An Ensemble Kernelized-based Approach for Precise Emotion Recognition in Depressed People

Bidyutlata Sahoo
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India
bidyutlata1@gmail.com (corresponding author)

Arpita Gupta
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India
arpitagupta2993@gmail.com

Received: 21 August 2024 | Revised: 21 September 2024, 10 October 2024, and 24 October 2024 | Accepted: 26 October 2024

 DOI: https://doi.org/10.48084/etasr.8785

ABSTRACT

As the COVID-19 pandemic created serious challenges for mental health worldwide, with a noticeable increase in depression cases, it has become important to quickly and accurately assess emotional states. Facial expression recognition technology is a key tool for this task. To address this need, this study proposes a new approach to emotion recognition using the Ensemble Kernelized Learning System (EKLS). Nonverbal cues, such as facial expressions, are crucial in showing emotional states. This study uses the Extended Cohn-Kanade (CK+) dataset, which was enhanced with images and videos from the COVID-19 era related to depression. Each of these images and videos is manually labeled with the corresponding emotions, creating a strong dataset for training and testing the proposed model. Facial feature detection techniques were used along with key facial measurements to aid in emotion recognition. EKLS is a flexible machine-learning framework that combines different techniques, including Support Vector Machines (SVMs), Self-Organizing Maps (SOMs), kernel methods, Random Forest (RF), and Gradient Boosting (GB). The ensemble model was thoroughly trained and fine-tuned to ensure high accuracy and consistency. EKLS is a powerful tool for real-time emotion recognition in both images and videos, achieving an impressive accuracy of 99.82%. This study offers a practical and effective approach to emotion recognition and makes a significant contribution to the field.

Keywords-COVID-19; depression; facial emotion recognition; ensemble learning; EKLS; machine learning; mental health

I. INTRODUCTION

The COVID-19 era presented unprecedented global challenges, affecting global mental well-being in addition to physical health [1]. In particular, there was a concerning increase in depression cases worldwide, which was attributed to the pandemic's isolation, fear, and uncertainty. Recognizing the importance of addressing mental health, particularly depression, has become critical [2]. The ability to assess emotional states quickly and accurately is critical to providing timely support. Facial expression recognition technology has emerged as a powerful tool in this context [3]. Facial expressions are powerful nonverbal cues that reveal an individual's emotional state. The potential for precise facial emotion recognition systems became clear during the pandemic. These systems promise to help evaluate emotional fluctuations in people suffering from depression, allowing timely interventions. The complexities of depression often make it difficult to identify and understand emotional states. Facial emotion recognition emerges as a noninvasive, objective method for evaluating emotional responses in this context.

This study uses the Ensemble Kernelized Learning System (EKLS), a versatile and powerful machine learning framework. To improve overall performance, EKLS seamlessly integrates a variety of machine learning techniques, including Support Vector Machines (SVMs), Self-Organizing Maps (SOM), kernel methods, Random Forest (RF), and Gradient Boosting (GB). Hyperparameter tuning is a critical step in managing model complexity and margin-of-error trade-offs. Feature extraction extracts facial features from input images or video frames, which are then used to compute similarity scores. Using kernel-based representations, these scores allow the ensemble to assess the similarity between input data and

training samples. Following that, the ensemble decision rules, such as majority averaging, consolidate base model predictions to achieve a precise emotion classification. Performance evaluation is the litmus test for the ensemble's effectiveness, rigorously evaluating its accuracy and mean delay on the validation or test set. Continuous fine-tuning ensures that the ensemble's hyperparameters and settings are constantly refined, resulting in consistently improved model performance. The motivation behind choosing the EKLS over other existing models is to leverage the strengths of multiple models, enhancing the accuracy and efficiency of emotion recognition. These principles are applied by EKLS in the practical application of emotion recognition in images and videos. Using its ensemble model, it detects facial features and predicts emotions in images. Each frame in a video is processed individually and recognized emotions are displayed. This comprehensive approach ensures accurate and real-time emotion recognition, making significant progress in addressing mental health challenges during these extraordinary times.

II. RELATED WORKS

Numerous psychological studies have emphasized the critical importance of emotional well-being as an integral component of overall well-being. A large proportion of the population in both developed and developing countries now suffer from various mental illnesses. The onset of the COVID-19 pandemic has exacerbated this situation, resulting in a disproportionate number of people in need of mental health care compared to available psychiatric or psychological resources. In [4], a novel method was proposed based on Convolutional Neural Networks (CNNs) to identify age, gender, and emotions from facial images [4]. Incorporating multitask learning allows for the simultaneous classification of emotions, age, and gender, improving overall efficiency. However, this study is somewhat limited by its relatively limited investigation of architectural efficacy and emotion embedding. In [5], an emotion-embedded autoencoder was presented that successfully captured nuanced emotional attributes at a high level [5]. This approach has merit because it can be applied to a wide range of emotional tasks. However, the encoder's performance may be influenced by the complexities and variations inherent in individual emotional expressions.

In [6], the importance of the ocular region in emotion perception was examined, which led to the improvement of facial feature extraction algorithms. Although this research provides valuable insights into emotion interpretation, its immediate applicability is limited by the lack of a precise mechanism to detect facial expressions. In [7], a real-world dataset of masked facial expressions was created in response to the COVID-19 pandemic. The strength of this approach is that it incorporated various masks, colors, and facial poses for comprehensive evaluation. However, potential constraints stem from dataset quantity, diversity, and bias. In [8], the importance of facial expressions as early indicators of emotional states was emphasized, calling for more research on facial expression-based methods. Limited data availability may result in overfitting, undermining the model's ability to generalize to previously unseen instances.

In [9], SVM classification was used in conjunction with Hidden Markov Models (HMM) for the facial recognition of six distinct emotions. However, this method's reliance on handcrafted features and dimensionality reduction methods may be insufficient to combat the curse of dimensionality. In [10], an SVM classifier was proposed for emotional recognition, incorporating Haar Wavelet Transform (HWT), Gabor wavelets, and Non-Linear Principal Component Analysis (NLPCA). However, determining the optimal combination of hyperparameters can be a time-consuming and complex process. In [11], a Raspberry Pi-assisted facial expression recognition system was presented with seven distinct emotions. The addition of new emotion classes aids in the development of a more comprehensive emotion identification system. However, the SVM's limitations in dealing with increasingly complex relationships within high-dimensional data may pose difficulties. In [12], the Facial Action Coding System (FACS) framework was presented, which breaks down facial gestures into action components, influencing future systems for facial expression recognition. However, models based on FACS may be prone to overfitting, especially when trained on limited or unbalanced datasets.

In [13], residual blocks were used with two successive convolution layers to mitigate the vanishing gradient problem and facilitate deep learning-based model training. Although residual blocks provide benefits such as addressing gradient vanishing, their two-channel design may limit the extraction of intricate features required for tasks such as image classification or object detection. In [14], a spatiotemporal CNN with nested Long Short-Term Memory (LSTM) units was proposed for emotion recognition. However, the complexity of the model and computational demands may impose constraints. In [15], a CNN with sparsity batch normalization was used for non-facial expression recognition. However, the applicability of this model to facial expression recognition may be influenced by biases in the training data. In [16], CNN and LSTM architectures were used to investigate temporal and spatial variations in facial expressions. The strength of this method is its ability to capture both spatial and temporal information, making it suitable for applications requiring the analysis of facial expressions over time. However, due to its complexity, data augmentation in this context may necessitate careful consideration.

In [17], a Local Binary Pattern (LBP)-based method was proposed for depression detection, utilizing LBP to extract texture features from facial images, which were then analyzed to identify signs of depression. The method involved face detection, preprocessing for normalization, feature extraction using LBP, and classification with SVM, followed by evaluation using as accuracy, sensitivity, and specificity. However, this approach had limitations, including sensitivity to noise, variations in lighting conditions, and facial expressions, which can impact the robustness and accuracy of depression detection. In [18], the VGG model was used for depression detection, employing the CK+ dataset. This approach achieved an accuracy of 95% and a precision of 92%, demonstrating the model's efficacy in accurately identifying depression-related features in the dataset. However, the limitation of this method lies in the VGG model's computational intensity and large

memory requirements. In [19], a ResNet-based model was proposed for depression detection. This approach yielded an accuracy of 94.889%, showcasing the model's capability to handle complex patterns associated with depression. Despite its high accuracy, the method's limitation is the computational complexity of ResNet. In [20], an R-CNN model was employed for the same task, achieving an accuracy of 76.23%. This approach, while effective to some extent, displayed a comparatively lower accuracy, indicating potential difficulties in capturing intricate depression-related features. The primary limitation of this method is the R-CNN model's sensitivity to variations in input data and its relatively slow processing speed. In [21], LSTM and KNN were combined for depression detection, attaining an accuracy of 94%. This hybrid approach leveraged the strengths of both models, providing a robust solution for detecting depression. However, a significant limitation is the complexity of integrating LSTM and KNN, which can result in longer training times and increased difficulty in model tuning, potentially limiting its scalability and ease of use in different application settings [22].

TABLE I. SUMMARY OF EXISTING WORKS

Model	Architecture	Key Features	Accuracy (%)
Temporal Relational Network (TRN)	Deep learning	Temporal relations in data	92.7
CNN + SVM	Hybrid (CNN + SVM)	Integration of CNN with SVM for enhanced classification	99.69
VGG	CNN	Deep architecture for feature extraction	95.0
ResNet	Residual network	Skip connections for deeper networks	94.889
R-CNN	Region-based CNN	Focused on object detection within images	76.23
LSTM + KNN	Hybrid (LSTM + KNN)	Sequence modeling with KNN for classification	94.0

Previous approaches to emotion recognition and depression detection have shown various shortcomings. CNN-based methods frequently face challenges in adequately exploring the effectiveness of architectural design and emotion embedding. Automated encoders that incorporate emotions encounter difficulties due to the intricate and diverse nature of human emotional displays. Methods that focus on particular facial areas lack accurate mechanisms for identifying facial expressions, and datasets obtained from real-world scenarios frequently have limitations concerning the number of samples, variety, and potential biases. The scarcity of data increases the likelihood of overfitting. Handcrafted features and dimensionality reduction techniques face difficulties due to the curse of dimensionality while adjusting hyperparameters can be a time-consuming and intricate process. SVM systems face challenges in effectively processing data with a large number of dimensions. Additionally, models that rely on the FACS may become overly specialized when trained on limited or biased datasets. Residual block designs, although they tackle gradient problems, can restrict feature extraction, and intricate models such as STC-LSTM require substantial computational resources. Training data biases have a direct impact on the suitability of specific models, whereas data augmentation can be intricate. Complex models such as VGG and ResNet have challenges related to high computing demands, memory usage, and slow inference speeds. Models such as R-CNN exhibit sensitivity to changes in input and have poor processing speeds. Hybrid techniques that combine LSTM and KNN have scaling challenges due to the complexity involved in their integration.

On the other hand, the Ensemble Kernelized Learning System (EKLS) provides a strong and flexible architecture that addresses many of these shortcomings. The EKLS system combines many machine-learning approaches such as Support Vector Machines (SVMs), Self-Organizing Maps (SOMs), kernel methods, Random Forest (RF), and Gradient Boosting (GB). This broad ensemble effectively captures distinct data patterns and improves the system's ability to generalize. By employing diverse kernel types and hyperparameters, EKLS can properly handle intricate data interactions. The utilization of Self-Organizing Maps (SOM) training and the computation of the kernel matrix allow the EKLS algorithm to effectively process data with a large number of dimensions. Ensemble decision rules, such as majority voting or averaging, enhance forecast accuracy and strengthen resilience. Consistent refinement ensures that EKLS maintains its efficiency and accuracy as time progresses. Moreover, its ability to analyze both picture and video data to identify emotions enhances its adaptability.

III. PROPOSED SYSTEM

Figure 1 describes the proposed approach. The extended Cohn-Kanade (CK+) dataset [23-24], a widely recognized benchmark for emotion identification research, was used. More precisely, a subset of 590 images was used to create and evaluate the proposed depression detection method. The images were meticulously chosen to ensure a varied portrayal of facial expressions that are pertinent to this study, encompassing those that indicate sad states. To provide a strong and rigorous training and evaluation procedure, the dataset was divided into separate training and testing sets using an 80:20 split. As a result, 472 images (80%) were used for training to enable the model to acquire and apply nuanced facial indicators linked to depression, while 118 images were used specifically for testing purposes.

Each image and video has an emotion label added by hand. The labeling process involved trained annotators, consensus discussions for ambiguous cases, and inter-rater reliability tests to assess the accuracy and consistency of the emotion labels. To detect faces, a Viola-Jones object identification framework was used and facial feature detection techniques extracted facial measurements. Facial feature vectors were built into the database by integrating measurements and emotion labels. For emotion recognition, users can choose between image and video input. For emotion prediction, an ensemble learning technique with several basic models was used. The system aimed to properly recognize emotions based on facial expressions and allow user input. However, significant constraints of the ensemble approach include dataset quality, differences in facial expressions, and computational complexity, which must be addressed for an effective and practical system.

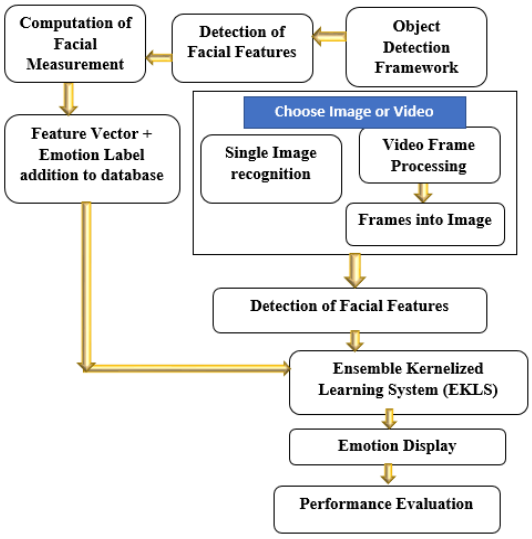

Fig. 1. Proposed block diagram.

An image can be represented as a 2D function of spatial coordinates (x, y) with intensity values:

$$Image = f(x, y) \quad (1)$$

A video is a sequence of frames indexed by time (t), each represented as a function of spatial coordinates (a, b)

$$Video = v(a, b, t) \quad (2)$$

Labels for emotions can be represented as variables or symbols, such as Depressive (D), Non-Depressive (ND), and Neutral (N). The emotion can be given as:

$$emotion = menu(Select\ emotion, D, ND, N) \quad (3)$$

To detect faces in images and videos, a Viola-Jones object detection framework with bounding boxes was used. The Viola-Jones object detection framework is a popular method for detecting objects in images, particularly faces. It efficiently computes feature responses using Haar-like features and an integral image representation. The Haar-like feature $H(a, b, t)$ and its weights $w[i]$ would operate on the video frames at spatial coordinates (a, b) and time t.

$$H(a, b, t) = \Sigma w[i] * I(a + w[i].x, b + w[i].y, t) \quad (4)$$

A bounding box is typically drawn around the detected region when an object is detected. The bounding box is defined by its top-left (x_{top}, y_{top}) and bottom-right (x_{bottom}, y_{bottom}) corners. The position and scale of the detected Haar-like feature can be used to calculate these coordinates.

$$BoundingBox: (x_{top}, y_{top}), (x_{bottom}, y_{bottom}) \quad (5)$$

The AdaBoost method is used by the Viola-Jones object detection framework to iteratively train a large number of weak classifiers, which are frequently basic decision trees based on the selected Haar-like features. The goal is to focus on the most discriminating characteristics for detecting objects in images and videos, such as faces. Within the framework, the AdaBoost training process can be represented follows.

A weak classifier, denoted as $h_t(a, b, t)$, is a straightforward decision rule based on a chosen Haar-like feature and threshold. It determines whether or not a specific region of an image or video frame contains the object of interest (e.g., a face).

$$h_t(a, b, t) = \begin{cases} +1 & \text{if } H(a, b, t) \geq Th(t) \quad \text{(Object detected)} \\ -1 & \text{otherwise} \quad \text{(No object detected)} \end{cases} \quad (6)$$

where $H(a, b, t)$ is the Haar-like feature response, and $Th(t)$ is the threshold for the t^{th} weak classifier.

At each iteration t of the training process, each training example (image patch or video frame) is assigned a weight $w_t(i)$. The importance of each example in the training set is represented by these weights. The AdaBoost algorithm finds the best combination of weak classifiers to minimize the weighted classification error. At iteration t, the objective function is:

$$\varepsilon_t = \Sigma_i w_t(i) * [h_t(a_i, b_i, t) \neq Emotion(a_i, b_i, t)] \quad (7)$$

$Emotion\ (a_i, b_i, t)$ is the true emotion label for training example i. The goal is to keep the weighted error rate ε_t as low as possible. The classification accuracy of the weak classifier $h_t(a, b, t)$ determines the weight α_t associated with it:

$$\alpha_t = 0.5 * \ln\left(\frac{(1 - \varepsilon_t)}{\varepsilon_t}\right) \quad (8)$$

The importance of the weak classifier in the final strong classifier is represented by this weight. The weights of the training examples are updated for the next iteration to give more weight to misclassified examples:

$$w_{t+1}(i) = w_t(i) * \exp\left(-\alpha_t * Emotion(a_i, b_i, t) * h_t(a_i, b_i, t)\right) \quad (9)$$

Correctly classified examples have lower weights, while misclassified examples have higher weights.

A weighted combination of the weak classifiers yields the final strong classifier:

$$H_{strong}(a, b, t) = sign\left(\Sigma t\ \alpha_t * h_t(a, b, t)\right) \quad (10)$$

which determines whether or not each region of the image or video frame contains the object of interest. The AdaBoost algorithm trains these weak classifiers iteratively before combining them into a strong classifier. When an object is detected, bounding boxes are drawn around the detected regions using this strong classifier.

Facial feature detection techniques are applied to extract the eyes, nose, mouth, etc., from the detected face regions. Then, relevant facial measurements are calculated from the detected facial features, such as eye-to-eye distance, nose width, etc. Using the coordinates and dimensions of their respective bounding boxes, the average width and height of the left and right eyes are calculated.

$$eyewidth = \frac{(lefteye(3) + righteye(3))}{2} \quad (11)$$

$$eyeheight = \frac{(lefteye(4) + righteye(4))}{2} \quad (12)$$

The indices 3 and 4 typically refer to the third and fourth elements of a data structure, which in this case would be the bounding box's dimensions (width and height).

The following facial measurements can be computed using the identified key facial features. The eye-to-eye distance is determined by calculating the Euclidean distance between the left and right eye keypoints, providing the interocular distance. The nose width is found by measuring the distance between the left and right nose keypoints, while the nose height is obtained by calculating the distance between the top and bottom nose keypoints. The eye width is measured as the distance between the outer corners of the eyes, and eye height is the distance between the top and bottom keypoints. Similarly, mouth width is the distance between the left and right mouth keypoints, and mouth height is calculated between the top and bottom mouth keypoints. The eye aspect ratio, useful for detecting whether the eyes are open or closed, is the ratio of the vertical to horizontal eye distances. The mouth aspect ratio is calculated as the ratio of the top two mouth keypoints to the bottom two, indicating the degree of mouth openness. The angle between the eyebrow keypoints can reveal facial expressions, while the pupil diameter is measured from the observed eye areas, varying with emotional states. Lastly, facial symmetry is determined by analyzing the differences between matching keypoints on both sides of the face.

Create a facial feature vector for each image and video by combining the calculated facial measurements with the corresponding emotion label. The calculated feature measurements are combined into a feature vector denoted as:

$$face\_feature = [eye\_width, eye\_height, nose\_width, nose_{height}, mouth_{width}, mouth_{height} \ldots \quad (13)$$

The computed facial feature vectors are stored along with their emotional labels in the database for future use. Now, allow the user to choose between image or video input for emotion recognition and provide it to the proposed EKLS model for emotion recognition and classification. EKLS employs an ensemble approach that combines multiple machine-learning models. SVM classifiers (SVM_k) with various kernel types (k) and parameters (C_k), SOM-based clustering and visualization, Kernelized Learning methods (KL), RF, and GB are included in the ensemble.

Each base model (M_i) in the ensemble is trained with a different kernel type and set of hyperparameters:

$$M_i = TrainModel\begin{pmatrix} Data, KernelType_i, \\ Hyperparameters_i \end{pmatrix} \quad (14)$$

Here, *Data* represents the training data, $Kernel\_Type_i$ is the kernel type for the i^{th} model, and $Hyperparameters_i$ are the hyperparameters specific to the i^{th} model.

The ensemble computes the kernel matrix (K) during SOM training to represent the similarity between data points in the feature space. The kernel matrix is built using the kernel function (K_{fn}) associated with the kernel type chosen:

$$K_{ij} = K_{fn}(Feature_i, Feature_j) \quad (15)$$

where $Feature_i$ and $Feature_j$ represent the feature vectors of data points i and j. To optimize overall performance, hyperparameter tuning involves adjusting hyperparameters. This is accomplished by resolving an optimization problem:

$$Hyperparameters_i^* = argmax\ Performance(M_i, Hyperparameters_i) \quad (16)$$

where $Performance(M_i, Hyperparameters_i)$ represents a performance metric such as accuracy. Facial features are extracted from the input image or video frames, resulting in a feature vector denoted as X.

Using kernelized representations, the ensemble computes similarity scores between the features of the input data (X) and the training samples (X_{train}):

$$Similarity_i = K_{fn}(X, X_{train}) \quad (17)$$

During feature extraction from images and videos, several challenges arise, including variations in lighting conditions, facial occlusions, and the diversity of emotional expressions among different individuals. These factors complicate the extraction of consistent and reliable features. To address these challenges, preprocessing techniques were used such as normalization and data augmentation, which helped improve the robustness of the extracted features, allowing for better model performance in diverse conditions.

To combine the predictions of the base models for accurate emotion classification, ensemble decision rules such as majority voting or averaging are used:

$$Emotion_{Prediction} = Ensemble_{Decision_{Rule}}(Predictions_{M1}, Predictions_{M2}, \ldots, Predictions_{Mn}) \quad (18)$$

where $Predictions_{Mi}$ represent the individual predictions of each base model M_i. The ensemble's performance was evaluated using metrics on the validation or test set, including accuracy and mean delay.

$$Accuracy(Emotion_{Prediction}, True_{Labels}) \quad (19)$$

$$MD = Mean\_Delay(Emotion_{Prediction}, True_{Labels}) \quad (20)$$

The ensemble continuously optimizes its hyperparameters and other relevant settings to improve model performance as provided earlier in (16). Extracting facial features that capture emotional expressions is required to recognize emotions in images. EKLS uses these features to determine the emotion in the image.

When it comes to videos, each frame is treated as a separate image, with facial features extracted. To track emotions throughout the video, similarity scores are calculated using training data. EKLS is then used to predict emotions in real-time for each frame, providing insights into how emotions evolve in the video. Performance metrics assess the system's accuracy and consistency in recognizing emotions in videos.

A. The Proposed Ensemble Kernelized Learning System (EKLS)

EKLS is a powerful and adaptable machine-learning framework that revolutionizes classification and regression tasks by seamlessly combining a variety of machine-learning techniques. Emotion recognition with EKLS can be broken down into several key steps, as shown in Figure 2.

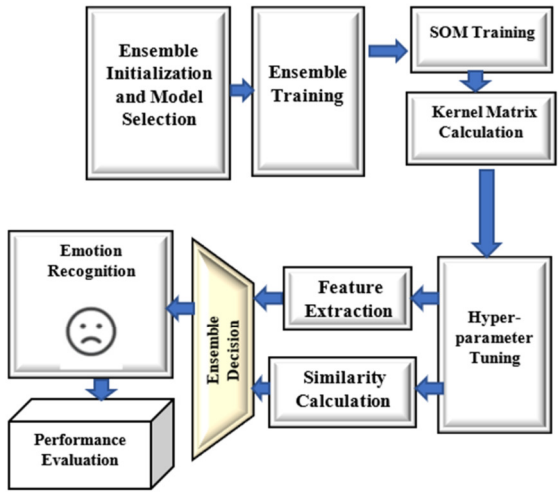

Fig. 2. Proposed EKLS architecture.

The journey begins with ensemble initialization and model selection, which brings together a variety of models such as SVM classifiers with various kernel types, SOM-based clustering, and ensemble methods. Following that, these models are individually trained with different kernel types and hyperparameters during ensemble training, fostering diversity and allowing the capture of unique data patterns. SOM training and kernel matrix calculation improve the process even further by calculating a kernel matrix that represents the similarity of data points in the feature space, leveraging the power of SOMs.

Hyperparameter tuning is critical because it aids in managing model complexity and margin-of-error trade-offs by fine-tuning parameters such as *C* for SVMs. Feature extraction meticulously extracts facial features from input images or video frames. These features are used to compute similarity scores, which allow the ensemble to assess the similarity of input data and training samples using kernelized representations. Following that, ensemble decision rules, such as majority voting or averaging, are used to consolidate the base models' predictions for precise emotion classification.

The EKLS ensemble employs similar principles in the practical application of emotion recognition in both images and videos. For images, it detects facial features and predicts emotions using its ensemble model, whereas for videos, each frame is processed individually. Recognized emotions are displayed for each frame, and a comprehensive performance evaluation is carried out, taking critical metrics such as accuracy and mean delay into account. EKLS integrates multiple models, each contributing distinct advantages to the overall system. For instance, CNNs excel at capturing spatial hierarchies in image data, while LSTMs are adept at recognizing temporal patterns in sequential data. By combining these models, EKLS leverages their complementary strengths: CNNs enhance feature extraction from facial images, while LSTMs analyze the temporal dynamics of emotional expressions over time.

Figure 3 illustrates the interaction between the various components of the ensemble, detailing how each model contributes to the overall emotion recognition process. EKLS creates an ensemble of base models with various kernel types and hyperparameter sets. This ensemble diversity is critical for capturing unique patterns in the data and improving the model's generalization ability. During SOM training, EKLS calculates a kernel matrix to represent the similarity between data points in the feature space. This kernelized representation can capture complex data relationships and is particularly useful when dealing with high-dimensional data.

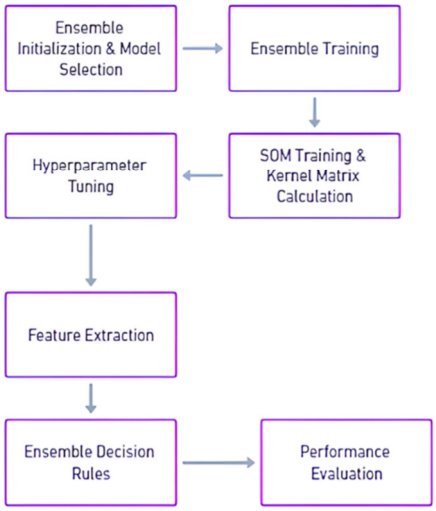

Fig. 3. Flowchart for EKLS architecture.

B. Algorithm of the Proposed Model

The EKLS algorithm is a robust framework that produces accurate emotion recognition results. Algorithm 1 describes how it works.

```
Algorithm: Ensemble Kernelized Learning
System (EKLS)
1:  Initialize ensemble models as an empty
    list.
2:  For each base model in base models:
      Train the model based on the type
      (e.g., SVM, SOM, etc.).
      Append the trained model to
      ensemble_models.
3:  For each model in ensemble_models:
3:    Fine-tune its hyperparameters.
4:  Extract facial_features from
    input_image_or_video_frames.
5:  For each model in ensemble_models:
      Calculate similarity_scores between
```

```
      facial_features and the model.
6:  Initialize ensemble_predictions as an
    empty list.
    For each model in ensemble_models:
      Predict emotion based on
      similarity_scores.
      Append the emotion prediction to
      ensemble_predictions.
7:  Combine individual model predictions
    using majority voting to get the final
    emotion prediction.
8:  Display the recognized emotion.
9:  Evaluate performance by comparing the
    final prediction to ground truth
    labels.
10: If continuous fine-tuning is needed:
      For each model in ensemble_models:
        Further fine-tune hyperparameters
        based on performance.
```

IV. RESULTS AND ANALYSIS

MATLAB was used for feature extraction, leveraging its in-built Computer Vision Toolbox and Machine Learning Toolbox. The experiments were carried out using MATLAB R2022b software on a personal computer with an Intel Core i5 processor. This setup provided a reliable computational environment for simulating and analyzing the data, ensuring smooth performance and accurate results during the experimental phase. When the user selects a video, the system prompts them to enter the number of frames they want to process. These frames are then recorded and saved for later analysis. This step is critical because it converts the video into discrete images for further processing, as shown in Figures 4 (Subject S130 as input) and 5.

Following the capture and storage of the frames, the system displays a dialog box to the user, allowing him to manually label each image with emotions such as Depressive (D), Non-Depressive (ND), or Neutral (N), as shown in Figure 6. This manual labeling creates the ground-truth emotional annotations for each frame, which are used for training and evaluation.

Fig. 4. Video into 34 frames (subject S130).

Fig. 5. Video into 36 frames.

Fig. 6. Emotion selection.

The next step is to start detecting facial components in each frame, as shown in Figure 7. This detection process aims to recognize key facial features such as eyes, nose, mouth, and others. This detection may be limited or incomplete at first.

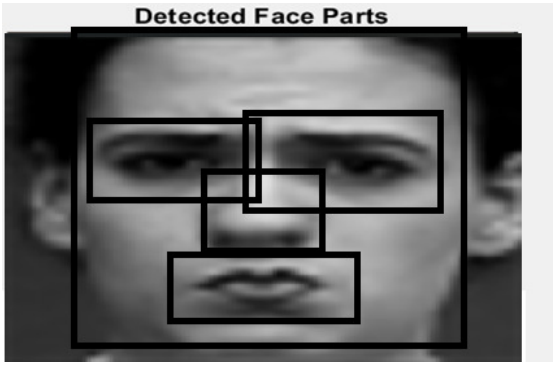

Fig. 7. Initialization of detected face parts frame-wise.

As shown in Figure 8, the system iteratively improves its detection of facial components across 36 frames during processing. For this purpose, the Viola-Jones object detection framework is used, which makes use of Haar-like features and integral image representations. For reference, bounding boxes are drawn around detected facial regions.

As the system detects facial components, it computes various facial measurements and features, as shown in Figure 9. Eye width, eye height, nose width, mouth width, and other relevant attributes are measured. These measurements are derived from the coordinates and dimensions of the bounding boxes surrounding detected facial components, as in (11, 12).

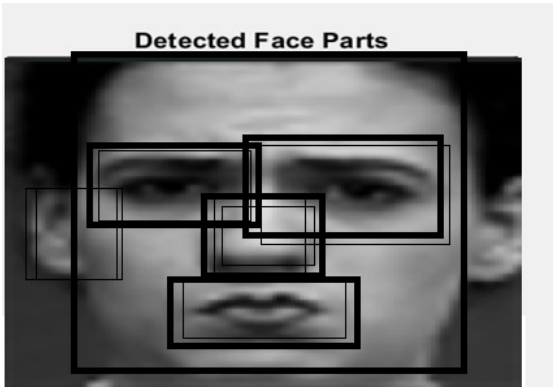

Fig. 8. Complete detection of face parts after 36 frames.

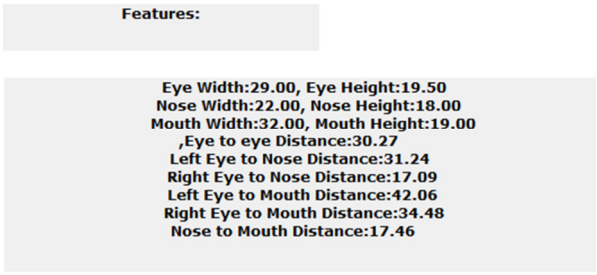

Fig. 9. Face features during the initialization of face detection.

However, Figure 10 depicts a critical stage in the evaluation process, where a subtle variation is introduced. The system maintains a focus on specific facial measurements that are particularly indicative of emotional expressions during the evaluation of facial emotion. These measurements include eye height and width, among others. Although these measurements are similar to the initialization stage (Figure 8), they may differ slightly for each frame because emotions are dynamic and can result in changing facial features. The eye height and eye width are two critical facial measurements that are closely examined during this phase. These measurements capture differences in the size and shape of the eyes, which can be very useful in determining emotional states. Although Figure 9 provided an initial snapshot of these measurements, Figure 10 shows how they can vary slightly from frame to frame as the system continuously evaluates and adapts to evolving emotional expressions. These facial measurements and features are critical in characterizing the emotional expressions in each frame. They provide useful information about how various emotional states manifest in facial expressions. The system then uses the calculated facial features and measurements to recognize emotions within each frame using EKLS and predict the emotional state expressed in each frame, resulting in an emotional label for each frame.

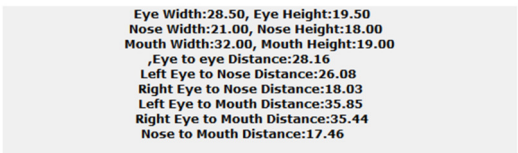

Fig. 10. Face features during the evaluation of facial emotion.

A. *Mean Delay*

Table I shows the mean delay. This metric is an important performance indicator for the emotion prediction system considered. The system's exceptional efficiency in rapidly predicting emotions based on input data, such as facial expressions, is highlighted by the system's impressively low mean delay of 0.3270 seconds. A low mean delay is critical in practical applications, particularly those involving human-computer interaction and real-time emotion recognition. It means faster responses, allowing for more natural and responsive interactions between users and systems.

TABLE II. MEAN DELAY METRIC ATTAINED DURING EVALUATION PER VARIOUS FRAMES

SNO	Frames	Mean delay(s)
1	1	0.0125
2	6	0.113
3	12	0.145
4	18	0.212
5	24	0.259
6	30	0.298
7	36	0.327

Figure 11 shows a mean delay plot, emphasizing the importance of observing the mean delay across different frames and shedding light on how the system behaves under various conditions. While the plot demonstrates that the mean delay does not increase in a strictly linear fashion as the number of frames increases, it emphasizes the system's consistency and reliability. Despite any fluctuations or nonlinear trends observed, the system consistently maintains a final mean delay of 0.3270 s.

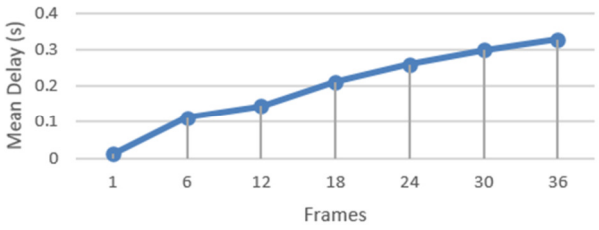

Fig. 11. Mean delay plot for various frames as provided in Table I.

B. *Accuracy*

Table II provides a detailed assessment of the accuracy performance of several techniques employed for the detection of emotions and depression, including TRN [8], CNN with SVM [4], VGG [18], ResNet [19], R-CNN [20], LSTM with KNN [21], and the proposed method. Accuracy results vary from 76.23% using the R-CNN model [20] to 99.82% using the proposed technique. In [4], the CNN+SVM technique reached a remarkable accuracy of 99.69%. The VGG [18] and ResNet [19] models demonstrated impressive accuracy rates of 95% and 94.889% respectively, indicating their strong ability to effectively process intricate data patterns. The R-CNN model [20], while successful, had a slightly lower accuracy of 76.23%, suggesting possible difficulties in capturing complex depression-related characteristics. The combination of LSTM and KNN [21] achieved an accuracy of 94%. Figure 12 allows

for a direct comparison of the accuracy performance among these models. The proposed method distinguishes itself with a remarkable accuracy of 99.82%, highlighting its better performance compared to the other techniques.

TABLE III. ACCURACY PERFORMANCE EVALUATION

Techniques used	Accuracy (%)
TRN [8]	92.7
CNN+SVM [4]	99.69
VGG [18]	95
ResNet [19]	94.889
R-CNN [20]	76.23
LSTM+KNN	94
Proposed method	99.82

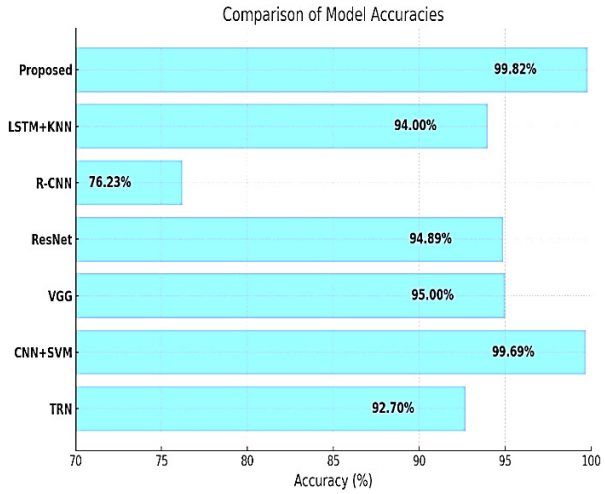

Fig. 12. Accuracy comparison.

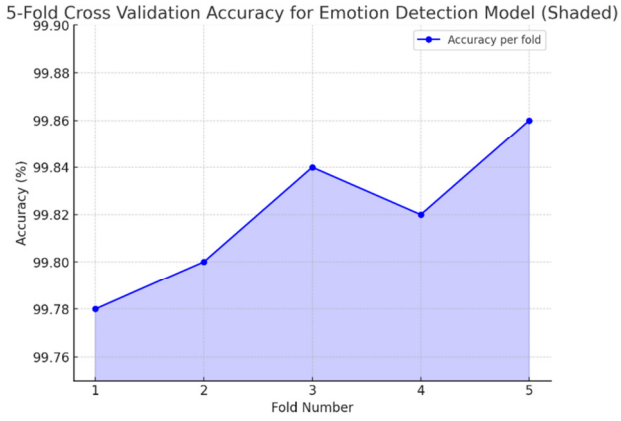

Fig. 13. 5-fold cross-validation accuracy.

Figure 13 represents the five-fold cross-validation accuracy for the proposed emotion detection model, with accuracies ranging from 99.78% to 99.86%. Each fold's performance is marked individually, showing slight variations, but overall the model performs consistently well across all folds. The shaded region under the curve highlights the minimal differences in accuracy between folds, indicating robust generalization with a nearly perfect performance of 99.82% on average. This level of accuracy demonstrates the model's strong capability to detect emotions across different validation sets, confirming its reliability and consistency.

V. CONCLUSION AND SUMMARY

This study presented a solid and novel approach to address the pressing issue of depression during the post-COVID-19 era. The proposed EKLS model for emotion recognition is an effective tool for determining emotional states. The model demonstrates superior performance in real-time emotion recognition, with an impressive accuracy rate of 99.82% and a mean delay of 0.3270 seconds, making it a valuable asset for mental health practitioners. The use of EKLS is not limited to depression, as it can be adapted and extended to address a wide range of mental health disorders, providing tailored solutions for a wider range of emotional difficulties. Collaborations with mental health professionals and institutions can help with the practical implementation of the model in clinical settings, ensuring that it reaches those who need it. The EKLS framework is designed to be scalable, allowing it to adapt to different contexts beyond the COVID-19 pandemic. Its modular architecture facilitates the integration of additional data sources, including real-time video feeds and other physiological signals. This adaptability makes EKLS suitable for applications in mental health monitoring, customer sentiment analysis, and even security systems to detect suspicious behaviors through emotion recognition.

This study paves the way for cutting-edge technology to play a pivotal role in addressing emotional well-being during difficult times in the ever-changing landscape of mental health support. Integration of advanced machine learning techniques and real-time applications has the potential to revolutionize the field of mental health support and emotional well-being assessment. Potential applications of the EKLS model include its use in telehealth platforms for real-time monitoring of patient emotional states, development of emotion-aware AI systems in customer service environments to enhance user experience, and implementation of educational technologies to assess student engagement and emotional well-being. Additionally, EKLS can be utilized in virtual reality environments to create immersive and emotionally responsive experiences.

ACKNOWLEDGMENT

The author would like to thank Arpitha Gupta for her constant support and feedback. Her remarks have made a significant contribution to this work.

REFERENCES

[1] A. Gupta, V. Jain, and A. Singh, "Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications," *New Generation Computing*, vol. 40, no. 4, pp. 987–1007, Dec. 2022, https://doi.org/10.1007/s00354-021-00144-0.

[2] H. M. Al-Dabbas, R. A. Azeez, and A. E. Ali, "Two Proposed Models for Face Recognition: Achieving High Accuracy and Speed with Artificial Intelligence," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13706–13713, Apr. 2024, https://doi.org/10.48084/etasr.7002.

[3] R. Kumar, S. Mukherjee, T. M. Choi, and L. Dhamotharan, "Mining voices from self-expressed messages on social-media: Diagnostics of

mental distress during COVID-19," *Decision Support Systems*, vol. 162, Nov. 2022, Art. no. 113792, https://doi.org/10.1016/j.dss.2022.113792.

[4] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, "An efficient deep learning technique for facial emotion recognition," *Multimedia Tools and Applications*, vol. 81, no. 2, pp. 1649–1683, Jan. 2022, https://doi.org/10.1007/s11042-021-11298-w.

[5] C. Zhang and L. Xue, "Autoencoder With Emotion Embedding for Speech Emotion Recognition," *IEEE Access*, vol. 9, pp. 51231–51241, 2021, https://doi.org/10.1109/ACCESS.2021.3069818.

[6] V. Ramachandra and H. Longacre, "Unmasking the psychology of recognizing emotions of people wearing masks: The role of empathizing, systemizing, and autistic traits," *Personality and Individual Differences*, vol. 185, Feb. 2022, Art. no. 111249, https://doi.org/10.1016/j.paid.2021.111249.

[7] B. Yang, J. Wu, and G. Hattori, "Facial expression recognition with the advent of human beings all behind face masks MUM2020," in *Proceedings of the 2020 ACM International Conference on Multimedia (MUM2020)*, 2020.

[8] A. Pise, H. Vadapalli, and I. Sanders, "Facial emotion recognition using temporal relational network: an application to E-learning," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 26633–26653, Aug. 2022, https://doi.org/10.1007/s11042-020-10133-y.

[9] S. Varma, M. Shinde, and S. S. Chavan, "Analysis of PCA and LDA Features for Facial Expression Recognition Using SVM and HMM Classifiers," in *Techno-Societal 2018*, 2020, pp. 109–119, https://doi.org/10.1007/978-3-030-16848-3_11.

[10] C. V. R. Reddy, U. S. Reddy, and K. V. K. Kishore, "Facial Emotion Recognition Using NLPCA and SVM," *Traitement du Signal*, vol. 36, no. 1, pp. 13–22, Apr. 2019, https://doi.org/10.18280/ts.360102.

[11] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services," *Information Sciences*, vol. 479, pp. 416–431, Apr. 2019, https://doi.org/10.1016/j.ins.2018.07.027.

[12] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, Apr. 2021, https://doi.org/10.1109/TAFFC.2018.2890471.

[13] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognition Letters*, vol. 120, pp. 69–74, Apr. 2019, https://doi.org/10.1016/j.patrec.2019.01.008.

[14] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018, https://doi.org/10.1016/j.neucom.2018.07.028.

[15] J. Cai, O. Chang, X.-L. Tang, C. Xue, and C. Wei, "Facial Expression Recognition Method Based on Sparse Batch Normalization CNN," in *2018 37th Chinese Control Conference (CCC)*, Wuhan, China, Jul. 2018, pp. 9608–9613, https://doi.org/10.23919/ChiCC.2018.8483567.

[16] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, Apr. 2019, https://doi.org/10.1109/TAFFC.2017.2695999.

[17] S. J. Park, B. G. Kim, and N. Chilamkurti, "A Robust Facial Expression Recognition Algorithm Based on Multi-Rate Feature Fusion Scheme," *Sensors*, vol. 21, no. 21, Jan. 2021, Art. no. 6954, https://doi.org/10.3390/s21216954.

[18] S. A. Hussein, A. E. R. S. Bayoumi, and A. M. Soliman, "Automated detection of human mental disorder," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Feb. 2023, Art. no. 9, https://doi.org/10.1186/s43067-023-00076-3.

[19] T. D. Pham, M. T. Duong, Q. T. Ho, S. Lee, and M. C. Hong, "CNN-Based Facial Expression Recognition with Simultaneous Consideration of Inter-Class and Intra-Class Variations," *Sensors*, vol. 23, no. 24, Jan. 2023, Art. no. 9658, https://doi.org/10.3390/s23249658.

[20] S. Kanjanawattana, P. Kittichaiwatthana, K. Srivisut, and P. Praneetpholkrang, "Deep Learning-Based Emotion Recognition through Facial Expressions," *Journal of Image and Graphics*, pp. 140–145, Jun. 2023, https://doi.org/10.18178/joig.11.2.140-145.

[21] D. Hebri, R. Nuthakki, A. K. Digal, K. G. S. Venkatesan, S. Chawla, and C. R. Reddy, "Effective Facial Expression Recognition System Using Machine Learning," *EAI Endorsed Transactions on Internet of Things*, vol. 10, Mar. 2024, https://doi.org/10.4108/eetiot.5362.

[22] A. B. Miled, M. A. Elhossiny, M. A. I. Elghazawy, A. F. A. Mahmoud, and F. A. Abdalla, "Enhanced Chaos Game Optimization for Multilevel Image Thresholding through Fitness Distance Balance Mechanism," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 14945–14955, Aug. 2024, https://doi.org/10.48084/etasr.7713.

[23] T. Kanade, J. F. Cohn, and Yingli Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, Grenoble, France, 2000, pp. 46–53, https://doi.org/10.1109/AFGR.2000.840611.

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101, https://doi.org/10.1109/CVPRW.2010.5543262.

AUTHORS PROFILE

Bidyutlata Sahoo earned her B.Sc. in Computer Science from Ramadevi Women's University in 2002 and her M.Sc. in Computer Science from Utkal University in 2004. She expanded her education with an MTech in Computer Science from IETE, Hyderabad. Currently, she is pursuing her Ph.D. in Computer Science and Engineering at KL University, Hyderabad. Her research interests include deep learning, machine learning, artificial intelligence, and computer vision.

Dr. Arpita Gupta received her Ph.D. from NIT, Tiruchirappalli, in transfer learning and her M.Tech. from IIIT, Trichy. She is an Associate Professor and HOD in the Department of Computer Science and Engineering, K L Deemed to be University, Hyderabad Aziz Nagar Campus. Her research work has been published in numerous peer-reviewed journals. She also has been an active reviewer for many journals. Her research interests include deep learning, machine learning, artificial intelligence, computer vision, and networks.