Polytechnic Journal

Polytechnic Journal

Volume 15 | Issue 1 Article 4

An Attention-Based Model for Recognition of Facial Expressions Using CNN-BiLSTM

Sushil Kumar Singh

Computer Engineering Department, Marwadi University, Rajkot, Gujarat, India,

sushilkumar.singh@marwadieducation.edu.in

Manish Kumar

Department of Computer Science, Seoul National University of Science and Technology, Seoul, S. Korea

Ikram Majeed Khan

Department of Computer Science, Coventry University, Priory St, Coventry, England, UK

A. Jayanthiladevi

Computer Engineering Department, Marwadi University, Rajkot, Gujarat, India

Chirag Agarwal

Information Technology Analyst at Tata Consultancy Services Ltd., Des Plaines, Illinois, USA

How to Cite This Article

Singh, Sushil Kumar; Kumar, Manish; Khan, Ikram Majeed; Jayanthiladevi, A.; and Agarwal, Chirag (2025) "An Attention-Based Model for Recognition of Facial Expressions Using CNN-BiLSTM," *Polytechnic Journal*: Vol. 15: Iss. 1, Article 4.

DOI: https://doi.org/10.59341/2707-7799.1849

An Attention-Based Model for Recognition of Facial Expressions Using CNN-BiLSTM

Data Availability Statement

The data supporting the findings of this study are publicly available and are included within this published article.

ORIGINAL ARTICLE

An Attention-based Model for Recognition of Facial Expressions Using CNN-BiLSTM

Sushil Kumar Singh [a,*], Manish Kumar [b], Ikram Majeed Khan [c], A. Jayanthiladevi [a], Chirag Agarwal [d]

[a] Marwadi University, Rajkot, Gujarat, India
[b] Seoul National University of Science and Technology, Seoul, South Korea
[c] Coventry University, Priory St, Coventry, England, UK
[d] Tata Consultancy Services Ltd., USA

Abstract

In recent studies, computer vision and human-computer interaction have focused heavily on facial expression recognition (FER). Traditional deep learning algorithms have faced significant challenges when processing images with occlusion, uneven lighting, and positional inconsistencies and addressing dataset imbalances. These issues often result in low accuracy, unreliable recognition rates, and slow convergence. To address the challenges posed by non-frontal visual features, this study proposes a hybrid model that fusion of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), augmented with a point multiplication attention model and Linear Discriminant Analysis (LDA). Data preparation incorporates median filtering and global contrast normalization to enhance image quality. The model then utilizes DenseNet and Softmax for image reconstruction, improving feature maps and providing essential data for classification tasks on the FER2013 and CK + datasets. We benchmark our proposed model against conventional models such as Convolutional Neural Networks- Long Short-Term Memory (CNN-LSTM), Depthwise Separable Convolutional Neural Networks- Long Short-Term Memory (DSCNN-LSTM), Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM), and Attention Convolutional Neural Network - Long Short-Term Memory (ACNN-LSTM), evaluating performance metrics including F1 score, accuracy, precision, and recall. The results demonstrate that our proposed model outperforms existing approaches, highlighting the effectiveness of incorporating attention paradigms, hybrid deep learning architectures, and advanced preprocessing methods for facial emotion detection. The non-parametric statistical test also analyzes it.

Keywords: Facial expression recognition, CNN, BiLSTM, Attention model, Data preprocessing, Global contrast normalization

1. Introduction

Facial expressions (FE), complemented by hand gestures and eye contact, form the cornerstone of non-verbal communication, playing a vital role in conveying emotions, thoughts, intentions, and mental states. The intricate connection between facial expressions and human emotions has long captured the attention of scientists, researchers, and academics, as it provides valuable insights into human behavior and social interactions. Recent advancements in machine learning have significantly contributed to research across diverse domains, including data protection, surveillance, emotion recognition, security, and natural disaster management, where understanding emotions enhances decision-making and response systems. In both psychology and computer vision, emotions are broadly categorized into two frameworks: categorical, which identifies discrete emotions such as happiness, sadness, neutrality, anger, fear, or surprise, and dimensional, which maps emotions along

Received 27 November 2024; accepted 16 January 2025.
Available online 20 February 2025

* Corresponding author.
E-mail addresses: sushilkumar.singh@marwadieducation.edu.in, sushil.sngh1@ieee.org (S.K. Singh), manish-08675@seoultech.ac.kr (M. Kumar), Khani72@coventry.ac.uk (I.M. Khan), a.jayanthiladevi@marwadieducation.edu.in (A. Jayanthiladevi), chirag_agarwal.2008@rediffmail.com, agarwal.chirag@tcs.com (C. Agarwal).

https://doi.org/10.59341/2707-7799.1849

continuous scales like valence (positive or negative affect) and arousal (intensity of the emotion) [1]. This dual perspective underscores the complexity of emotional interpretation and its applications in modern technology. Facial expression recognition (FER) is a rapidly evolving field within computer vision, gaining increasing prominence due to its vast array of real-world applications, ranging from healthcare and security to human-computer connection and affective computing. Over the past few decades, FER research has witnessed exponential growth, driven by advancements in machine learning and the growing availability of annotated datasets. Initially, challenges such as vanishing gradients, overfitting, and declining accuracy in deep networks hindered progress. However, Convolutional Neural Networks (CNNs) emerged as a foundational technology, excelling in feature extraction by capturing spatial hierarchies in images.

The introduction of Residual Neural Networks (ResNet) in 2015 marked a transformative milestone, effectively addressing performance bottlenecks by incorporating residual learning. This innovation allowed deeper networks to be trained without degradation, significantly improving FER accuracy and robustness. FER datasets often present additional complexities, including pose, illumination, resolution, and occlusion variations—where objects, accessories, or expressions obscure parts of the face. Such variability makes achieving consistent performance across diverse conditions particularly challenging, highlighting the need for more sophisticated models capable of generalizing effectively. The Attention Mechanism (AM) was introduced as a paradigm shift in neural network design to address these limitations. AM has revolutionized feature extraction and classification in FER by enabling models to selectively focus on the most salient features of an input image. Initially adopted in computer vision, it has since been adapted to other fields, such as natural language processing [NLP], where it underpins cutting-edge models like transformers. AM enhances network performance by prioritizing critical regions of an image, reducing redundancy, and preserving essential information that might otherwise be lost during processing [2]. Today, attention mechanisms are an indispensable component of modern neural architectures, solving optimization challenges and boosting performance in tasks like image synthesis, semantic segmentation, and machine translation. In the context of FER, studies leveraging attention mechanisms have yielded remarkable results, demonstrating improved classification accuracy, robustness to occlusions, and better handling of variations in facial features. As research continues to evolve, integrating attention with hybrid architectures and multimodal approaches holds immense promise for further advancing the accuracy and reliability of facial emotion recognition systems. FE is shown in Fig. 1.

FE is a vital component of human communication, providing critical insights into feelings, volitions, and mental states. Recognizing these expressions accurately is essential in Smart City Applications such as Smart Healthcare, Smart Hospitals, Smart Doctors and Patients, Security and Privacy, Human–Computer Interaction (HCI), and behavioral analysis [4–7]. However, FER faces challenges such as occlusion, inconsistent lighting, positional variations, and imbalanced datasets. While adequate to an extent, traditional deep learning approaches often struggle to achieve high accuracy and robustness under such conditions. Integrating CNNs with BiLSTM networks has shown promise in capturing spatial and temporal features in FER tasks. However, further advancements are needed to address the limitations of conventional models, such as the loss of crucial features and suboptimal handling of complex datasets. To overcome these challenges, attention mechanisms have emerged as a powerful tool, enabling models to focus selectively on the most relevant features of an image. Attention-based models can significantly enhance feature extraction and classification accuracy by emphasizing critical regions and mitigating the impact of redundant or noisy information. This study presents an attention-based CNN-BiLSTM model to leverage the strengths of spatial and temporal feature extraction while incorporating attention mechanisms for improved accuracy and robustness. By combining these advanced techniques, the objective of the model to push the boundaries of FER performance, addressing existing challenges and providing a reliable framework for real-world applications.

The main work contributions of this research are as follows.

- Propose an attention-based CNN-BiLSTM model for facial expressions.
- Leveraging the strengths of spatial and temporal feature extraction while incorporating attention mechanisms for improved accuracy and robustness.
- Discuss related work's limitations and how to address these limitations using the proposed attention-based CNN-BiLSTM model for facial expressions.
- Describe the proposed model performance are better compared to other existing research

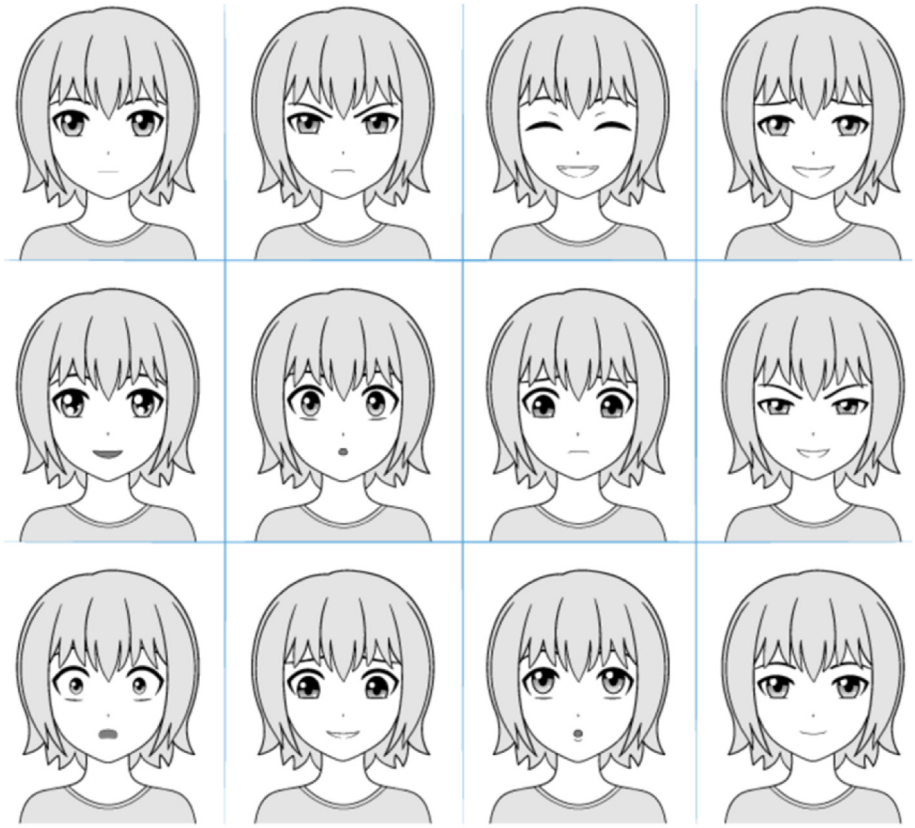

Fig. 1. Facial expressions [3].

studies based on standard parameters such as accuracy, precision, F1-score, and recall value.
- At the last, we discuss the limitations of the proposed model and how to mitigate them in future work.

The primary objective of this study is to develop and evaluate a CNN-BiLSTM-based model for predicting drivers' emotions, a critical task with implications for road safety, human-computer interaction, and intelligent transportation systems. Section 2 comprehensively reviews the existing literature, highlighting the advancements and limitations in facial expression recognition and emotion prediction, particularly in driving scenarios. Section 3 outlines the proposed methodology, detailing the integration of (CNNs for spatial feature extraction, BiLSTM networks for capturing temporal dependencies, and attention mechanisms for enhanced feature prioritization. Section 4 presents the experimental results, comparing the performance of the proposed model with existing approaches using standard metrics such as accuracy, precision, recall, and F1 score. This section also provides an in-depth analysis and interpretation of the results, emphasizing the model's strengths and addressing potential limitations. Finally, the study concludes in section 5 with a comprehensive summary of the findings, underscoring their relevance to the field, and offers recommendations for future research directions, such as exploring multimodal data integration and real-time emotion detection in dynamic driving environments.

2. Related works

In this section, we discuss related work as seminal contributions and provide comprehensive reviews of existing literature. We highlight the advancements and limitations in facial expression recognition and emotion prediction, particularly in driving scenarios, and highlight key considerations of the proposed work.

2.1. Seminal contribution

Many research studies have proposed models and architectures for facial expressions using deep learning and machine learning algorithms. All studies have various merits and demerits, so we discuss existing research studies' objectives, merits, and limitations and how we can mitigate these

limitations with the proposed work. To lessen variance in learning information linked to identification and expression, a suggested identification-Aware CNN (IA-CNN) model focused on identity and expression-sensitive contrastive losses [8]. An attention model combined with an end-to-end architecture was also implemented to improve recognition. The Region Attention Network (RAN), created to manage position and occlusion fluctuations and capture important face areas to enhance facial emotion recognition results, is another noteworthy development [9]. To detect expression-related areas, Region Aware Subnet (RASnet) was created to detect expression-related areas utilizing coarse-to-fine-granularity binary masks. Several attention mechanisms were used to learn discriminative characteristics, with each sub-branch of the hybrid attention branch focusing on a specific location [10]. To improve class separability, create attention maps, and fuse them into a comprehensive output, the Distract Your Attention Network (DAN) also included components like the Feature Clustering Network (FCN), Multi-head Cross Attention Network (MAN), and Attention Fusion Network (AFN) [11]. A Multiple Attention Network, including the Multi-Branch Stack Residual Network (MRN), Transitional Attention Network (TAN), and an Appropriate Cascade Structure (ACS) was presented in a different study. By focusing attention focuses on important face areas, our approach enhanced class separability. It has been discovered that attention mechanisms work well for concentrating on essential details in face expression identification tests. Using RGB and depth pictures from a dataset of 35 people (ages 20–25), an end-to-end network with attention mechanisms was evaluated [12]. Aggarwal et al. [13] comprehensively analyzed Hedge Funds in financial markets. They have applied ML algorithms to solve economic problems.

The findings were compared with benchmarks such as JAFFE, CK+, FER 2013, and Oulu-CASIA datasets. CNN face recognition is shown in Fig. 2.

To identify face emotions in situations with high occlusion or light, another research developed RCLnet, which used an attention mechanism with LBP feature fusion. Two branches are made up of RCLnet: a local binary feature extraction branch, and ResNet-CBAM, a branch for residual attention. Many datasets, such as FER 2013, FERPLUS, CK+, and RAF-DB, were used to verify the model. Another technique created to identify occluded regions, combine various representations from face regions of interest, and focus on un-occluded parts is attention-based CNN (ACNN) [14]. The RAF-DB dataset's occluded face photos were to be identified using an Enhanced CNN with Attention Mechanism (ECNN-AM). By employing a patch-based ECNN-AM and Global Gated Unit (GG-U) to weight global face representations automatically, the findings showed an accuracy of 86.2 %. FER2013 testing was used to present a Deep CNN with a Binary Attention Mechanism (BAM) that used regularization approaches to avoid overfitting and the Histogram of Oriented Gradients (HOG) for data preparation [14]. Furthermore, by identifying a horizontally symmetric region and using heterogeneous soft partitioning, the Symmetric Speed-up Robust Features (SURF) framework was established, which improved facial expression recognition by 7–8% on the Cohn-Kanade (CK+) and FER2013 datasets. Previous research used frontal and non-frontal picture poses from the FER2013 dataset to fine-tune the outputs of AlexNet-based Deep CNN models, yielding results that outperformed those of previous expression recognition systems [15,16]. Ultimately, a CNN-LSTM model with two layers was presented to extract rich information from crucial areas. It performed better on the FER2013 and CK + datasets than other approaches, such as CNN-ALSTM and

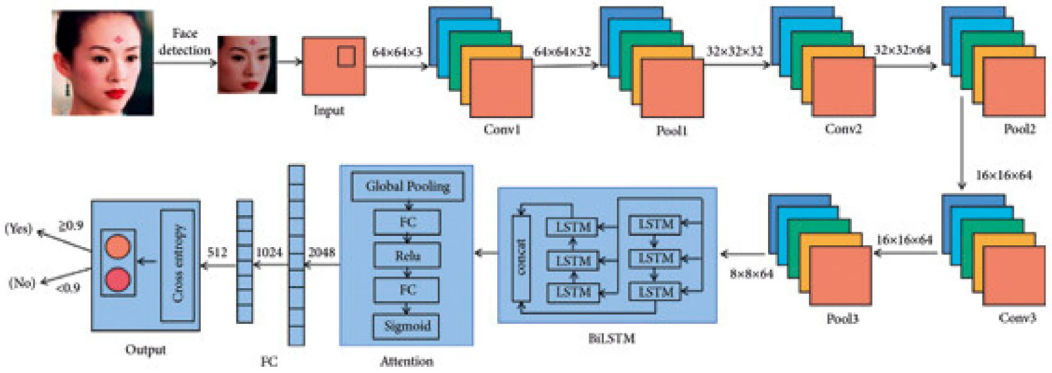

Fig. 2. CNN face recognition [4].

ACNN-ALSTM. Similarly, in order to avoid overfitting, a CNN-BiLSTM model was devised, and it proved to be more accurate than both CNN and CNN-LSTM models [17,19].

According to the seminal contribution, there are some limitations, such as limited handling of occlusions, dataset imbalances, loss of crucial features, inadequate robustness to variations, and lack of attention mechanisms that are based on standard parameters such as accuracy, precision, F1 Score, and recall value. To address these limitations, we propose an attention-based CNN-BiLSTM mechanism to leverage the strengths plus of spatial and temporal feature extraction while incorporating attention mechanisms for improved accuracy and robustness. By combining these advanced techniques, the objective of the model is to push the boundaries of FER performance, address existing challenges, and provide a reliable framework for real-world applications.

2.2. Key considerations of the proposed model

We discuss why the proposed model is essential at the present time and what its requirements are. There are various requirements for the proposed model.

- *Temporal Sequence Modeling:* Including Bidirectional Long-Term Memory (BiLSTM) is crucial to capture both forward and backward temporal dependencies in sequential data, as facial expressions evolve. It is essential to analyze dynamic facial expressions that may change gradually or react to a sequence of events. The model should be capable of handling variable-length sequences, such as video frames or time-series data from facial expressions.
- *Feature Prioritization:* The attention mechanism is implemented to focus on the most relevant parts of the face, such as the eyes and mouth, which are necessary for detecting emotions. The attention model must prioritize critical facial features and reduce the impact of irrelevant or noisy information (e.g., background or occlusions). The attention module can be applied at different levels: spatial (focusing on specific facial regions) and temporal (concentrating on essential frames in the sequence).
- *Real-Time Applications:* The model should be optimized for real-time performance, particularly for driver emotion recognition or human-computer interaction applications. If applicable, techniques for model compression (e.g., pruning, quantization) and optimization for edge devices (e.g., mobile phones or embedded systems) should be explored.
- *Security and Privacy:* Since facial expression recognition involves personal data, it is essential to address privacy concerns. The model should be developed with appropriate data anonymization and ethical guidelines, mainly used in sensitive applications such as surveillance or healthcare.

3. Proposed attention-based model

In this section, we present a comprehensive overview of the proposed Attention-based Model for Facial Expression Recognition (FER), particularly emphasizing image processing techniques and the methodological flow. While Convolutional Neural Networks (CNNs) have made significant strides in FER, they exhibit notable limitations when confronted with challenges such as blurry, obstructed, or differently posed images. These issues stem from the inability of traditional CNN-based models to effectively capture and interpret the rich, nuanced information contained within these types of images. In our previous research, we addressed these limitations by developing a CNN-LSTM hybrid model that enhanced expression recognition accuracy, particularly for frontal images, by integrating a point multiplication attention mechanism [18]. This approach significantly improved the model's ability to focus on relevant facial features, thereby increasing recognition performance. Building upon this foundation, the present study introduces a more advanced model that integrates CNN and Bidirectional Long Short-Term Memory (BiLSTM) networks, coupled with the point multiplication attention mechanism and additional complementary techniques. The CNN component of the model is responsible for extracting spatial features from facial images, while the BiLSTM model captures temporal dependencies, enabling better handling of dynamic facial expressions. The attention mechanism is employed to prioritize important facial regions, such as the eyes and mouth, which are crucial for emotion detection. As depicted in Fig. 3, the new model comprises four essential components: the CNN for feature extraction, the BiLSTM for temporal sequence modeling, the attention layer for dynamic feature weighting, and a reconstruction and classification layer for refining the extracted features and making final predictions. This enhanced model offers a more robust solution for recognizing facial expressions under various challenging conditions, making

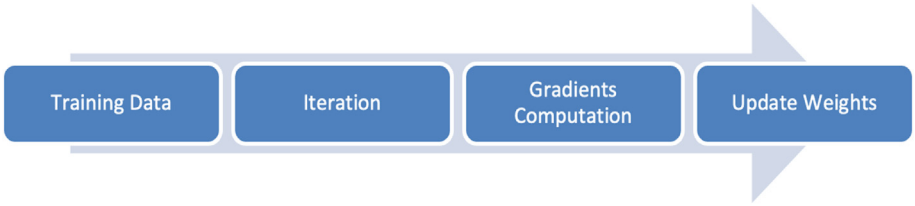

Fig. 3. Proposed attention-based model of processing images.

significant strides toward improving the accuracy and reliability of FER systems.

3.1. Data preprocessing

While the FER2013 dataset contains a much larger pool of images, the CK + dataset generally offers fewer high-quality images overall. For features to be considered high-quality and rich, effective preprocessing is crucial. In our previous work, we used a subset of 7074 high-quality images from the FER2013 dataset. However, we leverage the entire dataset in the present study to enhance the model's performance. Before processing, the images are resized to a uniform resolution of 128 by 128 pixels to ensure consistency, as datasets often contain images of varying sizes. To improve image quality, noise is removed using median filtering, which helps preserve essential details, such as facial edges, by eliminating unwanted artifacts like "salt and pepper" noise. Furthermore, Global Contrast Normalization (GCN) is applied to enhance the contrast of lower-quality images. This technique normalizes each image by subtracting its mean pixel value and dividing the result by its standard deviation, thereby improving the visibility of key facial features essential for expression recognition.

3.2. Training of CNN

Training large datasets presents significant challenges due to the vast number of images that need to be processed. Although techniques like Siamese networks and K-Nearest Neighbors (KNN) are widely used in such scenarios, we propose a novel center loss function further to improve the discriminative power of deeply learned features. This loss function works by minimizing the distance between the deep features of each class and their respective class centers, thereby enhancing the model's ability to distinguish between classes. To ensure efficient updates to class centers and reduce computational overhead, we employ mini-batch processing, accelerating the learning process while maintaining accuracy.

3.3. Network Architecture of the proposed model approach

We present a CNN-BiLSTM hybrid model that combines Linear Discriminant Analysis (LDA) with a point multiplication attention mechanism to achieve effective facial emotion identification. Hybrid model as Network Architecture is shown in Fig. 4.

3.4. A SoftMax-based classification layer

After abstract features are extracted from the local regions of facial images, they are fed into the BiLSTM network for sequence analysis. The CNN component, designed with a seven-layer architecture comprising four convolutional layers and three down-sampling layers, further refines the input by emphasizing the most significant features. This process enhances the model's ability to effectively capture critical facial information, improving its overall performance in facial expression recognition [18]. The proposed Structure with Attention Mechanism-based LSTM is shown in Fig. 5.

3.5. Bidirectional long Short-Term Memory (BiLSTM)

BiLSTM networks offer a significant advantage over traditional LSTMs by considering both forward and backward temporal dependencies. This bidirectional approach allows the model to effectively capture time-series information from images taken at different times or from various viewpoints [20]. The attention mechanism then receives the more refined sequence feature vectors generated by this bidirectional structure, enabling the model to focus on the most relevant information for improved recognition accuracy. BiLSTM − Proposed Structure is shown in Fig. 6.

3.6. Attention layer

The attention layer is crucial in improving the model's ability to focus on essential details, mainly when dealing with complex, blurry, or occluded

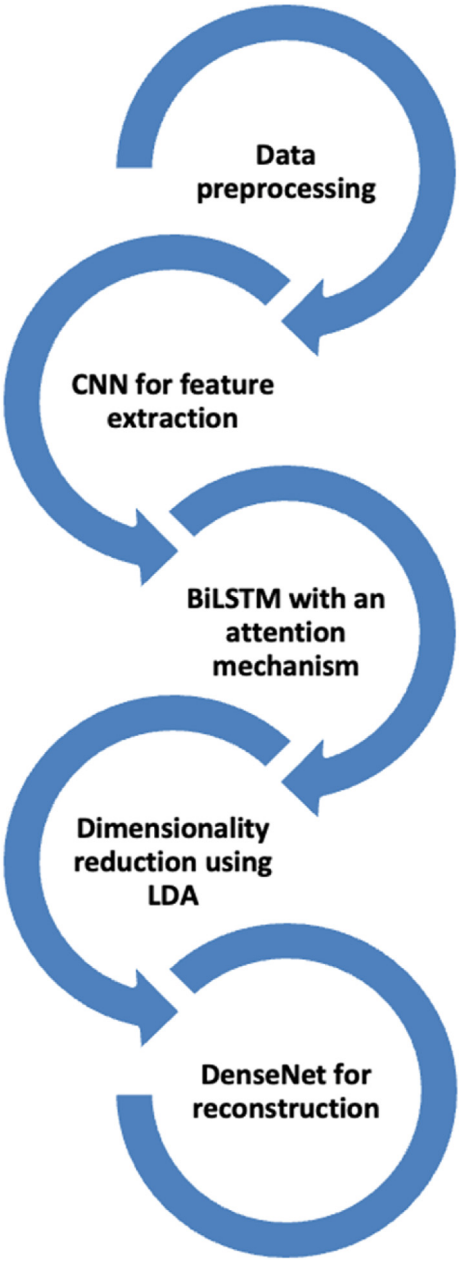

Fig. 4. Hybrid model.

images. In facial expression recognition, such images often present challenges where key facial features may be obscured or distorted, making it difficult for traditional models to classify emotions accurately. The attention layer addresses these challenges by dynamically emphasizing significant features—such as the eyes, mouth, or eyebrows—most indicative of emotional states. By assigning greater weight to these critical regions, the attention mechanism enables the model to differentiate between subtle emotional expressions better, improving its overall recognition accuracy. This mechanism is inspired by the human brain's ability to selectively concentrate on relevant stimuli while filtering out less important information. Similar to the way we focus our attention on specific aspects of a scene or object, the attention mechanism uses a weighted mean function to prioritize the most valuable features within an image. Doing so enhances feature extraction and optimizes the model's performance under challenging conditions, such as varying lighting, different poses, or partial occlusions. As a result, the model becomes more robust and efficient in recognizing and classifying facial emotions across a wide range of scenarios.

3.7. Dimensionality reduction and classification

As the number of features in a model increases, the complexity of the data also rises, making dimensionality reduction a critical step in preserving both accuracy and efficiency. High-dimensional data can lead to overfitting, increased computational costs, and slower processing times, all of which can degrade the model's performance. To address these challenges, we employ Linear Discriminant Analysis (LDA), a powerful technique for reducing the dimensionality of the data while retaining the most significant features. LDA works by projecting the data into a lower-dimensional space, maximizing the separation between different classes. It is imperative when dealing with complex datasets, such as those used in facial emotion recognition, where subtle distinctions between emotions must be captured effectively. By utilizing LDA, we can reduce the overall dimensionality and enhance the separation of the dataset's seven emotion classes—anger, disgust, fear, happiness, surprise, sadness, and neutrality. This improved separation facilitates a more accurate classification process, as the model can better distinguish between these emotions. Following this dimensionality reduction, the refined feature maps are passed through DenseNet, a deep learning architecture known for its efficiency in feature reuse and dense connections, allowing for more accurate reconstruction of facial features. Finally, a fully connected layer with a SoftMax activation function is applied for classification. This layer ensures the output corresponds to one of the seven emotion classes. SoftMax converts the model's raw output into probabilities that sum to one, thereby enabling the model to make confident predictions [21].

Overall, this combination of dimensionality reduction with LDA, followed by DenseNet reconstruction and SoftMax classification, enhances the model's ability to accurately recognize and classify

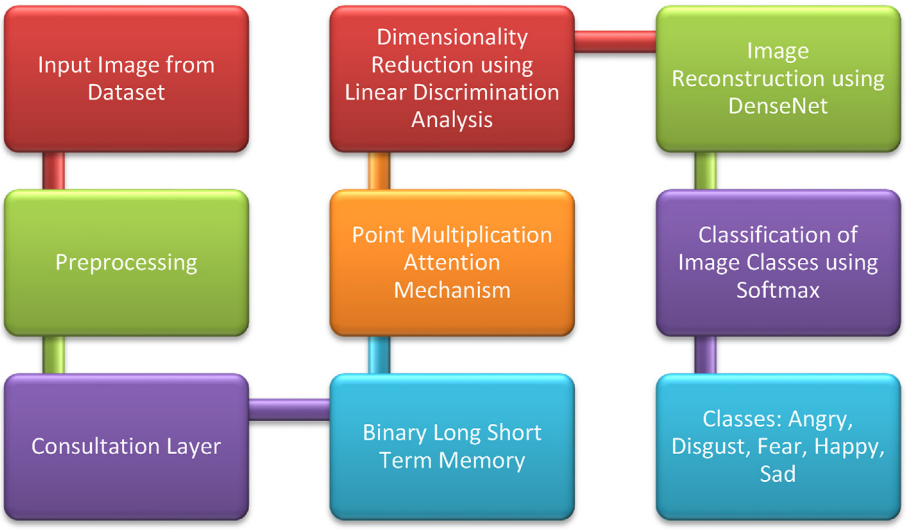

Fig. 5. Proposed structure with attention mechanism-based LSTM.

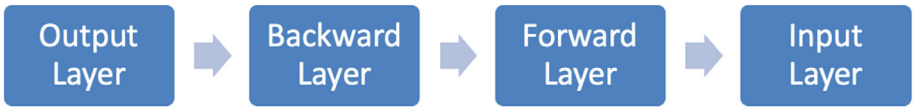

Fig. 6. BiLSTM — proposed structure.

facial emotions, even in complex, high-dimensional datasets. The SoftMax activation function effectively categorizes the expressions into their appropriate groups. To prevent overfitting and accelerate convergence, batch normalization is applied to every layer. This proposed model provides a comprehensive approach for improving facial expression detection, achieving better accuracy and performance metrics, even in challenging scenarios such as obscured or varying positions [22,23].

4. Results and performance analysis

This section comprehensively evaluates the proposed model's performance using standard performance metrics, including accuracy, precision, F1 score, and recall. These metrics are essential for assessing the model's ability to recognize and classify facial expressions across different datasets. To facilitate a robust analysis, we employ two widely used and well-established datasets—FER2013 and Cohn-Kanade+ (CK+), which contain images depicting seven distinct facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. These datasets were selected due to their diversity and the variety of facial expressions they represent, making them ideal for training and validating emotion recognition models. We provide a detailed description of both datasets, outlining their characteristics, the number of images, and the distribution of expressions within each dataset. Following this, we present the results of our experiments, highlighting the accuracy and recognition rates for each facial expression across both datasets. Our analysis includes a breakdown of how the model performs on individual expressions, offering insights into its strengths and potential areas for improvement. In addition to evaluating the model's overall performance, we examine the impact of each component of the proposed model. This includes an in-depth analysis of how the CNN, BiLSTM, attention mechanism, and dimensionality reduction techniques contribute to the model's effectiveness in recognizing facial emotions. We also evaluate the role of Long Short-Term Memory (LSTM) parameters, exploring their influence on the model's ability to capture temporal dependencies in the facial expression data.

Furthermore, we compare the proposed model's performance with existing state-of-the-art models in facial emotion recognition. This comparison is based on a thorough assessment of accuracy, precision, recall, and F1 score, focusing on how well the proposed model performs in recognizing facial expressions from the FER2013 and CK + datasets compared to previously published works. The results of this comparison provide valuable insights into the relative advantages of the proposed

approach. The method was implemented using MATLAB 2022a on a Windows 10 computer with an Intel i9 CPU and 8 GB of RAM, ensuring a robust and efficient computational environment for training and testing the model. The choice of MATLAB as the development platform, coupled with the computational resources available, allowed for seamless execution of the experiments and ensured reproducibility of the results.

4.1. FER dataset

In preparation for a Kaggle competition, the FER dataset [16] was carefully curated by sourcing grayscale images from Google. The dataset consists of 32,110 images, each with a resolution of 48 × 48 pixels. These images were selected to represent a wide range of facial expressions, providing diverse examples for training and testing facial emotion recognition models. Despite its size and diversity, the dataset is not without challenges. It contains significant noise, which can hinder the effectiveness of facial expression recognition models. Many images suffer from poor quality, including blurriness and low resolution, which can obscure crucial details necessary for accurate emotion classification.

Additionally, the dataset includes images that vary greatly in terms of facial pose, lighting conditions, and occlusions. Some faces are partially obscured, making extracting meaningful features for recognition even more challenging. These factors, along with variations in head tilt, facial orientation, and expression intensity, contribute to the complexity of the classification task. As a result, the FER dataset poses a considerable challenge for facial expression recognition systems, requiring robust preprocessing and feature extraction techniques to ensure high accuracy and reliability. Including such diverse and imperfect images in the dataset makes it an ideal testbed for evaluating the performance of emotion recognition models, particularly those capable of handling real-world conditions like varying angles, partial occlusions, and environmental noise. However, this also means that any model trained on this dataset must be highly adaptable, able to manage imperfections, and proficient in identifying emotions despite these challenges.

4.2. Cohn Kanade+ (CK+) dataset [21]

The dataset includes 123 subjects and a total of 593 images capturing various facial expressions. Out of these, 327 images represent seven distinct emotions. These images were resized to a uniform 48 × 48 pixel resolution, matching the resolution of the FER2013 dataset, ensuring consistency across the datasets. Additionally, several image augmentation techniques were applied, including flips, rotations, brightness adjustments, and saturation modifications. These augmentations were employed to increase the diversity of the dataset, helping to improve the robustness and generalization capability of the model.

4.3. Evaluating module effectiveness

There are four modules in the proposed Attention-based Model. We tested each by taking off a module at a time while leaving the categorization module in place to see how successful each was. Table 1 describes the recognition rates for these updated models.

Table 1 demonstrates that the recognition rate for CK + falls to 73.50 % and for FER2013 to 61.20 % without the feature extraction module. Removing the attention module results in lower recognition rates: 72.85 % for FER2013 and 76.94 % for CK+. Once the reconstruction module is removed, the rates for FER2013 and CK + are 75.34 % and 81.23 %, respectively. The recognition rates for FER2013 and CK + when the whole network is present are 81.42 % and 99.21 %, respectively. Recognition Rates for Updated Models are shown in Fig. 7.

4.4. Performance comparisons

An attention mechanism is incorporated into the proposed hybrid CNN-BiLSTM model to improve feature extraction and classification. Based on the accuracy, precision, recall, and F1 score, the model's performance was compared with that of other hybrid network models, such as CNN-LSTM, DSCNN-LSTM, and ACNN-LSTM. Performance for the FER dataset is shown in Table 2, and Performance on All Methods is described in Fig. 8.

The proposed model outperformed CNN-LSTM by 6.1 %, DSCNN-LSTM by 5.6 %, CNN-BiLSTM by 4.0 %, and ACNN-LSTM by 7.3 % with an accuracy of 99.21 % for the CK + dataset. Our model also performed better, as evidenced by its more excellent

Table 1. Recognition Rates for updated models.

Architecture	FER2013 Recognition Rate (%)	CK + Recognition Rate (%)
No feature extraction	61.20 %	73.50 %
No attention module	72.85 %	76.94 %
No reconstruction module	75.34 %	81.23 %
Complete network	81.42 %	99.21 %

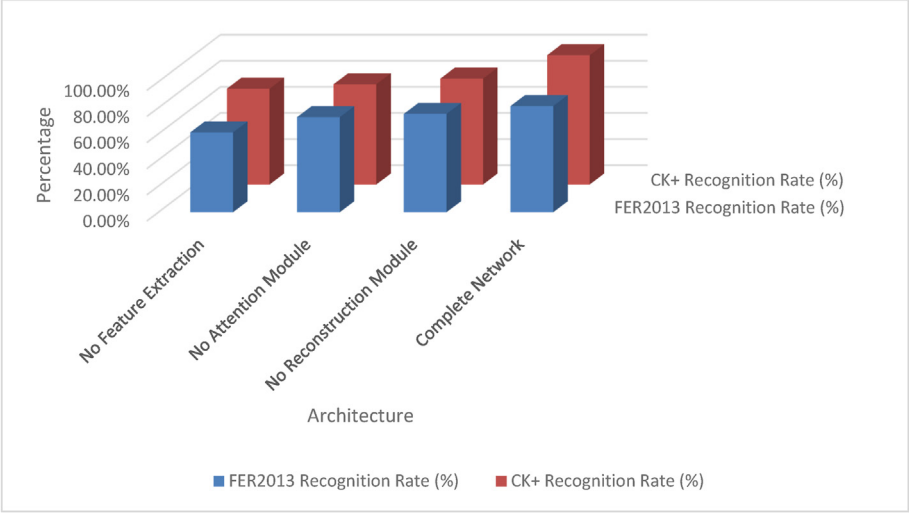

Fig. 7. Recognition rates for updated models.

Table 2. Performance for FER dataset.

Methods	Accuracy (%)	Precision (%)	F1 Score (%)	Recall (%)
CNN-LSTM	93.14 %	93.52 %	93.34 %	93.21 %
DSCNN-LSTM	93.67 %	93.98 %	93.82 %	94.12 %
CNN-BiLSTM	95.22 %	95.35 %	95.40 %	95.58 %
ACNN-LSTM	92.98 %	92.64 %	92.82 %	92.55 %
Proposed model	**97.34 %**	**96.84 %**	**96.96 %**	**97.21 %**

Table 3. Performance for FER 2013 dataset.

Methods	Accuracy (%)	Precision (%)	F1 Score (%)	Recall (%)
CNN-LSTM	92.47	91.84	93.12	92.68
CNN-SVM	91.95	92.16	90.78	91.22
CNN-BiLSTM	94.73	94.89	95.02	94.68
ACNN-LSTM	90.81	90.54	90.14	90.32
Proposed method	97.76	98.42	98.64	97.94

recall, F1 score, and accuracy values. The performance of the FER 2013 dataset is shown in Table 3, and the performance of all methods is described in Fig. 9.

The proposed model is analyzed using the non-parametric statistical test, the FRIEDMAN test. It is applied to performance tests at different iterations. This test provides an analysis similar to two-way ANOVA and returns a p-value of 0.003. It indicates the small p-value, which means the result is statistically significant.

4.5. Discussion

Facial expression recognition (FER) has made remarkable strides over the past decade, thanks to the rapid advancements in deep learning models, particularly those leveraging Convolutional Neural

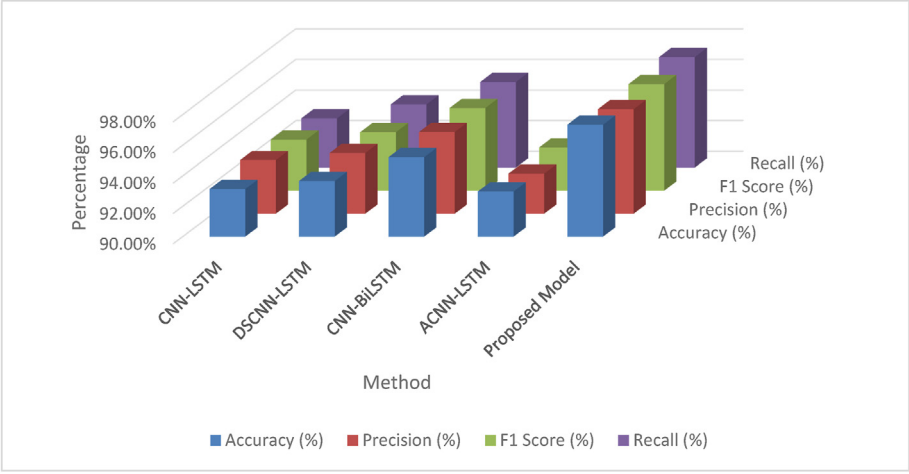

Fig. 8. Performance on all methods.

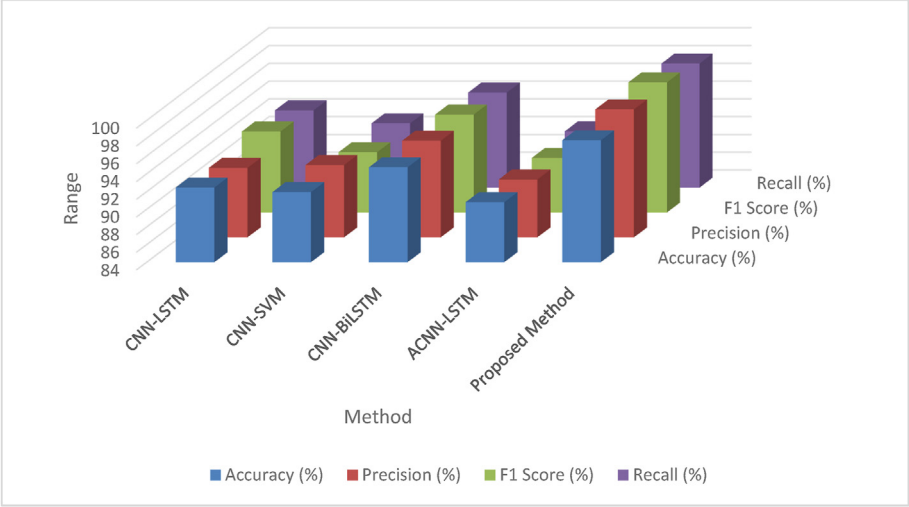

Fig. 9. Performance on all methods.

Networks (CNNs). These advancements have significantly improved the ability of systems to analyze and interpret facial expressions with greater accuracy and efficiency. In this study, we propose a novel approach that further enhances the performance of FER by combining several state-of-the-art techniques in our CNN-BiLSTM hybrid model, which is augmented with an attention mechanism. This combination enables our model to capture complex patterns and nuances in facial expressions more effectively, especially in challenging scenarios such as partial occlusions or varying facial poses. Our model incorporates multiple components, each contributing to improving recognition accuracy. The attention mechanism plays a key role by enhancing feature mapping, allowing the model to focus on the images' most relevant and informative features. This selective focus helps the model distinguish subtle differences between expressions, which can be challenging to capture in noisy or distorted images. We apply a median filter to improve further image quality, which removes noise and helps maintain edge clarity. This preprocessing step ensures the model works with cleaner, higher-quality images, improving its overall performance.

In addition to these preprocessing techniques, Global Contrast Normalization (GCN) is employed to normalize the images. GCN enhances the contrast of images, making it easier for the model to detect facial features across a range of lighting conditions. The CNN component of our model is responsible for feature extraction, where it automatically learns to identify important facial landmarks and other significant characteristics from the input images. Once these features are extracted, the BiLSTM network is used to analyze the temporal dynamics of the facial expressions, capturing both forward and backward dependencies in the image sequences. Finally, the softmax layer is employed for classification, effectively categorizing the extracted features into seven distinct emotion classes: happiness, sadness, surprise, fear, anger, disgust, and neutrality. By combining these various techniques into a single unified framework, our model can achieve improved accuracy in recognizing facial expressions, even in complex and variable conditions. This approach not only pushes the boundaries of traditional FER systems but also provides a robust solution for real-world applications in areas such as human-computer interaction, security, and healthcare.

When real-time applications are used, scalability for large datasets and deployment on resource-constrained devices could be expanded. We employ batch processing to ensure efficient updates to class centers and reduce computational overhead. This method can process data in smaller, manageable batches without compromising accuracy. It ensures that memory usage remains within practical limits, even for massive datasets. Also, hyperparameters can be optimized to improve the performance of the proposed model or equivalent classification technique.

5. Conclusion and future scope

This research presented a hybrid CNN-BiLSTM network mechanism model with a point multiplication attention paradigm for facial expression recognition, utilizing datasets such as FER2013 and CK+. The data preprocessing involved resizing the images to 128 × 128 pixels using median filters and

normalizing them with Global Contrast Normalization (GCN). To capture sequential information, the output from the CNN is passed through a Bidirectional LSTM (BiLSTM) to analyze both forward and backward temporal dependencies. The resulting features are then fed into the Attention Mechanism (AM) module for point multiplication. The attention map generated by the AM is further refined through dimensionality reduction using Linear Discriminant Analysis (LDA), producing an enhanced feature map that is utilized by the reconstruction module. A softmax layer is employed to categorize the expressions into seven distinct emotion categories for final classification. The performance of the proposed model was compared with existing models, including CNN-LSTM, DSCNN-LSTM, CNN-BiLSTM, and ACNN-LSTM. The results demonstrated that our approach outperforms these models regarding F1 score, accuracy, precision, and recall.

Future research studies will incorporate additional emotions, explore more diverse datasets, and refine the attention mechanism to enhance the model's effectiveness. This research objective to improve driver emotion recognition by integrating advanced deep-learning algorithms and optimized attention mechanisms, which will reduce accident rates and enhance road safety.

Ethics information

None.

Author contributions

Writing—review & editing, Sushil Kumar Singh; Writing—original draft, Sushil Kumar Singh; Methodology, Sushil Kumar Singh, Manish Kumar; Implementation, Manish Kumar; Validation, A. Jaynthiladevi; Resources, Chirag Agarwal, Sushil Kumar Singh; Visualization, A. Jayanthiladevi, Chirag Agarwal; Formal analysis, Sushil Kumar Singh, Ikram Majeed Khan; Supervision, Sushil Kumar Singh; Project Administration, Sushil Kumar Singh and A. Jayanthiladevi; Funding acquisition, Sushil Kumar Singh.

AI usage declaration

AI usage declaration in this scientific work, generative artificial intelligence (AI) was not used.

Acknowledgment and Funding

This research was supported by the Research Seed Grant funded by the Marwadi University, Rajkot, Gujarat (MU/R&D/22- 23/MRP/FT13). This research was collaborated by Marwadi University, Rajkot, Gujarat, India, Seoul National University of Science and Technology, Seoul, South Korea, Coventry University, Priory St, Coventry, England, UK, Tata Consultancy Services Ltd., USA.

Conflicts of interest

The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Wang S, Shuai H, Zhu L, Liu Q. Expression complementary disentanglement network for facial expression recognition. Chin J Electron 2024;33(3):742–52. https://doi.org/10.23919/cje.2022.00.351.

[2] Raman DR, Kumar V, Pillai BG, Rabadiya D, Patre S, Meenakshi R. Multi-modal facial expression recognition through a hierarchical cross-attention graph convolutional network. In: 2024 international conference on knowledge engineering and communication systems (ICKECS), Chikkaballapur, India; 2024. p. 1–5. https://doi.org/10.1109/ICKECS61492.2024.10616566.

[3] https://www.animeoutline.com/12-anime-facial-expressions-chart-drawing-tutorial/.

[4] Jeremiah SR, Ha J, Singh SK, Park JH. Articles privacy guard: collaborative edge-cloud computing architecture for attribute-preserving face anonymization in CCTV networks. Human-centric Comput Informat Sci 2024;14(43):1–16.

[5] Kumar, A., Singh, S. K., Ravikumar, R. N., Khanna, A., & Brahma, B. Fusion of DRL and CNN for effective face recognition. Inform Syst Design: AI and ML Appl, 129.

[6] Kurde A, Singh SK. Next-generation technologies for secure future communication-based social-media 3.0 and smart environment. IECE Trans Sens, Communicat Control 2024; 1(2):101–25.

[7] Singh SK, Kumar M, Khanna A, Virdee B. Blockchain and FL-based secure architecture for enhanced external intrusion detection in smart farming. IEEE Internet Things J 2024;12(3): 3297–304. https://doi.org/10.1109/JIOT.2024.3478820.

[8] C P S, Ml ASJ. Integrated facial expression recognition on occluded faces using feature fusion. In: 2024 3rd international conference on sentiment analysis and deep learning (ICSADL), Bhimdatta, Nepal; 2024. p. 451–6. https://doi.org/10.1109/ICSADL61749.2024.00079.

[9] Almulla MA. Facial expression recognition using deep convolution neural networks. In: 2024 IEEE annual congress on artificial intelligence of things (AIoT), Melbourne, Australia; 2024. p. 69–71. https://doi.org/10.1109/AIoT63253.2024.00022.

[10] Abu Mangshor NN, Ishak NH, Zainurin MH, Rashid NAM, Mohd Johari NF, Sabri N. Implementation of facial expression recognition (FER) using convolutional neural network (CNN). In: 2024 IEEE 15th control and system graduate research colloquium (ICSGRC), SHAH ALAM, Malaysia; 2024. p. 92–6. https://doi.org/10.1109/ICSGRC62081.2024.10691228.

[11] Yang D, Jiang S, Wang W. Research on facial expression recognition algorithm based on improved StarGAN network. In: 2024 2nd International Conference on signal processing and intelligent computing (SPIC), Guangzhou, China; 2024. p. 852–5. https://doi.org/10.1109/SPIC62469.2024.10691554.

[12] Wang R, Ji Y, Li J. ASD diagnosis using facial expression recognition and gaze estimation. In: 2024 3rd international joint conference on information and communication engineering (JCICE), Fuzhou, China; 2024. p. 192–6. https://doi.org/10.1109/JCICE61382.2024.00047.

[13] Aggarwal S. Comparative analysis of hedge funds in financial markets using machine learning models. Int J Comput Appl 2017;163(3).
[14] Li M, Jiang F, Zhang S. Research on facial expression recognition in the case of occlusion. In: 2024 5th international conference on computer vision, image and deep learning (CVIDL), Zhuhai, China; 2024. p. 328–33. https://doi.org/10.1109/CVIDL62147.2024.10603506.
[15] Chumachenko K, Iosifidis A, Gabbouj M. MMA-DFER: MultiModal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In: 2024 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), Seattle, WA, USA; 2024. p. 4673–82. https://doi.org/10.1109/CVPRW63382.2024.00470.
[16] Sambare Manas. FER-2013 Learn facial expressions from an image. November 2024. access on, https://www.kaggle.com/datasets/msambare/fer2013.
[17] Yang G, Wang G, Li Y, Yu W. ACBL: attentive CNN-BiLSTM model for trajectory prediction. In: 2024 43rd Chinese control conference (CCC), Kunming, China; 2024. p. 8243–8. https://doi.org/10.23919/CCC63176.2024.10662836.
[18] Meghana ML, Lakshmi KSV, Harshit NC, Vani KS. Cardiovascular disease detection in ECG images using CNN-BiLSTM model. In: 2024 IEEE international conference on information technology, electronics and intelligent communication systems (ICITEICS), Bangalore, India; 2024. p. 1–6. https://doi.org/10.1109/ICITEICS61368.2024.10625411.
[19] Yao X, Lv Z, Cao L, Jiang F, Shan T, Wang J. Parallel CNN-BiLSTM fault diagnosis method based on multi-domain transformation. In: 2024 IEEE international conference on sensing, diagnostics, drognostics, and control (SDPC), Shijiazhuang, China; 2024. p. 42–6. https://doi.org/10.1109/SDPC62810.2024.10707762.
[20] Xu J, Zeng P. Short-term load forecasting by BiLSTM model based on multidimensional time-domain feature. In: 2024 4th international conference on neural networks, information and communication engineering (NNICE), Guangzhou, China; 2024. p. 1526–30. https://doi.org/10.1109/NNICE61279.2024.10498827.
[21] Duan A, Raga RC. BiLSTM model with Attention mechanism for multi-label news text classification. In: 2024 4th international conference on neural networks, information and communication engineering (NNICE), Guangzhou, China; 2024. p. 566–9. https://doi.org/10.1109/NNICE61279.2024.10498894.
[22] Dino Hivi I, Abdulrazzaq Maiwan B. A comparison of four classification algorithms for facial expression recognition. Polytec J 2020;10(1):13. https://doi.org/10.25156/ptj.v10n1y2020.pp74-80.
[23] Hamad Zana O. Review of feature selection methods using optimization algorithm (Review paper for optimization algorithm). Polytec J 2023;12(2):24. https://doi.org/10.25156/ptj.v12n2y2022.pp203-214.