Enhancing English Learning Environments Through Real-Time Emotion Detection and Sentiment Analysis

Myagmarsuren Orosoo[1], Yaisna Rajkumari[2], Dr. Komminni Ramesh[3], Dr Gulnaz Fatma[4], Dr. M. Nagabhaskar[5], Dr. Adapa Gopi[6], Manikandan Rengarajan[7]

Mongolian National University of Education, Mongolia[1]
Assistant Professor, Department of Applied Sciences and Humanities & Management, NIT Delhi, India[2]
Assistant Professor of English, Chairperson, BoS Anurag Engineering College, Kodad, Suryapet District, Telangana, India[3]
Language Instructor, Dept. of English, Jazan University, Jazan, Saudi Arabia[4]
Associate Professor, Department of MBA, Mallareddy Engineering College (Autonomous), Main Campus, Maisammaguda, Hyderabad, India[5]
Associate Professor, Dept.of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India[6]
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India[7]

***Abstract*—Educational technology is increasingly focusing on real-time language learning. Prior studies have utilized Natural Language Processing (NLP) to assess students' classroom behavior by analyzing their reported feelings and thoughts. However, these studies have not fully enhanced the feedback provided to instructors and peers. This research addresses this issue by combining two innovative technologies: Federated 3D-Convolutional Neural Networks (Fed 3D-CNN) and Long Short-Term Memory (LSTM) networks and also aims to investigate classroom attitudes to enhance students' language competence. These technologies enable the modification of teaching strategies through text analysis and image recognition, providing comprehensive feedback on student interactions. For this study, the Multimodal Emotion Lines Dataset (MELD) and eNTERFACE'05 datasets were selected. eNTERFACE contains 3D images of individuals, while MELD analyzes spoken patterns. To address over fitting issues, the SMOTE technique is used to balance the dataset through oversampling and under sampling. The study accurately predicts human emotions using Federated 3D-CNN technology, which excels in image processing by predicting personal information from various angles. Federated Learning with 3D-CNNs allows simultaneous implementation for multiple clients by leveraging both local and global weight changes. The NLP system identifies emotional language patterns in students, laying the foundation for this analysis. Although not all student feedback has been extensively studied in the literature, the Fed 3D-CNN and LSTM algorithm recommendations are valuable for extracting feedback-related information from audio and video. The proposed framework achieves a prediction accuracy of 97.72%, outperforming existing methods. This study aims to investigate classroom attitudes to enhance students' language competence.**

Keywords*—Convolutional neural network; federated learning; LSTM; Natural Language Processing; SMOTE

I. Introduction

Emotions are important not just in human relationships, but also in interactions between humans and computers. Because this may have an impact on an individual's psychological state, such as concentrate, decision making, and task-solving abilities. Convolutional neural network that analyses visual data and provides an accurate information in image processing. CNN extract attributes and detects patterns in images by using linear equation ideas, namely convolution procedures. CNN can be programmed to handle both speech and extra signal data, even if their primary goal is to interpret images. The fusion of Natural Language Processing (NLP) and 3D-CNN techniques sparked an innovative method of English language learning. By harnessing three-Dimensional Convolutional Neural Network (3D-CNN) image recognition and Long Short Time Memory (LSTM) text analysis providing the overall feedback of student's interaction. 3D- CNN can be good at detecting video as well as audio recognition, recommendation systems, visual segmentation, medical imaging analysis, and processing of natural languages, brain–computer user interfaces and financial history data.

Akhtar, Ekbal, and Cambria [1] recognised significant improvements have been made to this field of study, each research approach has benefits and downsides. It remains difficult to evaluate the results they produce, owing to the usage of various datasets and feature extraction approaches. Even for the same datasets, different approaches typically yield varying degrees of appropriateness for feature sets. The most significant question, therefore, is not which approach is best, but whether the conclusions might be used more broadly. It is thus more profitable to try to improve the outcomes of each categorization. Ensemble learning aims to effectively enhance the overall efficiency of the network by mixing the outputs of several candidate systems. Usama et al. [2] proposed algorithms based on deep learning have demonstrated amazing performance in the areas of NLP and computer vision. As a result, there is still an increase of using text analytical techniques like convolutional and RNN analysis to extract meaningful information. One of the key factors contributing to these models' performance involves feature retrieval. Furthermore, characteristics were passed from a single layer to another inside the network, as well as from a single network to a different network. However, multilevel and multitype combination of features remain

explored in analysis of sentiment. So, this study utilizes the three datasets to demonstrate the benefits of extracting and combining multilevel and multitype features from various neural networks. Multilevel features come from many layers of an identical network, whereas multitype features are from varying network.

Caroppo, Leone, and Siciliano [3] proposed this technique can help blind people by capturing their emotions automatically to understand facial emotions, also robots to communicate flexibly with people for providing better service. It also can be used to extract a large number of opinions by tweets in social media users from conversational data in social networks in emotion recognitions in conversation. Some of the applications, such as assistive technology, security, medical, communications and robotics has been receiving attention of emotion recognition. Mental activities of the observed subject will be the result of outcoming different facial expressions. As a result, it is important to explore exclusive methods for automatic detection of facial emotions for older adults in order for creating intelligent systems capable of customizing, for example, the response of the circumstances. Soleymani, Pantic, and Pun [4] framed the network EEG and outermost physiological signals from the user's database capture the responses of emotional videos by CNN image processing into five classes namely happy, sad, disgust, angry, fear, and surprise. The accuracy rate of EEG signals achieved at 41.7%. However, this technique failed to improve the accuracy rate of classification in low feature level fusion of EEG signals and outermost physiological signals.

Wensong and Xiange [5] studies majorly concerns for education administrations and teachers has constantly been how to monitor students' emotional changes in real time during online instruction. An advanced machine learning network framework based on the fusion of several methods of attention and CNN is suggested, using the unique roles of global, part of speech, and position attentive methods in text processing. First, the traditional ChnSentiCorp_htl_all data set is used to investigate the blending features between the different types of concentration processes and CNN in order to determine how successful it is to combine the three attentiveness mechanisms with CNN. Li et al. [6] deep learning model is the main foundation for text categorization. On the other hand, important distinctive words and a strong contextual semantic link are features of online collaborative conversation. The reliability of results from classification may decrease if solely the deep learning approach is utilized for text classification because there may be inadequate knowledge of contextual semantic connections and ignoring of important feature words. As a result, this paper suggests a multi-feature integration model that extracts the context of the text features using the BiLSTM method, its local features using CNN, its average representation features using an average pooling model, and its text's word vector representation using BERT.

A multimodal emotion recognition technique in language learning is implemented for recognising emotion and sentimental analysis using Audio and video clips. To address this visual modality, firstly sampling the dataset by SMOTE technique. While CNNs are highly skilled at extracting characteristics from textual and visual data, NLP approaches allow for the analysis and comprehension of natural language. These two methods applied to enhance the language learning experience in real time by providing feedback to learners. In this research, investigated how real-time English language learning may be improved by the application of NLP and CNN approaches. Evolving optimization selects key audio and video features, which are then sent to the upgraded Federated distributed model for deep CNN classification. A weighted fusion method is used to combine the audio and visual modalities for improved emotion recognition performance. Finally, emotions are distributed for several decentralized systems.

The key contributions of this framework are summarised as follows:

- Using the SMOTE approach to address the problems with dataset imbalance. By oversampling and under sampling, it creates the minority classes.
- To determine the feelings and sentiments of several learners, a Federated 3D-CNN and NLP architecture is built. By use of federated learning and the decentralization of datasets among devices within a client-global server architecture, the study ensures that confidential information is safely sent while maintaining the privacy of individuals.
- The impact of picture categorization and processing the complete data sequence for feedback connection are increased when Deep Federated Learning 3D-CNN and LSTM Network are combined.
- The Federated CNN model identifies the six main emotion kinds that student's exhibit and gives administrators and students an overall evaluation.

The rest of the section is structured as: Section II examines the related work on audio-visual emotion identification. Section III refers to the problem statement. Section IV describes the proposed procedure in detail, followed by Section V which includes the results and discussion. Finally, Section VI summarises the findings of the proposed work with a conclusion.

II. Related Work

Imran et al. [7] proposed a cross-Cultural Polarity and detecting the emotion using LSTM. During the pandemic situation (COVID-19) a range of comparable feelings all over the nations and the judgments made by their individual governments are differentiable. Social media has been overwhelmed with messages about COVID-19, pandemics, lockdowns, and hashtags, both are good and bad in nature. Despite their geographic proximity, several adjacent nations responded differently to their neighbor countries. Some of the countries reacted as worry and animosity where some countries reacted as normal reaction. The goal of this study is to assess the reaction of individuals from varying backgrounds cultures and people's emotion about following actions done by various governments. The sentiment140 orientation analysis dataset is concerned in this study, it achieved the highest possible performance. This framework undergoes multimodal Long Short-Term Memory (LSTM) technique for analyzing

sentiment and emotion polarity. Yet this system classifies as only positive or negative emotions not recovers the exact emotions of social media users.

Kucherlapati and Varma Mantena [8] proposed a framework to analyze the emotions of students during seminars. When the seminar starts the admin would capture the images of students in the seminar by face expression recognition method. The graphical representation of inclined gradients is utilized for face detection, while the k-near-neighbors method is used for face identification, which has the best accuracy. The suggested system, in addition to presenting an overview of all attendance individuals, also includes a comment box where individuals can offer text-based input. Sentiment Analysis, when applied to feedback using the NLP, provides an accurate representation of the members' thoughts, expressed by a histogram indicating the total amount of members who chose good, negative, or neutral input. By using the IEMOCAP dataset, it possesses an accuracy of 97%. Although it predicts a high accuracy rate the overall output emotion of all students cannot be recognized. Only positive and negative emotions are predicted.

Chatterjee et al. [9] introduce a comprehensive approach that electronic home products analyze emotions. In the consumer field SER Speech Emotion Recognition was implemented through home products. Two datasets were used to predict the sentimental analysis are Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto Emotional Speech Set database (TESS) contains young and old voice samples. Mel-Frequency Cepstral Coefficient (MFCC) was applied to extract the accurate speech by pitch and frequency. The dataset examined the well robustness of movie scenes. A dimensional Convolutional Neural Network 1D-CNN is used to augment and classify accurate emotions. By using the valid dataset this approach gives 90.48%, 95.79% and 94.47% while classification. This improves the connection between the user and smart home assistance and offers more effective feedback. Yet it doesn't address a greater number of subjects such as voices of varying ages, gender categories etc.

Wang et al. [10] proposed a perspective computer simulation to recognise facial expressions in online education. Online teaching method opening is increased on pandemic situations COVID19. So, the efficiency and applicable reliability of education-based online classes has been arising question. This framework belongs to deep learning model based on 3D-CNN technique. During online classes students undergo different kinds of emotions. From the viewpoint of computer emotion recognition stimuli, the method face expression recognition algorithm with online courses is collide with this framework, student's face images are collected from e cameras. From this method online caretakers can identify the student's effectiveness. These input images get coordinate with Cohn -Kande dataset. The facial expressions of the students are analysed and classified into 8 different kinds of emotions. However, students' expressions may not fully capture their emotions when there are many students in the online class. Liliana [11] it recognizes the student's facial expressions inside the classroom. It will find out whether the student is in anger or in surprise mode. This paper uses the collection of images dataset Cohn Kanade (CK+) which is collected by faces for facial recognition experiment. The system performance gain average accuracy rate of 92.81% still low.

Above all framework works with the Centralized CNN network. The lack of emotion prediction when shadow falls on students individually. Only positive and negative classifiers shown as result doesn't provide any feedback to user in some of the framework. The Electroencephalography signals used for emotion recognition doesn't predicts the accurate emotions. Most of the framework are not flexible for both audio and video recognising dataset. The existing framework either works for audio or video. Some of the result doesn't put any concern for CNN image processing architect to gain the better result. But the proposed framework undergoes the best 3D-CNN technique for image preprocessing it examines the accurate emotions. By colliding both suitable audio MELD and eNTERFACE video datasets. By implementing Natural Language Processing (NLP) its transcripts the voices into linguistic patterns which made the framework more effective. Current emotion recognition frameworks using EEG and visual data face challenges with low accuracy and slow learning rates, often confusing intermediate emotion states like Afraid. Many operate unimodally, recognizing only visual emotions and ignoring spoken cues. This results in incomplete feedback for administrators and students, hindering the effectiveness of emotion recognition systems.

III. Problem Statement

According to the review mentioned above, all of the current frameworks have poor levels of accuracy and have rather slow learning rates for identifying video clips [12]. Using EEG data and visual modalities, early frameworks confused between the in-between state levels of (Afraid, Angry, Happy, Sad, Surprise, and Neutral). [13]. Occasionally, it is unable to accurately identify whether the pictures in the video clip are at an intermediate level of emotion (afraid or furious). Certain articles are in unimodal mode, meaning that spoken emotions are not detected; only visual emotions are recognized. The previous approach doesn't give administrators or students any comprehensive input. This possible disadvantage affects how well emotions are recognized and how much recognition is given. By extracting characteristics from audio and video levels, the suggested methodology seeks to close this research gap by accurately detecting high-level emotion states. Federated learning is implemented in this way, which improves learning speed. Through the use of deep federated 3D-CNN and LSTM methods, the feature achieves an accuracy rate of six state-level emotions.

IV. Proposed Materials and Methods

In this proposed framework, a deep learning method in federated 3D-CNN was implemented to forecast the learner's emotion, based on stacking layers. Thus, it results the implemented block performs an augmentation task on the inputted dataset images. Input dataset images are get sampled if they have any overfitting problem. The overfitting problem are detected by Smote technique. When the dataset images get pre-processed, they can be used to training and testing procedure of the network (i.e. recognition step). In the

training/testing step, a sequence of images eINTERFACE and MELD text datasets are given to the network. By using proposed federated method allocating different weights to local clients by the global system. So, it will get trained by the images and voice by NLP to detect emotions of the leaners. Last procedure for preprocessing is converting the input RGB Red Green Blue image into a grayscale intensity image. The first stage of preprocessing is increasing the minority class that is identifying and detecting the oversampling problems in MELD and eNTERFACE'05 dataset using SMOTE (Synthetic Minority Oversampling Technique). Each emotion expression comes under three levels of emotional intensity i.e., (Positive, Negative, Neutral) Emotion (Anger, Disgust, Fear, Joy, Neutral, Sad, Suprise) is final forecasting prediction of learner's emotion and provide overall feedback to admin and learners as shown in the Fig. 1.

A. Data Collection

Most of the framework multimodal emotion recognition have single-minded on MELD database as shown in the Fig. 2. In order to differentiate the contribution of this paper, the chosen databases are MELD and eNTERFACE'05. A valid MELD and eNTERFACE'05 is a multimodal database contains the collection of image dataset regarding seven different kinds of emotion Ho et al. [14]. The framework of pretrained input images and text are collected from two dataset used to classify the emotions by recognising audio and visual.

This dataset was improved and expanded to generate the Multimodal Emotion Lines Dataset (MELD). The conversation contexts in MELD are identical to those in Emotion Lines, but in addition to text, it also includes visual as well as audio components. MELD contains around 1400 exchanges and 13,000 words from the Friends television series. Several speakers took part in the conversations. Any one of these seven emotions assigned to each statement made in a discussion. Each speech in MELD additionally includes a sentiment annotation (good, negative, or neutral). The eNTERFACE dataset comprises examples of video sequences with the necessary information that have been segmented to the next level Nguyen et al. [15].

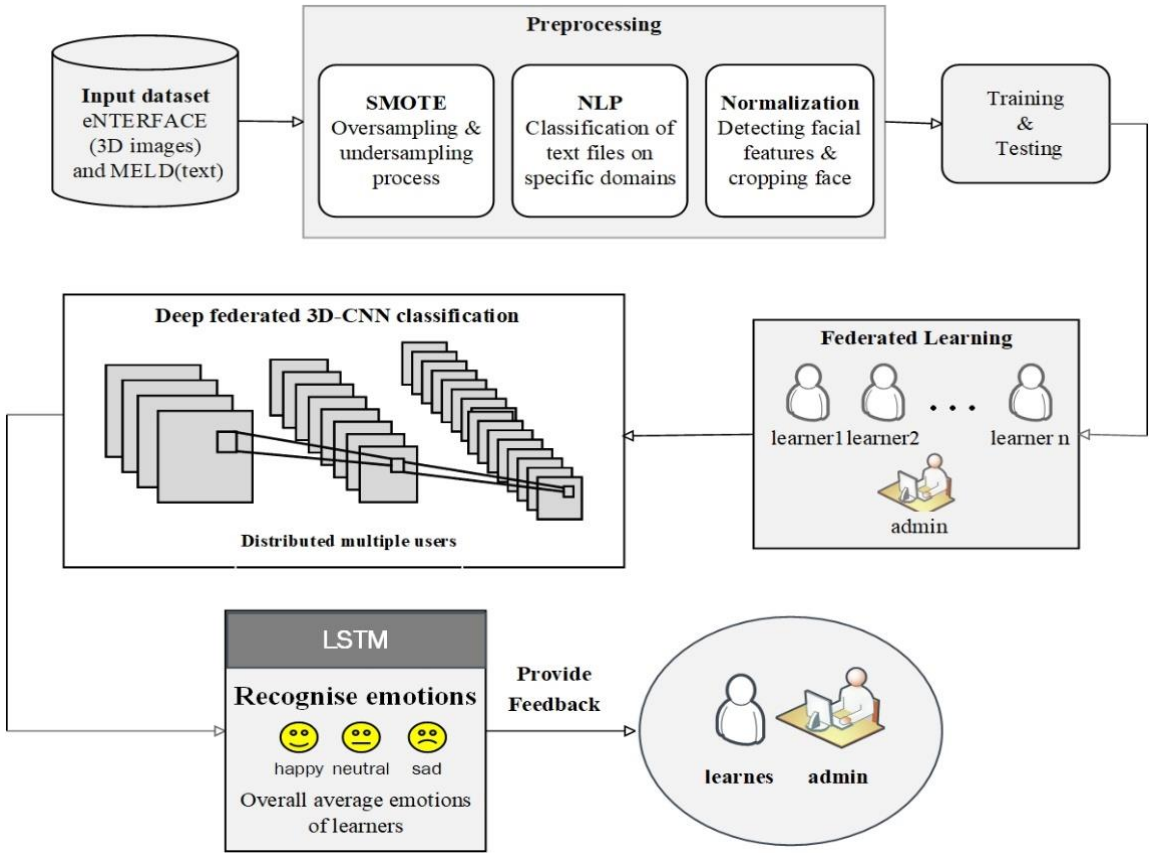

Fig. 1. Deep federated 3D-CNN architecture.

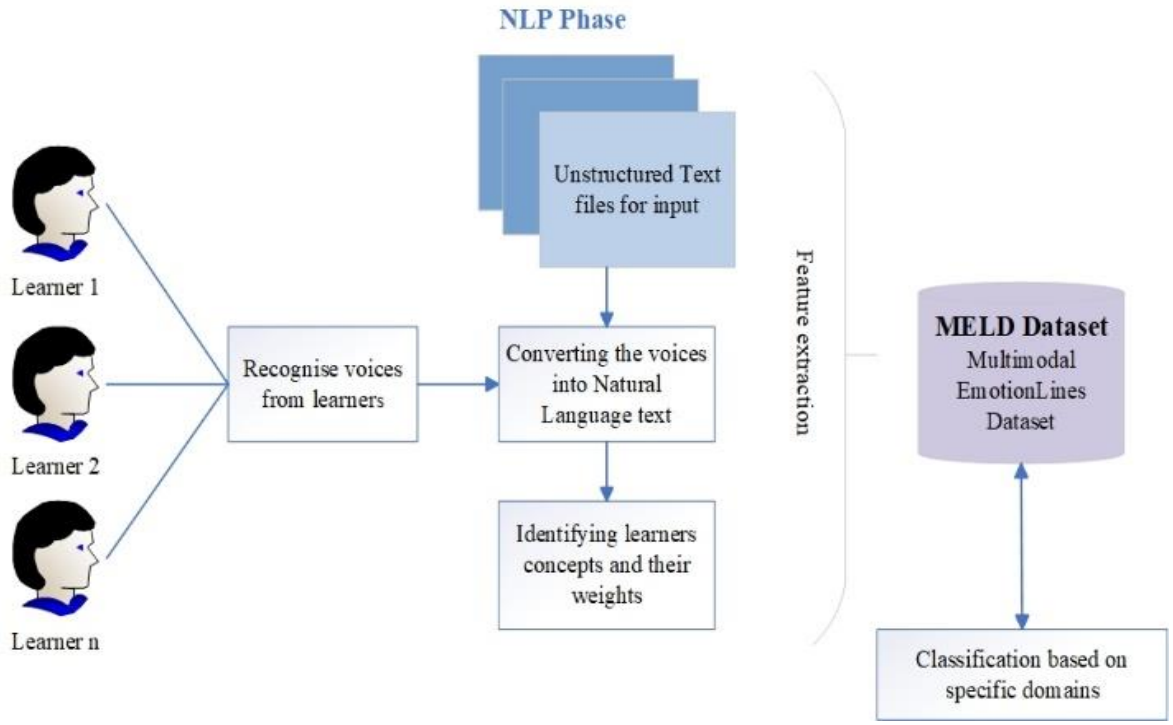

Fig. 2. Feature extraction of MELD dataset by experiment.

B. Data Pre-Processing

In this proposed framework, a deep learning method in federated CNN was implemented to forecast the learner's emotion, based on stacking layers. Thus, it results the implemented block performs augmentation task on the inputted dataset images. Input dataset images are get sampled if they have any overfitting problem. The overfitting problem are detected by Smote technique. When the dataset images get pre-processed, they can be used to training and testing procedure of the network (i.e. recognition step). In the training/testing step, a sequence of images and text MELD and eINTERFACE datasets are given to the network. The MELD dataset gets pre-processed by NLP method conversion of voices into natural language text and get classified by CNN network. Then allocating different weights to local clients by the global system. So, it will get trained by the images to detect emotions of the client. Last procedure for preprocessing is converting the input RGB Red Green Blue image into a grayscale intensity image.

The followed step of the preprocessing procedure is to extract the voices to text by NLP method and crop the image and the learners image get cropped and it will find the perfect match facial expression images in the dataset. This procedure keeps the CNNs face detection by prompting the eye aspect ratio, discarding all background information and patches of image from the background which are not matches to the facial expression. Denoised region which has already cropped delimits the image automatically. Final stage of preprocessing is the conversion of clients RGB coloured images into a black and white grayscale intensity picture. Because of RGB – Grayscale conversion, the facial features are extracted accurately. After the cropping procedure, different size of facial images is obtained. So, the cropped photos are got down sampled to 96×96 pixels using linear interpolation to remove variations in face size and maintain uniform pixel spacing. At last, the final step converts the fair coloured image into a grayscale image. Fig. 3 shows working of natural language processing.

C. Synthetic Minority Over-Sampling Technique - SMOTE

SMOTE Oversampling can predict accurate data imbalance. Under sampling deals class-imbalance problem in the MELD and eINTERFACE dataset. In the oversampling method, SMOTE can predict better accuracy rate in practical application. The pipeline of SMOTE which predicts new samples is as followed by the Fig. 4.

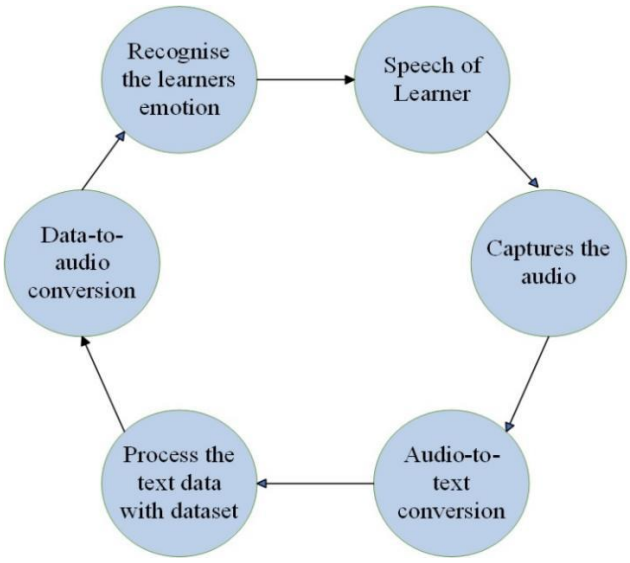

Fig. 3. Working of NLP (Natural Language Processing).

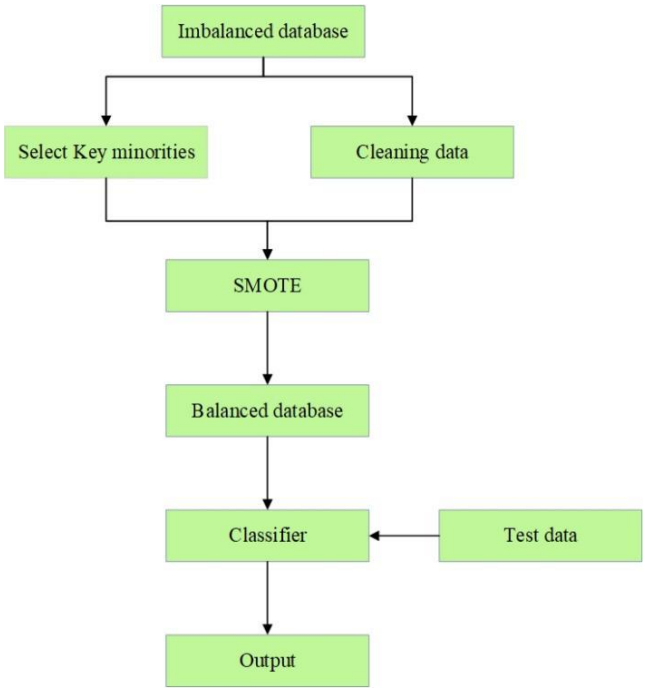

Fig. 4. Flowchart of SMOTE (Synthetic Minority Oversampling Technique).

The followed step of the preprocessing procedure is to crop the image and the clients image get cropped and it will find the perfect match facial expression images in the dataset. This procedure keeps the CNNs face detection, discarding all background information and patches of image from the background which are not matches to the facial expression. Denoised region which has already cropped delimits the image automatically. Final stage of preprocessing is the conversion of clients RGB coloured images into a black and white grayscale intensity picture. Because of RGB – Grayscale conversion, the facial features are extracted accurately. After the cropping procedure, different size of facial images is obtained. So, the cropped photos are got down sampled to 96×96 pixels using linear interpolation to remove variations in face size and maintain uniform pixel spacing. At last, the final step converts the fair coloured image into a grayscale image. Now by SMOTE oversampling improves the data imbalance Yan et al. [16].

D. Federated Learning in Image Classification

Federated learning is a machine learning technique in which an algorithm undergoes training over multiple iterative of independent sessions, each using its own dataset. This framework contrasts with standard centralized methods for machine learning which polyconnected with local datasets into a single training session, as well as assuming that local data samples are evenly distributed to all clients. Federated learning allows multiple clients to create a single, strong machine learned approach without sharing data, resolving crucial challenges like privacy of data, security purpose, access rights, and access to numerous of heterogeneous data. Its applications cover industries such as Internet of Things, telecommunications, defence and medicines. Federated learning technique trains the algorithm through multiple individual local devices, each using its own dataset. In this proposed federated framework two different datasets are classified to multiple clients. A set of users are interfaced to enumerate that return datasets.

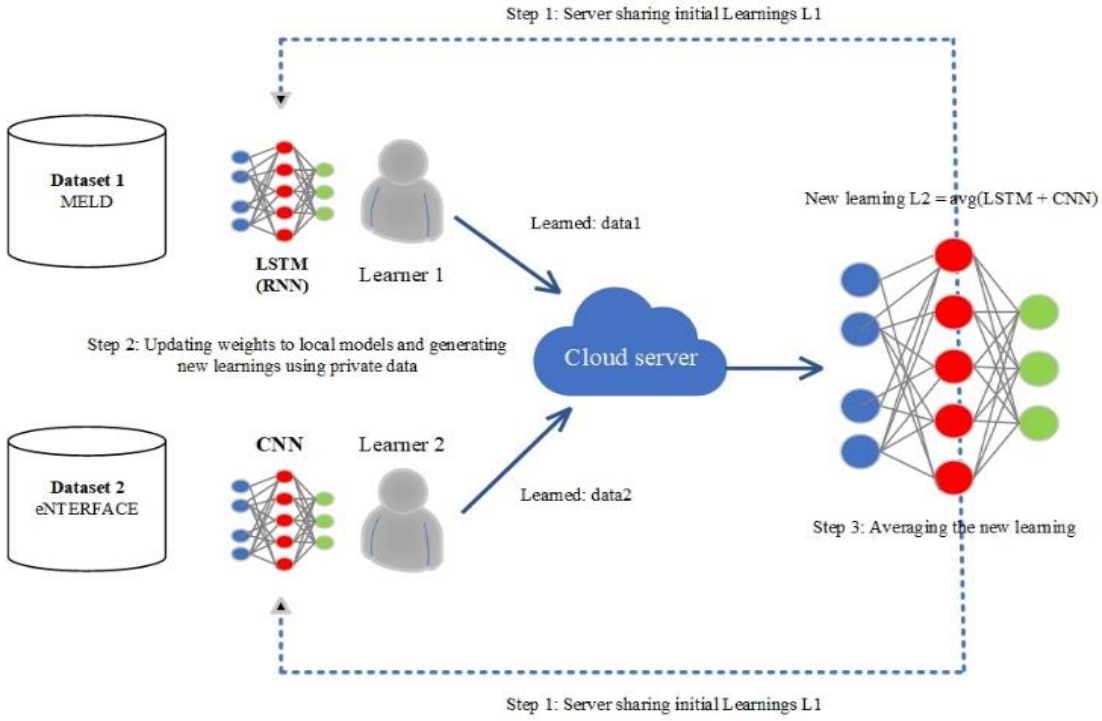

Fig. 5. Federated learning.

Their only purpose of federated system is to allow the selected subsets of the data for simulations. The main purpose of federated learning is decentralizing the local models by performing their own mandated technique. Dataset1 contains MELD and dataset2 contains eINTERFACE video and audio emotion recognising datasets. Dataset1 used by Learner 1 and dataset2 used by Learner 2 as shown in the Fig. 5. The independent local bodies interfaced the deep 3D - Convolutional Neural network but client1 process based on MELD dataset and client2 process based on eINTERFACE dataset. The initial stage is sharing the Learnings L1 i.e., sharing the local devices with two different datasets. Then updating the weights depending on the datasets and generating new learnings L2 using private data new learning L2 = avg (data1 + data2). Finally, by averaging the two learning the two datasets are combined in each learning as shown in Fig. 5. The iteration process is evoked in a convolution neural network. Each device performs the CNN technique with learned datasets.

E. Deep Federated CNN-Based Facial Extractions

Convolutional Neural Network is a widely used approach when it comes under any image processing or predictions and in performing any image-related task. This 3D-CNN network typically comprises several fundamental layers i.e., a sequence of frames that are repeated as necessary. In these layers, the convolutional layer plays a crucial role. In this layer, 3D-CNN filters systematically traverse the input image, computing values through a method as the dot product. This process works when filter moves horizontally and vertically across the large dataset input image. As already said, there will be a sequence of layers with the resulting values from the convolutional layer for the min and max pooling extraction process which are then passed to the pooling. A pooling filter is effectively reducing the size of information obtained from the output of the previous layer which is obtained from the initial image. Then the iterative process is repeated until it extracts the relevant features from the input image.

F. Long Short-Term Memory for Handling Entire Data Sequence

Long Short-Term Memory models are used for prediction and sentimental analysis and provide some feedback. By compiling a dataset of textual samples that have been annotated with the appropriate emotion (positive, negative, or neutral). After that, the raw data is processed. Tokenization of which divides texts into specific phrases. Recurrent neural networks (RNNs) of the long short-term memory type are excellent at preprocessing and predicting textual patterns of input. An embedding layer transforms words into vectors of numbers in the LSTM model, and a number of LSTM layers captures. After the model predicts an emotional response, feedback can be given by contrasting the emotion with the real emotion (if available). If the emotion predicted and the real-life emotion match, the model works as intended and no more intervention is required. The user can correct the forecast to offer feedback if it fails to reflect the real mood. By retraining on the updated data or adjusting the current factors, this feedback may be utilized for improving the model. All things considered, LSTM models for sentiment analysis offer an automated method for assessing and categorizing the sentiment of text input, and the model's predictions may be continuously enhanced and improved feedback. In the context of multilevel fusion in CNN-LSTM architectures, the integration occurs at different stages to leverage the complementary strengths of both Convolutional Neural Networks and Long Short-Term Memory networks. At the initial level, feature maps extracted by the CNN from image data are fed into the LSTM network, allowing the LSTM to capture spatial dependencies and patterns encoded in the visual information. This fusion enables the LSTM to learn contextual information from the images, enhancing its ability to make predictions based on the sequential nature of the data. Additionally, at a higher level, the output sequences generated by the LSTM are combined with features extracted from text data using CNNs. This fusion incorporates textual context into the model, enabling it to capture semantic relationships and linguistic nuances. By integrating information from multiple modalities at different levels, the multilevel fusion in CNN-LSTM architectures facilitates a comprehensive understanding of complex data, leading to improved performance in tasks such as sentiment analysis, emotion detection, and multimodal learning.

1) Model to represent the features fusion approach: After the model predicts a sentiment, feedback can be given. The algorithm worked as intended and no more steps are required if the forecast and the actual emotion match. A user can alter a forecast to offer feedback if it does not accurately reflect the emotion. Retraining the updated data or adjusting the current model parameters are two methods that may be employed to modify the model using this feedback. LSTM models offer an automated method for analysing and categorizing text data's emotions, and the model's predictions may be continuously enhanced and improved upon through feedback. Rather than approving CNN features to RNN in a sequential manner as performed in previous works, it individually learns CNN as well as RNN category features by using embedding videos and words as input for both CNN and RNN, then merging the two kinds of attributes to get multitype features fusion. Finally, a sentiment analysis into the combined set of features map. Three layers of convolution with various filter widths were employed within CNN. To obtain the final feature mapping from CNN, feed the word integrating to the convolution layer and record multilevel features following the maximum-pooling layer, as seen in Fig. 6. As seen in Figure, multilayer CNN and RNN are used to accomplish integrated multilevel and multitype feature fusion at the merged layer following the acquisition of multilevel feature fusion from CNN.

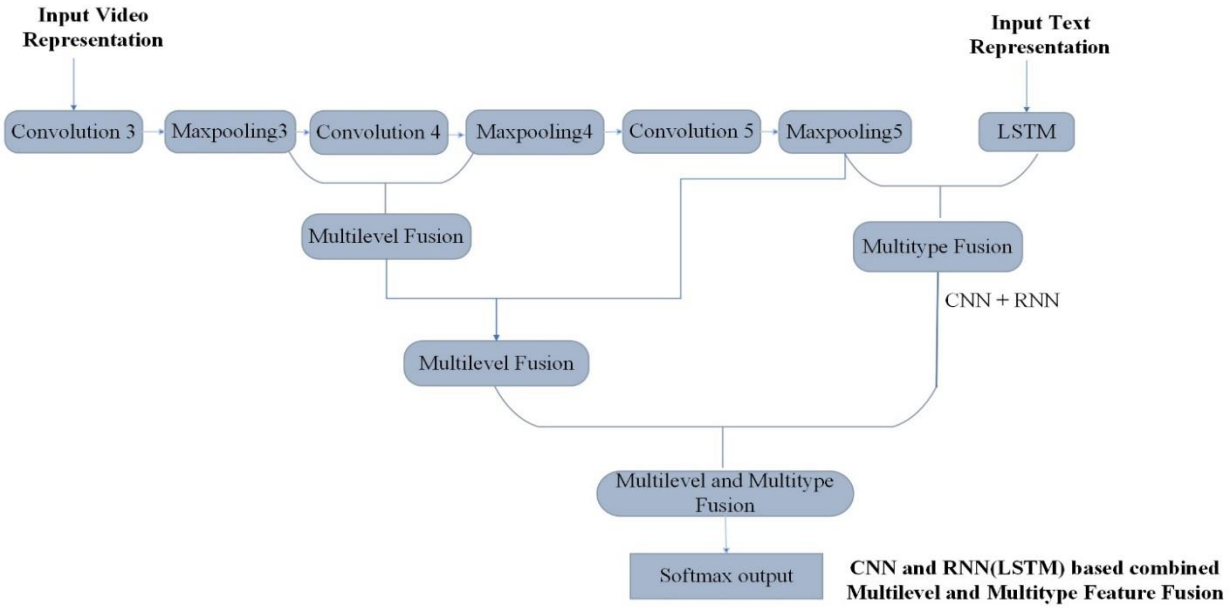

Fig. 6. Structured diagram of the proposed CNN and RNN (LSTM).

Consider an input of convolutional layer size of Weight(X) $X \times X \times D$ and Dout number of kernels with a spatial size of y with stride Z and amount of padding P, then the size of output volume can be determined by the following Eq. (1):

$$X'_{out} = \frac{X-Y+2P}{Z} + 1 \quad (1)$$

The number of iterations or repetitions process for convolutional and pooling method depends on the content of facial feature being predicted. By fine-tuning this 3D-CNN architecture, desired output can be predicted, making it an easy and effective tool for various image-related processing technique in this framework Ghosh et al. [17]. In this method, a 3D-CNN predict the key points of face which were first trained in this method Kumar et al. [18]. A cascading sequence of compression and pooling layers is used to carefully extract and fine tuning the information that is retrieved once features have been extracted from the input. If an activation map of size $X \times X \times D$, a pooling kernel of spatial size Y, and stride Z, then the size of output volume can be determined by the following Padding Eq. (2):

$$X'_{out} = \frac{X-Y}{Z} + 1 \quad (2)$$

This flattened representation is then modified to enable its smooth incorporation into a fully linked layer, which is a crucial point at which the model performs complex prediction tasks. The final output layer, the last phase of this complex process, presents a complete set of 68 facial key points that have been painstakingly extracted from the image, providing a sophisticated and in-depth comprehension of the underlying visual components. Hence the emotions are recognised by deep learning federated CNN model. Hence the result was predicted as Happy, Sad, Neutral, surprise, Disgust, Angry based on the expression delivered by user Mase et al. [19].

G. Functional Flowchart for Sentimental Analysis

The proposed framework, a deep learning method in federated CNN was implemented to forecast the client's emotion, based on stacking layers. Thus, it results the implemented block performs a pre-processing task on the inputted dataset images. Input dataset images are get sampled if they have any overfitting problem. The overfitting problem are detected by Smote technique. When the dataset of text files get pre-processed by Natural Language Processing (NLP), they can be used to training and testing procedure of the network (i.e. recognition step).

In the training/testing step, a sequence of audio and video MELD and eINTERFACE datasets are given to the network. Then allocating different weights to local clients by the global system. So, it will get trained by the images to detect emotions of the client. The last procedure for preprocessing is converting the input RGB Red Green Blue image into a grayscale intensity image. The first stage of preprocessing is increasing the minority class by identifying and detecting the oversampling problems in MELD and eNTERFACE'05 dataset using SMOTE (Synthetic Minority Oversampling Technique) and finally forecasting the emotions after CNN and LSTM classification by providing feedback to learners and admin as shown in the Fig. 7.

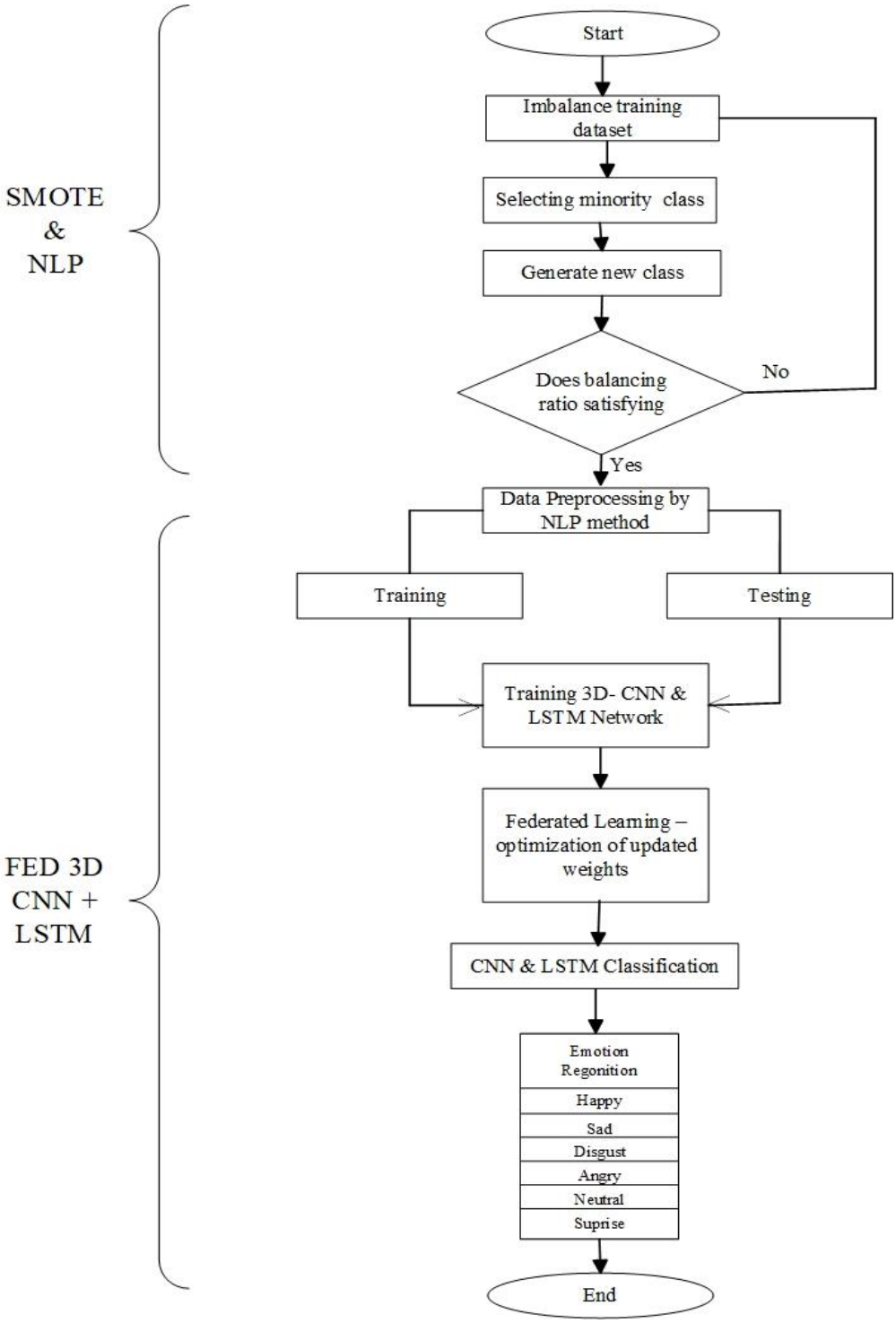

Fig. 7. The flow chart of the proposed SMOTE FED 3D- CNN.

V. Results and Discussion

In this proposed framework, a multimodal emotion recognition technique is implemented for recognising facial expressions using Audio and video clips. To address this visual modality, firstly sampling the dataset by SMOTE technique. The framework Federated 3D-CNN is implemented through Python and thus it predicts 97.72% accuracy in detecting emotions using the accurate predicting datasets MELD, eNTERFACE. And a sequence of keyframes from the image and define an aspect ratio to detect the transformation of sequence keyframes by training or splitting the data. Then the appropriate and exact facial emotion features are distributed by decentralized system i.e., to improve learning speed and kept the data very privacy at each learner and admin device. Once key features of the face have been recognised, they are passed for emotion recognition into the process of

optimized classifier. Evolving optimization selects key audio and video features, which are then sent to the upgraded Federated distributed model for deep Federated 3D-CNN classification. A weighted fusion method is used to combine the audio and visual modalities for improved emotion recognition performance. Finally, the average emotions are distributed with suitable feedback for several decentralized learner and admin systems.

A. Experimental Results

Setting up the eINTERFACE, AFEW and RAVDESS datasets, then performing preprocessing. These might include instances of labelled emotions in audio or face expressions. Divide the data into sets for testing and training to evaluate the model's performance. Preparing the dataset Mel-frequency cepstral coefficients (MFCCs) are features that may be derived from raw audio signals in order to study them.

Fig. 8 displays No of Samples per class before applying SMOTE (Synthetic Minority Over-sampling Technique), an analysis of the dataset revealed an imbalanced class distribution. In the binary classification problem at hand, Happy, Sad, Angry, Disgust, Surprise and Fear [number of samples]. The MELD dataset values are Anger=1109, Disgust=271, Fear=268, Joy=1743, Neutral=4710, Sad=683, Surprise=1205. The imbalance raised concerns about potential challenges in model training and classification performance. The decision to apply SMOTE was driven by the need to address this class imbalance systematically, ensuring a more representative and balanced dataset for subsequent analyses.

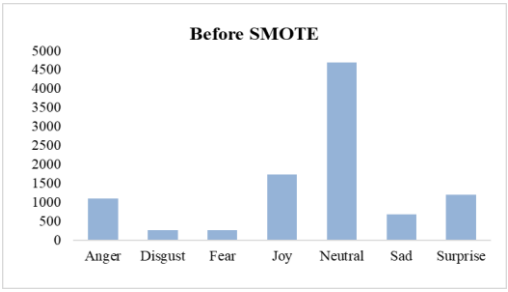

Fig. 8. No of samples per class before SMOTE in MELD dataset.

Fig. 9 displays No of Samples per class before applying SMOTE (Synthetic Minority Over-sampling Technique. The process of creating synthesis cases of the minority class increases the number of specimens per class following the use of SMOTE to rectify the imbalance of classes in a collection of data. By doing oversampling and under sampling process, SMOTE creates artificial examples across the boundary segments that link instances of minority classes that already exist. The median of all emotions is 1427. Eq. (3) it determines the SMOTE (synthetic minority oversampling technique).

$$X'_i = x_i + \lambda(x_j + x_i) \tag{3}$$

Eq. (4) explains about inverse document frequency (IDF),

$$IDF(T) = log\frac{n+1}{DF(T)+1} + 1 \tag{4}$$

Eq. (5) term frequency-inverse document frequency (TF-IDF),

$$TF - IDF(T) = TF(T) * IDF(T) \tag{5}$$

Eq. (6) F1 measure,

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \tag{6}$$

The objective is to improve the model's generalization to minority class trends while balancing the class distribution. The SMOTE parameters that are selected, including the appropriate degree of over-sampling, determine the precise rise in the total amount of samples per class. SMOTE helps lessen the effects of an unbalanced class distribution by injecting synthetic examples, which eventually leads to computerized learning frameworks that are more reliable and accurate.

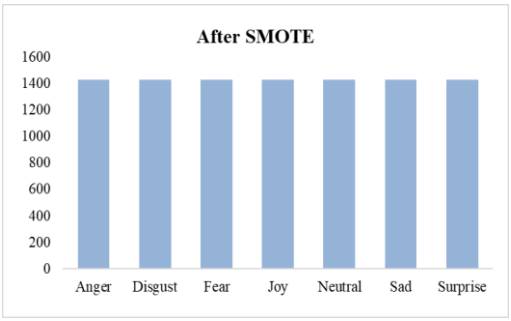

Fig. 9. No of samples per class after SMOTE in MELD dataset.

B. Training and Validation Accuracy

Using a weighted fusion method, determine the recognition accuracy for each emotion class. The suggested method fared well for identifying anger, joy, sorrow, and neutral mood. However, performances for deciding surprised and contempt were much lower. This finding is significantly influenced by the unbalanced data distribution reported in Table I. The training data contains the most video examples for annoyance, joy, and neutral, with only a few movies for disgust, anxiety, and surprised. Furthermore, human faces communicate powerful emotions like anger and delight. Fig. 9 compares the Training and Testing Accuracy for Client 1, 2 and 3 and shows how these networks learn and generalize across 100 epochs in different ways. With time, these losses diminish, suggesting that the model is acquiring up new skills and becoming more efficient. While the training accuracy increases more gradually over the course of the epochs, the testing accuracy also gradually increases like training accuracy and then on reaching further epochs.

TABLE I. AVERAGE ACCURACY OF MELD AND INTERFACE DATASET

Emotional class	Accuracy (%)
Anger	89.91
Disgust	94.56
Fear	54.67
Joy	93.87
Neutral	89.56
Sadness	95.81
Surprise	92.74

The training and testing accuracy of a Fed 3D-CNN + LSTM model and a CNN model for three individual learners are shown in Fig. 10. The figure shows how well each model predicts results using both data that it was trained on and data that it has never seen before.

Fig. 11 compares the Training and Testing Loss for Networks A, B, and C and shows how these networks learn and generalize across 100 epochs in different ways. With time, these losses diminish, suggesting that the model is acquiring up new skills and becoming more efficient. While the training loss declines more gradually over the course of the epochs, the testing loss begins at a greater value than the training loss and then abruptly declines before falling about epoch 20.

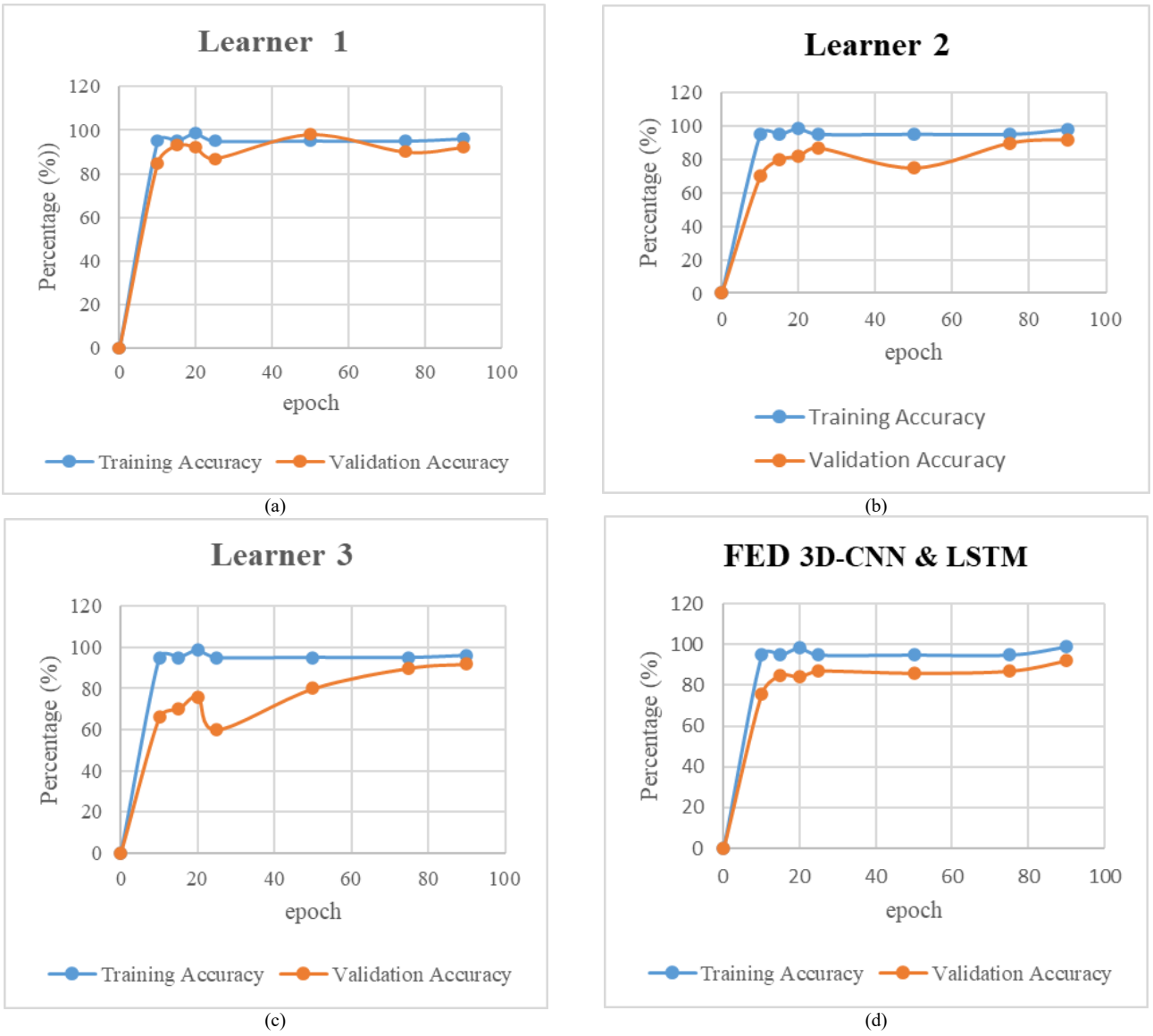

Fig. 10. Training and testing accuracy of CNN model for (a) Learner 1 (b) Learner 2 (c) Learner 3 and (d) Fed 3D-CNN + LSTM model.

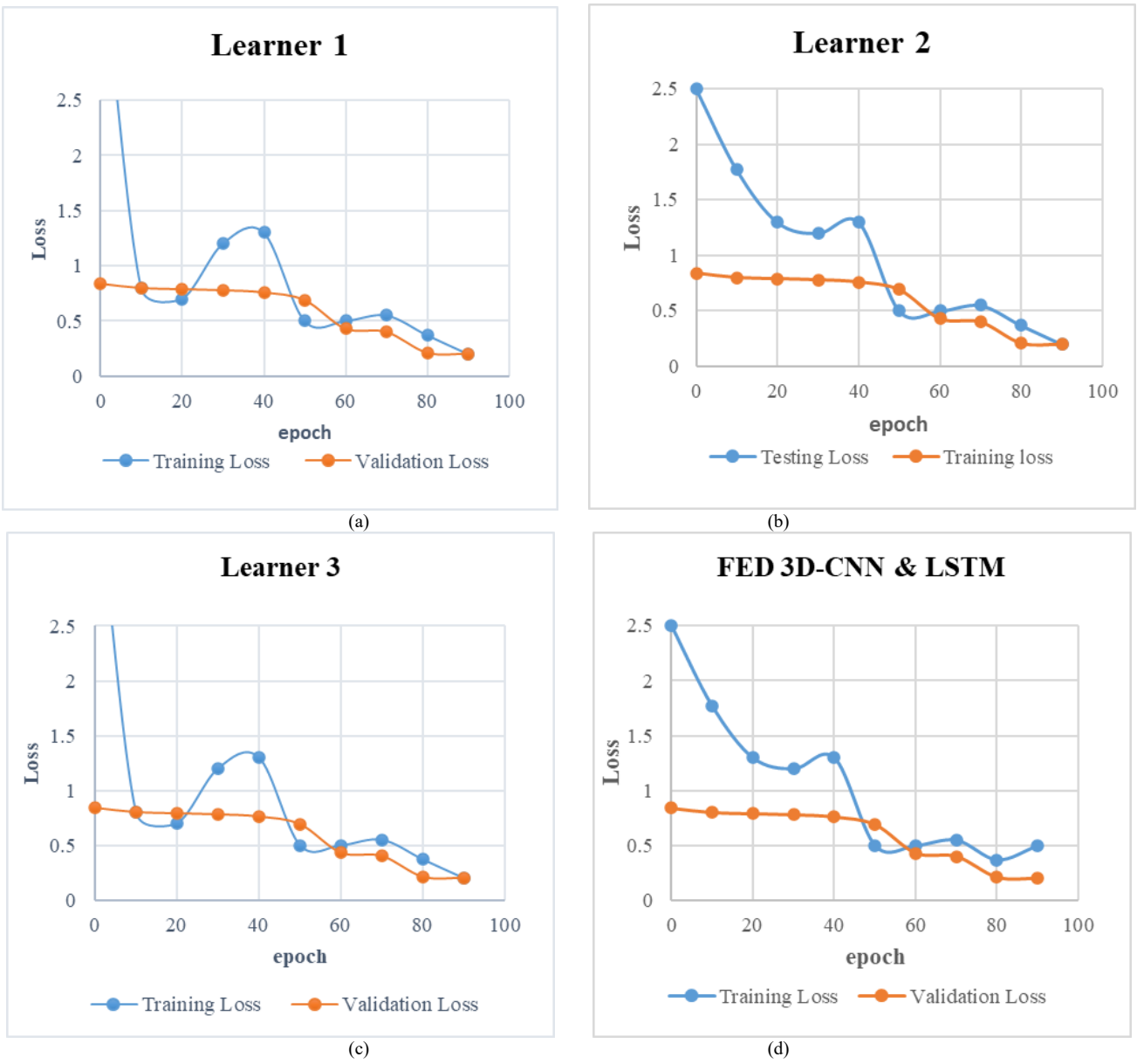

Fig. 11. Training and testing loss of CNN model for (a) Client 1, (b) Client 2, and (c) Client 3 and (d) Fed 3D-CNN + LSTM model.

C. Performance Evaluation

Metrics for performance assessment are crucial for evaluating machine learning models' efficacy and dependability quantitatively, especially when it comes to categorization tasks like diagnosing skin lesions. Below is a thorough description of a few measures employed in performance evaluations:

1) Accuracy: The proportion of accurate forecasts to all predicted outcomes is known as accuracy. When a collection of data is balanced, this measure works well. The results reported by this metric might not be accurate representations of how well the model performed when there's an overwhelming class in the data set is given in Eq. (7).

$$Accuracy = \frac{True\ Negative + True\ Positive}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \quad (7)$$

2) Precision: The deep learning algorithm's precision is a metric for determining how many anticipated positives are actually true positives. This statistic is helpful whenever the cost of a false positive is high for the efficacy of the model, like in the case of an email spam identification algorithm that is given in Eq. (8).

$$Precision = \frac{T*p}{T*p+F*n} \quad (8)$$

3) Recall: The Recall of the model in counting the number of positives out of all real positives is measured by recall. When False Negative is costly for model quality, such as in fraud detection models, this statistic is helpful and is given in Eq. (9).

$$Recall = \frac{T*p}{T*p+F*n} \quad (9)$$

4) F1-Score: The F1 score that is computed for this purpose assesses the correlation between the data's positive information and the classifier's predictions is given in Eq. (10).

$$F1\ score = \frac{2T*p}{2T*p+F*p+F*n} \quad (10)$$

To recognise the facial emotions flows through federated system. The experiment results show that RNN provides superior accuracy in terms of facial emotion recognition and it predicts 57-87% Mase et al [19]. Finalizing the emotional categories from social media. Emotions can be analysed by reacting, commenting for post, tweets etc. Love, happy, violence, sad and fear these following categories work under Flickr dataset which produce the accuracy rate. The Methods SVM on high level features of VGG-ImageNet, fine-tuning on pretrained models like RESNET, Places205-VGG16 and VGG ImageNet it predicts the accuracy 68% Gajarla and Gupta [20].

Table II visually represents the performance measures of the Federated 3D-CNN compared to traditional methods. The Fig. 12 illustration provides a clear and insightful overview of proposed model excels in metrics when compared it with the conventional approaches, emphasizing its superiority in emotion prediction. The effectiveness of deep learning techniques (GoogleNet and AlexNet) at recognizing facial movements, especially the presence of emotional content and the precise emotion character of such expressions, with an accuracy rate of 87% Giannopoulos, Perikos, and Hatzilygeroudis [21]. This work offers a comparison analysis of several approaches and algorithms that have been looked at for identifying emotions on people's faces using FERC (CNN-LSTM). This study has an accuracy rate of 78-96% by using the Viola Johnes Face Detection dataset Moolchandani et al [22].

TABLE II. COMPARING THE PERFORMANCE OF PROPOSED METHOD WITH EXISTING METHOD

Approach	Dataset	Accuracy (%)
Places205, ResNet-50, VGG	Flickr	67.68%
FLT+C3D	FACES Lifespan	74.38%
LSTM (STC-NLSTM)	SAVEE	93.45%
RNN	RECOLA	57.87%
FERC (CNN-LSTM)	Viola Johnes Face Detection	78.96%
LBP/Gabor + SRC	EmoReact	91.79%
DBN + MLP	AMFED and EmoReact	90.09%
CNN	AFEW, SAVEE 2016	89.57%
Resnet	CK+, Nimstin	73.30%
GoogleNet	CK+ and Oulu Casia	87%
Proposed Framework (SMOTE+FED 3DCNN+LSTM)	eNTERFACE and MELD	97.72%

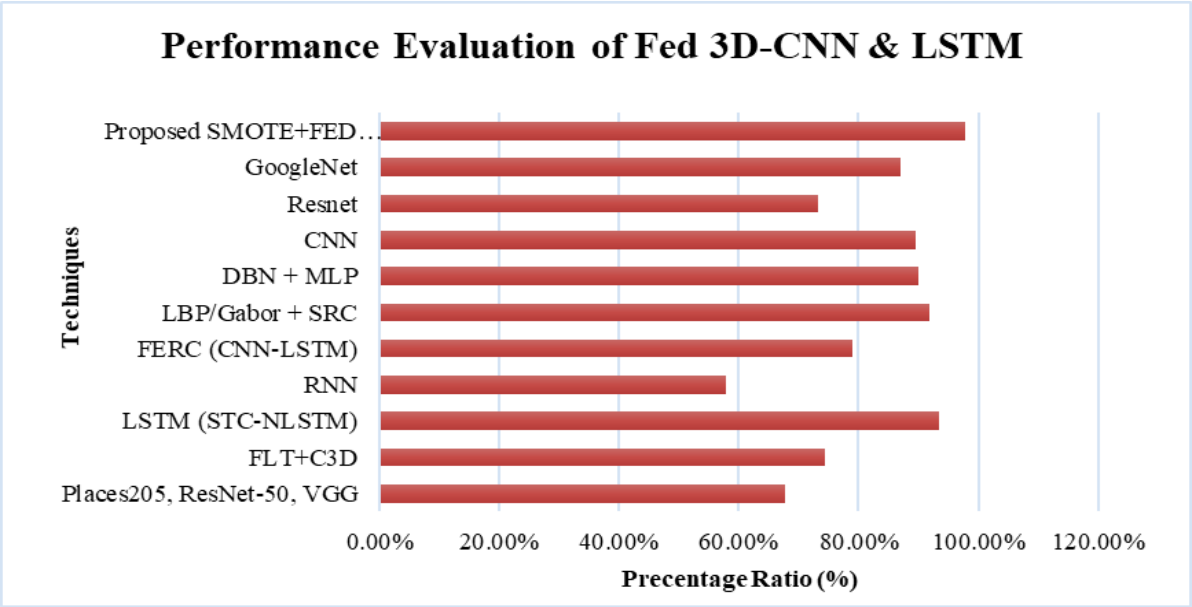

Fig. 12. Performance evaluation of Fed 3D-CNN with existing framework.

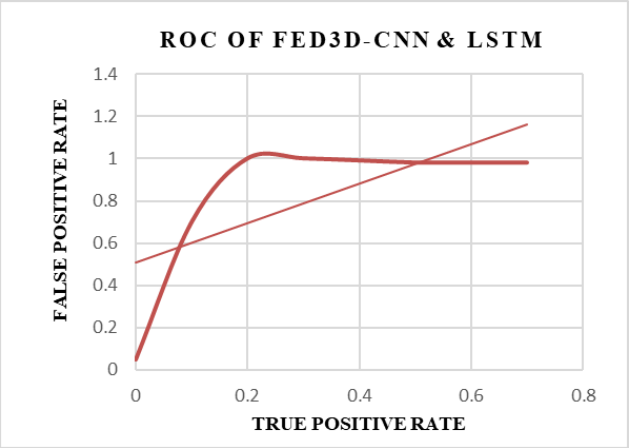

Fig. 13. ROC curve.

Fig. 13 shows ROC Curve. The ROC curve assesses binary arrangement methods' effectiveness by illustrating the compromise between sensitivity and specificity, with a sharper curve indicating higher model effectiveness.

D. Discussion

The suggested system achieves high emotion identification accuracy while maintaining data privacy by integrating Federated 3D-CNN with multimodal emotion recognition. It provides a complete method for improving emotion detection performance in decentralized learning environments by weighted integration of audio and visual modalities. Current algorithms suffer from low learning rates and low accuracy when it comes to recognizing emotions from video recordings, sometimes misinterpreting the same emotion states. Moreover, several frameworks suffer from imprecision in recognizing emotions, especially in transitional states like "afraid" or "angry," which may prevent them from providing administrators and students with comprehensive feedback [12]. Previous studies in emotion detection and sentiment analysis for language learning faced challenges with limited accuracy and real-time performance due to insufficient model complexity. They often struggled to integrate spatial and temporal data effectively, resulting in less effective emotional feedback and adaptive learning responses. [13]. To solve these shortcomings, the proposed architecture, on the other hand, combines auditory and visual aspects and uses federated learning to increase accuracy and learning speed. Using deep federated 3D-CNN and LSTM algorithms, which identify emotions at a high-level granularity of six emotion states, significantly improves the accuracy of emotion recognition. Even yet, there is still a chance that the recommended technique may encounter issues, such as scalability issues when working with large datasets and potential biases in emotion recognition. To address these limitations, future work should focus on studying other deep learning architectures, optimizing the federated learning process, and expanding the current understanding of emotion detection by integrating additional modalities.

VI. Conclusion and Future Work

This study presents a novel approach for enhancing language learning environments by utilizing real-time emotion detection and sentiment analysis through the integration of Federated 3D-CNN and LSTM networks. By leveraging these advanced technologies, the study addresses the limitations of traditional methods in providing comprehensive feedback on student interactions in classroom settings. The proposed framework accurately predicts human emotions, achieving a prediction accuracy of 97.72%, and offers valuable insights into students' emotional language patterns. The integration of text analysis and image recognition enables the modification of teaching strategies to better cater to individual student needs, ultimately aiming to enhance students' language competence.

Moving forward, several avenues for future research emerge from this study. Firstly, the proposed framework could be further validated and refined through longitudinal studies conducted in diverse educational settings to assess its scalability and generalizability. Additionally, incorporating real-time feedback mechanisms into the framework could enhance its utility in facilitating immediate adjustments to teaching strategies based on students' emotional states. Furthermore, exploring the application of the framework in other educational domains beyond language learning, such as STEM education or special education, could broaden its impact and applicability. Moreover, investigating the ethical implications and privacy concerns associated with deploying emotion detection technologies in educational settings is essential for ensuring responsible implementation. Finally, advancements in hardware capabilities and algorithmic developments may offer opportunities for optimizing the computational efficiency and performance of the framework, paving the way for more widespread adoption in educational practice.

References

[1] M. S. Akhtar, A. Ekbal, and E. Cambria, “How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble [Application Notes],” IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 64–75, Feb. 2020, doi: 10.1109/MCI.2019.2954667.

[2] M. Usama, W. Xiao, B. Ahmad, J. Wan, M. M. Hassan, and A. Alelaiwi, “Deep Learning Based Weighted Feature Fusion Approach for Sentiment Analysis,” IEEE Access, vol. 7, pp. 140252–140260, 2019, doi: 10.1109/ACCESS.2019.2940051.

[3] A. Caroppo, A. Leone, and P. Siciliano, “Facial Expression Recognition in Older Adults using Deep Machine Learning,” 2021.

[4] M. Soleymani, M. Pantic, and T. Pun, “Multimodal Emotion Recognition in Response to Videos,” IEEE Trans. Affective Comput., vol. 3, no. 2, Art. no. 2, Apr. 2020, doi: 10.1109/T-AFFC.2011.37.

[5] W. Wensong and S. Xiange, “Research on Text Multi-Feature Fusion Algorithm Based on AM-CNN,” J. Phys.: Conf. Ser., vol. 1924, no. 1, p. 012032, May 2021, doi: 10.1088/1742-6596/1924/1/012032.

[6] S. Li, M. Deng, Z. Shao, X. Chen, and Y. Zheng, “Automatic classification of interactive texts in online collaborative discussion based on multi-feature fusion,” Computers and Electrical Engineering, vol. 107, p. 108648, Apr. 2023, doi: 10.1016/j.compeleceng.2023.108648.

[7] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, “Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets,” IEEE Access, vol. 8, pp. 181074–181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[8] S. Kucherlapati and S. Varma Mantena, “A Face Recognition and Sentiment Analysis Activity System using Machine Learning Algorithm,” in 2022 International Conference on Edge Computing and

Applications (ICECAA), Oct. 2022, pp. 1346–1351. doi: 10.1109/ICECAA55415.2022.9936309.

[9] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-Time Speech Emotion Analysis for Smart Home Assistants," IEEE Transactions on Consumer Electronics, vol. 67, no. 1, pp. 68–76, Feb. 2021, doi: 10.1109/TCE.2021.3056421.

[10] W. Wang, K. Xu, H. Niu, and X. Miao, "Emotion Recognition of Students Based on Facial Expressions in Online Education Based on the Perspective of Computer Simulation," Complexity, vol. 2020, pp. 1–9, Sep. 2020, doi: 10.1155/2020/4065207.

[11] D. Y. Liliana, "Emotion recognition from facial expression using deep convolutional neural network," J. Phys.: Conf. Ser., vol. 1193, p. 012004, Apr. 2019, doi: 10.1088/1742-6596/1193/1/012004.

[12] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-Time Video Emotion Recognition based on Reinforcement Learning and Domain Knowledge," IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2020.

[13] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," SN Appl. Sci., vol. 2, no. 3, Art. no. 3, Mar. 2020, doi: 10.1007/s42452-020-2234-1.

[14] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," IEEE Access, vol. 8, pp. 61672–61686, 2020.

[15] D. Nguyen et al., "Deep cross-domain transfer for emotion recognition via joint learning," Multimed Tools Appl, Aug. 2023, doi: 10.1007/s11042-023-15441-7.

[16] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang, "A Parameter-Free Cleaning Method for SMOTE in Imbalanced Classification," vol. 7, 2019.

[17] T. Ghosh et al., "A Privacy-Preserving Federated-MobileNet for Facial Expression Detection from Images," in Applied Intelligence and Informatics, Springer, Cham, 2022, pp. 277–292. doi: 10.1007/978-3-031-24801-6_20.

[18] C. R. Kumar, S. N, M. Priyadharshini, D. G. E, and K. R. M, "Face recognition using CNN and siamese network," Measurement: Sensors, vol. 27, p. 100800, Jun. 2023, doi: 10.1016/j.measen.2023.100800.

[19] J. M. Mase, N. Leesakul, G. P. Figueredo, and M. T. Torres, "Facial identity protection using deep learning technologies: an application in affective computing," AI Ethics, vol. 3, no. 3, pp. 937–946, Aug. 2023, doi: 10.1007/s43681-022-00215-y.

[20] V. Gajarla and A. Gupta, "Emotion Detection and Sentiment Analysis of Images," 2020.

[21] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013," in Advances in Hybridization of Intelligent Methods: Models, Systems and Applications, I. Hatzilygeroudis and V. Palade, Eds., in Smart Innovation, Systems and Technologies. , Cham: Springer International Publishing, 2018, pp. 1–16. doi: 10.1007/978-3-319-66790-4_1.

[22] M. Moolchandani, S. Dwivedi, S. Nigam, and K. Gupta, "A survey on: Facial Emotion Recognition and Classification," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India: IEEE, Apr. 2021, pp. 1677–1686. doi: 10.1109/ICCMC51019.2021.9418349.