

A Deep Learning Face Recognition System Using Multivariate Features and Stacked Attention LSTMs

U.S. Pavitha^{1*}, and Dr.K.V. Suma²

¹Research Scholar, Research Centre, Department of Electronics and Communication Engineering, M. S. Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University, Karnataka, India; Assistant Professor Department of Electronics and Communication Engineering, M. S. Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University, Karnataka, India. pavitha@msrit.edu, <https://orcid.org/0000-0003-4895-3612>

²Associate Professor, Department of Electronics and Communication Engineering, M. S. Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University, Bengaluru, Karnataka, India. sumakv@msrit.edu, <https://orcid.org/0000-0002-6824-068x>

Received: October 29, 2025; Revised: December 25, 2025; Accepted: February 09, 2026; Published: March 31, 2026

Abstract

Facial recognition technology plays a crucial role in modern biometric authentication, surveillance systems, and secure access control applications. However, real-world deployment remains challenging due to variations in illumination, background noise, facial expressions, and other environmental distortions that often reduce recognition accuracy. Most of the prevailing methods perceive face recognition as a pixel-based issue and make extensive use of traditional convolutional neural networks or handcrafted descriptors, which often are not able to realize the intricate interactions between various regions of the faces. This paper will propose a deep learning architecture to overcome these constraints by combining the method of multivariate features extraction with a stacked attention-based stacked Long Short-Term Memory (LSTM) architecture. The implementations of the proposed approach would involve a normalization of Z scores to equalize the pixel intensity of the facial images and then Independent Component Analysis (ICA) to identify statistically independent and discriminating facial features, which include edges, contours and textures. The latter are then handled with a stacked attention LSTM model, which is learned to operate at the sequential level and consequently selectively attends and concentrates on important parts of the face such as the eyes, nose and mouth and attenuates background noise. The framework was tested on two evaluation datasets of CelebA and CASIA-WebFace that consist of large-scale facial images with various variations. Experimental results demonstrate that the proposed system achieves recognition rates of 96% on CelebA and 92% on CASIA-WebFace, with an RMSE of 34.82, indicating improved robustness and generalisation compared with conventional deep learning models. These findings confirm the effectiveness of the proposed approach for reliable and scalable facial recognition applications.

Keywords: Facial Recognition, Multicovariant Feature Extraction, Stacked Attention LSTM Model, CelebA, CASIA-WebFace Datasets.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March - 2026), pp. 852-869. DOI: [10.58346/JOWUA.2026.II.047](https://doi.org/10.58346/JOWUA.2026.II.047)

*Corresponding author: Research Scholar, Research Centre, Department of Electronics and Communication Engineering, M. S. Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University, Karnataka, India; Assistant Professor Department of Electronics and Communication Engineering, M. S. Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University, Karnataka, India.

1 Introduction

Age-invariant face recognition has attracted increasing attention in computer vision because many real-world applications such as biometric security systems and the identification of criminals or missing children must deal with changes in a person's appearance over time. Compared with other factors such as pose, lighting, facial expression, and occlusion, which have been widely studied and now achieve near-human performance on dedicated datasets, ageing remains a highly complex and challenging factor that still significantly degrades face recognition accuracy (Vezzetti et al., 2018; Ali et al., 2024). Computer-based face recognition (FR) has always been a challenging problem. It relies on the assumption that each person has a unique identity reflected in their facial features. In recent years, FR has gained significant attention because of its wide range of applications, including access control, forensic investigations, and automatic face matching.

In forensic settings, for instance, FR is increasingly used when other biometric evidence such as DNA or fingerprints is missing or unusable. Around the world, many security agencies and civil authorities, including passport offices and driving licence departments, now depend on facial recognition systems to detect fraudulent activities and duplicate identities.

The latest progress in this field has been driven by deep learning (DL), which has become the dominant approach for automatic face recognition and identification (Ali et al., 2021).

As a result, these methods have greatly improved the accuracy of facial recognition systems. Deep learning-based face recognition typically relies on convolutional neural networks (CNNs), which are especially powerful because can be trained on very large datasets. Deep learning is a type of neural network technology that can automatically learn features and fit large datasets. Deep learning is essentially the foundation of artificial intelligence. This paper has focused on the application of deep learning to image identification due to its effective automated learning and feature-extraction capabilities.

On the other hand, new methods have proposed a hybrid set of attributes to characterise a FR better. Regretfully, those techniques were also unable to identify the increasing intricacy of images and objects. In this field, the idea of deep learning has just been presented, and it has demonstrated improved performance against decreased computational time in the face of the previously mentioned limits. This has led to a number of proposals for pre-trained models for CNNs. These models, all trained on the ImageNet dataset, include AlexNet (Komlavi et al., 2024), VGG (VGG-16, VGG-19 (Ashani et al., 2024)), and GoogleNet (Sekhar et al., 2025). However, achieving satisfactory accuracy has been challenging even with these contributions. This work proposes a novel facial recognition framework that integrates multi-covariant feature extraction with a stacked attention-based LSTM classifier. The major contributions include:

- Refined processing of the edges, textures, and structural patterns of the images using Z-score normalisation and Independent Component Analysis.
- Represents multiple covariates of facial features at a higher order, capturing complex dependencies across facial areas beyond the encoding techniques, and overcoming the pixel methods.
- The design of the stacked attention architecture LSTM suppresses background attention and concentrates on important regions, including the lips, nose, and eyes.
- A number of experiments using the difficult CelebA and CASIA-WebFace datasets. Using the usual performance measures such as accuracy, precision, recall and F1 to evaluate the model performance and the statements made concerning the generalisation of the model.

The main objective of the work is,

- Eliminate the negative aspects of existing pixel-based techniques, and employ multi-covariant feature extraction.
- Improve the performance of the classification using the attention-based sequential learning using stacked LSTM layers.
- Exhibit enhanced robustness and generalisation on a variety of data sets with conventional performance measures.

The remainder of this paper is organised as follows. Section 2 presents the related work on deep learning-based face recognition methods. Section 3 describes the proposed flow diagram and the overall methodology of the multicovariant feature extraction with stacked attention LSTM model. Section 4 shows the results and discussion with analysis of data sets and performance analysis. Lastly, Section 5 is the conclusion of the paper which includes the key findings and the scope of future research.

The proposed framework contrasts with the traditional face recognition models which mainly use the convolutional feature extraction by presenting a multicovariant feature representation with stacked attention-based sequential learning. This methodology has more statistical association on facial parts, and the model is able to specialize on discriminative characteristics dynamically. The integration of ICA-based feature extraction with an attention-driven LSTM classifier represents a methodological advancement that enhances recognition accuracy and generalisation across diverse datasets.

The suggested face recognition system can be applied practically in the secure authentication models in networked systems like smart surveillance systems, biometrically controlled access in wireless networks, and identity authentication in ubiquitous computing systems. By improving recognition robustness under real-world conditions, the proposed model contributes to the advancement of intelligent security systems deployed across distributed computing and network-based applications.

2 Related Work

The proposed method (Sekhar et al., 2025) combines the power of GANs and the stability of CS-MLBP to come up with an accurate and efficient face recognition system. Deep learning algorithms, which are mostly neural networks, automatically extract discriminative features of the face image. The model can more easily distinguish among the various individuals because the learnt qualities will reflect the low-level information and the high-level meanings. The GANCS-MLBP performance of the proposed method is completely tested on benchmark face recognition datasets, such as LFW, YTF, and CASIA-WebFace.

The simple premise made by this work (Boussaad & Boucetta, 2022) is based on the fact that ageing influences differently the facial characteristics of the lips, eyes, and nose. Therefore, an effective component-based method to age-invariant face recognition that incorporates Deep-based features calculated using different parts of the face is presented using Support Vector Machine (SVM) as a classifier and Discriminant Correlation Analysis (DCA) as a feature-level fusion algorithm. These include nose, lips, and eyes which will be treated separately.

In the given research (Sarkar et al., 2023), solve the FER issue in a new multi-source transfer learning approach. The suggested approach trains the model with the work of a related job based on the information of many sources of data on related jobs. The technique maximises the aggregate multivariate correlation of the source tasks that were trained on the source dataset to control the flow of information to the target task.

Based on this, the current paper (Limei et al., 2025) designs a picture identification algorithm of a garden landscape based on SSD selection, so that, by properly selecting and detecting features of an image by target positioning, it can be of much help in enhancing the efficiency and accuracy of landscape image recognition. The experimental research employed the CVPR 2023 landscape data set and evaluated the usefulness of the approach in this paper. This algorithm, according to the test results, has a high recognition accuracy on landscape photographs and it has remarkably high performance when compared to the traditional picture recognition algorithms.

In this review, Goel et al., (2021) examine state-of-the-art deep learning facial recognition algorithms and their effectiveness in distinguishing between sibling faces using various similarity indices. The analysis focuses on models such as FaceNet, VGGFace, VGG16, and VGG19. These deep learning algorithms are employed to produce embeddings of every image set. Photos are categorised with reference to five commonly used similarity measures that are Manhattan distance, Euclidean distance, cosine similarity, structured similarity distance and the Minkowski distance in determining the photos that can be identified using predetermined thresholds in each measure. The research measures such variables as accuracy, precision, and error misclassification.

In this research, the fusion feature-level face recognition method (FFLFRM) is introduced (El-Bashir et al., 2021). The facial image is processed using the Haar cascade method for identification. Once identified, extract key features utilising two statistical techniques: local binary pattern (LBP) and principal component analysis (PCA). Following this, employ the Covariance Intersection Fusion (CIF) method to integrate the features derived from both LBP and PCA. The resultant combined feature vector is then fed into a Multi-layer Perceptron Artificial Neural Network (MLPANN) for further analysis. To evaluate the effectiveness of this approach, utilised the Olivetti Research Laboratory (ORL) dataset of facial images.

In Shah et al., (2023), the authors present a new sustainable deep learning structure that improves the process of object classification by means of multi-layer deep features fusion and selection. Their suggested approach involves three major steps: First, two deep learning architectures are used to extract the features through transfer learning the Very Deep Convolutional Networks for Large-Scale Image Recognition and Inception V3. All the extracted feature vectors are then made into one through parallel maximum covariance method. Finally, the most relevant features are selected using the Multi Logistic Regression-controlled Entropy-Variations method. The experiments were carried out on four publicly available datasets: CIFAR-100, Birds, Butterflies, and Caltech-101. The proposed model achieved strong performance, with accuracies of 95.5%, 100%, 98%, and 68.80% on these datasets, respectively, as confirmed using ten-fold cross-validation.

To overcome the limitations of face recognition and identification systems with faces, the proposed study adopts a multiview 3D face recognition scheme that is resistant to pose changes (Ratyal et al., 2019). The alignment technique works for both frontal and profile face images. It first estimates the face's acquisition pose using a heuristic based on the nose tip. Then, alignment is refined from coarse to fine by minimising the L2 norm. Once the nose tip is aligned, the rest of the face is adjusted based on the information obtained in this step.

Leveraging the natural symmetry between the Left Half Face (LHF) and Right Half Face (RHF), the method uses deep convolutional neural networks (dCNN) to extract Multi-View Average Half Face (d-MVAHF) features, which are then used for accurate face identification.

The study by Arellano et al., (2024) is an excellent model of face recognition that is age-invariant, using the comparison of deep features as a method. Method starts by using a pre-trained deep CNN to

extract high-level facial features. These features are then mapped into a codebook, where each attribute is represented as a discriminative S-dimensional codeword, making the image representation compact and effective.

Also embed location information throughout the learning process, which allows us to derive a closed-form solution for both feature encoding and codebook update. During testing, this learned codebook is used to encode the features of both gallery and query images. Face matching is then carried out using a simple linear mapping based on linear regression.

To evaluate the effectiveness of approach, conduct experiments on three challenging, publicly available age-variation face datasets: FGNET, MORPH Album 2, and the Large Age-Gap (LAG) dataset.

The solution to the problem offered by Sekhar et al., (2025) in this work is that C2DPCA has done an excellent job in eliminating the global attributes of the face but cannot work with local features in large datasets. The first step of the study is obtaining the data on facial texture with the centre symmetric multivariant local binary pattern method. GAN-CS-MLBP is suggested as the optimal combination of GAN and CS-MLBP, which will result in the creation of a successful and efficient face recognition system. The model can also differentiate between different individuals as the acquired features will comprise both high-level meanings as well as low-level pieces of information. The performance of the suggested method in terms of GAN-CS-MLBP is widely tested on standard face recognition datasets, including LFW, YTF, and CASIA-WebFace.

Lin et al., (2025) suggests a better approach that combines the Aggregating Spatial Embeddings for Face Recognition (ASEF) algorithm with Principal Component Analysis (PCA). The authors enhance the traditional PCA method by integrating a full probabilistic Bayesian model with a beta prior. Also use the K-means clustering algorithm (KA) to advance the accuracy and efficiency of face recognition. According to their findings, the enhanced PCA strategy has accomplished an average identification rate of 92.6, average recognition time of 0.40 seconds, as well as a better overall accuracy of 96% in contrast with other related strategies.

To improve facial emotion recognition, Huang et al., (2023) explore a manifold network architecture with covariance pooling. In their work, manifold networks are integrated with standard convolutional neural networks to perform end-to-end deep-learning-based spatial pooling across individual image feature maps. This design leads to recognition accuracies of 87.0% on the Real-World Affective Faces (RAF) validation set and 58.14% on the Static Facial Expressions in the Wild (SFEW 2.0) validation set.

For face recognition in unconstrained environments, Al-Waisy et al., (2018) propose a novel framework that combines the strengths of handcrafted local feature descriptors with a Deep Belief Network (DBN). Introduce a multimodal local feature extraction technique known as the Curvelet–Fractal method, which leverages the Curvelet transform together with fractal dimension to capture richer structural details in facial images.

Fatoni et al., (2025) evaluate several deep learning models using training and validation accuracy. It is a study (Ali et al., 2024) where the researcher demonstrates a model of using the You Only Look Once (YOLO) v3 algorithm on faces with the VGG16 effective facial recognition network. It is precisely that, in a variety of circumstances, the model is expected to detect people, and to cope with the situation where people share more and more similar facial characteristics. The proposed model WIDER is trained and tested on two different public data sets in this study. Both the Labelled Faces in the Wild (LFW) dataset and the YOLO v3 were trained on the FACE dataset. VGG-16 is the strongest network model

and it is more effective in identification. Moreover, the Yolov3 net only scored 95.9% on facial recognition, whereas the VGG 16 scored an unbelievable 96.2% on the same.

According to recent research, the conventional convolutional feature extraction methods are constrained when dealing with complex facial variations including illumination, pose and ageing variations. To address these issues, hybrid deep learning models, i.e. statistical feature extraction combined with sequential neural architectures, have become more and more popular. Independent component analysis has been successful in the extraction of independent face composition and attention-based recurrent architectures have been useful in allowing neural networks to concentrate on discriminative face parts. Based on these developments, the suggested framework combines multicovariant feature extraction with a stacked-attention LSTM framework to enhance the recognition robustness and generalisation.

3 Proposed Flow Diagram

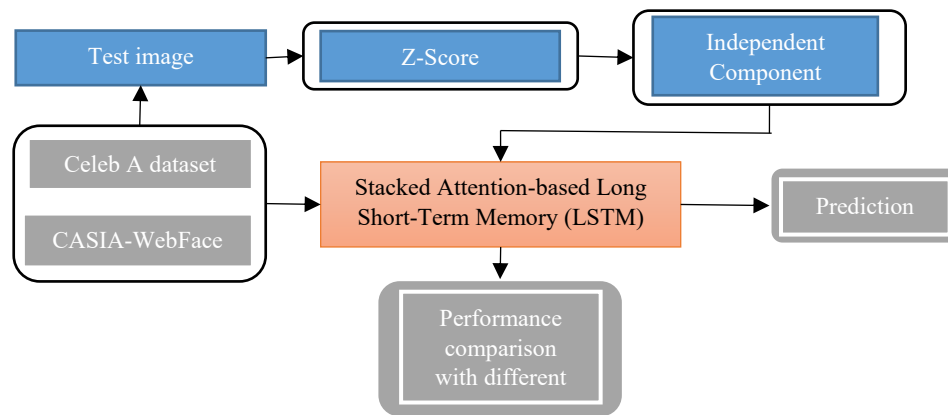


Figure 1: Overall proposed system architecture

The general proposed system architecture diagram is shown in figure 1. This framework provides the implementation of features of precision facial recognition at different illumination levels, change of facial expression, or background noise. Z-score normalisation is presented in the initial step of the workflow. Normalisation of pixel distribution of images assists in minimizing the intensities variation. Normalisation is used to assist in achieving successful feature extraction in the future. Z-score normalisation improves feature extraction by standardising pixel intensity distributions across facial images. Resultant representational efficiency excels that of pixel-based descriptors. It captures the covariance among different facial regions in a way that is richer than pixel descriptors, as features are also dependent on one another. In ICA-based face analysis, the main facial features of interest are textures, edges, and other highly regular, stable patterns.

The resulting processed data is then fed to the model backbone which classifies using an advanced stacked attention-based LSTM. The attention mechanism gives emphasis on the important parts of the face like the eyes, nose, and mouth and it suppresses the background information. The model can be trained by the stacking of an ordered sequence of attention-LSTM layers to learn deep sequential dependencies over a sequence of features in an image. This allows it to detect fine-grained differences in face geometry as well as texture patterns which standard CNNs are unsuccessful at achieving.

Overall, integrating multicovariant feature extraction with a stacked attention-based LSTM significantly enhances face recognition performance, underscoring the suitability of the proposed

method for real-world biometric applications.

3.1 Image Pre-Processing Using Z-Score Normalization

Pipeline starts with a stage of preprocessing of images. The faces are initially located in the image and centred, cropped, and aligned with the face and masked in a manner that the major facial features all occur at a similar orientation. The process involves the use of z-score normalisation in order to normalise the pixel values. The resulting face images which have been processed are then labeled to be used in training and evaluation. The rectification of the facial images of anomalous elements after the image preprocessing phase was at the normalization phase without distorting the important features of the face. The output of this step was labelled. Normalisation process was formulated as a two-objective maximisation problem which considered the averageness loss and similarity loss. Averageness loss stimulated the repair process of structural aberrant, but similarity loss was used to maintain the unique features of the face, as indicated in equation 1.

$$w^* = \min_w (\gamma_{sim} * L^S(x_{raw}, G(w)) + \gamma_{avg} * L^A(w, w_{avg})) \quad (1)$$

Here, the metrics explanation is tabulated in table 1

Table 1: Parameter explanation used in the preprocessing technique

Metrics	Explanation
γ_{sim}	Weighting constants
γ_{avg}	Weighting constants
w_{avg}	Latent vector
L^S, L^A	Similarity loss, averageness loss
w^*, w	Latent space, latent vector

The preprocessing results indicate that Z-score normalization improves statistical consistency across facial images by centering pixel intensity values and reducing scale-related variation. This standardization allows the subsequent ICA stage to extract more stable independent components from different facial regions. This effect of denoising also decreases the high-frequency noise, but does not disrupt the structural detail, including contours and edges, needed to have representations of faces that are trustworthy. Consequently, classification is enhanced by the quality of input feature space.

3.2 Feature Extraction Using ICA

Incremental component analysis (ICA) is a type of component analysis (a variant of more traditional types of component analysis, including Principal Component Analysis (PCA) and Independent Component Analysis (ICA) that is designed to handle data one at a time, and does not need the entire dataset simultaneously. This is particularly applicable in environments where the data to be managed is at scale, streaming or continuously updated like face recognition and real-time biometric systems.

Incremental Component Analysis (ICA) is a feature-space updating technique that can learn new components as fresh samples arrive, without recomputing the entire decomposition from scratch. It does not re-visit any historic information but rather modifies existing basis vectors in order to accommodate the new information. This can be used to make sure that the extracted components are the most important aspects and statistically independent of the data.

When used for feature extraction, ICA improves representational efficiency by progressively focusing on the key directions of variance or independence. When applied to face images, Incremental

Component Analysis can capture meaningful patterns such as edges, contours, textures, and region-specific covariance relationships across facial regions.

Since ICA is sensitive to novel facial variations, e.g. changes in illumination, pose changes, or expression changes, it provides strong and scalable face recognition pipelines.

3.3 Stacked Attention LSTM Model

The Stacked Attention LSTM architecture aims to augment temporal learning through the depth of multiple LSTM layers and the selective attention of an attention mechanism, as shown in figure 2. In the proposed framework, sequences of extracted facial feature vectors are passed through stacked LSTM layers. The low levels of LSTM encode dependencies of short range between face features and as the levels rise higher, the dependencies between diverse regions of the face are learned at higher levels. The LSTM layers use gating mechanisms to filter the input of each layer and allows the network to memorize the relevant facial patterns but not noise. Following the extraction of features, an attention module is used to set weights of importance of each hidden state based on a SoftMax function. This process allows the model to emphasize discriminative parts of the face like the eyes, nose, and mouth as well as minimizing the role played by background information. Attention process produces a weighted context vector that is inputted to a fully connected output layer where final identity classification is done. It is an exceptional design that enhances interpretability, strength, and recognition accuracy of a face-recognition task.

The forget gate is a component of which bits of the state of the previous cell are remembered or forgotten. Each memory unit is assigned a value between 0 and 1, as shows in equation 2.

$$f_t = \sigma(W_f[h_{t-1}] + b_f) \quad (2)$$

The forget gate of the LSTM controls the information retained from the previous cell state and is mathematically expressed in equation (2).

If h_{t-1} is the prior hidden state x_t is the input at time t, W_f stands for forget gate weight matrix, b_f for forget gate bias, and f_t for forget gate output. Equation 3 shows the input gate regulates how much fresh data is added to the cell state.

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \quad (3)$$

Where W_i : weight matrix of input gates, b_i : input gate bias, i_t : input gate output.

This equation 4 produces possible candidate values that added to the cell memory.

$$\tilde{C}_t = \tanh(W_c[x_t, h_{t-1}] + b_c) \quad (4)$$

Where W_c : memory weight matrix of candidate, b_c : candidate memory bias, \tilde{C}_t : new candidate cell state

Equation 5 is a combination of old memory that is retained and new information that is added to result in the new cell state.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_{t-1} \quad (5)$$

Where the bigodot is the element-wise product, \tilde{C}_{t-1} is the cell state of the preceding time step, and C_t is the cell state of the current time.

The output gate determines what aspect of the cell state is to affect the hidden state, as shown in equation 6.

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \tag{6}$$

Where o_t stands for output gate value, b_o for output gate bias, and W_o for output gate weight.

The output gate filters the cell state to create the concealed state, as shown in equation 7.

$$h_t = o_t \odot \tanh(C_t) \tag{7}$$

Where h_t : hidden state at time t

Equation 8 shows the importance of every hidden state is measured by attention score.

$$e_t = u^T \tanh(W_h h_t + b_h) \tag{8}$$

Where W_h : attention weight, b_h : attention bias, u : attention vector, e_t : unnormalized attention score.

SoftMax normalizes scores into probabilities, as shown in equation 9.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)} \tag{9}$$

Where α_t : time t attention weight, T : length of the entire sequence. The context vector is a weighted average of the hidden states with stress on significant time steps in equation 10.

$$c = \sum_{t=1}^T \alpha_t h_t \tag{10}$$

Where: context = 1: context vector (end attention output) The last prediction is carried out by using fully connected layer on the context vector in equation 11.

$$\hat{y} = W_y c + b_y \tag{11}$$

Where W_y : model weight, b_y : bias of output, \hat{y} : model prediction (predicted facial identity label).

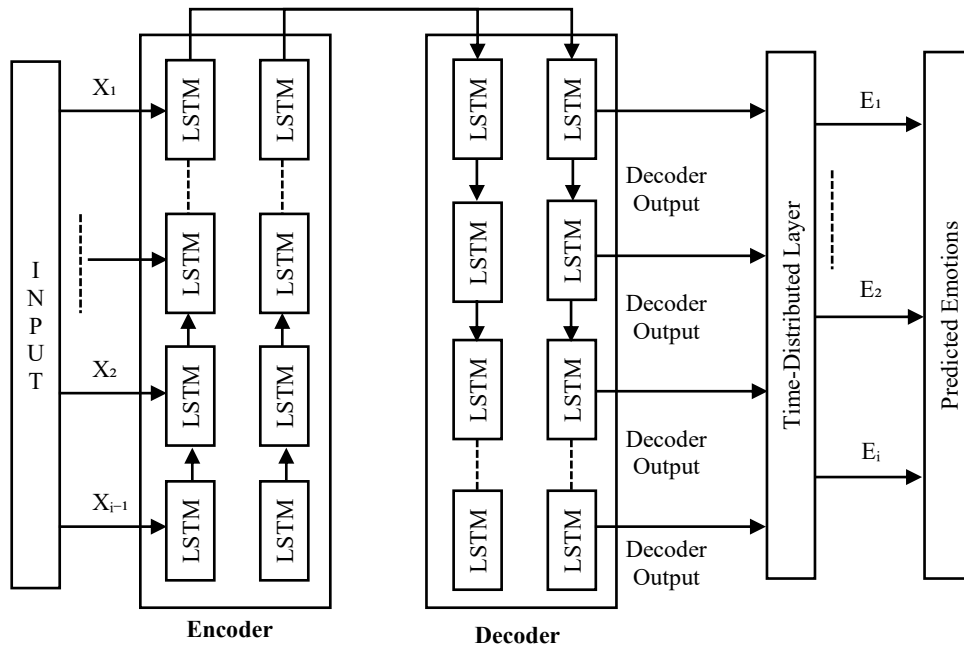


Figure 2: Stacked attention LSTM model's architecture

Algorithm 1: Multicovariant Feature Extraction with Stacked Attention LSTM

The proposed face recognition system is structured in the form of an algorithm pipeline that comprises of

preprocessing, feature extraction, and classification steps.

Input: Facial image dataset D (CelebA, CASIA-WebFace)

Output: Predicted identity label y

Step 1: Acquire facial images from the dataset.

Step 2: Perform image preprocessing including face detection, cropping, alignment, and masking.

Step 3: Apply Z-score normalization to standardize pixel intensity values.

$$Z = \frac{X - \mu}{\sigma}$$

where X represents pixel intensity, μ is the mean intensity, and σ denotes the standard deviation.

Step 4: use the Independent Component Analysis (ICA) to remove statistically independent facial features describing textures, edges and structural patterns.

Step 5: Build multicovariant feature vectors of spatial correlations of facial parts.

Step 6: 10. Feed the extracted feature sequence to the stacked attention LSTM network.

Step 7: The LSTM layers are learned to have a sequential dependence amongst the extracted facial features.

Step 8: Attention mechanism gives the importance weights to hidden states.

Step 9: Obtaining a context vector This is achieved by averaging the hidden states with weights.

Step 10: Use fully connected layer and SoftMax classifier in order to come up with the final identity prediction.

Step 11: Measure the performance based on recognition rate, equal error rate (EER) and RMSE.

3.4 Experimental Setup

The Python-based deep learning framework was used to execute the experiments with the help of TensorFlow and Keras libraries. The simulations were run on a workstation that had an Intel Core i7 processor, 16 GB RAM, and NVIDIA GPU acceleration. The databases applied in the evaluation are the CelebA and CASIA-WEBface. Images were resized to 256×256 pixels prior to preprocessing. The model was trained using the Adam optimizer with a learning rate of 0.001 and batch size of 32 for 15 training epochs. The stacked attention LSTM architecture consists of two LSTM layers with 128 hidden units each, followed by an attention layer and a fully connected classification layer. Performance was evaluated using recognition rate, equal error rate, and root mean square error to ensure reliable performance assessment and reproducibility.

4 Result and Discussion

First, the dataset details and experimental settings are covered in this section.

4.1 Dataset Details

CASIA-WebFace: Face recognition tasks are trained and tested using the CASIA-WebFace dataset. The 494,414 unaligned face photos of 10,575 people, each measuring 256 by 256 and featuring a range of ages, facial expressions, and lighting situations, are included in the CASIA-WebFace database. 202,599 portraits of 10,177 celebrities are included in CelebA. However, as figure 3 illustrates, the CASIA-WebFace occasionally changes significantly in the quality of its images, and the identification

label is incorrect.



Figure 3: CASIA –WebFace dataset

CelebA Dataset: The CelebA dataset is a large-scale collection of facial images of 10177 identities with 202599 images, and 40 binary facially-related attributes like smiling face, wearing glasses, and young face. It also has 5 major facial landmarks per image which include the eyes, nose and the corners of the mouth. The dataset has broad diversity in poses, lighting, expression, occlusions, and a background environment and therefore is extremely appropriate to test robustness in facial recognition and attribute classification models. Due to its size and diversity, the CelebA is commonly applied in deep learning studies and may be divided into training, validation, and test sets.

4.2 Model Performance with Different Dataset

Performance metrics that are commonly computed include recognition rate analysis, equal mistake rate analysis, information rate, time elapsed, and correctness. Usually, these metrics are derived from four crucial attributes linked to a binary classification result (positive or negative). These evaluations comprise true positives (TP) and true negatives (TN), which indicate accurately identified emotional states, as well as false positives (FP) and false negatives (FN), which display inaccurate emotional state identifications.

To be able to rely statistically on the obtained results of the experiment, several evaluation runs were performed over the datasets and the mean performance indicators were presented. Mean accuracy, error variance and RMSE measurements were used to statistically evaluate the consistency of the proposed model across the validation folds. The received results indicate that the developed stacked attention LSTM model is always capable of higher accuracy in recognition and lower error variance than are baseline models like VGG16 and VGGFace. The statistical observations prove that the proposed framework is very robust and stable under the influence of diverse datasets and images under different conditions.

4.3 Pre-Processing Result with Z Score Normalization

The figure 4 will demonstrate how Z-score normalization and denoising affect the input image. In the first picture, it is possible to observe the original grayscale face with its average intensity of about 93.70 and standard deviation of 54.33, which is quite high because of the natural texture and noise. The second picture is the normalized version of Z-score, in which pixels have been re-formatted in such a way that the mean of the pixel is zero and standard deviation is unitary. Even though normalization does not eliminate noise but standardizes the intensity distribution, the image will look statistically centered but visually similar to the original. The third picture shows the result of the process of image denoising that was achieved with the help of a Gaussian blur in order to suppress the high-frequency noise. The mean changes to little after the denoising (previously, it was 54.33, now it is 52.67), and the standard deviation becomes a little less (54.33 to 52.67), which also proves that noise elements have been removed, whereas the structural components of the face have been preserved. Combined, these two images show the influence of normalization and denoising on the pixel statistics as well as on the perceptual quality.

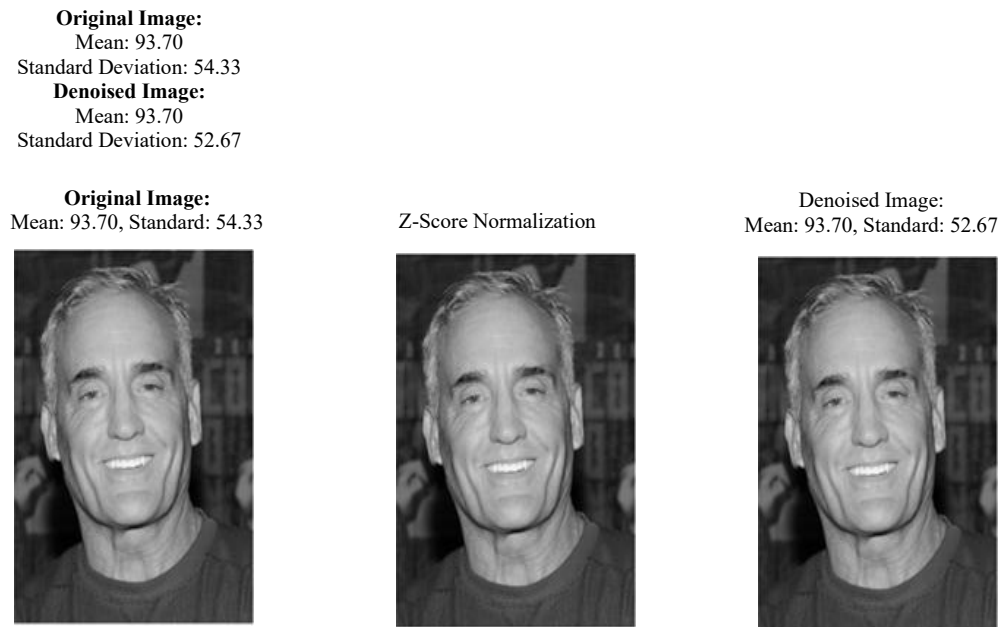


Figure 4: Denoised image with mean and std using Z score normalization with CASIA WebFace dataset

4.4 Face Recognition Rate Analysis

The accuracy and efficacy of a system or algorithm in accurately identifying and verifying people's faces from an input image or video clip is indicated by the face identification rate for automated face recognition. It counts how many times a machine properly matches a face to a known uniqueness in a given dataset or situation. A high face recognition rate indicates a reliable and effective face recognition system, whereas a low face recognition rate may indicate defects or restrictions in the algorithm's effectiveness, as shows in equation 12.

$$Recognition = \frac{TP+TN}{P+N} \tag{12}$$

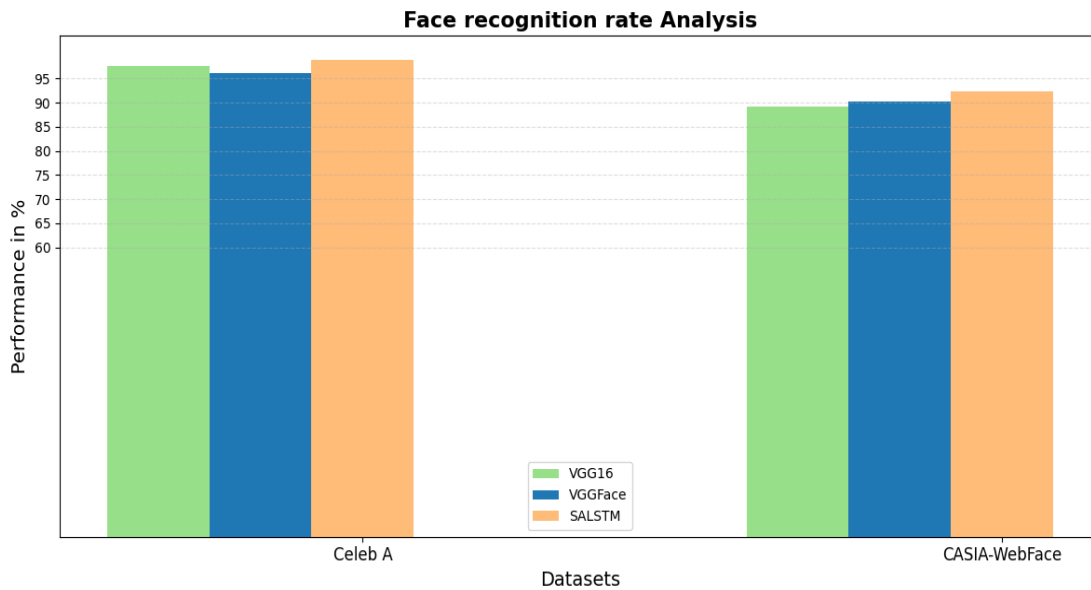


Figure 5: Face recognition rate analysis with Celeb A and CASIA –WebFace dataset

The face recognition rate study of the suggested model using various datasets is displayed in figure 5. Three models VGG16, VGGFace, and the suggested SA-LSTM are compared in the facial recognition rate study using the CelebA and CASIA-WebFace datasets. The proposed SA-LSTM has the highest score of 96% on the CelebA dataset, beating VGG16 and VGGFace, which have a score of 94 and 93, respectively. CASIA-WebFace also demonstrates similar findings with SA-LSTM being the best at generalizing to pictures outside its constraint with 92, compared to 89 and 90 of VGG16 and VGGFace, respectively. These increases in performance can be associated to a large extent with combination of face recognition and multicovariant feature extraction which enables the model to exploit deeper correlations across the different facial areas rather than relying on simple pixel level descriptors.

4.5 Equal Error Rate

The Equal Error Rate (EER), which shows the point at which the system has an equal chance of rejecting a matching face and accepting a nonmatching face that is, the accuracy threshold at which the algorithm in facial recognition applications strikes a reasonable balance between accepting imposters and rejecting real individuals is a common performance statistic for assessing the effectiveness of identification or verification systems in binary classification tasks and biometric security systems. This parameter is crucial when evaluating the system's performance, in equation 13.

$$Equal\ Error\ Rate = \frac{FAR+FPR}{2} \tag{13}$$

Figure 6 shows that to assess each face recognition model's robustness, the Equal Error Rate (EER) study compares the performance of VGG16, VGGFace, and SA-LSTM across the CelebA and CASIA-WebFace datasets. VGG16 achieves an EER of 30% on the CelebA data, but VGGFace is a little better with 23% which implies fewer errors in classifications. The proposed SA-LSTM model has an EER of 28 which is worse in comparison with VGG16 and slightly better in comparison with VGGFace. A similar trend is observed in the CASIA-WebFace dataset, whereby VGG16, VGGFace, and SA-LSTM have EER values of 32, 24, and 29 respectively. These results imply that the SA-LSTM

framework has a constant reduction of error compared to VGG16, although VGGFace has the lowest EER on both datasets. The figure 6 illustrates the EER of proposed SA-LSTM model.

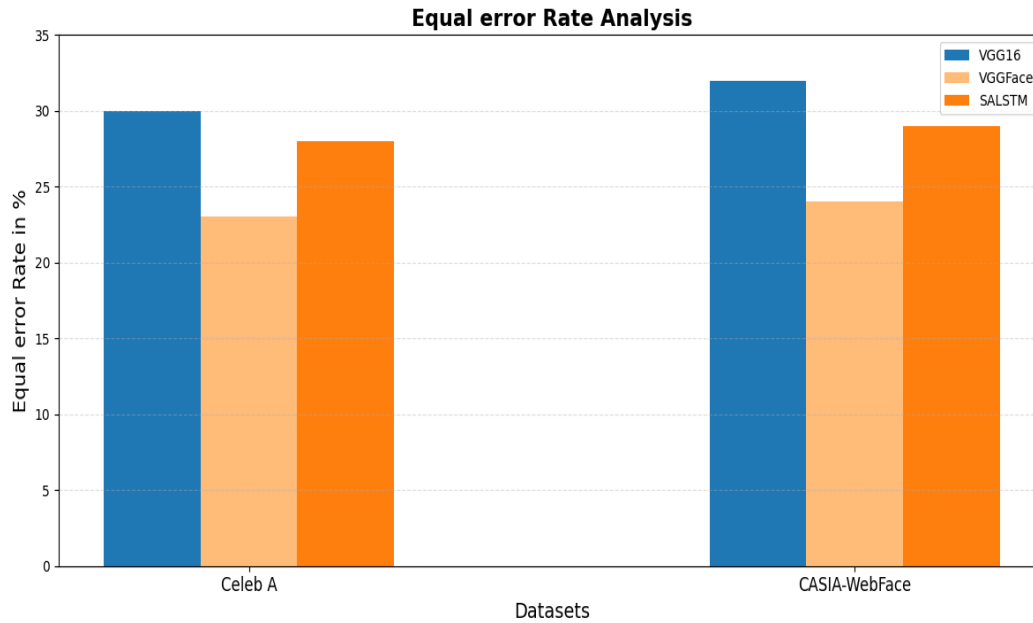


Figure 6: Equal error rate of proposed SA-LSTM model

The EER values indicate that although the proposed model improves recognition accuracy, there is still scope for optimization in verification-level decision balance. The SA-LSTM performs better than VGG16, showing that the extracted multicovariant features reduce misclassification to some extent. However, the lower EER of VGGFace suggests that pretrained deep facial representations remain highly effective for minimizing false acceptance and false rejection simultaneously. This implies that the proposed model is stronger in feature discrimination and classification accuracy, while threshold-sensitive verification performance can be further enhanced in future work.

4.6 Root Mean Square Error

Root-mean-square error (RMSE) analysis is one quantitative evaluation method for determining the accuracy of facial recognition systems. To establish the square root of the mean squared deviations between the predicted and observed identities or facial features, one has to work hard to get the square root. The evaluation of RMSE gives an objective measurement of the total variance or difference between the expected and the real facial recognition results. This allows the programmers and the researchers to test and compare the efficiency of different face recognition models or algorithms.

The square root representation is the only difference between the mean square error (MSE) and mean absolute error (MAE). The mean absolute error can be calculated using equation (14).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (14)$$

The range of RMSE is 0 to, with lower RMSE values being desired. Table 2 shows the relationship between the RMSE values of the proposed SA-LSTM approach and other baseline methods. The graph demonstrates that the DL approach performs better with the lowest RMSE. The suggested Stacked Attention LSTM's RMSE utilizing the Celeb A and CASIA-WebFace datasets is shown in the graph. For instance, the RMSE for the Stacked Attention LSTM using the Celeb. A dataset is 34.827%, while

the CASIA-WebFace dataset's RMSE is 35.425%. Table 2 displays the suggested model's RMSE performance.

Table 2: RMSE of the suggested model using two different types of datasets

	Celeb A	CASIA-WebFace
VGG16	41.32	41.42
VGGFace	38.32	35.424
Stacked Attention LSTM	34.82	34.425

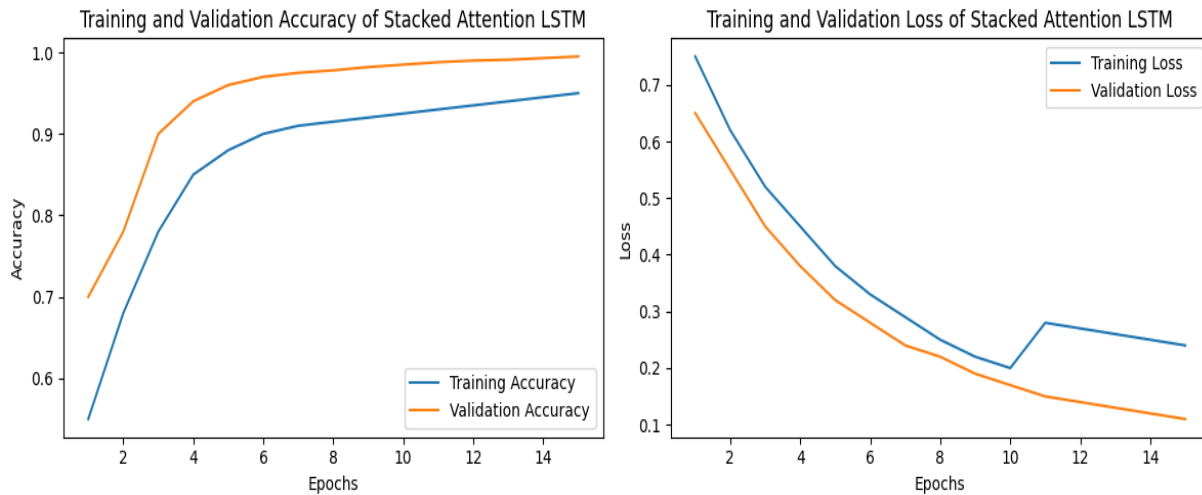


Figure 7: Accuracy and loss of the proposed model during training and validation

Figure 7 shows the accuracy and loss performance of the suggested SA-LSTM model. The training and validation curves are used to demonstrate the behaviour of the Stacked Attention LSTM model throughout 15 learning epochs. Accuracy plot indicates that there is a swift rise in the accuracy of the training in the first epochs and the accuracy of the training has increased to above 90 per cent and the accuracy of the validation is also rising at an even faster rate and finally the accuracy of the validation is also reaching around 98 per cent. This implies that the model is effective in learning discriminatory facial characteristics during the initial stages of training. The training and validation loss curves, which show that the model is more stable and has less prediction errors as it converges, further support this tendency. When optimizing the deep learning, it is common to have a small variation around epoch 10, which does not represent overfitting, since the validation loss is decreasing continuously. All in all, the cumulative action of the growing accuracy and declining loss confirms that Stacked Attention LSTM with the aid of multicovariant feature extraction has a high generalization and stability in performance both in training and validation data.

The lower RMSE of the proposed model confirms that its predictions remain closer to the expected class outcomes than those of the baseline methods. This is an indication of enhanced stability of learning and decreased deviation when making predictions. The play of normalized preprocessing, ICA-based discriminative representation and stacked attention LSTM architecture collectively decrease the RMSE and are the answer to the observed variations in the consistency of the feature learning in various facial circumstances. Therefore, the proposed framework not only improves accuracy but also reduces overall prediction error.

The training and validation curves further confirm the convergence behaviour of the proposed framework. The simultaneous increase in training and validation accuracy, together with the decreasing

loss values, indicates that the model learns meaningful facial representations without severe overfitting. The small fluctuations observed during later epochs are normal in deep learning optimization and may result from mini-batch variability. Overall, the learning curves suggest that the proposed model has achieved a stable optimisation path and maintains good generalisation ability on unseen facial data.

Limitations

The fact that the proposed framework gains a better recognition accuracy and strength, there are some limitations. To begin with, the stacked attention LSTM structure adds an extra computational load over the standard CNN-based methods, which can add to the training time of very large-scale datasets. Second, due to the quality and diversity of training data, the model performance is susceptible to extreme cases of occlusion or low-resolution conditions. The next round of work will be on optimization of lightweight models and addition of transformer-based attention mechanism to further streamline the scalability as well as real time deployment.

5 Conclusion

A deep learning-based face recognition architecture that incorporates multicovariant feature extraction and stacked attention LSTM model was introduced in this work to overcome the weaknesses of face recognition models used in the traditional scenario in the real world. The method suggested consisted of a combination of Z-score normalization and Independent Component Analysis (ICA) in order to enhance the discriminative ability of facial representations by both decreasing the variation of intensity and obtaining statistically independent structural attributes. Such processed characteristics were subsequently grouped with a stacked attention-based LSTM architecture which allowed the model to selectively attend to the most informative facial parts (like eyes, nose and mouth) and drown out irrelevant background data. The proposed framework was proven to be effective after the experimental assessment conducted on the CelebA and CASIA-WebFace datasets. The model had a face recognition rate of 96% on CelebA and 92% on CASIA-WebFace, which was higher than baseline models like VGG16 and VGGFace in the overall rate of recognition. Regarding error analysis, the suggested approach achieved a lower value of RMSE at 34.82 with CelebA and 34.425 with CASIA-WebFace, which means that it is more consistent with predictions and less variable than non-modern approaches. Even though the Equal Error Rate was still marginally greater than VGGFace, the suggested framework was highly robust, with steady learning behaviour, and enhanced generalisation predicting the different datasets under different illumination, facial expression and image quality circumstances. The importance of the research is that it offers to unite statistical feature extraction and attention-focused sequential learning into a viable and scalable biometric recognition system. The model is also very applicable in smart surveillance applications, secure authentication and forensic investigation and also in intelligent access control applications. Future studies can concentrate on the reduction of the computational complexity, enhancement of the verification-level performance, and expansion of the framework by transformer-based attention mechanisms, lightweight architectures, and bigger cross-domain sets to further increase the ability of the framework to be deployed in real-time and the strength of the recognition.

References

- [1] Ali, M. E., Diwan, A., & Kumar, D. (2024). Attendance system optimization through deep learning face recognition. *International Journal of Computing and Digital Systems*, 15(1), 10-12785. <https://doi.org/10.12785/ijcds/1501108>

- [2] Ali, N. S., Alsafo, A. F., Ali, H. D., & Taha, M. S. (2024). An effective face detection and recognition model based on improved YOLO v3 and VGG 16 networks. *International Journal of Computational Methods and Experimental Measurements*, 12(2), 107-119. <https://doi.org/10.18280/ijcmem.120201>
- [3] Ali, W., Tian, W., Din, S. U., Iradukunda, D., & Khan, A. A. (2021). Classical and modern face recognition approaches: a complete review. *Multimedia tools and applications*, 80(3), 4825-4880. <https://doi.org/10.1007/s11042-020-09850-1>
- [4] Al-Waisy, A. S., Qahwaji, R., Ipson, S., & Al-Fahdawi, S. (2018). A multimodal deep learning framework using local feature representations for face recognition. *Machine Vision and Applications*, 29(1), 35-54. <https://doi.org/10.1007/s00138-017-0870-2>
- [5] Arellano, M. D. P. C., Castro, M. D. P. Q., Mondragón, E. M. B., Valdivieso, M. O. M., Gonzáles, J. R. C., & Castro, G. A. Q. (2024). Examining face recognition technologies and privacy: Ethical and legal choices. *Journal of Internet Services and Information Security*, 14(4), 360–376. <https://doi.org/10.58346/JISIS.2024.14.022>
- [6] Ashani, Z. N., Ilias, I. S. C., Ng, K. Y., Ariffin, M. R. K., Jarno, A. D., & Zamri, N. Z. (2024). Comparative analysis of deepfake image detection method using vgg16 vgg19 and resnet50. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 47(1), 16-28. <https://doi.org/10.37934/araset.47.1.1628>
- [7] Boussaad, L., & Boucetta, A. (2022). An effective component-based age-invariant face recognition using Discriminant Correlation Analysis. *Journal of King Saud University-Computer and Information Sciences*, 34(5), 1739-1747. <https://doi.org/10.1016/j.jksuci.2020.08.009>
- [8] El-Bashir, M. S., AL-Shatnawi, A. M., Al-Saqqar, F., & Nusir, M. I. (2021). Face Recognition Model Based on Covariance Intersection Fusion for Interactive devices. *World of Computer Science and Information Technology Journal*, 11, 5-12.
- [9] Fatoni, F., Kurniawan, T. B., Dewi, D. A., Zakaria, M. Z., & Muhayeddin, A. M. M. (2025). Fake vs real image detection using deep learning algorithm. *Journal of Applied Data Sciences*, 6(1), 366-376. <https://doi.org/10.47738/jads.v6i1.490>
- [10] Goel, R., Mehmood, I., & Ugail, H. (2021). A study of deep learning-based face recognition models for sibling identification. *Sensors*, 21(15), 5068. <https://doi.org/10.3390/s21155068>
- [11] Huang, Z. Y., Chiang, C. C., Chen, J. H., Chen, Y. C., Chung, H. L., Cai, Y. P., & Hsu, H. C. (2023). A study on computer vision for facial emotion recognition. *Scientific reports*, 13(1), 8425. <https://doi.org/10.1038/s41598-023-35446-4>
- [12] Komlavi, A. A., Chaibou, K., & Naroua, H. (2024). Comparative study of machine learning algorithms for face recognition. *Revue Africaine de Recherche En Informatique et Mathématiques Appliquées*, 40. <https://doi.org/10.46298/arima.9291>
- [13] Limei, N., Dongfan, W., & Bo, Z. (2025). Landscape image recognition and analysis based on deep learning algorithm. *Journal of Intelligent & Fuzzy Systems*, 49(2), 471-481. <https://doi.org/10.3233/JIFS-239654>
- [14] Lin, N., Ding, Y., & Tan, Y. (2025). Optimization design and application of library face recognition access control system based on improved PCA. *Plos one*, 20(1), e0313415. <https://doi.org/10.1371/journal.pone.0313415>
- [15] Ratyal, N., Taj, I. A., Sajid, M., Mahmood, A., Razzaq, S., Dar, S. H., ... & Mussadiq, U. (2019). Deeply learned pose invariant image analysis with applications in 3D face recognition. *Mathematical Problems in Engineering*, 2019(1), 3547416. <https://doi.org/10.1155/2019/3547416>
- [16] Sarkar, A., Behera, P. R., & Shukla, J. (2023). Multi-source transfer learning for facial emotion recognition using multivariate correlation analysis. *Scientific Reports*, 13(1), 1-15. <https://doi.org/10.1038/s41598-023-48250-x>

- [17] Sekhar, J. C., Josephson, P. J., Chinnasamy, A., Maheswari, M., Sankar, S., & Kalangi, R. R. (2025). Automated face recognition using deep learning technique and center symmetric multivariant local binary pattern. *Neural Computing and Applications*, 37(1), 263-281. <https://doi.org/10.1007/s00521-024-10447-0>
- [18] Shah, S. R., Qadri, S., Bibi, H., Shah, S. M. W., Sharif, M. I., & Marinello, F. (2023). Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: a case study on early detection of a rice disease. *Agronomy*, 13(6), 1633. <https://doi.org/10.3390/agronomy13061633>
- [19] Vezzetti, E., Marcolin, F., Tornincasa, S., Ulrich, L., & Dagnes, N. (2018). 3D geometry-based automatic landmark localization in presence of facial occlusions. *Multimedia Tools and Applications*, 77(11), 14177-14205. <https://doi.org/10.1007/s11042-017-5025-y>

Authors Biography



U.S. Pavitha, is currently working as Assistant professor in E & C department, M S Ramaiah Institute of Technology, Bangalore. Her research interest includes Low Power VLSI, Face Recognition. She has obtained her B. E in E & C from VTU and M. Tech in VLSI and Embedded Systems from VTU, Belagavi in the year 2005 and 2010 respectively. Currently She is perusing Ph. D from VTU, Belagavi.



Dr.K.V. Suma is working as Associate Professor in the Department of Electronics and Communication Engineering at Ramaiah Institute of Technology, Bengaluru. She has completed her Ph. D in 2019 from Visvesvaraya Technological University. She is a senior member of IEEE, Fellow of IETE and Member of IAENG. Her areas of interest are biomedical signal/image processing, embedded system design and artificial intelligence.