IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 June 2024, accepted 5 July 2024, date of publication 15 July 2024, date of current version 24 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428336

RESEARCH ARTICLE

Triangular Region Cut-Mix Augmentation Algorithm-Based Speech Emotion Recognition System With Transfer Learning Approach

V. PREETHI AND V. ELIZABETH JESI
Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India

Corresponding authors: V. Preethi (pv9600@srmist.edu.in) and V. Elizabeth Jesi (jesiv@srmist.edu.in)

ABSTRACT Recently, spectrogram energy patterns that capture emotional information have demonstrated strong performance in the vocal-based image emotion detection challenge. The proposed augmentation technique, called Triangular Region cut-mix, is a novel system for emotion recognition. It utilizes voice image information to enhance classification accuracy by focusing on triangular regions instead of box regions, while minimizing information loss. This study utilizes an innovative approach that incorporates a triangular area to enhance the cutting or mixing of the input images, while preserving the information in order to create additional training examples. The dearth of information to enhance the accuracy of speech emotion recognition is therefore mitigated. In order to increase the amount of training data and enhance the precision of voice emotion recognition, a vanilla gradient technique is employed. The pitch attribution demonstrates the significance of a pixel to the human visual system. In contrast, transfer learning results in superior performance. Previous studies have not identified a model that achieves good performance in voice image emotion identification while using triangle region augmentation without sacrificing information. This limitation has been observed in earlier works. Constructing a proficient model for automatic emotion recognition is challenging in the absence of annotated data. We utilize raw, labeled audio data from kaggle's Ravdess dataset. Initially, we convert this data into a spectrogram, which serves as a representation of the audio image. We then apply image classification algorithms to classify the emotion. Additionally, we employ triangular region augmentation to expand the labeled training data. We conduct an assessment and evaluation of two distinct methodologies: 1) Transfer learning without augmentation; 2) Transfer learning with triangular region augmentation. We utilize a pre-trained VGG16 model that has undergone pre-training for image classification. Our model achieves an accuracy of 84.2% in detecting emotions in speech images. Experimental results demonstrates that the proposed system has achieved a 5.6% increase in accuracy compared to the baseline model without augmentation.

INDEX TERMS CNN, labelled training data, pre-trained model, Ravdess dataset, speech emotion recognition, tri-cut, tri-mix, triangular region augmentation, transfer learning, VGG16.

I. INTRODUCTION

Is it possible to convey one's emotions to someone through means other than verbal communication? There are a multitude of communication alternatives. Two prominent

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

viewpoints are visual perception and observation. In addition, discourse is a crucial tool that enables us to communicate with others while also understanding their profound emotions. Emotions strongly influence the vocal equilibrium during speech, potentially revealing hidden emotions within oneself. In contemporary times, the prevalence of heightened stress levels resulting from professional obligations or duties has become a prominent concern. Profound behavior reflects an individual's mental condition, which needs to be identified and addressed to prevent potentially severe consequences. In this paper, we aim to present a sentiment recognition system that utilizes voice-converted images.

The forefront of machine-based learning models for discourse-related objectives includes tasks such as sentimental analysis [1], discourse identification [2], and voice modulation [3]. Regrettably, some tasks lack sufficient quantities of substantial and authentic data. The discourse-related work acknowledges the negative impact of the model on performance and its limited ability to adapt to new knowledge as a drawback that needs to be addressed. To prevent model overfitting, one can employ techniques that utilize a larger amount of data [4]. The neural network's main concerns are the analysis of the input data and the model's ability to generalize. Increasing the amount of information is a common practice to reduce overfitting and enhance the predictive capabilities of the learning model. The emotional arrangement calculation examines and compares the effect of triangular area enlargement with or without expansion. Audio categorization of images is a vital subject that garners significant focus in the realm of audio image analysis.

Spectrograms with varying levels of granularity in SER display distinct energy patterns [5]. To identify the emotional tone in these energy patterns, three methods can be employed: employing an audio signal [6], analyzing the speech spectrogram using a structure called a grid or group [3], and considering the context [7]. As a means of contributing to our group, we make use of a spectrogram that converts a sound stream into a complete image. In order to enhance the accuracy of the model, we incorporate pitch attribution into a CNN model that focuses solely on the keep augment cut mix approach [8]. A spectrogram, commonly referred to as a Log-Mel spectrum, is a graphical representation of sound that includes additional information extracted from the original sound waveform. The Mel spectrogram is a graphical representation that shows the frequency content of a signal over time. These are the works that remained as motivation for this work.

For the purpose of enhancing the recognition and implementation of emotions in discourse, this review implements an area-level expansion of the discourse picture. Sound signs are time series that consist of pitch and rhythm. The range holds significant importance in both horizontal and vertical directions within the picture. Therefore, the process of expanding information is specifically targeted towards the distinct features of the vocal image, which differ from the speech signal, in order to gain a better understanding of the image and develop a more appropriate emotion recognition system. Two recently suggested methods for addressing visual impairments in PC vision are extraction [9] and irregular elimination [8]. By randomly extracting a triangular section from the image and comparing it to a threshold value, this improves the precision of picture classification.

We can verify our commitments by comparing our plan to our regular, regulated benchmarks. In the discourse sentiment evaluation technique, we achieve an accuracy of 78.6 percent using No-Aug, and an accuracy of 84.2 percent after augmentation using a pre-trained model. Here are some of our notable contributions and findings:

- To demonstrate, using the mAP@3 measure, that the Tri-augment outperforms the previous work in speech image emotion recognition,
- In order to obtain a meaningful score, the RGB procedure is used, which identifies the region containing the important data.
- Transfer learning is employed to consolidate information and enhance the rate at which our simulation recognizes emotions.

We also give a theoretical assessment of Tri-aug, which improves the performance of CNN architecture, the VGG16 model, and the role of state-of-the-art (SOTA) extension in classifying the sentiment of named data in discourse.

II. BACKGROUND AND MOTIVATION

Hataya et al. have proposed a differentiable data augmentation pipeline that uses adversarial learning to shorten the time intervals among the augmented data variations and the original data, allowing for faster policy search and proximity gradients for multiple discrete-parameter transformation operations [18]. A simple and successful assessment technique was developed based on AWS, thanks to the model's enhanced training dynamics Tian et al. [19].

Scientists examined the functioning of Auto-Augment and found that it has the potential to eliminate certain discriminatory knowledge acquired through the training image. Knowledge distillation, also known as a computer's output, is a method used to reduce the inaccuracy of supervision. It is a network training approach that requires the presence of a teacher, as described by Wei et al. [28]. The effectiveness of data augmentation techniques, specifically Auto Augment Flip and Crop, and Rand augment Flip and Crop, has been investigated for training recognition models in audio classification. Consequently, the utilization of particular data augmentation methods may enhance the ability of detection models to generalize more effectively as in Zoph et al. [9]. Data augmentation techniques have been shown to induce distribution shifts, leading to a decrease in inference performance on unenhanced data. The KeepAugment technology was proposed as a straightforward and efficient solution to this problem, utilizing a saliency map. KeepAugment enhances the uniformity of augmented images by utilizing a map of salient features. Gong et al. [8]. Instead of individually determining the size and probability of each operation,

they found that looking for a single distorted magnitude that impacts all operations was enough. Consequently, they implemented a narrower search field, leading to a significant reduction in the size of the resulting search space [16]. Cubuk et al.

The author of this paper examines the effective handling of noisy unlabeled samples, emphasizing the importance of the quality of noise, specifically the noise produced by modern data augmentation techniques like Rand Augment, in semi-supervised learning. The author asserts that their approach to self-training a noisy learner is more effective [33]. Xie et al. [33]. The author's solution to the problem of self-supervised voice identification relies on having access to inputs and results from speech applications, as well as word transcripts for the source domain speech. This paper discusses this approach. Through self-supervised training of a variation auto encoder on both input and application-oriented output data, they developed novel augmentation-based techniques for modifying speech without compromising transcripts and acquiring a latent model of speech for labeled data utilized for training Hsu et al. [2].

In this work, Voila and Jone's method of turning the picture into an integral image and performing feature extraction at multiple region levels is used by Proença et al. [12]. Two approaches, personal and impersonal, were employed to categorize acoustic data into distinct categories, such as energy-related or spectral-related. Each category was considered a separate perspective for recognizing emotions from speech, Zhang et al. [15]. The techniques of random erasing, Devries et al. [22], and cut out, as in Dark et al, are novel approaches to addressing occlusion in visual computing. As demonstrated in the subsequent illustration; in order to enhance the precision of image classification, a box region is arbitrarily eliminated from the source data. This enables the collection of nodes to focus on the complete source rather than a subset thereof. The methods all un-change the source's name, despite the differences in how they are processed.

Dang et al. [54] presented a unique data augmentation (DA) strategy, EMix for the SER problem. This method is straightforward but efficient. For the purpose of producing new data, the procedure involves combining pairs of selected samples taken from the initial data. When compared to their constructive counterparts, the mixtures that are formed will be more noisy or less unclear. A transformer-based network for the SER task was created and conducted experiments using two available datasets, namely IEMOCAP and Crema-D. Atmaja and Sasou [55] Glottal source extraction and silence removal are the two augmentation techniques that have been utilized for efficient speech emotion recognition in speaker-independent data.

Gong et al. [8] utilized the saliency map to identify significant regions on the initial images, to ensure that these informative regions were preserved during the augmentation process. Proença et al. [12] presented in this study were driven by this strategy, which served as the foundation for our work. In the beginning, we offer the idea of ''triangular integral feature'' to describe the properties of the triangular region. This idea is based on the concept of ''triangle mesh'' that is used in computer graphics. These works motivated to bring in triangular region augmentation to preserve more information while performing augmentation cutmix to overcome the limitations of data and transfer learning to generalize the model.

III. METHODOLOGY

A. DATA PRE-PROCESSING

The early action of information consists of two parts. The audio signal undergoes a process where the unspoken regions are eliminated, while the relevant sections are preserved, as an initial step. Improving the signal-to-noise ratio is the focus of the second stage. Power spectral density is utilized to convert time series frequency into intensity values. The sound signal is converted into a visual image and utilized as the input, eliminating the requirement for any supplementary feature extraction. Expert qualities like the Mel-frequency cepstral coefficients (MFCC) and Log-Mel spectrograms are often acquired as the main process. Numerous studies have shown that the Log-Mel spectral produces better results when used as an input for a convolutional neural network [10], and the body of literature supports this conclusion.

B. DATA AUGMENTATION

Data augmentation technology enhances algorithm performance by maintaining a consistent data distribution and minimizing variation [11]. When discussing data augmentation, our focus is on developing a system of categorization that is more efficient and better able to comprehend audio images. Data augmentation is a technique used to regularize the input of a deep neural network. It involves two types of operations: perturbing existing samples to create new samples, and intentionally expanding the dataset by adding extra samples to the original dataset. Alternative methods involve the fragmentation and blending of samples, but this does not lead to a substantial augmentation in the quantity of samples [10]. Fig.1 illustrates the frequency spectrum of speech associated with the emotion of happiness.

1) TRIANGULAR REGION AUGMENTATION (TRI AUG)

Providing a solitary input image is not an issue, augmenting the quantity of source input permits the generation of multiple input sources denoted by $z = D(z)$, where D signifies the precise source name and typically a pattern of random probability [8]. This method is employed for detecting both informative and non-informative areas. The main foundation of this detection system relies on utilizing essential information present within the designated region to identify abnormalities. The regions of an image that contain the least and most information can be identified by analyzing the variations in contrast between groups of pixels. The calculation of triangular-like regions in this context involves the utilization of integral images. The total number of pixels in a triangle can be determined by simultaneously analyzing the three values

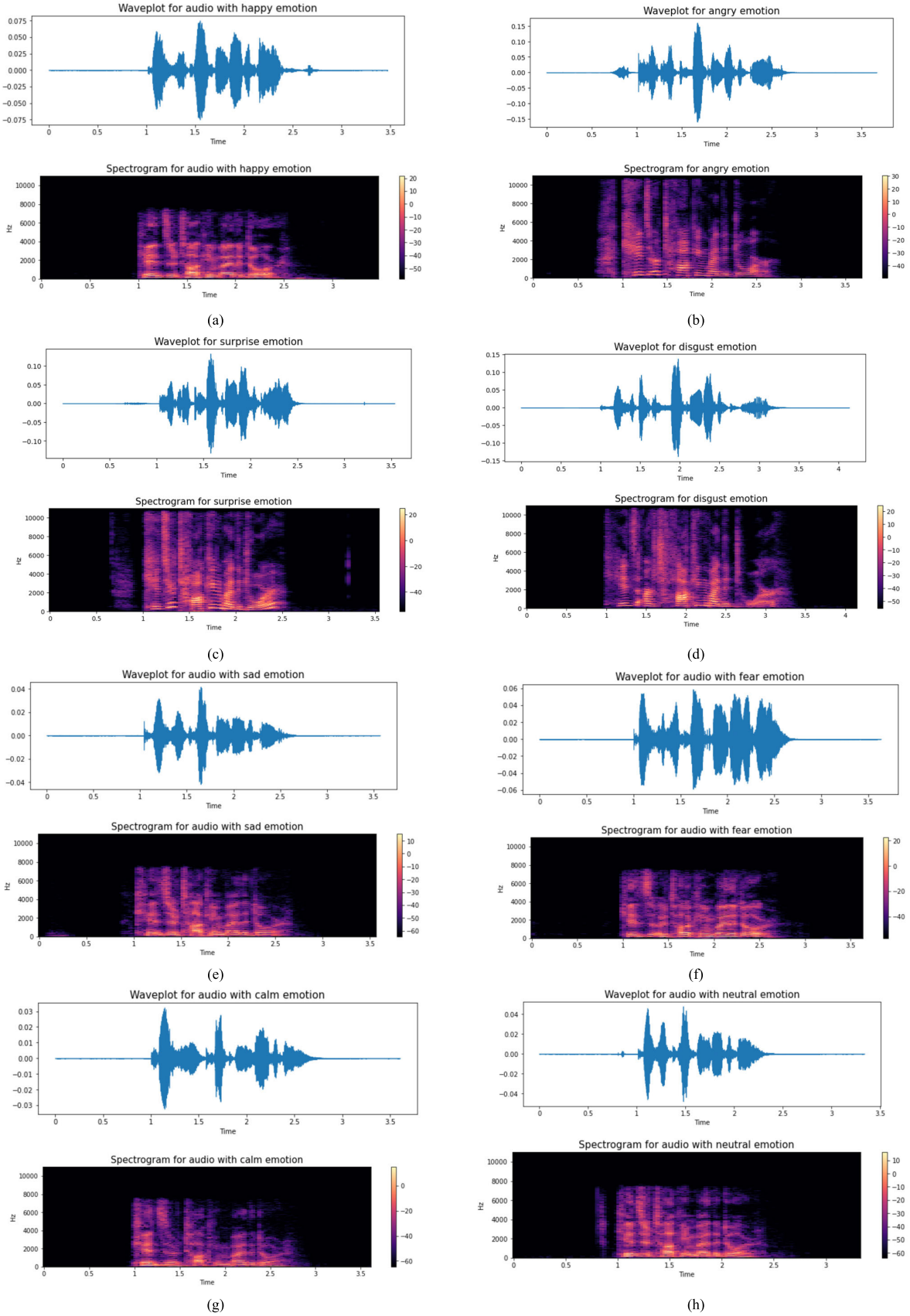


FIGURE 1. (a) – (h) Shows spectrogram images and wave plot of different emotions of Ravdess dataset.

present at each of the triangle's corners. The significance of every individual pixel in the composite picture is computed by the cumulative sum of the pixels preceding and above it in the source image.

Fig. 2 displays a depiction of a triangle where the sum of all adjacent pixels is included [12]. Region A is the region where, for all orientation the sum is calculated and chooses with minimal average intensity to be cut. Image assimilation is a versatile technique used to identify areas of relevance within an image and can be employed at various levels of detail. Integration, at its fundamental level, can be comprehended as the amalgamation of multiple distinct entities through the process of addition. The cumulative sum of the areas surrounding a given image, both vertically and diagonally, constitutes the integral image for each area. The overall image content is calculated using a mathematical algorithm carried out on a per-area basis.

Per the theory, triangular features in two-dimensional space are better able to represent emotional data than one-dimensional rectangular features [12]. Alternatively, we are optimistic that our diverse regions can work together to improve the classifier's ability to differentiate. The computation methodology and augmentation of four distinct orientation sections have been explained in detail below. This will be achieved using a single computational shot, which is essential for maintaining our ability to operate with real-time data.

The primary objective of this strategy's conceptual framework is the notion that a complete, integral visual can be recreated from a single scan of the original image. The integral image, denoted as ii, shares similar characteristics with image i.

It is computed by summing up the power of the areas located above and below each individual area. This is performed subsequent to the creation of an intensity image, denoted as i, using the parameters W and H. Subsequently, the resulting integral image ii is of dimensions W and H (x, y). Furthermore, it is important to highlight the method's resistance to changes in scale. The reason for this is that the computational complexity of calculating the size of big regions is equivalent to that of calculating the area of smaller regions. When considering the ability to perform the processing of information in real time, it is evident that this is a highly significant characteristic it possesses.

$$\mathrm{ii}(x,y) = \sum_{c=1}^{x} \sum_{r=1}^{y} i(c, r) \tag{1}$$

$$\mathrm{S}(x,y) = \mathrm{s}(x, y-1) + \mathrm{i}(x, y) \tag{2}$$

$$\mathrm{ii}(x, y) = \mathrm{ii}(x-1, y) + \mathrm{s}(x, y) \tag{3}$$

Next, the region with the least amount of information is identified and eliminated in a triangular shape. This approach is used to prevent the removal of additional information regions from the rectangular image when applying the random erasing technique [8], [13] for Tricut [14]. The Trimix algorithm utilizes an augmented image at the image level as input. This augmented image includes the pasted information region and applies distortion to the image [15], [16], [17], [18], [19].

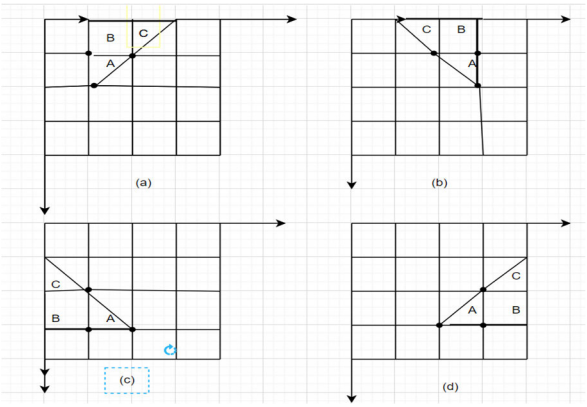


FIGURE 2. Representation of an integral image. The average intensity of the region labelled "A" is calculated. (a) nw (b) ne (c) sw (d) se.

C. MODEL PROPOSED

Fig. 3 depicts the structure for the recommended work. For illustrative purposes, we will utilize the Ravdess speech sample as our source of data, which will subsequently be transformed into a spectrogram image. In order to learn the classifier, it is necessary to utilize an image that has been labeled. Subsequently, we employ the triangular region augmentation technique, wherein the spectrogram is initially transformed into an integral image using the Voila-John's method. Then, we proceed to choose four triangular regions with distinct orientations, namely northwest (nw), northeast (ne), southeast (se), and southwest (sw). To determine a particular orientation, we employ a threshold and subsequently implement augmentation techniques such as random elimination with Tricut and cut mix using auto-augment for Trimix. We obtain Log-Mel characteristics based on the enhanced image and subsequently employ a technique for supervised learning to classify the emotion. This classification is performed using a pre-trained VGG16 model, which has already been trained on a labeled dataset using a convolutional neural network.

Fig.2 illustrates suggested areas to recognize emotional information, which are then contrasted with the other three positions to distinguish them from the initial types of suggested areas for detecting emotional information. The region A (1,2,3), B, and C is the region of an equilateral triangle, where A and C are the triangle parts that are identified as dark and light part for all orientation. The A (1,2,3) which is the dark part sum is calculated and checked for all orientations, and a region with minimal average intensity is chosen to be cut. For Tri Mix A (1, 2, 3), chooses maximal average intensity among all orientations to cut and mix it in another image. We calculate the brightness of each region using the contrast between the total brightness of the light and dark areas, which we propose to be triangular in shape. The dark part is chosen. This method is similar to that put forth by Viola and Jones.

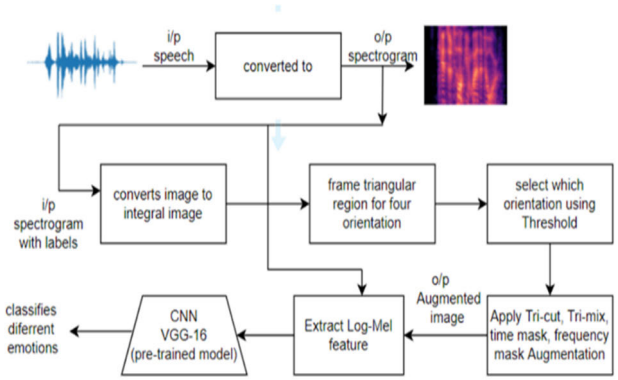


FIGURE 3. Shows a block diagram of the proposed model.

Algorithm 1 Modified Region Augment Algorithm: Triangular region data augmentation approach

Input: Given an input image with W × H and label pair(x,y).
Convert the image into an integral image using a triangular integral image.
Find the Calculus of triangular Integral image with a single scan using
a) ne triangular integral image
b) nw triangular integral image
c) se triangular integral image
d) sw triangular integral image
Compute the triangular region using simple array references given by
e) ti_{nw}(1) - (ii (3) - ii (2) -ti_{nw} (3))
f) ti_{ne}(1) - (ii (3) - ii (2) -ti_{ne} (3))
g) ti_{sw}(1) - (ii (3) - ii (2) -ti_{sw} (3))
h) ti_{se}(1) - (ii (3)-ii (2) -ti_{se} (3))
After finding the desired four oriented triangular regions calculate the RGB intensity
I (S, x, y) = $\sum_{(i,j)\in S} g_{i,j}(x, y)$, which calculates the summation of all three RGB color.
Compares the intensity for all regions with the threshold. The region with the lowest
Value will be selected
Apply tri-cut augmentation.
Check a region with a higher value than the threshold and paste it on an image-level
modified image and perform auto-augmentation.
Apply tri-mix augmentation
Repeat the process 2-8 for all images.
Returns Output: Augmented image.

Two rectangular triangles with four distinct orientations—top, left, down, and right—separate the equilateral rectangle. Each of these triangles is further split into two rectangular triangles, which consist of dark and light sub-regions.

$$ti_{sw}(\mathrm{x}, \mathrm{y}) = \sum_{r=1}^{x} \sum_{c=1}^{r} i(c, y - x + r) \quad (4)$$

$$ti_{se}(\mathrm{x}, \mathrm{y}) = \sum_{r=1}^{W-x} \sum_{c=1}^{r} i(W - c, y - W + x + r) \quad (5)$$

$$ti_{nw}(\mathrm{x}, \mathrm{y}) = \sum_{r=1}^{x} \sum_{c=1}^{r} i(c, y + x - r + 1) \quad (6)$$

$$ti_{se}(\mathrm{x}, \mathrm{y}) = \sum_{r=1}^{W-x} \sum_{c=1}^{r} i(W - c, y + W - x - r) \quad (7)$$

Considering, by definition, i(x, y) = 0 if x <= 0 or y <= 0 or x > W or y > H.

Method for scanning a single image, this article specifically discusses four types of recurrences that can be used to calculate integral images. The range from 2 to 7 serves as an example of how the computation of each significant data component necessitates the use of all adjacent values. This may prevent the removal of the most informative section of an image. To complete the integral images, it is imperative that you examine the images in a distinct sequence.

ne Triangular Integral Images

$$s_{ne}(\mathrm{x}, \mathrm{y}) = s_{ne}(\mathrm{x}, \mathrm{y} - 1) + \mathrm{i}(\mathrm{x}, \mathrm{y}) \quad (8)$$

$$ti_{ne}(\mathrm{x}, \mathrm{y}) = ti_{ne}(\mathrm{x} - 1, \mathrm{y} - 1) + s_{ne}(\mathrm{x}, \mathrm{y}) \quad (9)$$

nw Triangular Integral Images

$$s_{nw}(\mathrm{x}, \mathrm{y}) = s_{nw}(\mathrm{x}, \mathrm{y} - 1) + \mathrm{i}(\mathrm{x}, \mathrm{y}) \quad (10)$$

$$ti_{nw}(\mathrm{x}, \mathrm{y}) = ti_{nw}(\mathrm{x} + 1, \mathrm{y} - 1) + s_{nw}(\mathrm{x}, \mathrm{y}) \quad (11)$$

se Triangular Integral Images

$$s_{se}(\mathrm{x}, \mathrm{y}) = s_{se}(\mathrm{x}, \mathrm{y} + 1) + \mathrm{i}(\mathrm{x}, \mathrm{y}) \quad (12)$$

$$ti_{se}(\mathrm{x}, \mathrm{y}) = ti_{se}(\mathrm{x} - 1, \mathrm{y} + 1) + s_{se}(\mathrm{x}, \mathrm{y}) \quad (13)$$

sw Triangular Integral Images

$$s_{sw}(\mathrm{x}, \mathrm{y}) = s_{sw}(\mathrm{x}, \mathrm{y} + 1) + \mathrm{i}(\mathrm{x}, \mathrm{y}) \quad (14)$$

$$ti_{sw}(\mathrm{x}, \mathrm{y}) = ti_{sw}(\mathrm{x} + 1, \mathrm{y} + 1) + s_{sw}(\mathrm{x}, \mathrm{y}) \quad (15)$$

The triangular region can be computed using array references based on the previously described integral images, as depicted in Fig. 2. The formula $\mathrm{ti_{nw}}$ (1) − (ii (3) − ii (2) − $\mathrm{ti_{nw}}$ (3)) is used to determine the area of the triangle labeled A. This area is then compared to a threshold value and eliminated if it exceeds the threshold. The remaining triangular parts needed for augmentation may be obtained in the same manner. Furthermore, this procedure of utilizing rectangular triangles, as previously explained, allows for the acquisition of additional characteristics of the region. The cut mix as well as random erasing procedures are implemented according to the methodology described in [8]. Additionally, the image-level augmentation technique employed in this study adheres to the procedure outlined in [20]. Following the Tri-Aug region technique for augmentation, we then utilize the Keep augment algorithm as described in [8].

D. NETWORK OPERATING ENVIRONMENT

CNN has achieved remarkable success in various tasks due to its well-defined layers [21] and [22]. The process commonly involves the utilization of both pooling and convolution layers. This section provides a concise summary of these layers. We selected CNN primarily for its capacity to assess spatially invariant features using a small number of parameters [23], which we have found to be especially beneficial.

Following the inner product operation and the addition of the bias layer, the convolution layer convolves filters with the input. Each filter has distinct biases and weights. When training with weights as well as bias, there are actually only two parameters that need to be taken into account. Employing several filters on a single layer enables the acquisition of diverse attributes and characteristics. The model gains the ability to learn features that are unaffected by changes in spatial position by learning only a few parameters.

The pooling layer can effectively reduce the number of subsequent layers. The two most commonly used pooling methods are max pooling and average pooling. Max pooling is a process where the highest value within a specific window is selected, while average pooling is a process where the average value within the window is chosen. In this context, we employ max pooling to obtain contrast values. The Relu activation mechanism is utilized to reveal previously concealed layers within the model. The equation Rel(x) = max is utilized within the interval (0, x). The issue of gradient vanishing is avoided due to the nonlinearity of the expression. The function of softmax activation is utilized in the resultant layer to enable classification. This function allows you to reduce a vector in the range (0, 1) while ensuring that the total of all its components is equal to one. The formula for Softmax (s) is as follows:

$$Softmax(s)_i = \exp[(s)_i]/\sum_j^c \exp[(s)_j] \tag{16}$$

The model is trained using the categorical cross-entropy loss. The Adam optimizer is employed for executing stochastic gradient descent. The issue of the step size remaining constant regardless of the gradient's magnitude hinders conventional stochastic gradient descent. The Adam optimizer addresses this issue by utilizing a variable learning rate.

IV. DISCUSSION OF THE EXPERIMENT

This section focuses on evaluating the effectiveness of triangular area augmentation with transfer learning using standard datasets and comparing it to alternative methods. The procedure's outcome is outlined in Section A, while Section B, offers a comprehensive elucidation of tri-mix augmentation. Additionally, Section C delves into the transfer process of learning. In conclusion, Section D presents a comparison between the augmentation method and various state-of-the-art deep transfer learning techniques.

A. PROCEDURE FOR SETTING UP THE EXPERIMENT

In this part, we show that our triangular modified TriAug method is better than the current best data augmentation methods for a number of difficult deep learning tasks, such as supervised image grouping, emotion recognition, and transfer learning using models that have already been trained [1]. The saliency maps utilized for emotion recognition along with transfer learning are generated from low-resolution images [8].

Our method enhances the effectiveness and efficiency of previous region-level augmentation techniques, including cut-out [8], [21], cut-mix [14], [15], random erasing [13], time-frequency mask methodologies, and auto-augment [16]. Initially, we assess the total saliency values measured on all possible components for each image. We then employ the percentile value for our threshold to select the area of interest. We utilized the numerical value of 0.6. Incorporating this Ravdess speech image, it is also necessary to utilize the initial setting for cut-out, specifically the size of the cut-out paste-back [14].

1) DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) files, which contain annotated audio, were obtained from the well-known website Kaggle. The Ravdess dataset exclusively includes audio recordings for this project, specifically in the format of 16-bit, 48 kHz.wav files. The Ravdess Dataset consists of 1440 annotations performed by 24 male and female actors, with a total duration of one hour. Collectively, every actor contributed 60 files encompassing a range of emotions, including neutral, calm, happy, sad, angry, afraid, disgusted, and surprised. The dataset consists of seven distinct genres, excluding the neutral class, which is evenly distributed.

2) DATA PRE-PROCESSING

The waveform in the ravdess dataset is modified by clipping to eliminate silent regions and then adjusted to a length of 2 seconds through trimming or zero-padding. This is because the model requires a dataset of standardized size and an average time duration of 2 seconds. A spectrogram provides a visual representation of the spectral content of a signal, showing how it varies over time. The equation m = 2595 log10 (1 + f700) represents the Mel scale, where m represents the Mel and f represents Hertz. This equation is used to measure the range of frequencies that humans can hear and is input into our model. By using filter banks to avoid raw audio waveforms, one can generate the Mel spectrogram. Following this procedure, every sample assumes a 256 × 256 structure, denoting the presence of 256 filter banks along with 256-time steps per clip. The librosa library is employed for wavelet generation from every audio file and for converting every audio record into a spectrogram. Fig.1 (a)–(h) illustrates the spectrogram image and wave plot taken from the Ravdess speech sample for different emotions. The classification of our data is exclusively image-based. After creating spectrograms and wavelets, we use standard image pre-processing techniques to produce training and testing data for the model. Every image possesses a distinct resolution and size of 256 by 256 pixels with three color channels.

B. RAVDESS SPEECH AUGMENTATION

We apply our dynamic approach to the Ravdess [24] sample in order to improve two existing state-of-the-art augmentation schemes. These schemes are commonly referred to as tri-cut as well as tri-mix. Experiments are carried out utilizing the pre-trained VGG16 classifier. The training

conditions specified in [8], [22], and [24] were meticulously followed in this research project. Five independent and random tests are conducted to obtain a precise estimation of the overall mean. Three optimal outcomes are added up and regarded as a precise estimate. The spectrogram augmentation procedure utilizes randomly selected time-frequency masks to prevent overfitting and enhance the ability to adapt to unfamiliar speech image-based tasks. This approach effectively increases the training data samples for spectrograms. Enhanced versions of the initial spectrograms, which have been altered in terms of both time and frequency, are employed to enhance the training data. Promising outcomes are achieved through the application of spectrogram augmentation using the domain of audio-image recognition of emotions.

1) COMPARATIVE ANALYSIS OF AUGMENTATION WITH OTHER MODELS

When evaluating the importance of different techniques for enhancing data in speech, the mean average precision at 3 (mAP@3) is used as the evaluation metric [21], [26]. The expression can be simplified as the mean average precision at 3, which is equal to

$$\text{mAP@3} = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{\min(n,3)} P(k) \tag{17}$$

The formula represents the relationship between the quantity of scored audiovisual files within the test data (U), the precision at cut-off k (P (k)), and the number of estimations per audio file (n).

Table 1 demonstrates a performance improvement of 15.8% compared to the other augmentation method applied in previous work and a 5.2% improvement through transfer learning. Our suggested techniques achieved 84.2 percent and mAP@3s of 84.5 percent on the test set. Although both approaches are relatively easy to implement, a recent study [22] found that the selection of hyper-parameters in Cut-mix has a notable effect on performance. Contrarily, both the tri-cut and tri-mix methods in our proposed approach are not affected by minor variations in parameters. Fig. 4. shows the testing of the TriAug model on benchmark datasets such as IEMOCAP and EMODB with the RAVDESS dataset. The Ravdess dataset performs well on the TriAug model compared to other datasets.

C. UTILIZING TRANSFER LEARNING

It is a machine learning technique that entails constructing a model, then ''training'' it using a pre-existing dataset before applying the approach to a new task. Transfer learning offers the benefit of reducing training time while simultaneously improving performance. The VGG16 model utilized in this study had undergone prior training, leading to enhanced efficacy.

The model of VGG16 is a large-scale convolutional neural network (CNN) architecture that relies solely on data for its operation. This exceptional application makes image classification and organization effortless. When using the VGG16 model for various image classification tasks, it is customary to only train the fully connected layers. The reason for this is that the layers that are fully connected possess the highest amount of information. In addition, we improved the pre-trained methods as a replacement for the features obtained and as a means to utilize the knowledge gained from the previous networks. This enabled us to optimize the utilization of the knowledge we had already gathered. In order to optimize a base model that had previously undergone pre-training, we initially unfroze select upper layers of the model. Subsequently, we incorporated new layers and proceeded to train the model using the base model as a foundation.

We were able to fine-tune the more advanced representations of features in the base model to better align with the specific needs of the new job.

1) ANALYSIS OF RESULTS

For all of our tests, we assess the effectiveness of our Convolutional Neural Network (CNN) baseline that has been trained using the augmented model. During training, all models receive identical weight initialization and augmentations, including time mask, frequency mask, tri-cut, and tri-mix. Our research utilizes models that integrate transfer learning with spectrogram augmentation [4]. We generate and enhance two spectro-temporally modified duplicates of every original speech segment according to the augmentation criteria of each segment. During the training process, the system is trained in emotion recognition systems. In order to assess the impact of the augmentation technique on the experiment's outcome, we will analyze and contrast the results obtained from these strategies with and without the technique. The categorical cross-entropy loss is used as the objective function in every experiment. This grants us the capacity to train the models more effectively. The minimum number of channels in the first block in the VGG-16 framework is set to 32 during initialization. The location of such a setting may be found within the model. The method of training for the model employs PyTorch with the Adam optimizer, starting with an initial learning rate of 10^5 in addition to a batch size of 32. Following the initial eight epochs, the learning rate stays constant. However, after this point, it gradually decreases by half every alternate epoch. The model employs rectified linear unit (ReLU) activation procedures in each layer to expedite the learning process and improve the model's generalization capabilities. Layer-wise group normalization is employed to enhance the model's generalization capabilities and expedite the training process.

We conducted numerous experiments using the Ravdess database. During every study, our networks acquired the ability to accurately identify the training data to different extents. There were notable disparities in the accuracy of test results among different architectures. The precision of test data in cross-validation is determined by calculating the average accuracy over the five folds of the assessment. Over the course of 50 epochs, Fig. 5. shows the accuracy and error

TABLE 1. Applying proposed work in existing model.

Lable	Model	Augmentation	Accuracy	mAP @3
Ravdess speech[45]	VGG16	Noise, Strech, Speed, Pitch	66.8	68.7
Ravdess speech[proposed work]	Pretrained VGG16	Tri-cut,Tri-mix	84.2	**84.5**
Ravdess speech[proposed work]	VGG16	Tri-cut,Tri-mix	79	78.6

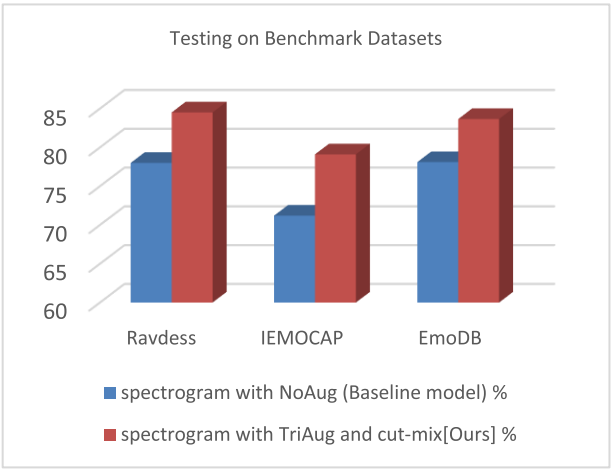
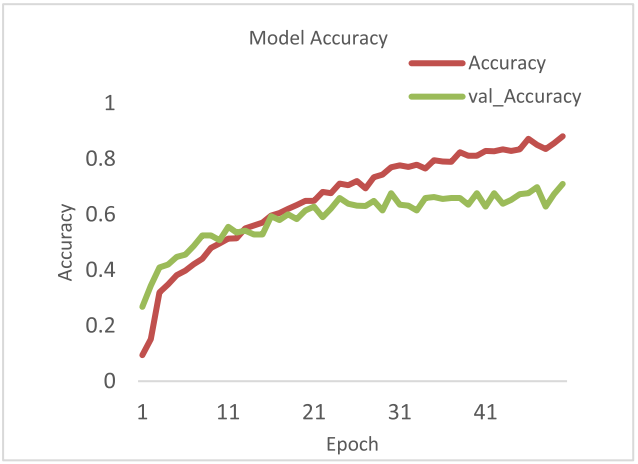


FIGURE 4. Shows the testing TriAug model on benchmark dataset.



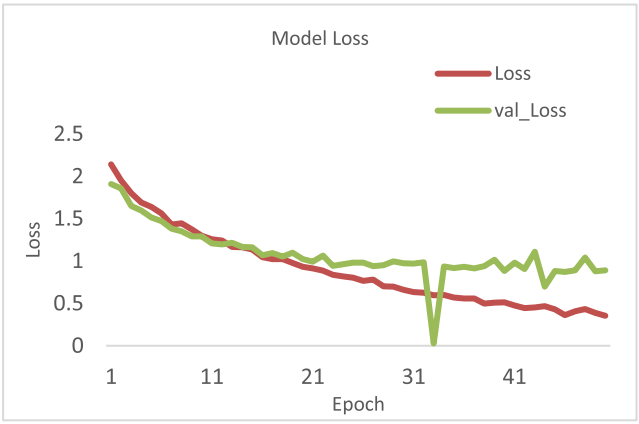


FIGURE 5. Shows train and test accuracy and loss for the SER NoAug model.

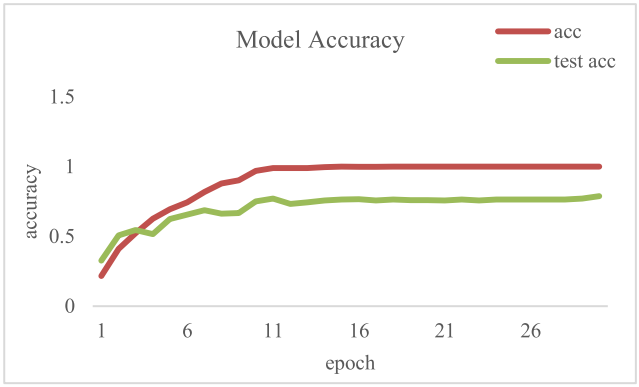
of the CNN models during the training and testing phases for each learning iteration before augmentation. The train set has an accuracy of 88 percent, whereas the test set has an accuracy of 71 percent. The train set experiences a loss of 0.35, while the test set experiences a loss of 0.88.

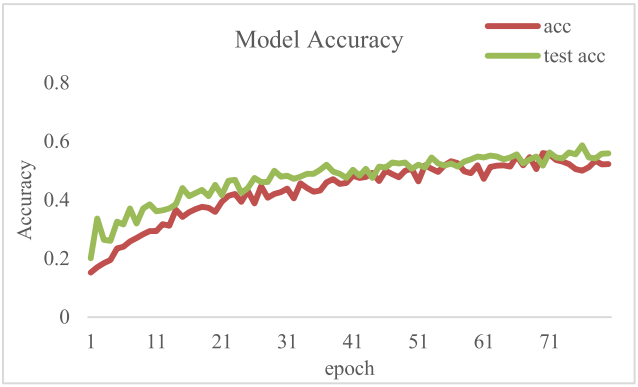
Fig. 6. shows how accurate the CNN models were and how much they lost at training and testing for every initial learning cycle shortly after fine-tuning to feed each learning cycle. This was demonstrated by the fact that the execution of dropouts led to a significant initial advancement in the overall efficiency of our network. A gradual improvement in accuracy was observed when dropout was utilized, which resulted in an increase of 71% to 78.84%. The effectiveness of the network was also improved by boosting the window size of the convolutional layer, which ranges from 16 to 20 pixels. This was done in order to optimize the layer's performance.

The CNN model that had been built on the increased information base arranged the identical test information with an accuracy of approximately 56%, and after some adjustments, it has increased to 78.6%, and its accuracy has increased even further. Consequently, this highlights the significance of expanding information in further developing CNN's executional capabilities. Fig.7. illustrates the precision as well as the loss of the CNN models as seen during the process of preparing and testing for each cycle of extension learning. This observation is made after the expansion process has been completed. Information expansion, in conjunction with increasing the count of parts in the first convolutional layer and making use of dropout, contributed to an 84% improvement in CNN's performance on this data set. The findings demonstrate the extent to which CNN's generalizability using the Ravdess data base is anticipated to be expanded to accommodate regularization operations. To put it another way, the CNN, as demonstrated by the p-esteem of 0.02, required less regularization. Fig. 8. illustrates the exactness and loss of the CNN models after they have been subjected to overtraining and testing for each and every emphasis on learning and tweaking. This is done after expansion and adjustment have been performed.

Another step was to train and test the architecture with various parameters, including average pooling, dropout with $p = 0.2$, and (20×20) kernels in the convolutional layer.

The best performance was measured over 50, 75, and 100 training epochs. Based on the findings, it was discovered that the accuracy of the test increased as the number of training epochs grew. With an accuracy of below 79 percent, the CNN that was trained on the initial database was able to correctly recognize the test data. On the other hand, the CNN that was trained on the augmented database was capable of identifying and grouping the test data with an accuracy of





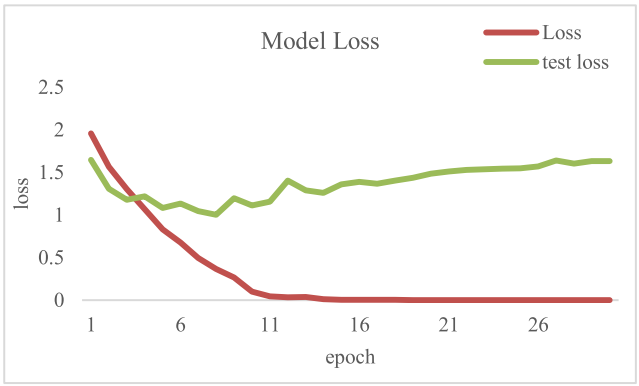


FIGURE 6. Shows train and test accuracy and loss after Noaug Fine-tuning.

FIGURE 7. Shows the train and test accuracy and loss after triaug.

roughly 84 percent. As a consequence of the augmentation, the CNN model's performance on the test data was substantially enhanced, resulting in this improvement. According to the Ravdess database, the classification accuracy of the network that was built on top of it was 84.2% after 80 training epochs had been completed. It was previously reported by DeVries and colleagues [22] that the whole detection accuracy of this CIFAR-100 sample was 78 percent.

Additionally, Kosaka [4] reported that the overall accuracy of the Ravdess speech dataset was 66.8 percent when noise, strech, speed, and pitch augmentation were used. It is clear from this result that the CNN and Tri-aug models that we developed as part of our research performed in a manner that was comparable to the previous model and achieved improvements over it. The database discovered a data shortage, which resulted in the identification of the models overfitting problems. The database result revealed that the model has overfitting issues. In other words, while the use of data augmentation improved the performance of the data, the use of dropout with the maximum effect of regularization had a negative impact on the performance of the CNN models. For the purpose of determining whether or not each emotion had a higher F1 score than other emotions, the F1 scores of the emotions were compared in the following manner:

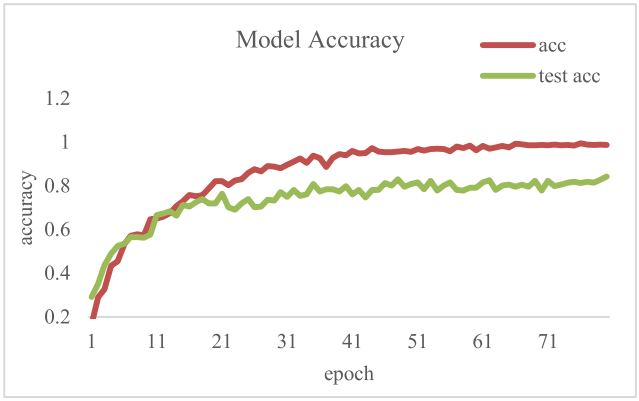
The F1 score is equal to two times the product of the

$$F1 = 2^{*}((\text{precision}^{*}\text{recall})/(\text{precision} + \text{recall})) \quad (18)$$

The recall and precision of recognition are both factors that are taken into consideration when calculating the F1 score. The recall factor determines the extent to which a specific emotion contributes to the false negatives of other emotions. Precision, on the other hand, is a measurement that determines the extent to which a specific emotion influences the false positives of other emotions. To put it another way, accuracy is a measurement of the amount of confusion that a particular emotion has introduced into a system. This is the proportion of other emotions that are identified as the emotions of interest within the system.

According to Table 2. the disgusted class was incorrectly categorized as fearful and as angry the majority of the time. This is because the F1 scores of the disgusted class are not lower than those of other emotions such as fear and sadness. Additionally, feelings of fear merged with disgust and surprise, as well as feelings of neutrality and sadness combined with both calm and disgust. It was determined that the sentiments of fear and sadness were the least accurate when measured against the other forms of emotion. Fig. 9. shows a comparative analysis of the proposed model with existing models. Table 5 compares accuracy with the baseline model and cifar-100.

In Ravdess databases, the F1 scores of the surprised classes performed significantly better in terms of classification accuracy in comparison to the F1 scores of other classes.



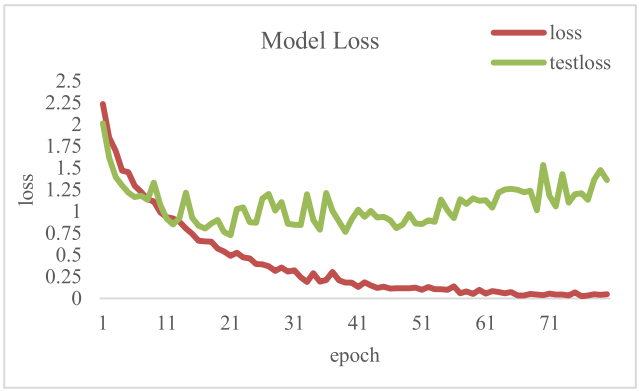


FIGURE 8. Shows train and test accuracy, loss with triaug fine-tuning.

This was the case when compared to the corresponding scores of other classes. The results presented in Table 3 show that the F1 score after the augmentation shows a significant improvement in all areas, with the exception of neutral. In addition, we are able to observe that the capacity to identify disgust as an emotion has demonstrated an improvement as a consequence of the enhancement. When compared to the Noaug model, this model has a higher recognition rate than the Noaug model. The recognition of calm, surprised, and angry emotions in speech signals was not especially challenging for CNN models trained with 50 training epochs prior to augmentation. However, it was challenging for CNN classifiers to learn with 80 training epochs after augmentation. However, after fine-tuning, the model improved in performance. In light of this, the enhancements made to the Ravdess database can be attributed, at least in part, to the increased availability of emotion instances, which made the process of emotion recognition simpler for users. Table 4 shows the comparison of existing augmentation methods with the proposed Triaug method.

When all of the findings of this study are considered together, they provide evidence that convolutional neural networks are effective in classifying the feelings that actors convey using their verbal utterances. Not only do the findings highlight the importance of data augmentation and dropout, but they also highlight the significance of training time and learning time in terms of enhancing the generalizability of CNN models.

The CNN models developed by DeVries and colleagues [22] were more complicated than the ones we developed because they utilized a greater number of convolutional layers and were trained for a longer period of time. In contrast, the performance of our CNN models was superior to that of theirs. Table 6 and Fig. 8 show details of the

TABLE 2. Shows the F1 score NoAug fine-tuned model.

	Precision	Recall	F1- score
angry	0.82	0.79	0.81
calm	0.78	0.95	0.84
disgust	0.76	0.76	0.73
fearful	0.81	0.6	0.66
happy	0.86	0.86	0.86
neutral	0.82	0.75	0.78
sad	0.69	0.66	0.67
surprised	0.83	0.9	0.88

TABLE 3. The F1 score after triaug fine-tuned model.

	Precision	Recall	F1- score
angry	0.95	0.8	0.87
calm	0.85	0.92	0.88
disgust	0.83	0.86	0.84
fearful	0.83	0.81	0.82
happy	0.86	0.8	0.83
neutral	0.77	0.81	0.79
sad	0.8	0.82	0.81
surprised	0.93	0.85	0.89

TABLE 4. Comparison of existing augmentation methods with the proposed triaug method.

Augmentation	Recognition Rate (%)
Generative Adversarial Networks (GANs) [56]	54.6
random time-frequency masks [30]	64.14
Signal-based Audio Augmentation (SA) [56]	50.7
SA with replacement of the majority class only (SAR_M) [56]	51.0
SAR adding only Background Noise (SAR_B) [56]	51.1
EMix-S [54]	77.6
EMix-NS [54]	77.32
EMix-N [54]	76.9
SA with replacement (SAR) [56]	51.2
SAR using only TS and PS (SAR_S) [56]	49.3
Tricut and Trimix [proposed work]	**84.2**

TABLE 5. Compares accuracy with the baseline model and CIFAR-100.

Model	Accuracy (%)	Cutout	cutmix
Baseline model	78	-	-
Imagenet	76.3(Resnet)	77.1	78.4
overall model [propose work]	**84.2(VGG16)**	**84**	**83.9**
Cifar-100[27]	78	-	-

TABLE 6. Comparative analysis of the proposed model with other existing model.

Models	IEMOCAP (%)	EMO-DB(%)	RAVDESS (%)
ACRNN [58]	64.74	82.82	-
ADRNN [59]	69.32	84.99	-
CB-SER [60]	72.25	85.57	77.02
Lightweight CNN [48]	77.01	92.02	-
Our Proposed Model	**79.1**	83.65	**84.2**

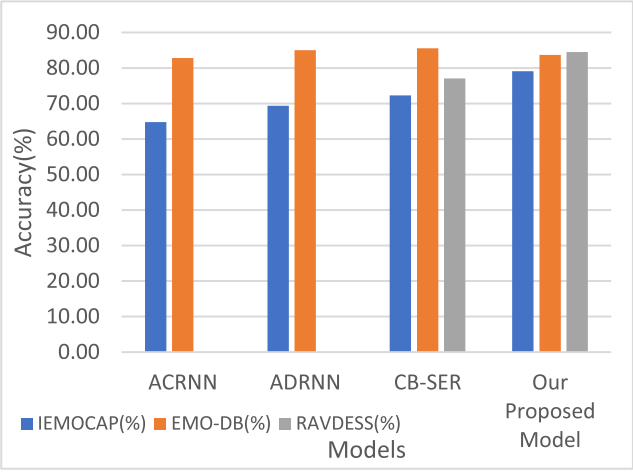


FIGURE 9. Shows comparative analysis of proposed model with existing models.

TABLE 7. Confusion matrix for tricut AUG.

EMO (Actual Lables)	A	C	D	F	H	N	S	SR
A	0.89	0	0.11	0	0	0	0	0
C	0	0.88	0	0	0	0.12	0	0
D	0.12	0	0.84	0	0	0	0	0.04
F	0.02	0	0.12	0.82	0	0	0	0.04
H	0	0.03	0	0	0.83	0.01	0	0.13
N	0	0.12	0	0	0.1	0.78	0	0
S	0	0.9	0.1	0	0	0	0.81	0
SR	0	0	0.03	0	0.1	0	0	0.87

Predicted Lables

A-Angry, C-Calm, D-Disgust, F-Fear, H-Happy, N-Neutral, S-Sad, SR-Surprise, EMO- EMOTIONS

comparison of existing models with the proposed model using benchmark datasets, in which IEMOCAP and RAVDESS performed well but EmoDB showed moderate performance.

TABLE 8. Confusion matrix for trimix AUG.

EMO (Actual Lables)	A	C	D	F	H	N	S	SR
A	0.88	0	0.12	0	0	0	0	0
C	0	0.88	0	0	0.11	0.01	0	0
D	0.1	0	0.84	0.04	0	0	0.02	0
F	0.03	0	0.12	0.82	0	0	0	0.03
H	0	0.04	0	0	0.83	0	0	0.13
N	0	0.2	0.02	0	0	0.78	0	0
S	0	0.1	0.09	0	0	0	0.81	0
SR	0.1	0	0	0	0.03	0	0	0.87

Predicted Lables

A-Angry, C-Calm, D-Disgust, F-Fear, H-Happy, N-Neutral, S-Sad, SR-Surprise, EMO- EMOTIONS

Tables 7 and 8 show the confusion matrix of Tricut and Trimix augmentation.

Findings also showed that increasing the size of the kernels for the convolutional layer had a regularizing effect on the models. This stopped the models from overfitting and made generalizations more accurate. In the same way, the results showed that the models could fix the problem of the training data not fitting well enough by increasing the number of training iterations and decreasing the size of the kernels in the convolutional layer at the same time. An evaluation of the levels of accuracy using the base model Noaug is shown in Table 4. It shows that things got better after Tri-Aug was added.

V. FINAL THOUGHTS

The purpose of this paper is to present the most extensive study that has been conducted to date, which investigates the application of CNN in speech image emotion classification.

This study utilizes a variety of data augmentation techniques. For the purpose of enhancing the effectiveness of speech image emotion categorization systems, we have shown that data augmentation strategies, in particular the direct utilization of the spectrogram, are extremely effective. In addition to this, we presented triangular augmentation, a novel approach to data augmentation that is not only straightforward but also highly efficient. The training of CNN, which is very good at identifying their surroundings and locating themselves, was aided by the introduction of the tri-cut and tri-mix techniques. Nevertheless, despite the fact that Tri-Cut Mix is easy to use and calls for a marginal amount of processing power, it is not without its drawbacks. When compared to the baseline Noaug method, the application of Tri-cut and Tri-mix to VGG-16 results in an increase of +6 percent in the accuracy of speech image emotion categorization. We were able to show that using Tri-Cut Mix with the vanilla model and other regularized models makes image classifiers more robust and less uncertain. There were many different data augmentation strategies that were utilized in the process of training CNNs. Experiments showed that using

fine-tuned CNNs in conjunction with a VGG16 pre-trained model led to an improvement in performance.

ACKNOWLEDGMENT

The authors would like to thank their supervisor, institution, and reviewers for their thoughtful comments and efforts toward improving their manuscript.

REFERENCES

[1] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, *arXiv:1912.02610*.
[2] W. N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 16–23.
[3] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An ensemble of convolutional neural networks for audio classification," *Appl. Sci.*, vol. 11, no. 13, p. 5796, Jun. 2021.
[4] GitHub. *mkosaka1/Speech Emotion Recognition: Using Convolutional Neural Networks in Speech Emotion Recognition on the RAVDESS Audio Dataset*. Accessed: Apr. 3, 2022. [Online]. Available: https://github.com/mkosaka1/Speech_Emotion_Recognition/blob/master/2-Copy1.%20Data_Augmentation.ipynb
[5] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6319–6323.
[6] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, "AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark," 2018, *arXiv:1807.03418*.
[7] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning 'BERT-like' self supervised models to improve multimodal speech emotion recognition," 2020, *arXiv:2008.06682*.
[8] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "KeepAugment: A simple information-preserving data augmentation approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1055–1064.
[9] B. Zoph, E. D. Cubuk, G. Ghiasi, T. Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 566–583.
[10] S. Wei, S. Zou, F. Liao, and W. Lang, "A comparison on data augmentation methods based on deep learning for audio classification," *J. Phys., Conf. Ser.*, vol. 1453, no. 1, Jan. 2020, Art. no. 012085.
[11] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 9885–9955, 2020.
[12] H. Proenca and S. Filipe, "Combining rectangular and triangular image regions to perform real-time face detection," in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 903–908.
[13] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.
[14] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.
[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[16] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*.
[17] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast AutoAugment," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 6665–6675.
[18] R. Hataya, J. Zdenek, K. Yoshizoe, and H. Nakayama, "Faster autoaugment: Learning augmentation strategies using backpropagation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 23, 2020, pp. 1–16.
[19] K. Tian, C. Lin, M. Sun, L. Zhou, J. Yan, and W. Ouyang, "Improving auto-augment via augmentation-wise weight sharing," 2020, *arXiv:2009.14737*.
[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.
[21] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. M. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.
[22] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[23] B. Z. J. L. S. Thornton, "Audio recognition using Mel spectrograms and convolution neural networks," 2019.
[24] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on CIFAR-10," *Unpublished Manuscript*, vol. 40, no. 7, pp. 1–9, 2010.
[25] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.
[26] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, "Pre-trained deep convolution neural network model with attention for speech emotion recognition," *Frontiers Physiol.*, vol. 12, Mar. 2021, Art. no. 643202.
[27] C. Summers and M. J. Dinneen, "Improved mixed-example data augmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1262–1270.
[28] L. Wei, A Xiao, L. Xie, X. Zhang, X. Chen, and Q. Tian, "Circumventing outliers of autoaugment with knowledge distillation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 608–625.
[29] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, May 2021.
[30] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 645–652.
[31] J. Li, W. Wu, D. Xue, and P. Gao, "Multi-source deep transfer neural network algorithm," *Sensors*, vol. 19, no. 18, p. 3992, Sep. 2019.
[32] R. S. Kumar, K. M. Muraleedharan, P. Vivek, and V. I. Lajish, "Study of nonlinear properties of vocal tract and its effectiveness in speaker modeling," *J. Acoust. Soc. India*, vol. 43, no. 2, pp. 116–124, 2016.
[33] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6256–6268.
[34] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, no. 6356, pp. 161–163, Jan. 1992.
[35] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, p. 7665, Nov. 2021.
[36] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Proc. Comput. Sci.*, vol. 112, pp. 316–322, Jan. 2017.
[37] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.
[38] S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition," *Sensors*, vol. 20, no. 22, p. 6688, Nov. 2020.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[40] K. Xu, B. Zhu, Q. Kong, H. Mi, B. Ding, D. Wang, and H. Wang, "General audio tagging with ensembling convolutional neural networks and statistical features," *J. Acoust. Soc. Amer.*, vol. 145, no. 6, pp. EL521–EL527, Jun. 2019.
[41] K. W. Cheuk, K. Agres, and D. Herremans, "NnAudio: A PyTorch audio processing tool using 1D convolution neural networks," in *Proc. ISMIR-Late Breaking Demo*, 2019, pp. 1–2.
[42] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[43] Y. Yamada, M. Iwamura, T. Akiba, and K. Kise, "Shakedrop regularization for deep residual learning," *IEEE Access*, vol. 7, pp. 186126–186136, 2019.
[44] X. Gastaldi, "Shake-shake regularization," 2017, *arXiv:1705.07485*.
[45] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[46] (2019). *Ebouteillon/Freesound Audio Tagging*. [Online]. Available: https://github.com/ebouteillon/freesound-audio-tagging-2019/blob/master/code/training-vgg16.ipynb
[47] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, Sep. 2020.
[48] S. Das, N. N. Lønfeldt, A. K. Pagsberg, and L. H. Clemmensen, "Towards interpretable and transferable speech emotion recognition: Latent representation based analysis of features, methods and corpora," 2021, *arXiv:2105.02055*.
[49] L. Luo and Y. Wang, "EmotionX-HSU: Adopting pre-trained BERT for emotion classification," 2019, *arXiv:1907.09669*.
[50] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021.
[51] S.-J. Lee and H.-Y. Kwon, "A preprocessing strategy for denoising of speech data based on speech segment detection," *Appl. Sci.*, vol. 10, no. 20, p. 7385, Oct. 2020.
[52] C. Zheng, C. Wang, and N. Jia, "An ensemble model for multi-level speech emotion recognition," *Appl. Sci.*, vol. 10, no. 1, p. 205, Dec. 2019.
[53] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech emotion recognition through hybrid features and convolutional neural network," *Appl. Sci.*, vol. 13, no. 8, p. 4750, Apr. 2023.
[54] A. Dang, T. H. Vu, L. D. Nguyen, and J.-C. Wang, "EMIX: A data augmentation method for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[55] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, Aug. 2022.
[56] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using GANs for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 171–175.
[57] H. K. Ali and Y. Khalil, "Energy conservation using voice recognition," *Ibrahim J. Inf. Eng. Appl.*, vol. 3, no. 11, pp. 59–62, 2013.
[58] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[59] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[60] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.

V. PREETHI was born in Chennai, India, in 1980. She received the B.E. degree in computer science and engineering from Madras University and the M.Tech. degree in pervasive computing technologies from Anna University, Chennai, in 2012. From 2014 to 2019, she was an Assistant Professor with the Computer Science Department, GKM affiliated to Anna University, for three years, and the SRM Institute of Science and Technology, Chennai, for three years. From 2019 to 2023, she was a Research Associate with the SRM Institute of Science and Technology. She is currently an Assistant Professor with the SRM Institute of Science and Technology. Her research interests include artificial intelligence, machine learning, deep learning, and adhoc networks.

V. ELIZABETH JESI was born in Chennai, India, in 1971. She received the M.C.A. degree in computer applications from Madras University, in 1994, the M.S. degree in computer science and engineering from the SRM Institute of Science and Technology, in 2011, and the Ph.D. degree in computer science and engineering, in 2020.

Since 1995, she has been a Lecturer with the Computer Science Department in various colleges, such as the Auxilium College, Vellore; the Karunya Institute of Technology, Coimbatore; and the Jaya Engineering College, Chennai. Since 2000, she has been an Associate Professor with the SRM Institute of Science and Technology. Her research interests include design and analysis of algorithm, machine learning, and image processing.