IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 3, 2021

Emotion Detection in Speech:ACNN Neural Network Approach

C.Prabhavathi, P.Vijaya Kumari, K.Chandrasekhar,D.Raghunath Kumar Babu, S.Ghouhar Taj

Assistant Professor(Adhoc) CSE Department,JNTUACEP, prabhavathi1231@gmail.com
Assistant Professor(Adhoc)CSE Department,JNTUACEP, vijaya.jntuacep@gmail.com
Assistant Professor(Adhoc)CSE Department,JNTUACEP,chandra507shiva@gmail.com
Assistant Professor(Adhoc)CSE Department,JNTUACEP, raghunath.d29@gmail.com
Assistant Professor(Adhoc)CSE Department,JNTUACEP,sgtaj786@gmail.com

Abstract—Speechemotionrecognitionisanareaofresearch dedicated to identifying and categorizing emotions expressed throughspeech.Itspurposeistocomprehendandinterpretthe emotional content conveyed in spoken words, leveraging signal processing techniques, feature extraction algorithms, and machine learning models. The ultimate aim is to apply this knowledge in diverse applications such as human-computer relation, affective computing, and mental health diagnosis. Although the complexity and variability of emotional speech present significant challenges, recent years have witnessed notable progress through the development of advanced algorithmsandtheutilizationofextensivetrainingdatasets.Here we proposed a CNN network pattern for speech emotion recognitionforthebenchmarkdatasetsandgottheaccuracyof 88% for the convolutional neural network model.

Keywords—CNN,Neuralnetworks,Speech,emotionintelligence, datavisualization

I. Introduction

Non-verbalcommunicationplaysacrucialroleinhuman interactions. Apart fromtheliteralmeaning conveyedthrough spokenlanguage,thewaywordsarespokencarriessignificant information. The same spoken text can have multiple interpretations basedonthemanner inwhichit isexpressed. The term 'really' in the English language has versatile applications.It canbe employedto inquire aboutsomething, display admiration, indicate skepticism, or assert a strong statement. Merely understanding the textual content of a spoken phrase is insufficient for accurately interpreting its meaning.Emotionrecognitioninspeechholdsvariousfuture implementations. One such application involves enhancing speechunderstandingbyusingemotionrecognitionasatool. Traditionally, emotion has been considered a disruptive element that hampers the comprehension of spoken text. However, by recognizing and isolating emotions within speech,itmaybepossibletoenhancetheexecutionofspeech comprehension systems.

Multimediapatternidentificationisannewinnovationthat enables the extraction and analysis of large volumes of multimedia information from video and audio sources. In current years, deep learning techniques, particularly using machine learning with deep neural networks, have been extensivelyappliedtoaddressdifferentidentification problems. However, the challenge lies in the fact that individuals indicate emotions in individual method, and distinguishing between these emotions based on unclear features is a difficult problem, even for humans.

Conventionalmethodsforaddressingthisissueentailthe extraction of basic characteristics and training machine learning models based on these extracted features. These techniqueshave beenconsideredfuturistic forlongtime,but selectingappropriatefeaturestoextractisachallengingtask, and optimizing the results can be protracted.

CNNshaveemergedastheleadingapproachincomputer vision applications and have garnered attention in diverse fields. These networks comprise various elements and particularly designed to learn hierarchical spatial features through the utilization of backpropagation algorithms. Understandingtheidea,advantages,andlimitationsofCNNs is crucial to fully leverage their potential and improve the execution of the recognition of the emotion model.

The proposedimplementation, authorsdeveloped asystem that uses neural networks, specifically CNNs, for emotion recognitioninspeech.Sincetheproject involvesclassification, a CNN is the natural choice. The model is trained to detect eight human emotions likely happy, sad, neutral, angry,calm,disgust, surprised and fearful along with determiningthegenderofthespeaker.ByutilizingCNNs,we aimtoleveragetheirabilitytoautomaticallylearnhierarchical features and upgrade theoverall executionof the recognition in emotion system.

The prefer model is done with the datasets fromRAVDESS and SAVEE. Finally, the efficiency of the trained model is observed by testing against live voice.

The foremost purpose of the research is to: a)To understand speech recognition and its fundamentals. b) To Collect the datasets on Speaker emotion recognition. c) Developthe algorithm for feature extraction. d) Developthe algorithm for Classification. e) Compare the proposed methods with existing works.

Followingaretheremainingsections:SectionIIdealswith studies and researches made by some of the certified researchersthroughoutthe world onspeechemotiondetection and related works. We have used and improved our project modelbasedontheirworkandimplementationsmethods.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 3, 2021

Section IIIprovides the basic architecture for the emotion detectionusingthespeechprocessingtechnique.Thischapter alsodealswiththeexplanationofblockdiagramwhichisused for implementation with the necessary evaluation models. Section IV deals with the final step of our project, in this section we discuss about the accuracy and losses of training and testing model that we have implemented with the validationslossovertime.Wewillalso havelookatthemodel classificationreports.InSectionVweconclude our work by explaining the challenges that are present in speech emotion detection and how we have tried to overcome them with the future scope of this project.

II. RELATEDWORK

In their research, Dias Issa et al. described for emotion identification using sound files. They extracted various features. They achieved accuracies of 71.61% for the RAVDESS dataset, 86.1% and 95.71% for different subsets of the EMO-DB dataset[1]. Harshawardhan worked using MFCC features and an LSTM algorithm. They obtained an 84.81% accuracy [2]. Deepak Bharti and Poonam Kukana presentedawithMSVM(MultipleSupportVectorMachine) classifier. Theyachieved a high accuracyrateof97% onthe RAVDESS dataset using feature extraction (GFCC) and featureselection(ALO)techniques. Onexisting datasets,they obtainedanaccuracyof79.48%withfeatureextractionusing MFCC [3].

AnushaKoduruetal.focusedonpre-processingaudio samples by removing noise using filters.Their results showed accuraciesof70%withSVM,85%withdecisiontree,and65% withLDA[4].AuthorsproposedadeeprecurrentNeural Networkprocessforlearningemotionvariations.Their methodology wasappliedtothedatasetcalledRAVDESS, achieving correctness over 80% [5].

Christyetal.achievedanaccuracyof78.20%usingCNN on the RAVDESS dataset [6]. Ting-Wei Sun useda hybrid algorithm and this algorithm achieved high accuracies on FAU and eNTERFACE databases [7].

Latifetal.developedalatestdatabaseinUrduspeechand evaluated the performance of a model using SVM classifier [8]. Jalal et al. proposed and compared bimodels, CNN and attentionandbi-LSTMandattention,foremotionrecognition [9]. Satya et al. aimed to provide an extensive survey that highlights the requirements of speech and vision systems, considering both hardware and software aspects. [10].

Abdelhamidetal.concentratedondevelopinganoveldata boosting technique to enhance the emotions database by introducing additional illustrative through the controlled inclusion of noise selection [11]. In a related study, Their approachemployedmelfrequencylogspectrogramtoextract relevant evidence from the emotional speaker database and employed a 2D DCNN for analysis [12].

Middya et al. conducted an extensive investigation into fusion on a model-level techniques to choose the most effective multimodal model for emotion recognition[13]. Bakhshi et al. Introduced CyTex, a revolutionary speech-to-image transformation technique that leverages the basic frequency of every speech shape to directly convert the raw speechsignalintovisuallytexturedimages[14].Bagadietal. conductedastudytoexaminetheinfluenceofmeta-heuristic for feature selection methods on speech-based emotion identification[15].Zhongresearchaimsto accomplishspeech emotion recognition in Chinese by machine learning using CNN [16].

III. PROPOSED FRAMEWORK

Theaudiofilesarepreprocessedbyaddingnoiseand shifting time, pitch and speed to improve the model's ability to generalize. Features are then extracted using the MFCC method and stored in a csv file. The final step involves building a Convolution Neural Network (CNN) model for classification.TheinputfeaturesareMFCCsandthemodelis trained to classifythe audio files into 8 different emotions.

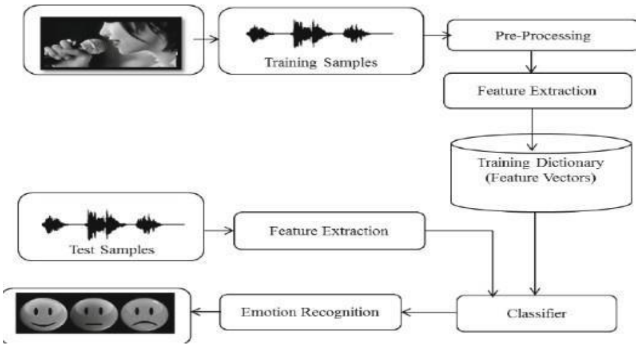


Figure1: BlockDiagramforModel

Detailofdatasets:
RAVDESS-Datasetincorporateover1500auditoryfolder input starting 24 distinct actors. Twelve malesand Twelve femaleswherever these performers record little auditory in dissimilarfeelings. Figure 2representsthe Ravdess Dataset Visualization

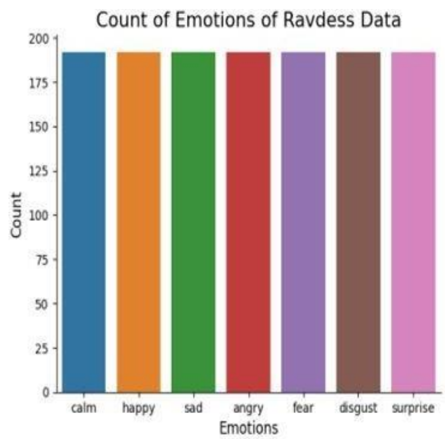


Figure2: RavdessDatasetVisualization

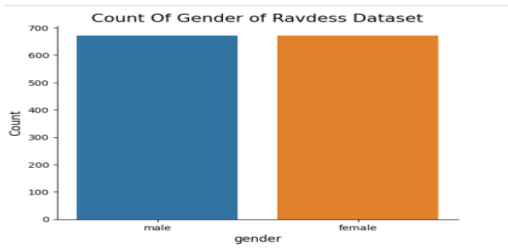
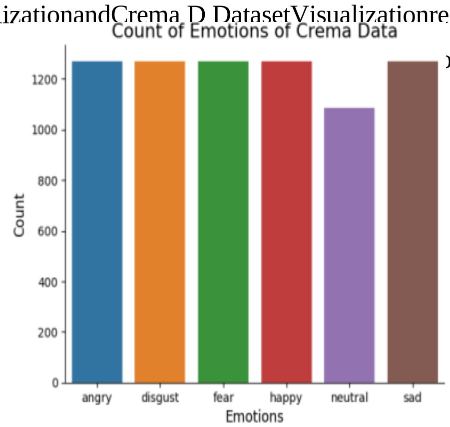


Figure3: GenderofRavdessDatasetVisualization

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 3, 2021

The audio files are named in a consistent manner where the 7th character reflects the various emotions they represent. Figure 3 and 4 depicts the Gender of Ravdess Dataset VisualizationandCrema D DatasetVisualizationrespectively.



Pre-processing: Toenhancethemodel'sabilitytogeneralize, we create additional synthetic data samples by making slight modificationstoouroriginaltrainingset.Foraudiodata, theseperturbationsincludenoiseinjection,timeshifting, pitch alteration, and dash modification. The aim is to make our model fixed to such variations and improve its abilityto generalize across different conditions. We add two types of noisestomakeourmodeltrainingmoreefficient.Weuse whitenoiseand Gaussiannoisesforthis purpose.

FeatureExtraction: Toenableourmodeltostudyfromthe auditory files, the next phase is to extract the features from them. For feature extraction, we utilize the Python library called LibROSA, which is widely employed for audio analysis. This library offers a range of tools and functions specifically designed to process audio data effectively. The feature is being extracted and put in a csv file. Feature extractionistheprocessofidentifyingandselectingthemost relevantanddescriptivecharacteristicsorattributesofaset ofdataand transformingthemintoanewsetoffeatures that canbeusedinfurtheranalysisormodeling.Itisacrucialstep inmanymachinelearningalgorithmsandbenefitsinrefining the performance and accuracyofthe models. Proposed flow diagram is showed in figure 5.

MFCCisusedforfeatureextractionhere.

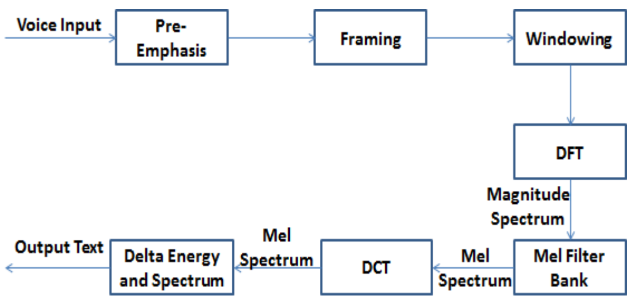


Figure5:MFCCblockdiagram

TheMFCCareasetofstructuresusedinspeechandaudio. ThestepstocomputeMFCCsinclude:

1.Pre Emphasis: Amplifying the high occurrence componentsoftheindicationbyapplyingahigh-pass filter.

$$(t)=(t).x(t-1) \quad (1)$$

Wherex(t)istheinputsignalandaisapre-emphasis coefficient.

2. Windowing:Dividingthesignalintooverlapping frames and applying a window function.

$$(n)=(n).x(n) \quad (2)$$

Wherew(n)isthewindowfunctionandx(n)istheinput signal.

3. Spectralanalysis:Computingthepowerspectrumof eachframeusingtheFastFourierTransform(FFT).

$$\sum(n)N-1n=0.e^{\frac{-j2\pi kn}{N}} \quad (3)$$

WhereX(k)theFourierTransformofx(n),Nisthesum of models, and k is the frequency index.

4. Mel-scale transformation: Applying a non-linear transformation to the power spectrum to model the way human ear perceives different frequencies.

$$(f)=2596.\log_{10} \quad (1+\frac{m}{700}) \quad (4)$$

Wherem-frequency(Hz)andH(m)isthecorresponding Melfrequency. Next,thepower spectrumis mappedto Mel-scale using triangular overlapping filters. The equation is:

$$(k)=\sum_{m=0}^{p-1} h_{(k)}\cdot X^2(k) \quad (5)$$

Where$h_{(k)}$is the triangular filter,$X^2(k)$ is the squared magnitude of X (k), and P is the number of filters.

5. Cepstralanalysis: Taking the logarithmofthe Mel-scaled power spectrum and applying the Discrete CosineTransform(DCT)toconvertittothecepstral domain.

$$(i)=\sum_{m=0}^{p-1} (k).\cos(\frac{\pi ik}{P}) \quad (6)$$

Wherec(i)isthei-thecepstralcoefficientandPisthesum of Mel-scale coefficients.

6. Cepstralcoefficients:Selectingthe first Ncoefficientsas theMFCCs,whereNisaparameterthatcanbeadjustedbased on the desired level of detail. (i) = (i) where i = 1,2,…..N and number of desired cepstral coefficients represent in N.

- **CNN Architecture**

The Convolution Layer applies learnable filters to small windows of the input matrix, producing a 2-dimensional activation matrix that captures visual features. The Completely Connected Layer connects all participations to neurons, allowing for more global interactions. The Final Output Layer predicts the likelihood of each image belongingtodifferentclasses.Modelbuildingandtuningis an inefficient process, starting with a simple architecture and gradually adding complexity. The best-performing modelachieved a validationaccuracyofslightlyover85%.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 3, 2021

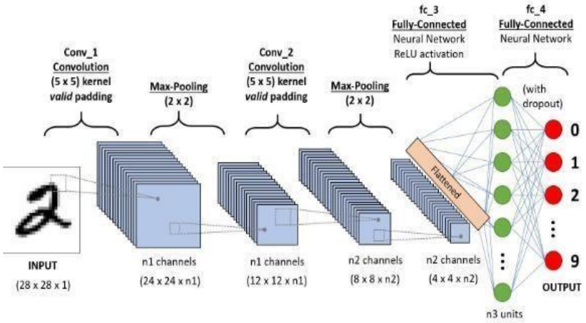


Figure6:CNNArchitecture

An algorithm for linear classification is support vector machine. The following can be used to condense the mathematical steps in an SVM algorithm:

1. Inputdatarepresentation:

LetXbead-dimensionalfeaturespaceandYbethetarget spacewhere$y = -1,+1$.The data is representedas asetofn samples$(x1,y1),(x2,y2),......,(xn,yn)$where xiisad-dimensional feature vector and yi i is the target value.

2. Constructingtheoptimizationproblem:

This can be created as a minimization with constraints. The objectiveistofindtheweightswandbiasbsuchthat: maximize:$\frac{1}{||W||}$

Subjectto:$(WXi+b)$ i=1,2,....,$n$where$||W||$is theEuclideannormofw.

3. Solvingtheoptimizationproblem:

Using a quadratic programming (QP) solver, the optimization problem can be resolved. The solution to the problem gives the values of w and b that define the hyper plane

4. Makingpredictions:Givenanewsamplex, itsclasscan be predicted as:

$$y=(^{w}+b) \quad \frac{}{cx} \tag{7}$$

Wheresign(x)returns+1$if x$≥0and−1$if x$<0

5. Informationthatcannotbedetachedlinearly:Datathat cannot be separated linearly can be translated into a high-dimensionalspacewherealinearboundarycanbefoundusing a kernel method.

TheReLUisanactivationfunctionthatoutputstheinput value if it is positive and 0 otherwise. ReLU is computationally efficient associated to further activation functions, which have more complex formulas and higher computational costs. ReLU is also advantageous as it is not affected by vanishing gradients, unlike Sigmoid and Tanh, whichcanslowdownlearninginanetwork.Itsformula,f(x) =max(0,x),ensuresthattheoutputrangeisfrom0toinfinity. ReLUiswidelyemployedinneuralnetworks,particularlyin Convolutional Neural Networks (CNNs), and is often the default choice for an activation function. Figure 6 and 7 representsthe graphofthe ReLUactivationand graphofthe sigmoid activation function respectively.

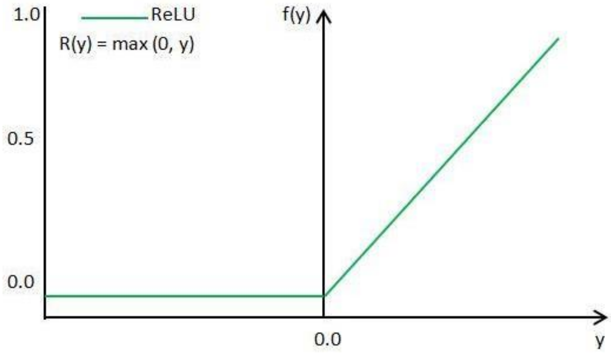


Figure7: ReLUactivationfunction

ReLU offers computational efficiency and faster training/operation due to its simple arithmetic. It promotes sparsity, which means manyweights inthe network become zero, leading to compact models with enhanced predictive abilityandreducedoverfitting.Inasparsenetwork,neurons aremorelikelytofocusonimportantfeaturesoftheproblem, resulting in more meaningful processing. For example, in a face detection model, certain neurons may specialize in identifying specific facial components; remaining inactive when irrelevant features are present.

The logistic function is a commonly used example of a sigmoid function, defined by the formula:

$$(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1} = 1-(-X) \tag{8}$$

Sigmoid functions are typically monotonic and have bell-shapedfirst derivatives.Thecumulativedistributionfunctions ofvariousprobabilitydistributions,suchastheerrorfunction and the arctan function, are also sigmoidal. A sigmoid function is characterized by horizontal asymptotes and exhibits convexity for values less than a certain point, and concavity for values greater than that point, often 0.

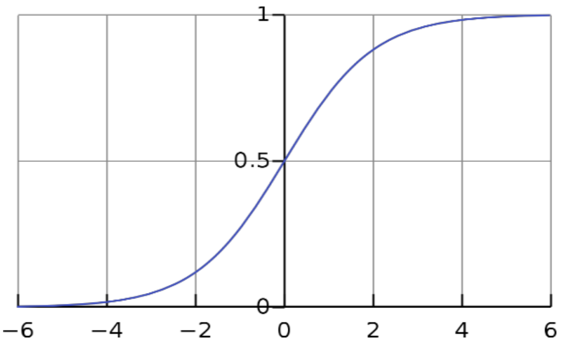


Figure8:Sigmoidactivationfunction.

Classification

The given signal will be analyzed by the machine and it willbeclassifiedintothefollowing7classes.Theoutputmay be either one of these will be displayed according to given emotion. Angry, calm, fearful, happy, sad,disgust,Surprise.

EvaluationMetrics

Thegoalofassessmentistoidentifyasmanyinstancesas possible from a population for a screening technique; hence false negatives should be kept to a minimum at the cost of increasing false positives. As a result, three main measurements must be established: accuracy (ACC), false positiverate(FPR)andtruepositiverate(TPR).Inmedical

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 10, Iss 3, 2021

language,thefirst parameterisreferredtoassensitivity (SEN) and is written as Equation:

$$TPR=SEN=\frac{TP}{P} \quad (9)$$

whereTPstandsfortruepositivesandPisforpositiveevents. The estimation of the second period, false positive amount, expressed as Equation:

$$FPR=\frac{FP}{N} \quad (10)$$

Thepopulation'scumulativenumberofnegativeoccurrences isN,whiletheproportionoffalsepositivesisFP,andnumber oftruenegativesamplesisN.Thisstatistic,ontheotherhand, is better understood as the ratio of genuine negatives to real negatives, known in medical language as the specificity (SPEC), which is given as Equation:

$$TNR=SPEC=\frac{TN}{N}=1-FPR \quad (11)$$

Finally, accuracy determines the stability between actual positives and accurate negatives. Figure 9 shows the Evaluation Matrix of the proposed method.

$$ACC=\frac{(TP+TN)}{(P+N)} \quad (12)$$

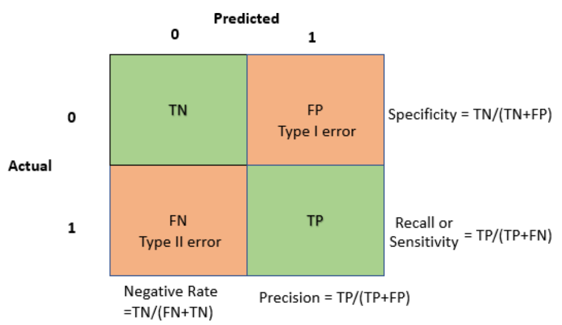
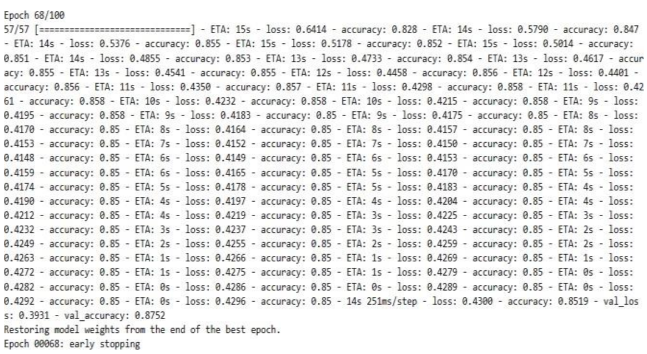


Figure9: EvaluationMatrix

IV. RESULTS ANDDISCUSSIONS

TherecommendedtechniquewasverifiedontheRAVDESS database and accomplished a correctness of 71% with a f1-scoreof0.71.Themethodoutperformstheexistingmethods [5] and [7] which have accuracyrates ofmorethan80% and 78.20% respectively. The error rate of the proposed method was0.5. Figure10depictstheEpoch havinganaccurateness of71%. And a f1-scoreof0.71.The modelhas beenworked for the following dataset and obtained model is having an accuratenessof71%andaf1-scoreof0.71.TheAccurateness Arc and The Defeat Arc are shown in the Figure 11 and 12 respectively.



```
Epoch 68/100
57/57 [==============================] - ETA: 15s - loss: 0.6414 - accuracy: 0.828 - ETA: 14s - loss: 0.5790 - accuracy: 0.847
- ETA: 14s - loss: 0.5376 - accuracy: 0.855 - ETA: 15s - loss: 0.5178 - accuracy: 0.852 - ETA: 15s - loss: 0.5014 - accuracy:
0.851 - ETA: 14s - loss: 0.4855 - accuracy: 0.853 - ETA: 13s - loss: 0.4733 - accuracy: 0.854 - ETA: 13s - loss: 0.4617 - accur
acy: 0.855 - ETA: 13s - loss: 0.4541 - accuracy: 0.855 - ETA: 12s - loss: 0.4458 - accuracy: 0.856 - ETA: 12s - loss: 0.4401 -
accuracy: 0.856 - ETA: 11s - loss: 0.4350 - accuracy: 0.857 - ETA: 11s - loss: 0.4298 - accuracy: 0.858 - ETA: 11s - loss: 0.42
61 - accuracy: 0.858 - ETA: 10s - loss: 0.4232 - accuracy: 0.858 - ETA: 10s - loss: 0.4215 - accuracy: 0.858 - ETA: 9s - loss:
0.4195 - accuracy: 0.858 - ETA: 9s - loss: 0.4183 - accuracy: 0.85 - ETA: 9s - loss: 0.4175 - accuracy: 0.85 - ETA: 8s - loss:
0.4170 - accuracy: 0.85 - ETA: 8s - loss: 0.4164 - accuracy: 0.85 - ETA: 8s - loss: 0.4157 - accuracy: 0.85 - ETA: 8s - loss:
0.4153 - accuracy: 0.85 - ETA: 7s - loss: 0.4152 - accuracy: 0.85 - ETA: 7s - loss: 0.4150 - accuracy: 0.85 - ETA: 7s - loss:
0.4148 - accuracy: 0.85 - ETA: 6s - loss: 0.4149 - accuracy: 0.85 - ETA: 6s - loss: 0.4153 - accuracy: 0.85 - ETA: 6s - loss:
0.4159 - accuracy: 0.85 - ETA: 6s - loss: 0.4165 - accuracy: 0.85 - ETA: 5s - loss: 0.4170 - accuracy: 0.85 - ETA: 5s - loss:
0.4174 - accuracy: 0.85 - ETA: 5s - loss: 0.4178 - accuracy: 0.85 - ETA: 5s - loss: 0.4183 - accuracy: 0.85 - ETA: 4s - loss:
0.4190 - accuracy: 0.85 - ETA: 4s - loss: 0.4197 - accuracy: 0.85 - ETA: 4s - loss: 0.4204 - accuracy: 0.85 - ETA: 4s - loss:
0.4212 - accuracy: 0.85 - ETA: 4s - loss: 0.4219 - accuracy: 0.85 - ETA: 3s - loss: 0.4225 - accuracy: 0.85 - ETA: 3s - loss:
0.4232 - accuracy: 0.85 - ETA: 3s - loss: 0.4237 - accuracy: 0.85 - ETA: 3s - loss: 0.4243 - accuracy: 0.85 - ETA: 2s - loss:
0.4249 - accuracy: 0.85 - ETA: 2s - loss: 0.4255 - accuracy: 0.85 - ETA: 2s - loss: 0.4259 - accuracy: 0.85 - ETA: 2s - loss:
0.4263 - accuracy: 0.85 - ETA: 1s - loss: 0.4266 - accuracy: 0.85 - ETA: 1s - loss: 0.4269 - accuracy: 0.85 - ETA: 1s - loss:
0.4272 - accuracy: 0.85 - ETA: 1s - loss: 0.4275 - accuracy: 0.85 - ETA: 1s - loss: 0.4279 - accuracy: 0.85 - ETA: 0s - loss:
0.4282 - accuracy: 0.85 - ETA: 0s - loss: 0.4286 - accuracy: 0.85 - ETA: 0s - loss: 0.4289 - accuracy: 0.85 - ETA: 0s - loss:
0.4292 - accuracy: 0.85 - ETA: 0s - loss: 0.4296 - accuracy: 0.85 - 14s 251ms/step - loss: 0.4300 - accuracy: 0.8519 - val_los
s: 0.3931 - val_accuracy: 0.8752
Restoring model weights from the end of the best epoch.
Epoch 00068: early stopping
```

Figure10: Epoch

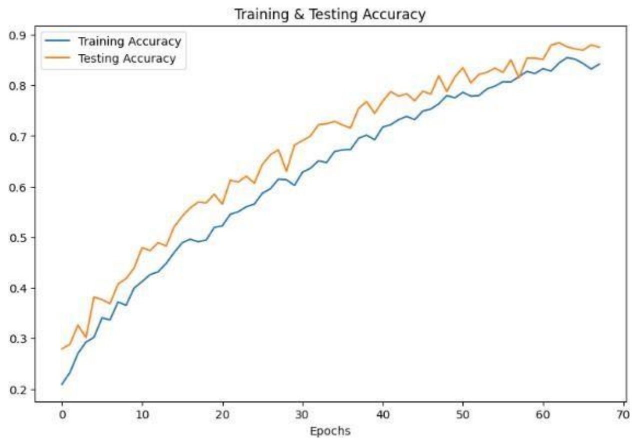


Figure11: AccuracyArchofthe CNN Model

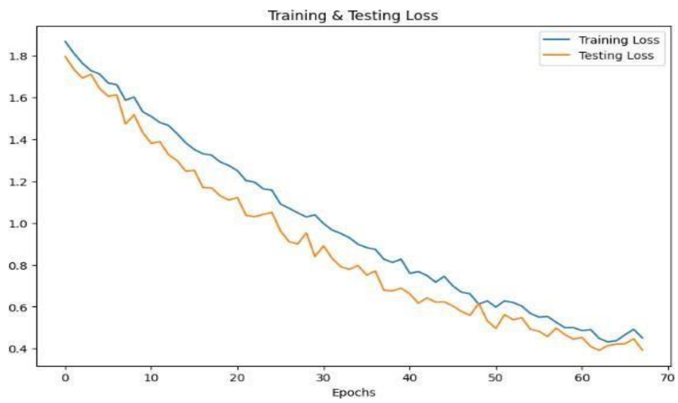


Figure12: LossArchofthe CNN Model

ConfusionmatrixforCNNmodelisrepresentedinthefigure 13 with six emotion details.

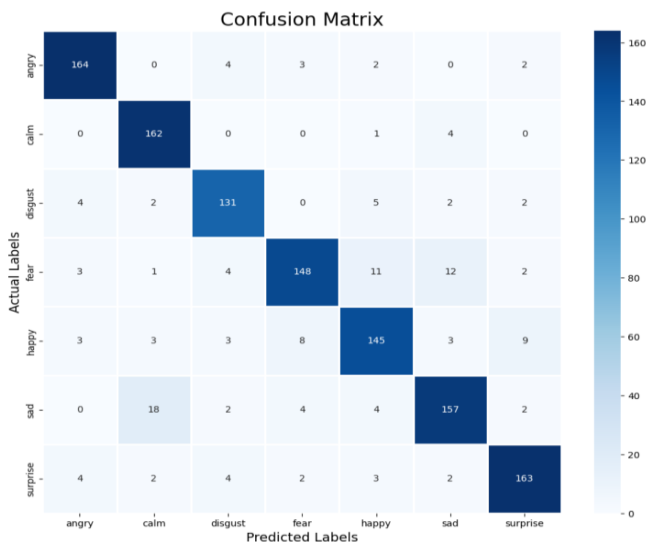


Figure13: Confusionmatrixfor CNNmodel

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 10, Iss 3, 2021

The Classification reportfor CNN model and Classification matrix for SVM is represented in the figure 14 and 16 respectively.

	precision	recall	f1-score	support
angry	0.92	0.94	0.93	175
calm	0.86	0.97	0.91	167
disgust	0.89	0.90	0.89	146
fear	0.90	0.82	0.86	181
happy	0.85	0.83	0.84	174
sad	0.87	0.84	0.86	187
surprise	0.91	0.91	0.91	180
accuracy			0.88	1210
macro avg	0.88	0.89	0.88	1210
weighted avg	0.88	0.88	0.88	1210

Figure14: ClassificationreportforCNNmodel

Comparison result with proposed method with existing techniques with the newest findings in the field as shows in table 1.

TABLE I. COMPARISON BETWEEN PROPOSED AND EXISTINGWORK

Method/Algorithm	Dataset	Accuracy	Error
DeepC-RNNapproach[5]	RAVDESS	80%	20%
Convolutional neural network[7]	RAVDESS	78.2%	21.8%
Residual Convolutional NeuralNetwork(R-CNN)	FAU	85.8%	14.2%
CNN[Proposed]	RAVDESS	88.42%	11.58

V. CONCLUSIONSANDFUTURESCOPE

The proposed model was built using both SVM and CNN algorithms and was tested on the RAVDEES dataset. The highest accuracy of 85% was achieved using the CNN algorithm. There is scope for improvement by adding more datatothedataset and increasing robustnessbyadding more noise.ThemodelhasbeenbuiltusingSVMandCNN.When the model is done using SVM we obtained an accuracy of 72%. When the model is built using the CNN the accuracy we obtained is 85%. The dataset used is RAVDEES. When we done it with the other datasets the accuracy went below 70%. So we went ahead with RAVDEES data with CNN algorithm.

In future work researchers can increase the accurateness of the archetypal by adding a higher size of data which we cannot be able to do due to limitations of the ability of our device. Also we can make the model more robust adding a more noise to the dataset.

REFERENCE

[1] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotionrecognitionwithdeepconvolutionalneuralnetworks.Biom edicalSignal Processing and Control, 59, 101894.

[2] Bhandari,S.U., Kumbhar,H.S.,Harpale,V. K.,&Dhamale,T. D. (2022). On the Evaluation and Implementation of LSTMModel for Speech Emotion Recognition Using MFCC. InProceedings of International Conference on ComputationalIntelligenceandDataEngineering(pp.421-434). Springer,Singapore

[3] Bharti, D., & Kukana, P. (2020, September). A hybrid machinelearning model for emotion recognition from speech signals. In2020 International Conference on Smart Electronics andCommunication (ICOSEC) (pp. 491-496). IEEE.

[4] Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Featureextractionalgorithmstoimprovethespeechemotionrecognit ionrate. International Journal of Speech Technology, 23(1), 45-55.

[5] Kumaran,U.,RadhaRammohan,S.,Nagarajan,S.M.,&Prathik, A. (2021). Fusion of mel and gammatone frequency cepstralcoefficientsforspeechemotionrecognitionusingdeepC-RNN. InternationalJournalofSpeechTechnology,24(2),303-314

[6] Christy,A.,Vaithyasubramanian,S.,Jesudoss,A.,&Praveena, M. D. (2020). Multimodal speech emotion recognition andclassification using convolutional neural network techniques.International Journal of Speech Technology, 23(2), 381-388.

[7] Sun, T. W. (2020). End-to-end speech emotion recognition withgender information. IEEE Access, 8, 152423- 152438.

[8] M.A.Jalal,R.Milner,andT. Hain,"Empiricalinterpretation ofspeechemotionperceptionwithattentionbasedmodelforspeeche motion recognition," Proc. Interspeech 2020, 2020.

[9] Aggarwal,A.,Srivastava,A.,Agarwal,A.,Chahal,N.,Singh,D.,Aln uaim, A. A., ... & Lee, H. N. (2022). Two-way featureextractionforspeechemotionrecognitionusingdeeplearning . *Sensors*, *22*(6), 2378.

[10] Yadav, S. P., Zaidi, S., Mishra, A., & Yadav, V. (2022). Surveyon machine learning in speech emotion recognition and visionsystems using a recurrent neural network (RNN). *Archives ofComputational Methods in Engineering*, *29*(3), 1753-1770.

[11] Abdelhamid,A.A.,El-Kenawy,E.S.M.,Alotaibi,B.,Amer,G.M.,Abdelkader,M.Y.,Ibrahi m,A.,&Eid,M.M.(2022).Robustspeech emotion recognition using CNN+ LSTM based onstochasticfractalsearchoptimizationalgorithm. *IEEEAccess*, *10*, 49265-49284.

[12] Bhangale, K., & Mohanaprasad, K. (2022). Speech emotionrecognition using mel frequency log spectrogram and deepconvolutional neural network. In*Futuristic Communication andNetworkTechnologies:SelectProceedingsofVICFCNT2020* (pp. 241-250). Springer Singapore.

[13] Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning basedmultimodal emotion recognition using model-level fusion ofaudio–visualmodalities.*Knowledge-BasedSystems*,*244*,108580.

[14] Bakhshi, A., Harimi, A., & Chalup, S. (2022). CyTex:Transforming speech to textured images for speech emotionrecognition. *Speech Communication*, *139*, 62-75.

[15] Bagadi, K. R., & Sivappagari, C. M. R. (2023). An evolutionaryoptimization method for selecting features for speech emotionrecognition. TELKOMNIKA (Telecommunication ComputingElectronics and Control), 21(1), 159-167.

[16] Zhong,Z.(2023,January).SpeechemotionrecognitionbasedonSV M and CNN using MFCC feature extraction. In InternationalConference on Statistics, Data Science, and ComputationalIntelligence (CSDSCI 2022) (Vol. 12510, pp. 445-452). SPIE.

[17] Kukreja, V., & Dhiman, P. (2020, September). A Deep NeuralNetworkbaseddiseasedetectionschemeforCitrusfruits.In20 20Internationalconferenceonsmartelectronicsandcommunication (ICOSEC) (pp. 97-101). IEEE.

[18] Dhiman,P.,Kukreja,V.,Manoharan,P.,Kaur,A.,Kamruzzaman, M.M.,Dhaou,I.B.,&Iwendi,C.(2022).Anoveldeeplearningmodelf ordetectionofseveritylevelofthediseaseincitrusfruits.Electronics, 11(3), 495.

[19] A. Panwar, R. Yadav, K. Mishra, and S. Gupta, "Deep learningtechniquesfortherealtimedetectionofCovid19andpneumo

niausingchestradiographs,"inEUROCON2021-19thIEEE

ISSN PRINT 2319 1775 Online 2320 7876

Volume 10, Iss 3, 2021

InternationalConferenceonSmartTechnologies,Proceedings,2021, pp. 250–253, doi: 10.1109/EUROCON52738.2021.9535604

[20] C.Bhatt,I.Kumar,V.Vijayakumar,K.U.Singh,andA.Kumar,“The state ofthe art ofdeep learning models in medical scienceand their challenges,” Multimed. Syst., vol. 27, no. 4, pp. 599–613, 2021, doi: 10.1007/s00530-020-00694-1

[21] Sharma, N., Chakraborty, C., & Kumar, R. (2022). Optimizedmultimedia data through computationallyintelligent algorithms.Multimedia Systems, 1-17