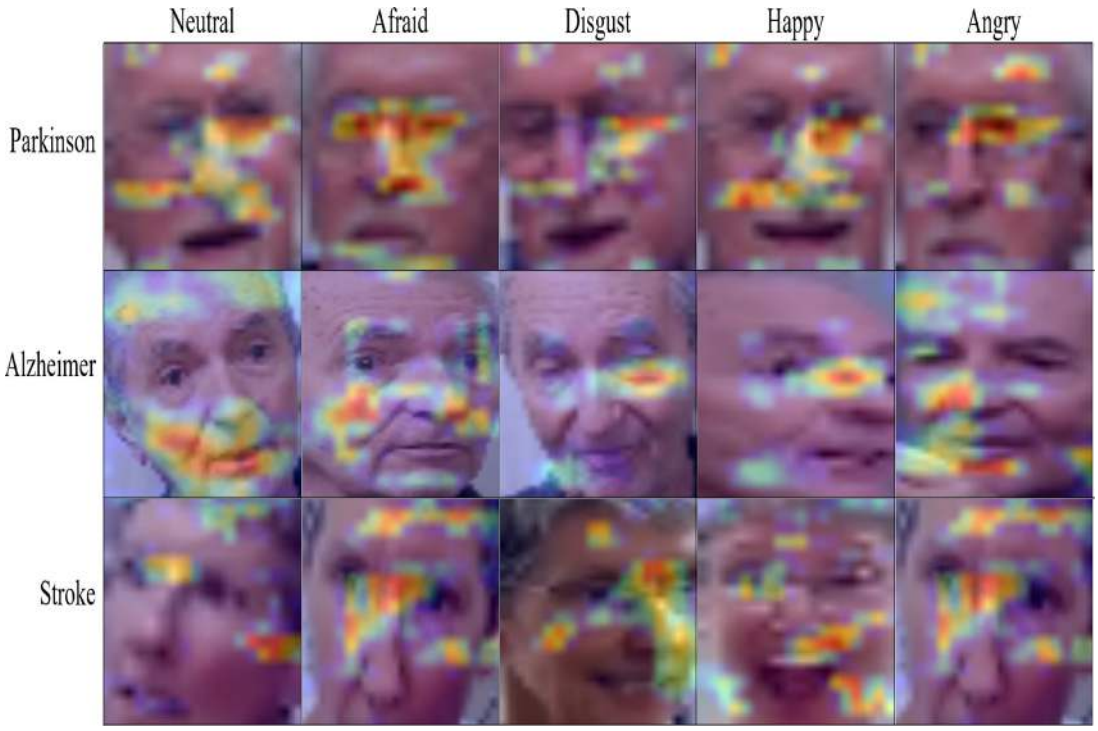

Fig. 8. Visualization of model attention on facial regions of patients in real-world scenarios while diagnosing neurological disorders across different expressions.

Fig. 9. Results of real-time testing: The initial row portrays the results corresponding to individuals diagnosed with Parkinson's disease. Conversely, the subsequent two rows illustrate the test outcomes associated with individuals suffering from Alzheimer's disease and Stroke.

state-of-the-art models on all three datasets. Table 1 highlights that a combination of channel and spatial attention mechanisms led to a gradual improvement in accuracy. This can be attributed to the model's enhanced capability to extract accurate and high-level facial features. However, it should be noted that the model size also increased alongside the number of trainable parameters within the convolutional layers. The introduction of additional dense layers resulted in overfitting. The inference time exhibited an ascending trend in Table 1 due to the augmented number of feature extraction layers, each requiring a specific duration for processing. Following extensive experimentation, the proposed model, incorporating the CBAM and six convolutional layers with two dense layers, achieved notable performance with a testing accuracy of 73.2%. Moreover, it attained precision and recall values of 73.4% and 73.5%, respectively, on real patients' data. Notably, the total size of the model is achieved at only 3 MB, comprising a mere 0.9 million parameters. The inference time per frame was measured at

Table 2
Experiments on all datasets using various state-of-the-art generalized and specialized FER DL architectures and their impact on patient's data performance.

Model	Attention Mechanism	KDEF			RAVDESS			Patients			Parameters (M)
		Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	
ResNet-101 [50]	SAM	75.0	77.1	77.0	78.5	78.0	78.3	38.0	38.0	38.0	–
ResNet-101 [50]	CAM	77.0	77.6	78.0	79.3	79.5	79.6	39.5	39.8	39.0	–
ResNet-101 [50]	CBAM	83.2	83.0	83.0	83.9	84.0	83.8	42.0	41.2	43.0	43.9
ResNet-50-V2 [43]	SAM	78.2	78.0	79.0	80.0	80.0	80.0	43.0	43.9	43.0	–
ResNet-50-V2 [43]	CAM	81.0	82.0	80.8	82.9	82.5	82.6	43.9	43.0	44.4	–
ResNet-50-V2 [43]	CBAM	84.5	82.0	83.5	84.0	84.0	83.9	44.5	43.0	44.0	14.8
Inception-V3 [44]	SAM	49.0	53.5	50.0	55.0	55.0	55.3	20.0	20.0	20.5	–
Inception-V3 [44]	CAM	52.5	52.0	53.0	62.0	62.0	62.0	22.0	22.0	21.5	–
Inception-V3 [44]	CBAM	57.0	56.7	58.0	65.0	65.3	65.9	25.4	25.3	25.0	25.1
Res-Next [46]	SAM	31.7	29.4	33.4	35.3	33.4	36.2	05.0	05.0	04.0	–
Res-Next [46]	CAM	35.2	34.5	47.4	42.0	43.0	41.0	06.3	05.0	06.7	–
Res-Next [46]	CBAM	38.5	36.0	41.5	45.0	45.0	45.0	08.0	08.5	07.5	34.4
DenseNet [51]	SAM	20.0	22.3	17.6	24.0	25.0	23.0	03.0	03.0	03.5	–
DenseNet [51]	CAM	24.2	23.4	25.7	29.0	29.1	29.0	04.0	05.0	04.5	–
DenseNet [51]	CBAM	27.8	27.0	26.6	33.0	33.0	32.8	05.5	05.0	05.5	8.1
MobileNet [52]	SAM	80.6	80.3	80.8	83.3	83.3	83.2	44.2	44.5	43.0	–
MobileNet [52]	CAM	82.5	81.2	82.0	83.9	83.6	83.5	44.2	44.5	44.6	–
MobileNet [52]	CBAM	83.6	86.0	83.5	84.0	84.0	84.0	44.5	44.5	44.3	4.1
H-attention [53]	–	88.5	–	–	–	–	–	–	–	–	–
Transfer Learning [54]	–	–	–	–	57.0	–	–	–	–	–	–
Aural Transformers [55]	–	–	–	–	62.1	–	–	–	–	–	–
Orthogonal-attention [56]	–	89.0	–	–	–	–	–	–	–	–	–
L-CNN [57]	–	93.0	93.0	93.0	–	–	–	–	–	–	–
Joint-Attention [58]	–	–	–	–	58.2	–	–	–	–	–	–
Hit-mst [23]	–	–	–	–	87.4	–	–	–	–	–	–
PIDViT [36]	–	–	–	–	89.3	88.3	89.0	53.1	53.0	53.2	86.0
MTAC (SwinTrans) [34]	–	–	–	–	90.3	90.0	91.1	55.1	54.8	55.0	84.7
H-attention [35]	–	–	–	–	86.4	87.2	87.0	52.3	52.9	54.1	69.63
Proposed	SAM	89.0	89.6	90.1	92.0	92.3	92.0	67.0	67.5	67.0	–
Proposed	*CAM*	*93.0*	*92.2*	*93.5*	*94.0*	*94.0*	*94.0*	*69.0*	*69.2*	*69.3*	-
Proposed	CBAM	**94.2**	**94.3**	**94.2**	**95.0**	**95.3**	**94.8**	**73.2**	**73.4**	**73.5**	**00.9**

0.723 s, rendering the model easily deployable on resource-constrained devices. Furthermore, Table 2 presents the evaluation results of various state-of-the-art models, ranging from deep models to lightweight models, including MobileNet, ResNet v1, and ResNet v2. It shows these architectural designs were initially developed for general purposes and not specifically tailored for facial feature extraction. Nonetheless, we assessed the performance of these models on our specified datasets. However, the outcomes revealed that these models did not exhibit high accuracy.

4.5. Real-time performance

Fig. 9 Shows the results obtained from real-time testing of the proposed model. The performance evaluation involved analyzing images of patients at early stages of Parkinson's disease under treatment and individuals diagnosed with early-stage Alzheimer's disease and stroke. The outcomes demonstrated the model's ability to recognize and classify the patient's facial expressions accurately, with a few exceptions observed in challenging scenarios. In instances where drastic changes in facial angle or appearance occurred, the model occasionally misclassified the expressions. For instance, an example from the Alzheimer's disease group depicted a patient expressing disgust but misclassified as angry or happy. Similarly, a stroke patient exhibiting a neutral expression was mistakenly classified as neutral. These particular cases highlight the challenges faced by the model when encountering significant variations in facial characteristics. Attention maps were visualized for each class to provide further insights into the model's attention mechanism, as illustrated in Fig. 8. These attention maps reveal the regions of the image where the model focused its attention on extracting relevant features. In most cases, the proposed model successfully captured and emphasized relevant facial features associated with the respective expressions. However, it is worth noting that the model seemed to extract less valuable features for specific instances, such as the disgusted expression in stroke patients, explicitly focusing on mouth-related attributes. These findings demonstrate the effectiveness of the proposed model in accurately recognizing facial expressions in real-time testing. The model's attention maps offer valuable insights into its decision-making process by highlighting the regions that are deemed significant for expression recognition. Nevertheless, it is crucial to acknowledge the challenges posed by extreme variations in facial angles or appearances, which can impact the model's recognition accuracy in certain scenarios.

4.6. Comparison with state of the art

For a fair evaluation of the proposed method, we have compared the performance of our framework with both generalized DL models and specialized FER models, as detailed in Table 2. These models include generalized architectures such as ResNet [50], Inception Network [44], and lightweight models like MobileNet [52], as well as specialized attention-based FER models like those by Liu et al. [36,53,56], and Zhang et al. [34]. The table offers a detailed comparison of various state-of-the-art FER architectures, evaluated across multiple datasets, including KDEF, RAVDESS, and a patient-specific dataset. Among these models, MobileNet with CBAM and ResNet-50-V2 with CBAM performed notably well, achieving high metrics in accuracy, precision, and recall. For instance, MobileNet with CBAM achieved significant results with a relatively small number of parameters (4.1 million), while ResNet-50-V2 with CBAM also showed strong performance across the datasets. In the context of patient data, several models demonstrated varied performance. For example, ResNet-101 with CBAM achieved an accuracy of 42.0%, while MobileNet with CBAM reached an accuracy of 44.5% on the patient dataset and attention-based specialized models like Zhang et al. [34] achieved comparatively good results but these models are more computationally expensive. Furthermore these models still face challenges in handling the variability and complexity of patient data, which is often characterized by diverse expressions and suboptimal conditions. Our proposed model with CBAM significantly outperforms other models on the patient dataset, achieving the highest accuracy (73.2%), precision (73.4%), and recall (73.5%), with a low parameter count of only 0.9 million. In addition, to show the model's fair capabilities and avoid any bias of the proposed model performance, provided a ROC curve on patient data which shows the highest area under the curve (AUC) of 0.73 in Fig. 10 for the proposed model. These performances demonstrate the proposed framework's superior capability in accurately FEA on patient data while maintaining efficiency and showcasing the effectiveness and suitability of our model for ND FEA, highlighting its superiority over the state-of-the-art models, and suggesting its utilization in real-world scenarios. Despite these results, the proposed model has limitations. Its performance can still be affected by extreme conditions such as very poor lighting and occlusions, which are common in real-world scenarios involving patients. Additionally, while the model is efficient, further optimizations on a larger scale NDs

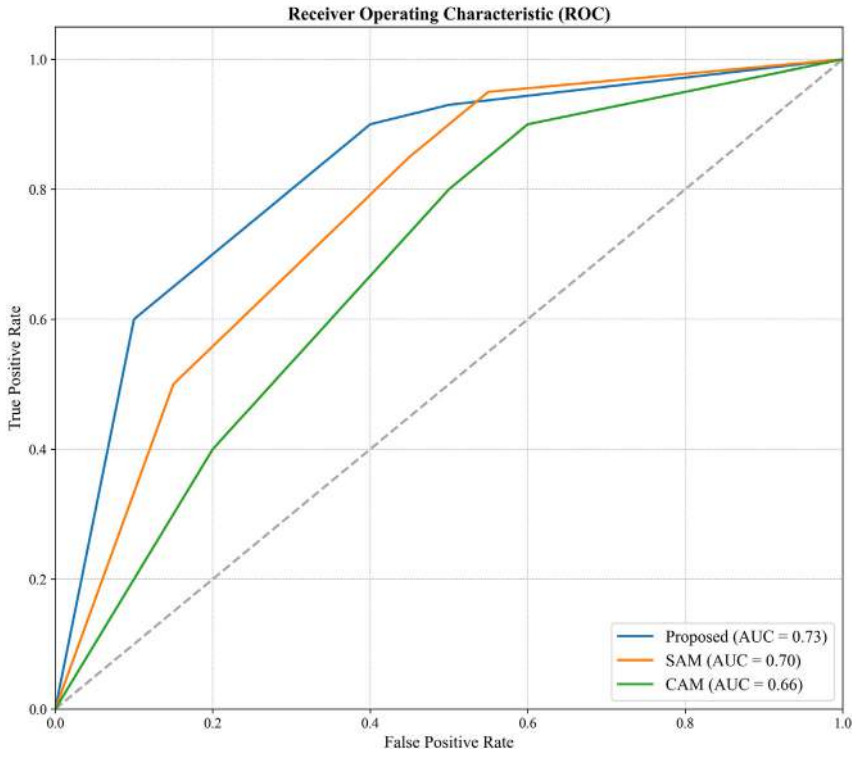

Fig. 10. ROC of the Proposed framework and runner-up models on patients data.

patient database are needed to ensure its robustness and applicability in real-time healthcare systems. Future research should focus on advanced lightweight DL-based pre-processing techniques to enhance robustness and specific model quantization and pruning techniques to reduce the computational complexity for real-time applications.

5. Conclusion and future directions

This study presents an FEA approach for NDs such as Alzheimer's and Parkinson's. The proposed approach offers a CBAM-based DL network that efficiently and accurately recognizes emotions along with small-scale real NDs patients dataset for cross-validation. The proposed method demonstrates convincing performance through extensive experimentation on real collected patients' data, achieving accuracy of up to 73.2%. This high level of accuracy is attributed to the model's ability to extract the most relevant features and conduct deep FEA, facilitated by the combination of a CNN and the attention-based CBAM module. Furthermore, the lightweight nature of the model, with a size of only 3MB, enables its practical deployment on resource-constrained devices. This study's findings highlight the proposed method's superiority over existing FEA approaches. By eliminating the dependence on heavy models and irrelevant handcrafted or DL-based features, the proposed approach demonstrates significant improvement in accuracy and efficiency. These advancements have promising implications for the effective diagnosis and treatment of ND patients, enabling healthcare professionals to obtain reliable and timely insights into patients' emotional states. Despite these impressive results, the proposed model has limitations. Its performance can still be affected by extreme conditions such as very poor lighting and occlusions, which are common in real-world scenarios involving patients.

Future work should consider expanding the dataset to encompass a broader range of patients and developing more robust and lightweight DL models. This includes employing advanced pre-processing techniques to enhance the model's robustness under challenging conditions. Specific model quantization and pruning techniques should also be explored to further compact the model, reduce computational complexity and resource usage. Furthermore, multimodal integration for enhanced emotion recognition and conducting longitudinal studies to track disease progression can improve early detection and treatment. Additionally, investigating model transferability, considering ethical implications, and addressing biases are crucial to ensure responsible deployment in healthcare settings.

CRediT authorship contribution statement

Muhammad Munsif: Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Muhammad Sajjad:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition. **Mohib Ullah:** Writing – review & editing, Investigation, Formal analysis. **Adane Nega Tarekegn:** Validation, Formal analysis. **Faouzi Alaya Cheikh:** Supervision, Investigation, Conceptualization. **Panagiotis Tsakanikas:** Writing – review & editing, Validation, Formal analysis. **Khan Muhammad:** Validation, Project administration, Investigation, Formal analysis.

Declaration of competing interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The European Union funded this research through the Horizon 2020 Research and Innovation Programme in the context of the ALAMEDA (Bridging the Early Diagnosis and Treatment Gap of Brain Diseases via Smart, Connected, Proactive and Evidence-based Technological Interventions) project under grant agreement No. GA 101017558.

References

[1] D. White, A.M. Burton, Individual differences and the multidimensional nature of face perception, Nat. Rev. Psychol. 1 (5) (2022) 287–300.

[2] L. Xiao, X. Wu, S. Yang, J. Xu, J. Zhou, L. He, Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis, Inf. Process. Manage. 60 (6) (2023) 103508.

[3] A. Flechsenhar, P. Kanske, S. Krach, C. Korn, K. Bertsch, The (un) learning of social functions and its significance for mental health, Clin. Psychol. Rev. (2022) 102204.

[4] L. Ricciardi, F. Visco-Comandini, R. Erro, F. Morgante, M. Bologna, A. Fasano, D. Ricciardi, M.J. Edwards, J. Kilner, Facial emotion recognition and expression in Parkinson's disease: an emotional mirror mechanism? PLoS One 12 (1) (2017) e0169110.

[5] J. Lin, Y. Chen, H. Wen, Z. Yang, J. Zeng, Weakness of eye closure with central facial paralysis after unilateral hemispheric stroke predicts a worse outcome, J. Stroke Cerebrovasc. Dis. 26 (4) (2017) 834–841.

[6] C. Ferrari, S. Berretti, P. Pala, A. Del Bimbo, Measuring 3D face deformations from RGB images of expression rehabilitation exercises, Virtual Real. Intell. Hardw. 4 (4) (2022) 306–323.

[7] H. Tanaka, R. Umeda, T. Kurogi, Y. Nagata, D. Ishimaru, K. Fukuhara, S. Nakai, M. Tenjin, T. Nishikawa, Clinical utility of an assessment scale for engagement in activities for patients with moderate-to-severe dementia: additional analysis, Psychogeriatrics 22 (4) (2022) 433–444.

[8] M.O. Bertelli, L. Salvador-Carulla, K.M. Munir, M.L. Scattoni, M.W. Azeem, A. Javed, Intellectual developmental disorder and autism spectrum disorder in the WPA next triennium mainstream, World Psychiatry 19 (2) (2020) 260.

[9] P. Kormas, A. Moutzouri, Current psychological approaches in neurodegenerative diseases, in: Handbook of Computational Neurodegeneration, Springer, 2022, pp. 1–29.

[10] J. Kerr-Gaffney, L. Mason, E. Jones, H. Hayward, J. Ahmad, A. Harrison, E. Loth, D. Murphy, K. Tchanturia, Emotion recognition abilities in adults with anorexia nervosa are associated with autistic traits, J. Clin. Med. 9 (4) (2020) 1057.

[11] C. Luo, N. Sanger, N. Singhal, K. Pattrick, I. Shams, H. Shahid, P. Hoang, J. Schmidt, J. Lee, S. Haber, et al., A comparison of electronically-delivered and face to face cognitive behavioural therapies in depressive disorders: A systematic review and meta-analysis, EClinicalMedicine 24 (2020) 100442.

[12] M. Verma, S.K. Vipparthi, G. Singh, S. Murala, LEARNet: Dynamic imaging network for micro expression recognition, IEEE Trans. Image Process. 29 (2019) 1618–1627.

[13] F.Z. Canal, T.R. Müller, J.C. Matias, G.G. Scotton, A.R. de Sa Junior, E. Pozzebon, A.C. Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, Inform. Sci. 582 (2022) 593–617.

[14] Q. Wang, T. Su, R.Y.K. Lau, H. Xie, DeepEmotionNet: Emotion mining for corporate performance analysis and prediction, Inf. Process. Manage. 60 (3) (2023) 103151.

[15] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: A review, Electron. Notes Theor. Comput. Sci. 343 (2019) 35–55.

[16] M.M. Krishna, B. Duraisamy, J. Vankara, Independent component support vector regressive deep learning for sentiment classification, Measurement: Sensors 26 (2023) 100678.

[17] Y.-K. Li, Q.-H. Meng, Y.-X. Wang, H.-R. Hou, MMFN: Emotion recognition by fusing touch gesture and facial expression information, Expert Syst. Appl. 228 (2023) 120469.

[18] G. Karpagam, B. Balasarath, J.Y. Nicholas, R. Lokesh, S.S. Rahul, S. Sarkar, Facial emotion detection using convolutional neural network algorithm, Int. J. Adapt. Innov. Syst. 3 (2) (2022) 119–134.

[19] Z. Pang, G. Yang, R. Khedri, Y.-T. Zhang, Introduction to the special section: convergence of automation technology, biomedical engineering, and health informatics toward the healthcare 4.0, IEEE Rev. Biomed. Eng. 11 (2018) 249–259.

[20] I.M. Revina, W.S. Emmanuel, A survey on human face expression recognition techniques, J. King Saud Univ.-Comput. Inform. Sci. 33 (6) (2021) 619–628.

[21] M. Zhu, D. Shi, M. Zheng, M. Sadiq, Robust facial landmark detection via occlusion-adaptive deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3486–3496.

[22] D. Zeng, R. Veldhuis, L. Spreeuwers, A survey of face recognition techniques under occlusion, IET Biometr. 10 (6) (2021) 581–606.

[23] X. Xia, D. Jiang, HiT-MST: Dynamic facial expression recognition with hierarchical transformers and multi-scale spatiotemporal aggregation, Inform. Sci. (2023) 119301.

[24] C. Xu, C. Yan, M. Jiang, F. Alenezi, A. Alhudhaif, N. Alnaim, K. Polat, W. Wu, A novel facial emotion recognition method for stress inference of facial nerve paralysis patients, Expert Syst. Appl. 197 (2022) 116705.

[25] J. Liao, Y. Hao, Z. Zhou, J. Pan, Y. Liang, Sequence-level affective level estimation based on pyramidal facial expression features, Pattern Recognit. 145 (2024) 109958.

[26] Y. Ye, Y. Pan, Y. Liang, J. Pan, A cascaded spatiotemporal attention network for dynamic facial expression recognition, Appl. Intell. 53 (5) (2023) 5402–5415.

[27] C.G. Kohler, G. Anselmo-Gallagher, W. Bilker, J. Karlawish, R.E. Gur, C.M. Clark, Emotion-discrimination deficits in mild alzheimer disease, Am. J. Geriatr. Psychiatry 13 (11) (2005) 926–933.

[28] V. Bevilacqua, D. D'Ambruoso, G. Mandolino, M. Suma, A new tool to support diagnosis of neurological disorders by means of facial expressions, in: 2011 IEEE International Symposium on Medical Measurements and Applications, IEEE, 2011, pp. 544–549.

[29] A. Dantcheva, P. Bilinski, H.T. Nguyen, J.-C. Broutart, F. Bremond, Expression recognition for severely demented patients in music reminiscence-therapy, in: 2017 25th European Signal Processing Conference, Eusipco, IEEE, 2017, pp. 783–787.

[30] A. Dapogny, C. Grossard, S. Hun, S. Serret, J. Bourgeois, H. Jean-Marie, P. Foulon, H. Ding, L. Chen, S. Dubuisson, et al., JEMImE: a serious game to teach children with ASD how to adequately produce facial expressions, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, IEEE, 2018, pp. 723–730.

[31] B. Jin, Y. Qu, L. Zhang, Z. Gao, Diagnosing Parkinson disease through facial expression recognition: video analysis, J. Med. Internet Res. 22 (7) (2020) e18697.

[32] H. Liu, C. Zhang, Y. Deng, T. Liu, Z. Zhang, Y.-F. Li, Orientation cues-aware facial relationship representation for head pose estimation via transformer, IEEE Trans. Image Process. 32 (2023) 6289–6302.

[33] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, Y.-F. Li, TransIFC: invariant cues-aware feature concentration learning for efficient fine-grained bird image classification, IEEE Trans. Multimed. (2023).

[34] Z. Zhang, X. Tian, Y. Zhang, K. Guo, X. Xu, Enhanced discriminative global-local feature learning with priority for facial expression recognition, Inform. Sci. 630 (2023) 370–384.

[35] H. Tao, Q. Duan, Hierarchical attention network with progressive feature fusion for facial expression recognition, Neural Netw. 170 (2024) 337–348.

[36] Y.-F. Huang, C.-H. Tsai, PIDViT: Pose-invariant distilled vision transformer for facial expression recognition in the wild, IEEE Trans. Affect. Comput. (2022).

[37] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, CVPR 2001, Ieee, 2001, p. I.

[38] H. Liu, T. Liu, Y. Chen, Z. Zhang, Y.-F. Li, EHPE: Skeleton cues-based gaussian coordinate encoding for efficient human pose estimation, IEEE Trans. Multimed. (2022).

[39] H. Liu, T. Liu, Z. Zhang, A.K. Sangaiah, B. Yang, Y. Li, Arhpe: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction, IEEE Trans. Ind. Inform. 18 (10) (2022) 7107–7117.

[40] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, J. Wang, MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation, IEEE Trans. Multimed. 24 (2021) 2449–2460.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[42] D. Han, J. Kim, J. Kim, Deep pyramidal residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5927–5935.

[43] S. Zagoruyko, N. Komodakis, Wide residual networks, 2016, arXiv preprint arXiv:1605.07146.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[45] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[47] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.

[48] M. Pourebadi, L.D. Riek, Facial expression modeling and synthesis for patient simulator systems: past, present, and future, ACM Trans. Comput. Healthc. (HEALTH) 3 (2) (2022) 1–32.

[49] S.R. Livingstone, F.A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PLoS One 13 (5) (2018) e0196391.

[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[51] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[52] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.

[53] Y. Liu, J. Peng, J. Zeng, S. Shan, Pose-adaptive hierarchical attention network for facial expression recognition, 2019, arXiv preprint arXiv:1905.10059.

[54] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J.M. Montero, F. Fernández-Martínez, Multimodal emotion recognition on RAVDESS dataset using transfer learning, Sensors 21 (22) (2021) 7665.

[55] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J.M. Montero, F. Fernández-Martínez, A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset, Appl. Sci. 12 (1) (2021) 327.

[56] J. Chen, L. Yang, L. Tan, R. Xu, Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition, Pattern Recognit. 129 (2022) 108753.

[57] M. Munsif, M. Ullah, B. Ahmad, M. Sajjad, F.A. Cheikh, Monitoring neurological disorder patients via deep learning based facial expressions analysis, in: Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops: MHDW 2022, 5G-PINE 2022, AIBMG 2022, ML@ HC 2022, and AIBEI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Springer, 2022, pp. 412–423.

[58] E. Ghaleb, J. Niehues, S. Asteriadis, Joint modelling of audio-visual cues using attention mechanisms for emotion recognition, Multimedia Tools Appl. 82 (8) (2023) 11239–11264.