IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2195-2204
Journal homepage: http://iieta.org/journals/ts

Analysis of Face Emotion Identification and Recognition Using State-of-the-Art Deep Learning Models

Ghanshyam Prasad Dubey[1], Praveen Kumar Mannepalli[2], Santosh Sahu[1], Akash Saxena[3*], Konda Hari Krishna[4], Gobi Natesan[5], Kapil Joshi[6]

[1] Department of CSE, Amity School of Engineering & Technology, Amity University Madhya Pradesh, Gwalior 474005, India
[2] Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, India
[3] Department of Computer Science, Compucom Institute of Technology and Management, Jaipur 302022, India
[4]Department of CSE, School of Computing, Mohan Babu University, Tirupati 517102, India
[5] Department of CS and IT, JAIN (Deemed-to-be University), Bangalore 562112, India
[6] Department of CSE, Uttaranchal Institute of Technology, Uttaranchal University, Dehradun 248007, India

Corresponding Author Email: akash27jaipur@gmail.com

https://doi.org/10.18280/ts.420430

Received: 10 August 2024
Revised: 7 January 2025
Accepted: 15 April 2025
Available online: 14 August 2025

Keywords:
Facial Emotion Recognition (FER), deep learning, convolutional neural network (CNN), transfer learning, pre-trained models, FER-2013

ABSTRACT

The application of several DL approaches for FER has been a focus of many researchers worldwide. Many Deep Learning algorithms are available to do this task, but the absence of large online datasets prevents any of them from achieving better accuracy. Researchers have decided to transfer learning as a solution to this problem after increasing prediction accuracy. This research presents a comparative examination of different DL pre-trained models for face emotion identification that may be applied in different real-world scenarios. Using deep learning approaches, this research endeavour gives a comparative assessment of recognising emotions of faces. The study aims to compare and assess the performance of two advanced DL models: transfer learning pre-trained models and CNNs. The well-known FER-2013 dataset, which consists of various emotion classes for classification, was utilised for this study. Phases of automated emotion identification, including data preprocessing, categorization, and visualisation, are presented. The models were assessed in the experimental work according to their F1-score, recall, accuracy, and precision. The CNN model gets 75.47% accuracy for the facial expression recognition, which outperforms other transfer learning models. The report summarises previous research on emotion recognition with deep learning models and compares and contrasts various methods. The study's results can help people who study and work in areas of data processing, DL, and facial emotion detection.

1. INTRODUCTION

An important part of interpersonal communication, emotions are basic components of human nature. Humans use body language, facial gestures, and voice as nonverbal ways to convey their emotions. Among the many components of Emotion Detection, facial expression analysis has received the greatest amount of attention and research. Recognising emotions from facial expressions has recently attracted the attention of the fields of psychology, psychiatry, and mental health. The ability to automatically identify emotions based on facial expressions is crucial for many applications, including smart homes, healthcare systems, the diagnosis of emotional disorders in individuals with autism spectrum disorders and schizophrenia, human-computer interaction (HCI), human-robot interaction (HRI), and social welfare programs that leverage HRI. Researchers have taken an interest in FER due to its potential usefulness in a variety of contexts [1].

The fundamental objective of FER is to assign emotional states to different facial expressions. There are two primary processes in the traditional FER: emotion recognition and feature extraction. The images need to be further processed by resizing, cropping, and generally improving their quality. Face detection involves cropping the face area after removing the background and non-facial regions. The primary function of a conventional FER system is to feature extraction by a processed picture; the techniques now in use include DWT, LDA, etc. [2-4]. Ultimately, by categorizing the retrieved features, typically employing an NN and other ML approaches, emotions are understood [5, 6]. After that, the collected features are sent to ML classifiers, like KNN [7], HMM [8], SVM [9], DT, and NB [10]. The key obstacles to enhancing the system performance of traditional vision-based FER include labor-intensiveness, error-proneness, dependence on domain experts for critical feature selection and classification, and independent feature extraction and classification [11]. In order to provide somewhat greater accuracy, the researcher looked into DL-based FER techniques. Several researchers used CNN-based techniques after being impressed by the recent success of DL approaches [12-15].

Recently, DNNs, especially CNNs, have been more interested in FER because of their integrated feature extraction approach from images. Several efforts to address FER issues have been reported by CNN [16-20]. Nevertheless, the CNN with only a few layers was taken into account by the extant FER methods, even though its deeper model is significantly more effective at other image-processing tasks [21]. The obstacles associated with FER may be the cause of this. A picture with a decent amount of resolution is first required for emotion identification to examine data with a lot of dimensions. Secondly, the classification task is ultimately complicated by the extremely low difference in features that results from varying emotional states. Conversely, an exceedingly deep CNN is composed of an immense quantity of concealed convolutional layers. In the CNN, the process of training an enormous quantity of concealed layers becomes tedious and fails to generalize effectively. Moreover, accuracy does not increase beyond a certain threshold as a result of the vanishing gradient problem, which is exacerbated by the mere addition of layers [22, 23]. In order to improve accuracy, many changes and methods are implemented to a DCNN architecture and training method [24-26]. A widely employed pre-trained DCNN models are VGG-16 [27], Resnet-50 [28], Resnet-152 [29], Inception-v3 [30], and DenseNet-161 [31]. Although deep models may be trained, it requires a substantial amount of data and computational capacity. This project aims to examine FER models that use DL, CNN, and TL [32], and secondly, to develop a superior model. After improving the overall performance of a DL model, this work builds a model for FER using CNNs that have been tuned for hyperparameters. The dataset FER2013, sourced by Kaggle, has around 32,298 grayscale pictures depicting diverse facial expressions on humans. Anger, fear, disgust, sadness, happiness, surprise, and neutral are seven different categories of emotions into which labelled data is classified using the Multiclass classifier. The CNN algorithm will be used to conduct emotion identification on each of the two sets, the training and testing sets of a classified dataset. The veracity of a model with respect to the FER2013 dataset will be assessed based on the results obtained. The term "training set" is used to describe the dataset or collection of samples utilized to train a model. The test set, sometimes called a subset or collection of samples, is employed to evaluate how well the trained model performed.

1.1 Research motivations and contribution

The increasing need for efficient and successful FER systems, which have a wide range of uses in fields including marketing, healthcare, security, and HCI, is what spurred this study. Despite the advancements in deep learning, there remain significant challenges in achieving high accuracy and robustness in FER models across varied environments and datasets. While CNNs have demonstrated promising results, the potential of Transfer Learning (TL) models—especially pre-trained models—has not been fully explored for FER tasks. This study is motivated by the need to explore and compare these two approaches to identify the most suitable method for real-world applications, offering insights that could lead to more accurate, scalable, and adaptable FER systems. The following research contribution of this work as:

• This work aims to create a DL model capable of reliably identifying a human face's mood using TensorFlow analysis of facial aspects through experimental research.

• Creating a CNN model-based FER technique that effectively handles problems using transfer learning.

• Create a deep learning model utilizing the FER2013 dataset, hyper-parameters, and CNN.

• Data balance was performed on the pre-processed data. Following validation on the test set, the DL model was trained using the train set.

• I Improving the performance of matching facial expression classification and identification is achieved by continuous adversarial training of CNN, which allows for deep extraction of features from the processed expression dataset.

• A proper training size dataset and hyper-parameter adjustment using CNN are also part of the experiment to enhance performance.

• An evaluation of the suggested method's efficacy in relation to existing emotion recognition techniques.

1.2 Novelty of this work

The novelty of the research lies in its comprehensive comparison of CNN and TL models for Facial Emotion Recognition (FER). While both approaches are well-explored in the domain, the paper differentiates itself by evaluating the performance of these models on multiple datasets, thus providing deeper insights into their generalization ability across different FER tasks. Furthermore, the study innovates by incorporating various pre-trained models in transfer learning, contrasting them against traditional CNN architectures, and analyzing the trade-offs in accuracy, computational efficiency, and real-world applicability. This comparative evaluation contributes to optimizing FER systems by offering a detailed understanding of how these methods can be leveraged for better performance in diverse environments.

2. LITERATURE REVIEW

This literature review summarizes the most current results in the field of FER. The paper delves into the methods used for face detection, the structure of the models used for feature extraction and classification, and the accuracy of FER achieved by the researchers mentioned. The field of FER has garnered significant academic attention.

Le Nhu et al. [33] present a novel method for FER through the integration of a deep learning (DL) model and a weighted average of face regions. The first phase involves facial and landmark recognition using DL. After training a CNN to detect emotions across all facial subregions, a weighted evaluation set is applied to each set of identifications and combined with the predicted outcome. With an accuracy rate of 74.14%, the experimental findings using the FER2013 dataset show that this suggested method outperforms advanced methodologies.

To address these challenges, Manohar et al. [34] propose a hybrid deep learning (HDL) model. SEI datasets are collected and pre-processed by artifact removal and filtering techniques. Optimal feature selection is performed using the Adaptive Search-based Deer Hunting (AS-DH) algorithm to enhance learning performance. The HDL model, which combines Recurrent Neural Networks (RNN) and Dense Neural Networks (DNN) enhanced by AS-DH, achieves high accuracy in classifying emotions. The proposed AS-DH-HDL method outperforms various existing HDL methods, showing an improvement in accuracy at a learning rate of 85%.

Alzahrani [35] develops a novel BIPFER-EOHDL model.

The hyperparameters of the EfficientNetB7 model are selected through the use of the Evolutionary Optimization (EO) algorithm. The final technique for recognizing and categorizing facial emotions employs the MA-BLSTM model. Compared to other contemporary methods, the simulation results show that BIPFER-EOHDL achieves superior FER outcomes.

Sharma and Yadav [36] explore how a CNN-based DL strategy can enhance the feature-based emotion recognition capabilities of CNNs. Their proposed method identifies the same seven FACS emotions, even though previous research did not use optimization techniques to classify different facial reactions like the 7 emotions of FACS.

Agarwal and Susan [37] investigate the use of deep pre-trained networks for recognizing emotions from veiled faces through transfer learning. They enhance several pre-trained models, including AlexNet, EfficientNet-B0, Inception-v3, ResNet-50, and Xception, using the seven-emotion category benchmark of the FER 2013 dataset. The Inception-v3 model outperforms all other DL models, as well as SVM and ANN, two machine learning models, in FER from masked faces.

Babajee et al. [38] utilize a CNN approach to study DL-based facial expression recognition. Training and testing are conducted using a system that employs a labeled dataset containing over 32,298 images, which includes various facial expressions. This ongoing effort has achieved a 79.8% accuracy rate in identifying all seven basic human emotions without the use of optimization methods.

Ayyalasomayajula et al. [39] detail an MMCNN capable of emotion detection and identification in nearly real-time. To make an emotional prediction, the final classifier considers results from the speech synthesizer and the micro-expression detector. The emotion class is confirmed if the Berkeley Expressivity Questionnaire is passed. The report concludes with results from an accuracy test of the method.

Saranyaraj and Arcot [40] assess the efficiency and effectiveness of transfer learning algorithms, such as ResNet-50, VGG-16, VGG-19, etc. Their experiments demonstrate that AI systems analyzing emotions are far more effective when using transfer learning methods. Upon rigorous experimentation and evaluation on the FER2013 dataset, their findings reveal that VGG-16 demonstrates the highest accuracy.

Table 1 provides some related work on FER using DL models. The table offers a concise comparison of various studies on FER using DL models. Each study utilizes different datasets and methods, resulting in varied accuracies and findings. While some studies focus on hybrid models and optimization techniques, others emphasize transfer learning and multimodal approaches. Common limitations include the need for further optimization, testing on diverse datasets, and real-world applications. Future work is suggested in expanding these models' applicability and exploring additional algorithms and techniques.

The authors' work contributes to an existing body of research by offering a detailed comparison of multiple Transfer Learning (TL) models for FER and providing unique insights into the effectiveness of hyperparameter optimization for CNN-based architectures. Unlike previous studies, the paper emphasizes the role of fine-tuning pre-trained models such as VGG-16, ResNet-50, and Inception-v3, showcasing how their hyperparameters can be optimized to enhance FER accuracy. The research highlights key advancements in utilizing TL for FER from masked faces and presents a novel framework that integrates multiple CNN-based models for improved emotion classification. By evaluating the performance of these models against the FER2013 dataset, the study not only provides empirical evidence of TL's superior performance but also refines the understanding of model configurations and training strategies that yield better results, distinguishing it from prior works.

Table 1. Related work for facial expression recognition using deep learning models

Reference	Methodology	Result	Limitation	Future Work
[33]	DL model with weighted average of face-regions	Accuracy: 74.14%	Less Accuracy.	Further optimization and testing on other datasets
[34]	HDL using AS-DH algorithm	Accuracy: 85%	Individuality of Facial Expression Not Detected.	Expansion to other datasets and real-world applications
[35]	BIPFER-EOHDL model using EfficientNetB7 and MA-BLSTM	Superior FER results	Lacks Optimization Approach.	Investigation of other bioinspired algorithms
[36]	Deep learning with CNN for FACS emotion identification	Demonstrates effectiveness	Less Accuracy, Complex Network	Integration of optimization techniques
[37]	Transfer learning with AlexNet, EfficientNet-BO, Inception-v3, ResNet-50, and Xception	Inception-v3 outperformed other models	Not All Models Perform Well	Enhances performance of emotion analysis
[38]	CNN algorithm	Accuracy: 79.8%	Less Accuracy.	Application of optimization methods for improved accuracy
[39]	Multimodal CNN (MMCNN)	Near real-time emotion detection	Limited exploration of the design system	Validation with larger datasets and different modalities
[40]	Transfer learning (Resnet-50, VGG-16, VGG-19)	VGG16 shows highest accuracy	Not perform well with hyperparameters	Exploration of other transfer learning models

3. METHODOLOGY

FER, a subfield of computer vision, involves the complex task of identifying emotions from facial images. The suggested technique uses the FER2013 dataset as its main source of data collection in an effort to improve FER's accuracy. Following this, data preprocessing techniques like grayscale conversion, normalization, and resizing are applied to enhance image

quality and consistency. The processed data is then split into training, validation, and testing sets to facilitate robust model training and evaluation. In the classification stage, a transfer learning (TL) model is utilized, enabling the reuse of pre-trained deep learning architectures to effectively perform FER. The final step is to evaluate the model by measuring its performance using metrics such as the F1 score, recall, accuracy, and precision. The entire methodology is structured into primary stages, as illustrated in Figure 1.

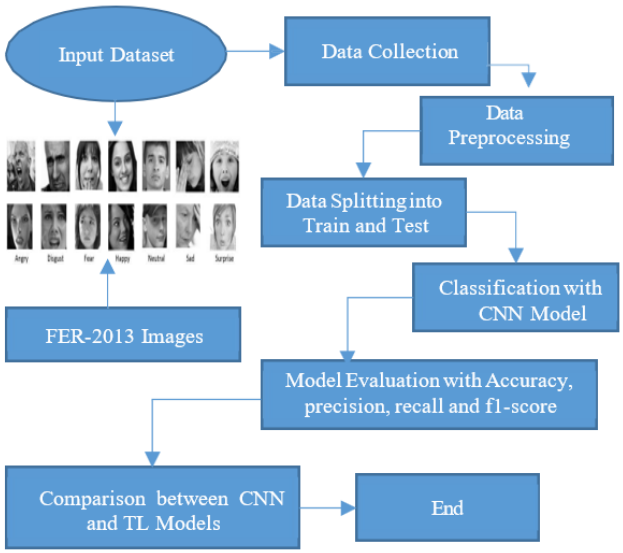

Figure 1. Flowchart for identification of facial emotions

3.1 Data collection

An initial stage entails the acquisition of input images by a FER 2013 dataset, a dataset that was expressly developed for FER in 2013 [30]. The reason this dataset was chosen is that deep networks need a large database to be trained effectively. When the input picture changes lighting, emotion, or attitude, the Gabor filter is used to fix the local distortions.

There are around 32,298 images in this collection, and each one represents one of seven different emotions: joy, wrath, surprise, fear, disgust, sorrow, or neutrality. There are headshots in both staged and candid poses. Figure 2 displays a selection of photos taken from the FER2013 database.

Figure 2. Sample images from FER2013 [30]

3.2 Data pre-processing

A dataset called FER2013 was used for the thesis. Nearly 32,298 grayscale pictures of human face expressions are part of the dataset, which is derived from Kaggle. These images are categorized into 7 categories: anger, revulsion, happiness, fear, sorrow, astonishment, and neutral. Each image in a database has been size-normalized to 48×48 pixels, and it is exclusively grayscale. Instances are viewed using the head method, and the Pandas library is employed to import the dataset.

In data preprocessing, abnormalities such as missing or incorrect numbers, illogical data structures, or data types are removed from the dataset. This generated high-quality data of high quality and facilitated the execution of an efficient data analysis. Before normalization, Figure 3 displays a bar chart including all 7 emotion classes found in a FER2013 dataset. A dataset's x-axis shows the total number of classes, while a y-axis displays the sample size. Therefore, we have resampled the whole dataset and fairly distributed 8,500 image examples across all of the classes. Following data normalization, Figure 4 shows a bar chart depicting the 7 emotion classes contained in the FER2013 dataset.

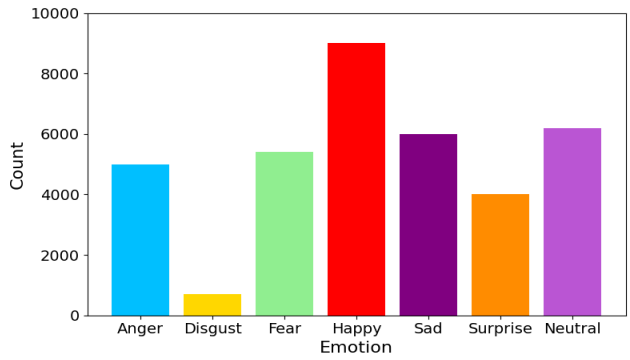

Figure 3. Bar graph of the FER2013 dataset's seven emotion classes

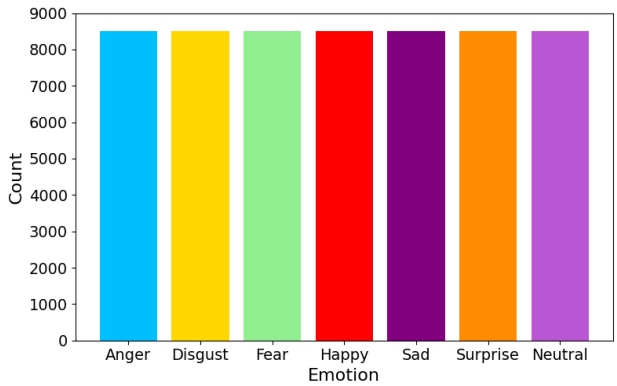

Figure 4. Bar graph of the FER2013 dataset's seven emotion classes after data balancing with 8500 samples in each emotion class

3.3 Dataset train-test split

The scikit-plot package and the train set divide function were employed to divide an entire dataset into 7 train sets and seven test sets of varying sizes. We considered training set sizes between twenty percent and eighty percent. Training the DL model using varying sized training sets, beginning at (20, 30, 40, 50, 60, 70, 80) percent of the complete dataset, was the next stage. A training set size that produced the best accuracy for the DL model was identified by analysing the many results that were obtained for every training set size.

3.4 Classification with CNN

One of the most well-known types of DNNs is the CNN.

Convolution derives its name from the mathematical operation of linearly connecting matrices. The assessment of visual pictures makes extensive use of this kind of ANN. In NLP, recommendation systems, classification, segmentation, medical image analysis, and video and image identification, they play an important role in brain-computer interfaces. Impressive achievements were achieved by picture data applications, including computer vision, NLP, and the biggest image classification dataset (Image Net) [12].

What a magnificent design CNN has. An input layer, two hidden layers (the convolutional and pooling layers, and an output layer, respectively), and a CNN are displayed in Figure 5 [13]. After training, the model receives input from the user, and the CNN processes it to provide assessed outputs, as shown in the figure.

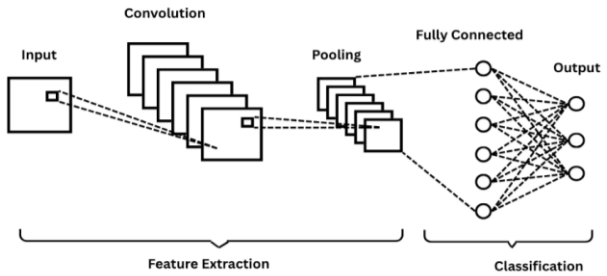

Figure 5. Schematic diagram illustrating the architecture of basic CNN [14]

3.4.1 Layers of CNN

Following is a list of the three primary CNN layers (concerning Figure 5 up top):

(1) Convolutional Layer

The input pictures' various attributes are first extracted by the convolutional layer. The mathematical procedure of convolution with an input image is performed by a filter of size M×M (where M denotes the height and width of the square filter, such as 3×3 or 5×5) in this layer. As it moves across the input picture, the filter calculates a dot product with the relevant parts of the image. This process culminates in the creation of a feature map that draws attention to critical picture elements like edges and corners. Subsequent layers use this feature map as a foundation for learning a wide range of attributes from the original image.

(2) Pooling Layer

After a convolutional layer, there comes the pooling layer. Reducing the computational expenditure by shrinking the convolved feature map is the main goal of this layer. To do this, we work on each feature map separately and reduce the connections between layers. There are several kinds of pooling operations that are defined by the methods utilised.

Max Pooling states that the feature map produces the largest element. By utilising average pooling, one may ascertain the average of every element inside a pre-sized picture segment. Min Pooling takes the feature map and pulls out the smallest part. By combining the elements in the specified area, Sum Pooling determines their total value. Combining the Convolutional and Fully Connected (FC) Layers is a typical use case for the Pooling Layer.

(3) Fully Connected Layer

To connect neurons in other layers, the FC layer uses weights and biases in addition to neurons. The output layer often follows the final few layers in a CNN. The data becomes progressively more complicated beyond this point because of the large number of hidden layers with different weights for every neuron's output. This is where all the data processing and reasoning happens.

Lastly, input pictures are transferred to the FC layer after being flattened from the layers above. After that, the mathematical functional operations are often carried out by passing the flattened vector through a few additional FC levels. The classifying procedure starts at this point.

3.4.2 Hyper-parameter tuning of CNN model with fine-tune

One of ML's hyperparameters that may influence how the model learns. Various additional parameters, including node weights, are set throughout the training process. For various datasets and models, a model's efficacy can be enhanced by utilizing a variety of hyperparameters. This set of hyperparameters is for optimizing a model, as the name suggests [15].

(1) Learning Rate (0.001): The value of this hyperparameter dictates the weight that newly acquired data will carry over older data. A failed optimisation might occur if this hyperparameter's value is significant, as the model would potentially pass over the minimum. Conversely, a slow learning rate will cause convergence to take longer than expected.

(2) Batch Size (32): An instruction set is broken down into numerous sections to make learning it faster. There will be an increase in the learning time and the need for additional memory to perform matrix multiplication if the group size is greater than. The error computation will contain a greater amount of noise if the group size is reduced.

(3) Number of Epochs (100): The epochs in DL stand for whole data cycles that can be learnt. Epochs are particularly important throughout an iterative learning process. Assuming the validation error drops, increasing the number of epochs is acceptable. Reducing the number of epochs is recommended if a validation error does not change after a certain number of epochs. It's frequently called "early stopping."

(4) Number of Layers: The performance of CNN models is enhanced by adding more layers.

To facilitate generalization, we employed dropouts at regular intervals, as this is a CNN model. The ELU was selected as an activation function in a Batch Normalization procedure. This activation function is based on ReLU and is responsible for determining the smoothness of the function when the inputs are negative. Its superior performance in this case and its ability to fix the ReLU expiration problem make it the preferred choice over LeakyReLU. Also chosen for use is the 'he_normal' kernel initializer, which is more suited for ELU. We ensured that we utilised a reduced amount of dropout, as dropout typically contributes to cacophony. The padding value was assigned to 'SAME' during the model training process, as the input and output sizes are identical.

3.5 Justification

The choice of methods and parameters in the development of the CNN model appears to be based on established practices in deep learning. For instance, the use of the FER2013 dataset is justified by its size, diversity, and suitability for training deep networks, given their reliance on large datasets to learn effectively. The decision to employ data preprocessing, such as normalization and resampling to balance class distributions, ensures uniform representation of all emotion classes, which enhances the model's generalization ability. Furthermore, the

selection of hyperparameters, including a learning rate of 0.001, batch size of 32, and 100epochs, reflects a standard approach aimed at optimizing convergence speed and minimizing overfitting. The use of ELU as an activation function, combined with the "he_normal" kernel initializer, addresses challenges like the vanishing gradient problem and ensures better initialization for deeper networks. Similarly, incorporating dropouts and batch normalization fosters model regularization and training stability.

The selection of CNN over advanced models like EfficientNet or Vision Transformers (ViTs) is justified by its simplicity, computational efficiency, and suitability for the FER2013 dataset. Using optimised hyperparameters such as a learning rate of 0.001, batch size of 32, ELU activation, and the 'he_normal' initialiser, CNNs succeeded in achieving greater performance when it came to extracting localised features, which are vital for emotion identification. Techniques like dropout regularization and batch normalization further improved generalization and stability. These factors highlight CNN's effectiveness for this task, while advanced architectures could be explored for potential future improvements.

4. RESULTS ANALYSIS AND DISCUSSION

To evaluate the proposed FER-2013 approach, this part uses CNN and TL on a benchmark dataset. First, a method for constructing an experimental environment and the baseline data sets is discussed. Another way of providing evidence for the effectiveness of a recommended model is by comparing it to the benchmark datasets alongside those found from other methodologies.

4.1 Performance indicators

One should consider many factors in the analysis of performance to understand to what extent the proposed approach will estimate the facial emotions. These are quite informative about the behavior of the model and the capability of a model to differentiate among different emotions. The proposed approach has anticipated the following performance metrics [40].

Accuracy: This is a measure of the accuracy of the test set in correctly identifying different categories of facial expressions. Overall, the information that is provided allows for assessing the model's efficiency in several different emotions. It may be relevant to express the accuracy formula as Eq. (1).

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

Precision: It is a ratio of the number of expected positive outcomes for a specific emotion class to the actual positive results that have been correctly predicted. The accuracy formula is Eq. (2).

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

Recall: This is the extent to which a particular emotion class performs in terms of correctly labelling the positive instances relative to the number of positive instances. The accuracy formula is Eq. (3):

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

F-measure: By finding the harmonic mean of memory and accuracy, it measures how well they are balanced. False positives and false negatives are both included in this single score. The accuracy formula is Eq. (4):

$$\mathrm{F1} = \frac{2 * (\mathrm{precision} * \mathrm{recall})}{\mathrm{precision} + \mathrm{recall}} \quad (4)$$

Confusion matrices, which are sometimes called error matrices, are tables that show how an algorithm has performed while running. The actual instances of each class are included in the rows of the matrix, while the predicted instances are represented in the columns, or the inverse is also true. To determine F1 scores, recall, accuracy, and precision it is used. The confusion matrix makes use of the following terminology [41]:

True Positive: Data points with expected and actual classes of 1 are good examples of this.

False Positive: A data point is considered to have this attribute when its expected attribute is one and its actual attribute is zero.

True Negative: This feature is shown by a data point when its expected and observed classes are both zero.

False Negative: This occurs when a data point's actual class is one and its projected class is zero.

Loss: This loss, which happened during training and is also known as training loss, shows that there was a mistake with the data utilized for training. This metric measures the overall number of mistakes made by the training set throughout the model training process.

4.2 Results of the CNN model

The following Table 2 displays the classification results of the CNN model for FER according to precision, recall, accuracy, and f1-score measures. CNN model gets 73%, 72% and 75% performance, respectively.

Table 2 summarises the outcomes of the CNN model's evaluation on the FER-2013 dataset, which shows how well it recognises emotions. The majority of predictions across all emotion classes were right, as the model obtained an accuracy of 75.47 percent. The 72% precision rate shows that the model is good at reducing FP, which means that the majority of the time, the emotions predicted are correct. The fact that the model was able to properly identify the majority of the real positive events (73% recall) lends credence to its capacity to distinguish between the predicted class of emotions. Lastly, the F1-score of 73 confirms the model's overall dependability in classification tasks by balancing recall and precision. Despite the fact that more optimisation might enhance its results, these performance measures show that the model is balanced and capable of emotion identification.

Table 2. Classification results of the CNN model

Parameters	Performance
Accuracy	75.47
Precision	72
Recall	73
F1-score	73

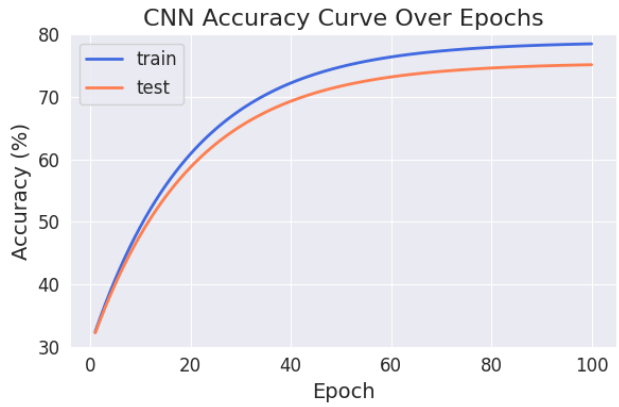

Figure 6. Emotion recognition CNN model accuracy graphs

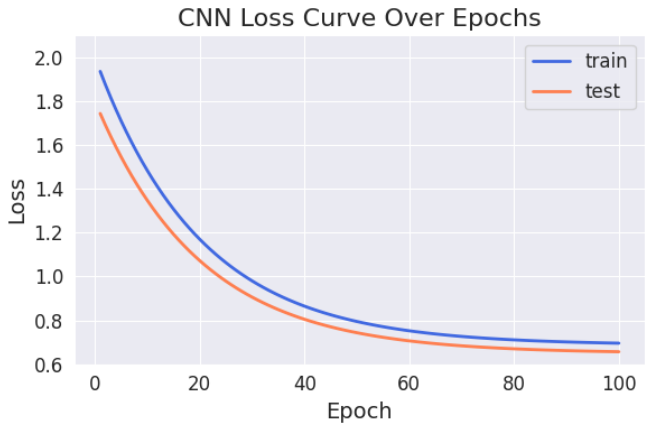

Figure 7. Emotion recognition CNN model loss graphs

Figures 6 and 7 exhibit the outcomes, showing the accuracy and loss of the CNN model that was built for emotion recognition. A model's accuracy is shown on the y-axis, and the number of epochs is shown on the x-axis. On the loss and accuracy graph, the blue line represents the training set's accuracy, while the orange line represents the test set's accuracy. The accuracy of an epoch's history has been steadily rising, reaching 75.47% on the test set and 78.84% on the training set. The best result was achieved by the DL model with a batch size of 32, which improved its overall accuracy. A group of optimisers called Nadam was used. The history of increasing declines in loss over the epochs shows that it has achieved a loss of 68.71% on the training set and 64.92% on the test set.

The CNN model's emotion classification capabilities are thoroughly evaluated using the Confusion Matrix in Figure 8. In this matrix, real emotions are represented by rows and predicted emotions are represented by columns. Elements that lie off the diagonal represent misclassifications, whereas those that lie on the diagonal represent accurate predictions. Notably, the model performs well with emotions like Disgust (1798 correct predictions), Happy (1551 correct predictions), Surprise (1590 correct predictions), and Neutral (1402 correct predictions), which suggests that these emotions are well-represented and easy for the model to identify. However, it struggles with Anger (1286 correct predictions), which is frequently misclassified as Disgust, Fear, and other emotions. The Fear class (994 correct predictions) is also problematic, with misclassifications primarily as Happy and Sad, suggesting that Fear shares visual features with these emotions in the dataset. Similarly, Sad (1096 correct predictions) faces misclassifications as Fear and Surprise, indicating that these emotions may have overlapping facial expressions, which confuse the model. The frequent misclassifications point to areas where the model could be improved, such as better differentiating between emotions with similar facial features, like Fear and Sad, or enhancing the recognition of Anger. Addressing these misclassification patterns through data augmentation, model refinement, or feature enhancement could lead to improved accuracy in future iterations.

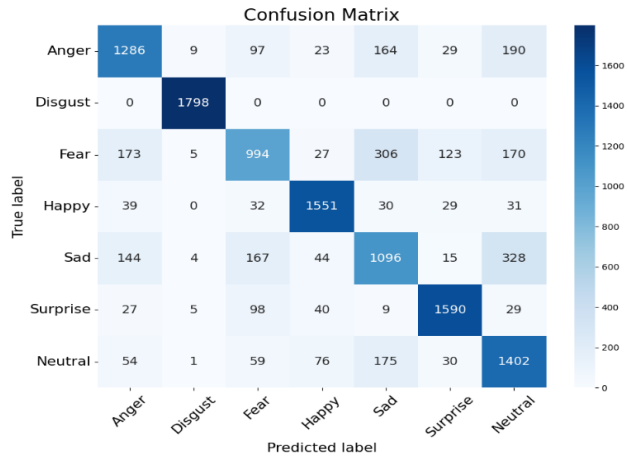

Figure 8. Confusion matrix of the emotion classes with CNN models

4.3 Comparative analysis

The following Table 3 displays a comparison of existing TL (ResNet-50 [42], SqueezeNet [43], and AlexNet+Inception V3+ResNet50 (Ensemble) [44]) and CNN models for emotion recognition on the FER-2013 dataset according to the accuracy parameter.

Table 3. Comparison between the existing TL and CNN models

Models	Accuracy (%)
CNN	75.47
ResNet-50 [42]	73.40
SqueezeNet [43]	61.09
AlexNet+Inception V3+ResNet50	73.56

The comparison in Table 3 highlights the performance of the CNN model alongside several transfer learning (TL) models, including ResNet-50, SqueezeNet, and an ensemble of AlexNet, Inception V3, and ResNet-50. While these models were likely chosen for their established performance in image classification tasks, the paper does not explicitly justify their selection. ResNet-50 is known for its deep residual architecture, which addresses vanishing gradient issues and facilitates the learning of complex features, making it a logical candidate for emotion recognition tasks. SqueezeNet, with its lightweight architecture, is designed for computational efficiency, offering a contrast in resource utilization compared to heavier models. The AlexNet-Inception V3-ResNet50 ensemble likely aims to combine the strengths of multiple architectures to improve generalization and accuracy. However, the lack of a detailed rationale for these choices leaves it unclear whether other models, such as EfficientNet or Vision Transformers, were considered and how their inclusion might have impacted the results. A more thorough explanation

of the criteria for selecting these TL models would strengthen the study's comparative analysis [44].

The accuracy comparison of various TL and CNN models for FER on the FER-2013 dataset reveals distinct performance differences. The CNN model stands out with the highest accuracy at 75.47%, indicating its robustness in handling the dataset's complexities. ResNet-50, a deep residual network, follows closely with 73.40%, showcasing its effective feature extraction capabilities despite its deeper architecture. The ensemble method combining AlexNet, Inception V3, and ResNet-50 achieves an accuracy of 73.56%, slightly surpassing ResNet-50 alone, highlighting the benefits of leveraging multiple models to enhance performance. SqueezeNet, designed for efficiency with a smaller model size, achieves 61.09%, which, while lower than the others, demonstrates a reasonable trade-off among accuracy and computational efficiency. These results underscore a trade-off between model complexity, computational demands, and accuracy in FER tasks, with CNN providing superior accuracy on the FER-2013 dataset.

4.4 Discussion

While the study demonstrates promising outcomes on the FER2013dataset, the model's performance on real-world images with diverse lighting conditions, occlusions, or variations in head poses remains uncertain. Real-world scenarios often introduce significant challenges such as these, which can hinder the model's generalization. To address these issues, the models can be adapted by incorporating domain adaptation techniques, where the model is fine-tuned on more diverse and challenging datasets that better reflect real-world conditions. Additionally, data augmentation methods, such as introducing controlled variations in lighting, head poses, and occlusions, can help the model become more robust to these real-world variations. Fine-tuning the pre-trained models on such augmented datasets would further improve their ability to handle the complexity and variability seen in real-world FER tasks.

Although transfer learning models like ResNet-50 and SqueezeNet are evaluated in the study, the paper lacks an in-depth analysis of the benefits and limitations of using transfer learning for FER tasks. It is crucial to assess whether these pre-trained models, which were originally trained on large-scale image classification datasets such as ImageNet, are well-suited for FER, given the differences in the nature and characteristics of facial expression data. For instance, models pre-trained on general object datasets may not have learned features that are specific to the nuances of facial expressions, potentially affecting their performance on FER tasks. A detailed discussion on this aspect would help readers understand the trade-offs involved in transfer learning—such as leveraging pre-trained knowledge versus training a model from scratch on a domain-specific dataset—and how these trade-offs impact the model's generalization, accuracy, and efficiency for FER applications.

5. CONCLUSION AND FUTURE WORK

FER is a rapidly expanding area with promising medical and computer vision applications. It has been gaining momentum in recent years. This thesis presents a DL model for FER that optimises the model's hyperparameters using a CNN. Testing the model on the FER2013 dataset yielded an accuracy of 75.47 percent, proving that hyperparameter adjustment improved the model's performance. This demonstrates the promise of deep learning for emotion identification and lays the groundwork for studies to improve and generalise these models in the future.

There are a number of constraints that need to be addressed in future research, even if this study shows promise in FER. Overfitting is a real possibility due to the FER2013 dataset's limited size, which is particularly problematic when dealing with complicated deep learning models such as CNNs. The model's performance may also be influenced by the lack of diversity in the dataset, which may not account for variations in lighting conditions, head poses, or occlusions often encountered in real-world scenarios. Additionally, the transfer learning models used in this study, such as ResNet-50 and SqueezeNet, may not be ideally suited for FER, as they were pre-trained on general image classification tasks, and their features may not fully capture the nuances required for accurate emotion recognition.

To get around these restrictions, researchers may train the model on a bigger, more varied dataset that more accurately depicts the range of facial emotions seen in real life. Furthermore, exploring the use of advanced architectures, such as transformers or multimodal approaches combining facial expressions with other cues like speech or body language, could lead to more robust models. It would also be beneficial to evaluate the model's performance on a wider range of datasets to test its generalizability and effectiveness in different environments. Finally, investigating domain adaptation and fine-tuning strategies could further enhance the model's ability to perform well under varied real-world conditions.

REFERENCES

[1] Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S., Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, 10(9): 1036. https://doi.org/10.3390/electronics10091036

[2] Bendjillali, R.I., Beladgham, M., Merit, K., Taleb-Ahmed, A. (2019). Improved facial expression recognition based on DWT feature for deep CNN. Electronics, 8(3): 324. https://doi.org/10.3390/electronics80figu30324

[3] Liew, C.F., Yairi, T. (2015). Facial expression recognition and analysis: A comparison study of feature descriptors. IPSJ Transactions on Computer Vision and Applications, 7: 104-120. https://doi.org/10.2197/ipsjtcva.7.104

[4] Ko, B.C. (2018). A brief review of facial emotion recognition based on visual information. Sensors, 18(2): 401. https://doi.org/10.3390/s18020401

[5] Hebri, D., Nuthakki, R., Digal, A.K., Venkatesan, K.G.S., Chawla, S., Raghavendra Reddy, C. (2024). Effective facial expression recognition system using machine learning. EAI Endorsed Transactions on Internet of Things. https://doi.org/10.4108/eetiot.5362

[6] Ullah, S., Jan, A., Khan, G.M. (2021). Facial expression recognition using machine learning techniques. In 2021 International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkey, pp. 1-6. https://doi.org/10.1109/ICEET53442.2021.9659631

[7] Murugappan, M., Mutawa, A.M., Sruthi, S., Hassouneh, A., Abdulsalam, A. (2020). Facial expression classification using KNN and decision tree classifiers. In 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, pp. 1-6. https://doi.org/10.1109/ICCCSP49186.2020.9315234

[8] Wang, J., Wang, S., Ji, Q. (2014). Early facial expression recognition using hidden markov models. In 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, pp. 4594-4599. https://doi.org/10.1109/ICPR.2014.786

[9] Bellamkonda, S., Gopalan, N.P. (2018). A facial expression recognition model using support vector machines. IJ Mathematical Sciences and Computing, 4: 56-65. https://doi.org/10.5815/ijmsc.2018.04.05

[10] Mathur, S., Gupta, S. (2024). An enhanced edge detection using laplacian gaussian filtering method from different denoising images. International Journal of Intelligent Systems and Applications in Engineering, 12(18s): 13-323. https://ijisae.org/index.php/IJISAE/article/view/4975.

[11] Rohilla, V., Chakraborty, S., Kumar, R. (2022). Deep learning based feature extraction and a bidirectional hybrid optimized model for location based advertising. Multimedia Tools and Applications, 81(11): 16067-16095. https://doi.org/10.1007/s11042-022-12457-3

[12] Gogić, I., Manhart, M., Pandžić, I.S., Ahlberg, J. (2020). Fast facial expression recognition using local binary features and shallow neural networks. The Visual Computer, 36(1): 97-112. https://doi.org/10.1007/s00371-018-1585-8

[13] Mathur, S., Gupta, S. (2023). Classification and detection of automated facial mask to COVID-19 based on deep CNN model. In 2023 IEEE 7th Conference on Information and Communication Technology (CICT), Jabalpur, India, pp. 1-6. https://doi.org/10.1109/CICT59886.2023.10455699

[14] Abdelsamad, S.E., Abdelteef, M.A., Elsheikh, O.Y., Ali, Y.A., Elsonni, T., Abdelhaq, M., Alsaqour, R., Saeed, R.A. (2023). Vision-based support for the detection and recognition of drones with small radar cross sections. Electronics, 12(10): 2235. https://doi.org/10.3390/electronics12102235

[15] Mohan, P., Chang, H. T. (2023). Deep learning based on face emotion recognition using an artificial neural network. In 2023 International Conference on Artificial Intelligence and Power Engineering (AIPE), Tokyo, Japan, pp. 19-23. https://doi.org/10.1109/AIPE58786.2023.00012

[16] Rath, A., Das Gupta, A., Rohilla, V., Balyan, A., Mann, S. (2022). Intelligent smart waste management using regression analysis: An empirical study. In International Conference on Emerging Technologies in Computer Engineering. Cham: Springer International Publishing. Springer, Cham, pp. 138-148. https://doi.org/10.1007/978-3-031-07012-9_12

[17] Mollahosseini, A., Chan, D., Mahoor, M.H. (2016). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, USA, pp. 1-10. https://doi.org/10.1109/WACV.2016.7477450

[18] Pranav, E., Kamal, S., Chandran, C.S., Supriya, M.H. (2020). Facial emotion recognition using deep convolutional neural network. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp. 317-320. https://doi.org/10.1109/ICACCS48705.2020.9074302

[19] Kumar, S.G., Sunny, S., Sayed, A., Jyothidasan, A., Nanda, V., Trinity, J., Namakkal-Soorappan, R. (2022). Chronic reductive stress modifies ribosomal proteins in Nrf2 transgenic mouse hearts. Free Radical Biology and Medicine, 192: 73. https://doi.org/10.1016/j.freeradbiomed.2022.10.125

[20] Naz, S., Aslam, M., Sayed, A. (2023). Prevalence of anemia and its determinants among the rural women of khyber pakhtunkhwa-pakistan. Annals of Human and Social Sciences, 4(4): 42-50. https://doi.org/10.35484/ahss.2023(4-IV)04

[21] Argurio, P., Fontananova, E., Molinari, R., Drioli, E. (2018). Photocatalytic membranes in photocatalytic membrane reactors. Processes, 6(9): 162. https://doi.org/10.3390/pr6090162

[22] Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review, 53: 5455-5516. https://doi.org/10.1007/s10462-020-09825-6

[23] Khan, A.R. (2022). Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges. Information, 13(6): 268. https://doi.org/10.3390/info13060268

[24] Thomas, J. (2024). Optimizing bio-energy supply chain to achieve alternative energy targets. arXiv Preprint arXiv: 2406.00056. https://doi.org/10.52783/jes.3176

[25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[26] Thomas, J. (2024). Optimizing nurse scheduling: A supply chain approach for healthcare institutions. arXiv Preprint arXiv: 2407.11195. https://doi.org/10.52783/jes.3175

[27] Shibu, N., Sunny, S., Rajkumar, A., Sayed, A., Kalaiselvi, P., Namakkal-Soorappan, R. (2022). N-acetyl cysteine administration impairs EKG signals in the humanized reductive stress mouse. Free Radical Biology and Medicine, 192: 71-72. https://doi.org/10.1016/j.freeradbiomed.2022.10.122

[28] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv: 1409.1556. https://doi.org/10.48550/arXiv.1409.1556

[29] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[30] Shafiq, M., Gu, Z. (2022). Deep residual learning for image recognition: A survey. Applied Sciences, 12(18): 8972. https://doi.org/10.3390/app12188972

[31] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of The

IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 1251-1258. https://doi.org/10.1109/CVPR.2017.195

[32] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[33] Le Nhu, H., Dang, H.V., Xuan, H.H. (2023). Facial emotion recognition by combining deep learning and averaged weight of face-regions. In 2023 15th International Conference on Knowledge and Systems Engineering (KSE), Hanoi, Vietnam, pp. 1-4. https://doi.org/10.1109/KSE59128.2023.10299482

[34] Manohar, K., Sravani, K., Reddy, U.S. (2023). Enhanced speech based human emotion identification by hybrid deep learning and optimal features with meta-heuristic algorithm. In 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), Bangalore, India, pp. 1-6. https://doi.org/10.1109/ICCAMS60113.2023.10525931

[35] Alzahrani, A.A. (2024). Bioinspired image processing enabled facial emotion recognition using equilibrium optimizer with a hybrid deep learning model. IEEE Access, 12: 22219-22229. https://doi.org/10.1109/ACCESS.2024.3359436

[36] Sharma, S., Yadav, S. (2023). Facial emotion classification in emotional intelligence using deep learning techniques. In 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballar, India, pp. 1-6. https://doi.org/10.1109/ICDCECE57866.2023.10150842

[37] Agarwal, A., Susan, S. (2023). Emotion recognition from masked faces using inception-v3. In 2023 5th International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, India, pp. 1-6. https://doi.org/10.1109/RAIT57693.2023.10126777

[38] Babajee, P., Suddul, G., Armoogum, S., Foogooa, R. (2020). Identifying human emotions from facial expressions with deep learning. In 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, pp. 36-39. https://doi.org/10.1109/ZINC50678.2020.9161445

[39] Ayyalasomayajula, S.C., Ionescu, B., Trifan, M., Ionescu, D. (2022). A multimodal deep learning approach to emotion detection and identification. In 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, pp. 000135-000142. https://doi.org/10.1109/SACI55618.2022.9919496

[40] Saranyaraj, D., Arcot, P. (2024). A comparative study on emotion analysis using transfer learning. In 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp. 1-6. https://doi.org/10.1109/ESCI59607.2024.10497381

[41] Kaur, N., Kaur, K. (2023). Facial emotion recognition using deep learning: Advancements, challenges, and future directions. Research Square, pp. 1-17. https://doi.org/10.21203/rs.3.rs-3244446/v1

[42] Hossin, M., Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2): 1. https://doi.org/10.5121/ijdkp.2015.5201

[43] Gupta, S., Kumar, P., Tekchandani, R.K. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 82(8): 11365-11394. https://doi.org/10.1007/s11042-022-13558-9

[44] Sahoo, G.K., Das, S.K., Singh, P. (2022). Deep learning-based facial emotion recognition for driver healthcare. In 2022 National Conference on Communications (NCC), Mumbai, India, pp. 154-159. https://doi.org/10.1109/NCC55593.2022.9806751