Two-Phase Transfer Learning Framework for Automated Depression Classification in the Elderly via Facial Expression Recognition

Muhammad Daffa Zahrandika Wibisono[1], Marizuana Mat Daud[2]*, Wan Mimi Diyana Wan Zaki[3]
Institute of Visual Informatics, National University of Malaysia, 43600, Bangi, Selangor Darul Ehsan, Malaysia [1, 2]
Faculty of Engineering & Built Environment, National University of Malaysia, 43600, Bangi, Selangor Darul Ehsan, Malaysia [3]

***Abstract*—Automatic detection of depression in the elderly through Facial Expression Recognition faces a fundamental challenge in the form of domain shift due to skin deformation and facial structural changes due to aging, such as ptosis and deep wrinkles. This study proposes a Two-Phase Transfer Learning framework that integrates high-density facial landmark point extraction (468 points using MediaPipe) with a hybrid spatiotemporal CNN-BiLSTM-VGG19 architecture to address these challenges. Phase I training was conducted on a standard facial dataset to obtain fundamental feature representations, followed by a fine-tuning process in Phase II using a geriatric facial dataset. Experimental results show that the CNN-BiLSTM-VGG19 architecture is highly robust, exploiting deep facial wrinkles as informative texture features. The model successfully achieved 91.42% accuracy on 70-year-old older adults. Furthermore, hyperparameter evaluation confirmed that the Stochastic Gradient Descent (SGD) optimizer combined with a low learning rate of 0.0005 was the most optimal configuration. This balance effectively prevented catastrophic forgetting during domain adaptation, while also achieving a clinical sensitivity recall rate above 96%. Comprehensively, this study demonstrates that the texture-biased CNN-BiLSTM-VGG19 model offers a robust, non-invasive, and highly efficient depression screening instrument for implementation in elderly care facilities.**

***Keywords*—Elderly depression; Facial Expression Recognition; transfer learning; VGG19; texture bias; spatiotemporal network**

I. Introduction

The phenomenon of population aging in Southeast Asia has become a serious concern, with the elderly population growing by 41.5% in the past two decades [1]. This trend aligns with conditions in Malaysia, where the population growth rate for those aged 65 and over has reached 6.1%. This significant demographic shift is correlated with an increase in mental illness, particularly depression, with a prevalence of 19.1% among the elderly [2]. A major challenge in current medical management is the high number of undetected cases [3], often caused by a decline in natural emotional expression in the elderly that masks clinical signs [4].

Conventional screening methods such as the PHQ-9 and GDS-15 questionnaires, although commonly used, are susceptible to subjective self-report bias [5]. To overcome this limitation, a facial analysis based approach is crucial, as Major Depressive Disorder (MDD) fundamentally alters the dynamics of an individual's emotional valence. The affective computing literature confirms that depression triggers negative affect potentiation, where patients tend to display predominantly sad and angry expressions, and Positive Affect Attenuation, or anhedonia, which limits the patient's ability to express happiness [6]. Therefore, Facial Expression Recognition (FER) technology does not simply perform a binary classification, but rather estimates the probability of depression based on a shift in patterns from relaxed, positive, and neutral expressions to predominantly negative affect.

More specifically, this classification analysis relies on the interpretation of different facial features between healthy and depressed groups. Referring to the findings of Girard [6], the Non Depression classification is strongly influenced by the detection of authentic happiness signals and dynamic facial muscle engagement; the absence of these signals statistically increases the predicted probability of depression due to anhedonia mechanisms. Conversely, for the 'Depression' class classification in the geriatric population, it emphasized that visual indicators are not limited to melancholic sadness but often manifest as irritability or agitation that mimics anger [6]. Therefore, an FER system must be able to integrate the dominant anger features in older adults as predictors of depression, rather than simply transient anger.

However, the application of precision FER technology to older adults faces significant technical gaps, particularly related to data availability. Existing public datasets are predominantly young adult faces and validated in controlled environments [7]. Consequently, artificial intelligence models trained on standard datasets such as CASME II or CK+ often fail to distinguish clinical emotional features from natural aging features. Permanent facial wrinkles, for example, are frequently misinterpreted as negative expressions.

Based on these challenges, this study proposes a two-phase transfer learning framework to classify depression from Non Depression in older adults. Phase I aims to extract basic emotional features from an extensive young adult dataset to build a foundation for expression recognition. Subsequently, Phase II refines the model's sensitivity through domain adaptation on an older adult facial dataset, ensuring the system can accurately distinguish between signs of pathological depression and characteristics of biological aging.

*Corresponding author

II. Literature Review

A. FER Using Deep Learning

Deep Learning represents a cutting-edge computational paradigm in artificial intelligence that models high-level abstractions on data using multi-layered processing architectures, or deep architectures [8]. Unlike traditional machine learning methods that rely heavily on hand-crafted features, Deep Learning's fundamental strength lies in its ability to perform end-to-end representation learning [9], where the model automatically discovers intrusive structures and non-linear patterns directly from raw data without human intervention [10].

In the context of Facial Expression Recognition (FER), this capability is crucial because the human face has extreme visual variability [11]. Deep Learning is able to capture subtle nuances of facial deformation and is invariant to changes in lighting and pose [12], making it a state-of-the-art standard for deciphering the complexity of human emotions that are difficult to quantify by conventional algorithms [13].

B. Spatio-Temporal and Sequential Model for FER

The development of spatio-temporal and sequential models emerged as a response to the limitations of conventional CNNs that only process spatial information on static images, where the integration of recurrent architectures such as Long Short-Term Memory (LSTM), Bidirectional LSTM, or Gated Recurrent Unit with CNNs allows the system to capture the dynamics of facial expression changes as well as the temporal relationships between frames in a video sequence [14][15]. This hybrid architecture works by extracting spatial features from each frame using a CNN, which are then sequenced and processed by recurrent layers to learn long-term dependencies between frames, so that the model can distinguish emotions based on the evolution of expressions over time [16].

Spatial features extracted by CNN represent static characteristics in each frame, but temporal information obtained from LSTM, BiLSTM allows the model to understand the context of movement and expression transitions that are crucial for distinguishing similar emotions but have different dynamic patterns [17]. The combined implementation of CNN with BILSTM can combine deep visual information with temporal sequence context simultaneously, thus enabling the model to learn expression change patterns from both time directions to improve recognition accuracy [18].

III. Methodology

The proposed methodology, as it shows on Fig. 1, begins with dataset collection, followed by feature extraction to obtain crucial facial landmarks, including the pre-processing stage. In the core stage, the method applies a two-phase transfer learning strategy: the first phase involves fine-tuning the reduction model to adapt general features, while the second phase focuses on specific training according to the intended task. Finally, the model's performance is evaluated based on its ability to recognize patterns and produce accurate predictions.

A. Dataset

To build a robust classification sample against variations in facial expressions, this study prepared a comprehensive dataset consisting of a total of 10,432 facial images, divided equally into 5,216 samples for the Depression class and 5,216 for the Non Depression class. This dataset is an amalgamation of various standard data sources, including CASME II for micro-expressions, and RAVDESS, TFEID, and CK+ for macro-expressions, as described in Table I. In determining this binary labeling, we adopted protocols validated by Girard et al. [6] and Lee [19], which showed that facial emotion profiles have a strong statistical correlation with psychological states. Individuals with depression tend to exhibit Negative Affect Potentiation (predominance of sadness and anger/irritability), while Non Depression is associated with the ability to exhibit relaxed, happy, and neutral expressions. This emotional valence-based label mapping methodology aligns with the approach of [20], which demonstrated the effectiveness of Depressed and Non Depressed labels based on emotional intensity tendencies in training Deep Learning models.

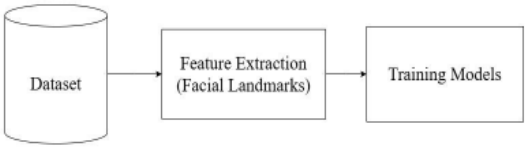

Fig. 1. Process flow of the proposed method.

This integration of multiple datasets is designed to bridge the complexity of facial dynamics, which vary from the macro to the micro level. RAVDESS plays a crucial role in capturing high-level temporal dynamics, as this dataset records not only pose statistics but also subjects speaking and singing across a wide spectrum of emotions, allowing the model to learn natural and complex expression transitions. CASME II, on the other hand, complements this analysis by capturing micro-expressions. This data presents a different challenge because expressions appear for very short durations and low intensity, yet are crucial because they represent the involuntary "emotional leaks" that are often key indicators of hidden signs of depression.

TABLE I. Technical Specifications and Dataset Sources

Dataset	Expression	Validator Institution & Key Characteristics
CASME [21]	Micro-Expression	Chinese Academy of Sciences (CAS). Recorded at high speed (200 fps) to capture involuntary micro-muscle changes.
RAVDESS [22]	Macro-Expression	Ryerson University, Canada. Provides high temporal dynamics variations through speech and song modalities.
TFEID [23]	Macro-Expression	Brain Mapping Laboratory, Taiwan. Focuses on subjects with Asian backgrounds for accurate demographic representation.
CK+ [24]	Macro-Expression	Carnegie Mellon University (CMU). The gold standard with precise FACS annotations for basic muscle movement patterns.

B. Feature Extraction

To transform raw visual information into structured data representations, this study implements a geometry-based feature extraction framework. Unlike holistic pixel-based approaches that process the entire image intensity spectrum and are highly

susceptible to environmental disturbances such as lighting fluctuations, this method focuses exclusively on the topological configuration of the face. This approach was strategically chosen to address two major challenges in facial expression analysis of elderly individuals: 1) the need to separate emotional features from skin texture and aging artifacts (such as permanent wrinkles), and 2) the urgency to capture involuntary micro movements that often go undetected in low-resolution representations.

1) High-density facial landmark detection: The crucial stage in the high-density facial landmark detection process is implemented using the MediaPipe Face Mesh Architecture. This cutting-edge framework was specifically chosen to replace conventional methods such as the sixty-eight-point Dlib Model due to its capability to reconstruct a massive three-dimensional topology consisting of four hundred and sixty-eight landmarks, as presented in Fig. 2. The fundamental justification for using this high-density topology is based on two primary technical arguments. The first argument emphasizes the increased spatial resolution essential for detecting microexpressions.

Conventional standard models have been shown to lack adequate spatial coverage of critical facial areas such as the glabella between the eyebrows and the nasolabial folds. In contrast, this new architecture provides a highly dense tessellation in these crucial areas. This level of density is essential for capturing the subtle movements of facial action units, such as action unit four, which represents the contraction of the eyebrow-lowering muscle, and action unit twelve, which represents the lip corner puller, both of which clinically serve as key indicators in identifying depressive states. The second argument relates to a high degree of pose invariance.

This framework's ability to comprehensively estimate three-dimensional spatial coordinates allows the system to consistently maintain topological consistency. This robust spatial tracking remains optimally operational even when the research subjects exhibit head pose variations up to forty-five degrees, a dynamic condition often encountered during naturalistic clinical interviews.

In the experimental setup, MediaPipe hyperparameters were strictly configured to guarantee data integrity at the frame level, as presented in Table II.

TABLE II. MEDIAPIPE HYPERPARAMETER CONFIGURATION

Parameter	Value	Description
static_image_mode	TRUE	This parameter is explicitly enabled to process each video frame as an independent entity. This approach prevents temporal bias from the tracking algorithm and ensures that landmark detection is based purely on the visual features of the current frame. This is crucial for preventing information leakage when training CNN architectures on shuffled datasets.
max_num_face	1	Limiting detection to just one face aims to focus computing resources exclusively on the primary subject. This configuration effectively eliminates background noise from other, irrelevant individuals in the frame.
min_detection_confidence	0.5	This threshold is set empirically to achieve an optimal balance between sensitivity (recall) and precision. This value ensures that the facial topology mesh will only be generated when facial features are clearly visible, significantly reducing the occurrence of false positives in ambiguous visual frames.

2) Geometry-based ROI normalization and preprocessing: After coordinate extraction, a custom preprocessing workflow is applied to normalize the Region of Interest (ROI) and eliminate external variables irrelevant to emotion, specifically lighting inconsistencies and skin tone bias. As implemented in our algorithm, this process follows three systematic steps:

a) Grayscale projection: The input image is converted to a grayscale spectrum (cv2.COLOR_BGR2GRAY). This step neutralizes chromatic lighting artifacts (e.g., the yellow tint of room lights) and removes skin color information, forcing the model to rely solely on structural deformations.

b) Structural overlay: A topological mask is drawn over a grayscale background using a Tessellation technique, connecting 468 landmarks with 1-pixel-thick, high-contrast green lines (BGR: 0, 255, 0). This effectively turns the image recognition problem into a geometric pattern recognition task.

c) Spatial standardization: The face is cropped based on the extreme coordinates of the landmarks, with 20 pixels of padding added to preserve edge detail. The resulting image is then resized to a uniform 224 x 224 pixel dimension to meet the input layer requirements.

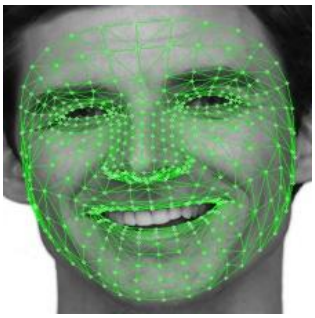

Fig. 2. Visualization of the results of feature extraction pre-processing.

A visual implementation of this framework is demonstrated in Fig. 2. The image shows the transformed result where the original color information has been removed to neutralize skin tone bias. On top of this base layer, the topological structure of the face is rendered using geometric tessellation. As seen in the details of the mouth and eye area in Fig. 2, this mesh precisely tracks facial muscle deformations such as the contraction of the Zygomaticus major during smiling, verifying that the input features passed to the classification model are pure structural movement patterns that are invariant to both subject identity and lighting conditions.

C. Training Models: Comparative Deep Transfer Learning Framework

This study applies a strategically designed Two-Phase Transfer Learning framework to build a robust depression classification model in the elderly population. The training approach is hierarchically structured to address the key challenge of labeled data scarcity in the geriatric domain.

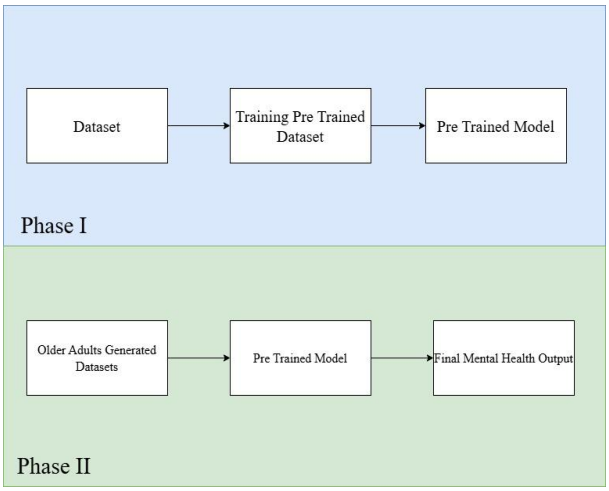

Fig. 3. Model Phase I and II.

As shown in Fig. 3, Phase I focused on Source Domain Benchmarking, where we conducted an intensive comparative study between backbone architectures such as VGG19 to learn universal spatiotemporal feature representations. Phase II focused on Target Domain Adaptation, where the best model resulting from Phase I was adapted through a fine-tuning mechanism using a synthetic elderly face dataset.

1) Phase I: Phase I served as a foundational stage, where we conducted a series of experiments to build the most optimal spatiotemporal representation model. In this stage, the model was trained from scratch to distinguish micro and macro-expression patterns associated with Depression Sad, Angry) and Non Depression (Happy, Neutral) using a standard composite dataset.

a) Dataset: To ensure robust model generalization and mitigate potential biases, Phase I of this study utilizes a Composite Source Dataset. This dataset is an amalgamation of four internationally recognized benchmark repositories, as seen in Fig. 4: Casme II (micro-expressions), Ravdess (dynamic Macro-expressions), Tfeid (Asian Facial Morphological Features), and CK+ (Action Unit Precision Validation). The integration of these diverse sources provides a comprehensive representation of facial dynamics across various temporal and morphological scales.

Fig. 4. Dataset collection from RAVDESS, CASME II, TFEID and CK+.

The acquisition and utilization of these datasets were conducted in strict adherence to ethical research protocols and privacy standards. Access to the CASME II repository was secured through a formal institutional application process. For RAVDESS, TFEID, and CK+, data were obtained under open source licenses, which involved the completion of formal consent agreements and compliance with institutional citation mandates. These datasets have been clinically and scientifically validated by their respective provider institutions, ensuring that subject privacy is preserved and that the research aligns with international ethical guidelines for automated mental health screening.

Following the amalgamation, class balancing was performed to maintain an equitable distribution between classes. The final dataset for Phase I comprises 10,432 facial image samples, symmetrically partitioned into 5,216 samples for the 'Depression' class and 5,216 for the 'Non Depression' class. For model development and empirical evaluation, a systematic 80:20 data split ratio was implemented. Consequently, 8,346 images were allocated for the training phase, while 2,086 images were strictly reserved for the testing and validation set. This explicit partition addresses the requirement for statistically significant performance evaluation on unseen data.

All samples underwent a standardized preprocessing pipeline: conversion to grayscale format, application of a high-density topological dense mesh structure (green), and rescaling to a uniform dimension of 224 x 224 pixels. This structured input format is specifically engineered to optimize compatibility with the CNN-BiLSTM-VGG19 architecture, facilitating the extraction of invariant geometric facial features while eliminating confounding color variations.

b) Trained pre-train model: The training process in Phase I was designed through a multi-stage experimental approach (multi-stage benchmarking) to thoroughly validate each architectural component.

c) Preliminary ablation study: Spatial vs. Spatiotemporal: Before finalizing the main feature extraction architecture, a preliminary ablation study comparing spatial and spatiotemporal approaches was conducted to validate the importance of integrating the temporal module. This test was designed for a direct performance comparison between a purely spatial model represented by CNN-VGG19 and a hybrid spatiotemporal model based on CNN-BiLSTM-VGG19. In the former, the model works exclusively by relying on static spatial feature extraction capabilities. In contrast, the hybrid approach integrates a Bidirectional Long Short-term Memory Layer or BiLSTM specifically designed to comprehensively capture temporal dependencies between frames. This series of initial experimental results specifically aimed to provide empirical evidence that the addition of the BiLSTM module is able to demonstrate a very significant improvement in the accuracy of microexpression recognition when compared to purely spatial-based methods.

d) Backbone architecture benchmarking: After validating the superiority of the hybrid architecture, the next step focused on selecting the most optimal primary spatial feature extractor. This experiment was specifically designed to

answer the fundamental question of whether the density of texture extraction in VGG is superior in capturing microexpressions related to Depression. To answer this question, this study compares two variants of the hybrid architecture with very contrasting characteristics. The first variant, CNN-BiLSTM-VGG19, is built on a stack of nineteen convolutional layers with a small kernel size of 3 X 3. This architectural design allows the model to learn a very detailed hierarchical feature representation ranging from simple edge extraction to complex facial texture patterns. In the context of Depression detection, VGG19 demonstrates comparative advantages in extracting very fine texture nuances such as the depth of forehead wrinkles and shadows around the eyes.

Through synergy with BiLSTM, these static textural features are transformed into a continuous time sequence, enabling the model to not only detect the presence of wrinkles but also understand the dynamics of their depth and fading over time. The hypothesis proposed for this variant is that VGG19's sensitivity to pixel-level details makes it a very strong candidate for processing low-resolution inputs and micro-feature extraction.

e) Hyperparameter grid search optimization: To ensure a comprehensive comparison, all models were trained using the rigorous grid search scheme by testing various parameter combinations to find the best convergence configuration. This testing included optimizing Adam for fast convergence and SGD for generalization stability, as well as testing learning rates across a wide spectrum of values, including 0.01, 0.05, 0.001, 0.005, 0.0005, and 0.0001. The training duration was also set at a maximum limit of 100 epochs by implementing an early stopping mechanism with a tolerance limit of 10 epochs to prevent overfitting and stop the training process if there is no improvement in the validation loss.

f) Pre-trained model: The final result of Phase I is the best-performing pre-trained model selected based on the highest validation accuracy and lowest loss metric across all combinations of experiments. This selected VGG19 model stores basic knowledge about universal facial expression dynamics within its synaptic weights. This model is then stored and will serve as an intelligent initiator for the transfer learning process in Phase II to replace the less efficient random weight initialization. Next, this VGG19 model is combined with the BiLSTM architecture to continuously optimize expression recognition. This combination is implemented because VGG19 has high reliability in extracting detailed spatial features, while BiLSTM is very effective in capturing the temporal dependencies of a series of bidirectional emotional transitions. Therefore, the combination of the two can produce a system that understands the dynamics of facial changes much more comprehensively and accurately.

g) Phase II: Phase II is the specialization stage where the general knowledge about facial expressions acquired in Phase I is transferred to the target domain, namely the elderly population. This stage is crucial to address the domain shift problem, where models trained on young faces often fail to adapt to the complex morphological characteristics of aging faces, such as deep wrinkles and decreased skin elasticity (ptosis). Instead of training the model from scratch, Phase II applies a transfer learning strategy to refine the model's sensitivity to the nuances of Geriatric faces.

h) Older adults dataset: A major challenge in analyzing depression in the elderly is the lack of public datasets containing high-quality micro- and macro-expressions in Geriatric subjects. To address this without sacrificing precise emotion labels, Phase II utilized the Synthetic Elderly Dataset, generated through a generative transformation of all source datasets from Phase I, such as Casme II, Ravdess, Tfeid, and CK+.

Algorithmically, the U-Net model in Face Re-Aging Network (FRAN) receives input in the form of a 5-channel tensor. This tensor consists of the original RGB image to be aged, as well as two single-channel age maps, each representing the input current age and the target age (in this case, set to 75 years old). The pixel values in these age maps are normalized to a range of 0 to 1 to represent a continuous age interval [25].

To ensure experimental consistency and statistical rigor, the total volume of the Phase II dataset was maintained at 10,432 facial image samples, mirroring the distribution of Phase I (5,216 samples for both Depression and Non Depression classes). This consistency allows for a direct comparative analysis of the model's performance before and after the domain adaptation to geriatric features. In line with the established methodology, a stratified 80:20 split ratio was applied: 8,346 images were utilized for fine-tuning the hybrid architecture, while 2,086 images were allocated for the terminal testing and validation phase.

A fundamental advantage of the FRAN architecture, crucial for microexpression analysis, is its ability to preserve the identity and spatial geometry of the original face. Instead of generating a completely new facial image, which is prone to identity loss, the U-Net model predicts per-pixel RGB deltas or offsets. These aging deltas are then added to the original RGB input image to produce the final aged image. This mechanism, combined with the U-Net's skip connections, ensures that the subject's high-resolution facial features and emotional expressions are not distorted. FRAN has proven robust in adding realistic aging textures, such as deep wrinkles, sagging skin (ptosis), and changes in cartilage proportions, such as in the ear and nose areas, while maintaining temporal consistency, head pose, and lighting variations.

This approach effectively produces geriatric twins that retain the original emotional semantics of young subjects while capturing realistic and challenging facial aging features. These generative transformation specifications are strategically applied to various datasets to meet comprehensive testing objectives. In the Aged-CASME II and CK+ datasets, the transformations are specifically used to train a classification model to recognize and distinguish the complex interactions between static wrinkles inherent in older adults and dynamic wrinkles triggered by microexpressions. Modifications to the Aged-RAVDESS dataset aim to simulate older subjects with high expression intensity or who are speaking, a representation highly relevant for supporting video analysis of real-world clinical interviews. Furthermore, the transformations in the Aged-TFEID dataset play a crucial role in generating synthetic

data of Asian older adults aligned with the study's target demographic, significantly reducing racial bias in the model's final inference capabilities.

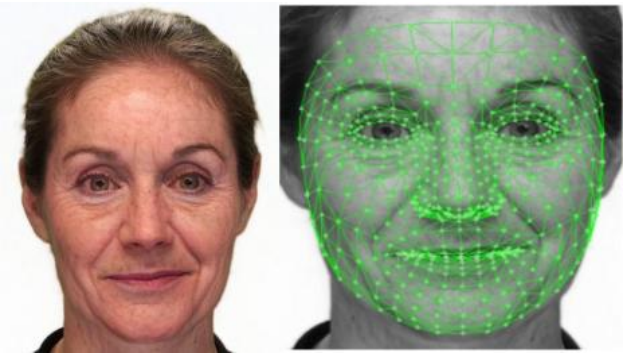

Fig. 5. Left: [Fig. 5(a)] - Synthetic elderly face in color | Right: [Fig. 5(b)] - Grayscale face with green mesh.

To visualize the results of the generative transformation and data preparation process, Fig. 5 presents a composite example from a sample of the Synthetic Elderly Dataset. Fig. 5(a) displays the raw output of the FRAN U-Net model with a target age of 75, clearly demonstrating the addition of aging features such as deep nasolabial folds and sagging skin around the eyes. Fig. 5(b) shows the same image after the standard preprocessing steps described in Section III-B, where the color information has been normalized to grayscale and a high-density MediaPipe mesh topology is overlaid on the face. These geometric and textural representations in Fig. 5(b) serve as the actual input to the CNN-BiLSTM model during the domain adaptation of the fine-tuning phase.

2) Pre-trained model CNN-BiLSTM: The training mechanism in Phase II applies a transfer learning via a fine-tuning strategy. We reload the weights from the best-performing pre-trained model whose performance has been validated in Phase I. This training strategy is designed to prevent damage to learned spatial features (Feature destruction) through the following steps:

a) Partial freezing: The initial convolutional layers (lower layers) in the backbone are frozen. These layers are responsible for detecting basic features such as edges and geometric shapes that are universal and do not change significantly with aging. Fine-tuning high-level layers: the higher-level convolutional layers and the entire temporal module (BiLSTM) are unfrozen. This allows the model weights to adapt specifically to the skin texture of older adults and the dynamics of movement that may be slower or stiffer.

b) Learning rate: The learning rate is significantly reduced compared to Phase I. This reduction aims to prevent catastrophic forgetting, ensuring that the model only makes subtle weight adjustments without forgetting basic knowledge of facial expressions.

Similar to Phase I, we again compared the performance of the VGG19architectures on this synthetic dataset. The goal was to test the robustness of each architecture's adaptation regarding VGG's texture features that were more robust in handling visual noise caused by aging.

3) Mental health output: The final result of Phase II is the Specialized Elderly Depression Classifier. This model is an evolution of the Phase I model, which already had a specific "intuition" for elderly faces. The output layer uses a Softmax activation function that generates probability distributions for two diagnostic classes: Depression and Non Depression.

The final classification is based on an interpretation of emotional valence validated by the clinical literature [26][27]:

a) Depression: Classified based on the probability dominance of persistent negative expression features, particularly Sadness and Anger. This reflects the phenomenon of negative affect potentiation in people with major depression, where subjects tend to maintain negative affect for longer.

b) Non Depression: Classified based on the stability of a relaxed neutral expression and the presence of the happiness feature. The dominance of this feature reflects the absence of negative interpretation bias and healthy emotional regulation function.

This final model is ready to be used for inference on real clinical data as a non-invasive and objective initial screening tool for mental health in older adults.

IV. Results and Discussion

This section presents a comprehensive analysis of the performance of a Deep Learning architecture developed for depression detection in the elderly population using a Two-Phase Transfer Learning scheme. Experimental evaluations are systematically designed to address the research questions, starting from the development of a baseline model in the source domain to testing the model's robustness in the target domain, which has complex aging features.

A. Phase I: Source Domain Benchmarking (Training and Classification)

The first phase of this research focused on building a robust model baseline. The model was trained using a Composite Dataset encompassing a variety of universal micro and macro expressions. The primary goal of this phase was to identify the optimal combination of optimizer and learning rate that achieves optimal convergence, which will serve as the smart initialization for the adaptation process in Phase II.

1) Phase I: Classification results (Testing Metrics): A comprehensive quantitative evaluation was conducted on 12 different experimental scenarios on the testing set of the source domain. Table III, below, presents a complete summary of model performance based on hyperparameter variations.

Best Model Identification: Based on the comprehensive data in Table III, the SGD configuration with a Learning Rate of 0.0005 in No 8 was determined as the best model. This model dominates with an Accuracy of 96.21% and an F1-Score of 0.9621, demonstrating perfect consistency in recognizing the Depression and Non Depression classes, and significantly outperforming all Adam variations.

TABLE III. COMPLETE RECAPITULATION OF PHASE I CLASSIFICATION RESULTS (CNN-BILSTM-VGG19).

No	Optimizer	Learning Rate	Accuracy	Precision	Recall
1	Adam	0.0001	0.7411	0.8225	0.7411
2	Adam	0.0005	0.686	0.7923	0.686
3	Adam	0.001	0.7891	0.8431	0.7891
4	Adam	0.005	0.8337	0.8578	0.8337
5	Adam	0.01	0.7301	0.7447	0.7301
6	Adam	0.05	0.5	0.25	0.5
7	SGD	0.0001	0.5427	0.6295	0.5427
8	SGD	0.0005	0.9621	0.9628	0.9621
9	SGD	0.001	0.908	0.9178	0.908
10	SGD	0.005	0.93	0.935	0.93
11	SGD	0.01	0.8509	0.8747	0.8509
12	SGD	0.05	0.9521	0.9536	0.9521

2) Training dynamics analysis: The varying classification results in Table III above were further analyzed by reviewing the loss and accuracy dynamics during the training process. This analysis revealed the scientific rationale for the superiority of the combination of SGD and LR 0.0005.

A comparative analysis between the Stochastic Gradient Descent (SGD) and Adam optimizers reveals significant differences in stability characteristics. Based on the first through sixth experiments, the Adam optimizer tends to exhibit instability on this dataset. Although Adam is known for its fast initial convergence rate, this method often gets stuck in sharp minima (local minima), resulting in poor model generalization. For example, using a Learning Rate of 0.001, despite high training accuracy, the testing accuracy only reached 78.91%.

In contrast, the SGD optimizer, particularly in experiments 8, 9, 10, and 12, consistently demonstrated high performance with accuracy above 90%. SGD's more gradual weight update mechanism allows it to find flat minima (sloping minimum points). Convergence solutions located in these gentle areas are more robust, allowing the model to maintain its accuracy even when faced with new variations in the testing dataset.

In addition to optimizer selection, analysis of the Learning Rate precision demonstrates that this parameter is a key determinant of model success or failure. In the case of underfitting with a learning rate that is too small, namely Learning Rate < 0.0005, as in experiment 7, namely the CNN-BILSTM SGD Model 0.0001, the model accuracy dropped to 45.06%. This occurs because the model fails to learn complex features within the specified limit of 100 epochs. On the other hand, the overshooting phenomenon occurs when the learning rate is too large, such as Learning Rate > 0.01, as seen in experiment 6, namely the model with the Adam optimizer 0.05 and the model with the SGD optimizer 0.01, which experienced a drastic decrease in performance.

An overly aggressive learning rate causes the gradient step to "jump" the optimal point, namely the lowest loss point, and ultimately damages the model's weight structure. Therefore, the LR value of 0.0005 in the SGD Model in Experiment 8 is the perfect balance point (sweet spot). This value proved to be large enough to help the model escape local minima, but still small enough to perform precise fine-tuning, as evidenced by the lowest Validation Loss achieved with a value of 0.2419 among all configurations.

B. Phase II: Target Domain Adaptation (Augmented Elderly Dataset)

After successfully building a robust baseline model in Phase I using a Universal Dataset, the research proceeded to the Transfer Learning stage to adapt the model to a more specific target domain. In this phase, the optimal weights (pre-trained weights) from the CNN-BiLSTM-VGG19 model trained in Phase I were reloaded and fine-tuned.

The dataset used in this stage was the Augmented Elderly Dataset, which represents facial images of 70-year-olds. The purpose of this augmented data (U-Net generation) was to train the model to recognize emotional features amidst visual noise typical of aging, such as forehead wrinkles, drooping eyelids (ptosis), and sagging skin texture, while maintaining the previously learned baseline feature representation.

1) Phase II: Classification results (Testing Metrics): A comprehensive evaluation of the model's adaptation performance was conducted using a 70-year-old elderly person as the test set. Table IV below presents a complete summary of the results of fine-tuning experiments with various optimizers and learning rate variations to determine the best adaptation configuration.

TABLE IV. COMPLETE RECAPITULATION OF PHASE II CLASSIFICATION RESULTS (DATASET OF ELDERLY PEOPLE 70 YEARS AND OVER - CNN-BILSTM-VGG-19).

No	Optimizer	Learning Rate	Accuracy	Precision	Recall
1	Adam	0.0001	0.7411	0.8225	0.7411
2	Adam	0.0005	0.686	0.7923	0.686
3	Adam	0.001	0.7891	0.8431	0.7891
4	Adam	0.005	0.8337	0.8578	0.8337
5	Adam	0.01	0.7301	0.7447	0.7301
6	Adam	0.05	0.5	0.25	0.5
7	SGD	0.0001	0.5427	0.6295	0.5427
8	SGD	0.0005	0.9621	0.9628	0.9621
9	SGD	0.001	0.908	0.9178	0.908
10	SGD	0.005	0.93	0.935	0.93
11	SGD	0.01	0.8509	0.8747	0.8509
12	SGD	0.05	0.9521	0.9536	0.9521

C. In-Depth Analysis: Model Adaptation to Elderly Faces

Based on the quantitative data in Table IV, the Transfer Learning process from Phase I to Phase II exhibits complex dynamics. The selection of the best model is not based solely on accuracy but also considers the sensitivity metric (Recall), which is crucial for medical diagnosis.

In determining the best model in Phase II, although the SGD configuration with a learning rate of 0.001 in Experiment 9

achieved the highest marginal accuracy of 92.19 per cent, this study determined SGD 0.0005 in Experiment 8 as the most optimal final model. This decision was based on clinical considerations prioritizing recall or sensitivity metrics.

The SGD 0.0005 configuration produced the highest recall value of 0.9271, outperforming the SGD 0.001 configuration at 0.9080. In the context of a medical decision support system for depression detection, false negatives, or the system's failure to detect a patient who is actually ill, carry a much higher risk of fatality than false positives. Therefore, a model with high positive detection capability is an absolute priority. Furthermore, with an F1-score of 0.9153, the SGD 0.0005 configuration demonstrates a harmonious balance between accuracy and sensitivity. This demonstrates that the model is capable of adapting to the wrinkled texture of elderly faces without losing its discriminatory ability.

The success of the 0.0005 SGD model in achieving accuracy above 91 per cent also validates the effectiveness of the transfer learning strategy, particularly when fine-tuning with a low learning rate. Using a learning rate of 0.0005 allows the model to make micro-adjustments to its convolutional weights. Through this mechanism, the model relearns to interpret aging lines on elderly faces as representing texture features, rather than simply visual noise, while retaining the structural facial knowledge learned in Phase I. Conversely, using too high a learning rate proved highly destructive, as evidenced by the catastrophic forgetting phenomenon in Experiment 11 with an SGD setting of 0.01. In this configuration, the model's accuracy plummeted to 50.00 percent, equivalent to the probability of a random guess in binary classification. This complete failure indicates that the overly aggressive gradient update step during the fine-tuning process damaged the optimal pre-trained weight structure from Phase I. As a result, the model lost all memory of facial feature representations and failed to recognize any patterns in the elderly data set.

Furthermore, this experiment also revealed the limitations of the Adam optimizer when faced with domain shift phenomena. Overall, testing with the Adam optimizer from Experiments 1 to 6 yielded significantly lower recall performance than SGD. As a representative example, the Adam configuration of 0.0005 in Experiment 2 achieved a fairly good accuracy of 89.07 percent with a high precision of 93.77 percent, but its recall value remained at 0.8370. This finding indicates that Adam tends to be conservative and often predicts the majority class, namely the Non Depressed class, to avoid prediction errors. This behavior ultimately makes the model less sensitive to capturing subtle manifestations of depressive emotions behind facial skin deformation in the elderly.

Concluding Phase II, this comprehensive analysis confirmed that the CNN-BiLSTM-VGG19 architecture combined with the SGD optimizer and a learning rate of 0.0005 is the most robust configuration for transferring expression detection knowledge to the geriatric domain. This model has successfully overcome the challenges of visual complexity in facial data augmentation of 70-year-old elderly with excellent sensitivity, and is certainly ready to undergo robustness testing on more extreme data sets in the next stage.

V. Conclusion

Based on the experimental results and analysis, this study concludes that the CNN-BiLSTM-VGG19 architecture is proven effective in handling domain shift in elderly faces, with an accuracy of 91.42% and a recall of 0.9271 on the 70-year-old dataset, making it a safe screening tool with minimal risk of false negatives. The model's performance is supported by an optimal configuration through a fine-tuning strategy using the SGD optimizer with a learning rate of 0.0005, a setting that has proven to be the most stable in preventing catastrophic forgetting while providing better generalization compared to the Adam optimizer or the use of high learning rates that are prone to triggering model collapse.

References

[1] ASEANstats. (2023). Ageing ASEAN: Shifting Demographic Structure [Infographic].

[2] Jalali, A., Ziapour, A., Karimi, Z., Rezaei, M., Emami, B., Kalhori, R. P., Khosravi, F., Sameni, J. S., & Kazeminia, M. (2024). Global prevalence of depression, anxiety, and stress in the elderly population: a systematic review and meta-analysis. BMC Geriatrics, 24(1), 809.

[3] Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 Validity of a Brief Depression Severity Measure. Journal of General Internal Medicine, 16(9), 606-613.

[4] Cui, L., Li, S., Wang, S., Wu, X., Liu, Y., Yu, W., Wang, Y.,

[5] Tang, Y., Xia, M., & Li, B. (2024). Major depressive disorder: hypothesis, mechanism, prevention and treatment. Signal Transduction and Targeted Therapy, 9(1).

[6] Parikh, A., Sadeghi, M., Richer, R., Rupp, L. H., Schindler-Gmelch, L., Keinert, M., Hager, M., Capito, K., Rahimi, F., Egger, B., Berking, M., & Eskofier, B. M. (2024). <i>Exploring Facial Biomarkers for Depression through Temporal Analysis of Action Units

[7] Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S., & Rosenwald, D. P. (2013). Social risk and depression: Evidence from manual and automatic facial expression analysis. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013).

[8] Pot, A., Carstensen, L.L. A generated image repository of aging faces. Sci Data 12, 1610 (2025). https://doi.org/10.1038/s41597-025-05909-6.

[9] Ekundayo, O. S., & Ezugwu, A. E. (2025). Deep learning: Historical overview from inception to actualization, models, applications and future trends. Applied Soft Computing, 181, 113378. https://doi.org/10.1016/j.asoc.2025.113378

[10] Mumuni, A., & Mumuni, F. (2025). Automated data processing and feature engineering for deep learning and big data applications: A survey. Journal of Information and Intelligence, 3(2), 113–153. https://doi.org/10.1016/j.jiixd.2024.01.002

[11] Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput Sci. 2021;2(6):420. doi: 10.1007/s42979-021-00815-1. Epub 2021 Aug 18. PMID: 34426802; PMCID: PMC8372231.

[12] Singh, S., & Prasad, S. V. A. V. (2018). Techniques and challenges of face recognition: A critical review. Procedia Computer Science, 143, 536–543. https://doi.org/10.1016/j.procs.2018.10.427

[13] Zhalgas, A., Amirgaliyev, B., & Sovet, A. (2025). Robust Face Recognition Under Challenging Conditions: A Comprehensive Review of Deep Learning Methods and Challenges. Applied Sciences, 15(17), 9390. https://doi.org/10.3390/app15179390

[14] Yazici, A., Kucukyilmaz, T., Dokeroglu, T., Sharipbay, A., Lee, M.-H., & Tyler, B. (2026). State-of-the-art multimodal emotion recognition: A comprehensive survey and taxonomy. Intelligent Systems with Applications, 30, 200642. https://doi.org/10.1016/j.iswa.2026.200642

[15] Gao, Y., Xiao, Z., Gong, Z., Huang, S., & Zhu, H. (2025). Spatiotemporal Deformation Prediction Model for Retaining Structures Integrating ConvGRU and Cross-Attention Mechanism. Buildings, 15(14), 2537. https://doi.org/10.3390/buildings15142537

[16] Hayat MT, Allawi YM, Alamro W, Sultan SM, Abadleh A, Kang H, Zreikat AI. A Hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)-Attention Model Architecture for Precise Medical Image Analysis and Disease Diagnosis. Diagnostics (Basel). 2025 Oct 23;15(21):2673. doi: 10.3390/diagnostics15212673. PMID: 41225966; PMCID: PMC12608630.

[17] Amadi, Christian & Odii, Juliet & Ofoegbu, Christopher & Okpalla, Chidimma. (2023). Emotion Detection Using a Bidirectional Long-Short Term Memory (BiLSTM) Neural Network. International Journal of Current Pharmaceutical Review and Research. Vol 4, no 11. 1718-1732.

[18] Yadav, Srijana & Mangalampalli, S.. (2025). Deepfake defense: Combining spatial and temporal cues with CNN–BiLSTM–transformer architecture. PLOS One. 20. 10.1371/journal.pone.0334980.

[19] Najia, Mechichi & Faouzi, Benzarti. (2025). An Enhanced Hybrid Model Combining CNN, BiLSTM, and Attention Mechanism for ECG Segment Classification. Biomedical Engineering and Computational Biology. 16. 10.1177/11795972251341051

[20] Lee, T., Baek, S., Lee, J., Chung, E. S., Yun, K., Kim, T. S., & Oh, J. (2024). A Deep Learning Driven Simulation Analysis of the Emotional Profiles of Depression Based on Facial Expression Dynamics. Clinical Psychopharmacology and Neuroscience, 22(1), 87–94.

[21] Sugiyanto, S., Purnama, I. K. E., Yuniarno, E. M., Anggraeni, W., & Purnomo, M. H. (2024). Depression Classification Based on Facial Action Unit Intensity Features Using CNN-Poolingless Framework. International Journal of Intelligent Engineering and Systems, 17(5), 172–187.

[22] Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., & Fu, X. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLOS ONE, 9(1), Article e86041. https://doi.org/10.1371/journal.pone.0086041

[23] Livingstone, S. R., & Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A highly validated multimodal database of emotional speech and song. PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

[24] Chen, J. W., & Yen, N. S. 2007. Taiwanese Facial Expression Image Database (TFEID). Brain and Consciousness Research Center, National Chengchi University.

[25] Lucey, P., et al. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 94-101

[26] Garcia, R., et al. 2022. FRAN: Face Re-aging Network for Realistic Geriatric Transformation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2415-2424.

[27] DOI : 10.1109/CVPR52688.2022.

[28] Li, M., Li, J. X., Han, J. M., Liu, X. H., Gao, X. Z., Chen, L. M., Zhou, Z. H., & Zhou, H. L. 2026. The Socio-emotional preference task in major depressive disorder: ERP evidence of social appraisal dysfunction. BMC Psychiatry, 26(31). https://doi.org/10.1186/s12888-025-07684-5