International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-03, JUNE 2021

Performance analysis of various classifiers on IEMOCAP database for Robust Emotion Recognition

Divya Gupta, Research Scholar, Jagan Nath University,Jaipur, India, fromdivya81@gmail.com

Poonam Bansal, Professor,GGSIPU University,Delhi, India, pbansal89@gmail.com

Kavita Choudhary, Associate Professor, Jayoti Vidyapeeth University,India,Kavita.yogen@gmail.com

Abstract Because emotions are transmitted through a mixture of verbal and nonverbal channels, understanding expressive human communication necessitates a simultaneous investigation of speech and gestures. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a recently gathered acted, multimodal, and multi speaker database from USC's SAIL lab. The audiovisual content includes around 12 hours of video, voice, face movements and text transcriptions. It consists of dyadic sessions in which actors perform improvisations or scripted scenarios which have been chosen intentionally to evoke emotions. In this research study various classifiers are used on text transcript in order to calculate the precision, accuracy, recall, f1-score also expressed in terms of confusion matrices. It is found that the accuracy will be nearly 100% with deep learning using lugwig classifier. The results are present in order to prove above.

Keywords —IEMOCAP, Sail lab, machine learning, deep learning

I. INTRODUCTION

The emotional condition of the subjects, which is transmitted through both speech and gestures, is one of the most intriguing paralinguistic signals presented during human contact [1]. The tone and energy of the voice, facial emotions, body posture, head position, hand gestures, and gaze are all mixed in an unusual way during real human communication. These communication routes must be investigated collaboratively if strong emotional models are to be established and deployed. Natural language is frequently reflective of one's emotional state [2].

As a result of its numerous applications in opinion mining, recommender systems, health-care, and other fields, emotion recognition has grown in prominence in the field of NLP. Strapparava and Mihalcea tackled the challenge of detecting emotions in news headlines [3]. To address the difficulty of textual emotion recognition, a variety of emotion lexicons have been developed.

Because of the increased public availability of conversational data, emotion recognition in conversation (ERC) has just recently attracted interest from the NLP community [4]. ERC can be **used** to evaluate social media talks. It can also help with real-time dialogue analysis, which can be useful in court proceedings, interviews, e-health services, and other situations [5].

In contrast to vanilla emotion identification of sentences/utterances, ERC ideally necessitates context modelling of the individual utterances [6]. This context is based on the temporal sequence of statements and can be attributed to the preceding utterances.

In comparison to recently published works on ERC, both lexicon-based and modern deep learning-based vanilla emotion recognition approaches fail to perform well on ERC datasets because these works ignore conversational factors such as the presence of contextual cues, the temporality of speaker turns, or speaker-specific information [7].

In this research study various ML classifier along with impact of Deep learning Ludwig classifier is used to calculate the various factors on text transcript of IEMOCAP in order to calculate the precision, accuracy ,recall, f1-score also expressed in terms of confusion matrices.

II Speech Recognition and Emotion Detection

Automated speech recognition (ASR), a computer speech recognition, or a word-to-text recognition is a facility for manipulation of the voice in a written format by a software [8]. Although it is sometimes mistaken with voice recognition, the goal of voice recognition is on translating speech from an oral to a written format, whereas voice

 DOI : 10.35291/2454-9150.2021.0325

recognition just aims at identifying the voice of a particular user.

Since the commencement of the 1962 release of Shoebox, the IBM has played a key role in the recognition of speech. The system was able to distinguish 16 distinct words and was able to advance Bell Labs' original work from the 1950s. However, IBM has not stopped here, but has throughout the years continued to develop, introducing Voice Type Simply Speaking in 1996 [9]. The programmed has a 42,000-word language, a spelled dictionary of 100,000 terms and supplemented English and Spanish.

Many software and devices for voice recognition are available but increasingly advanced ones incorporate AI and machine learning [10]. They combine audio and voice signals into the grammar, syntax, structures and composition to comprehend the human speech and process it. Ideally, you will learn as you go – with each encounter evolving replies.

The greatest sort of systems also allows organizations, from voice and subtleties to brand identification, to create and adapt technologies to their own requirements [11]. Examples include:

- Specific weighing phrases that are often uttered (such as product names or industry jargon) are improved by improving accuracy, beyond any phrases present in the basic vocabulary.
- Speaker labelling: Call or tag the contributions of each speaker in a multi-stakeholder discourse.
- Training in acoustics: attend the business's acoustic aspect. Train the system to adjust to an acoustic environment and speaker styles (such as ambient noise at the contact centre) (like voice pitch, volume and pace).
- Filtering profanity: Use filters to identify certain words or phrases and to cleanse the output of speech.

Algorithms

Human voice fluctuations have challenged progress. It is one of the most complicated fields in computer science – languages, mathematics and statistics. Speaking recognizers consist of a handful of components, such as input, extracting features, feature vectors, decoder, and output of words [12]. To determining the proper output, the decoder uses acoustic models, a dictionary of pronunciation and language model.

Different methods and techniques of computing are employed to detect voice in text and to enhance transcription accuracy. Below are some of the most widely utilized approaches for concise explications:

Natural language processing (NLP): While NLP is not necessarily a certain voice recognition method, it is the field of artificial intelligence that focuses on the interaction between humans and machines via voice and text. In their Systems many mobile devices feature voice search recognition — for example, Siri — or make messaging more accessible [13].

Models of Hidden Markov (HMM): Markov Hidden Models build on the Markov chain model, which argues that the likelihood of a particular state depends not on its past conditions, but on the present state. While the Markov Chain Model may be used for observable events like text inputs, hidden Markov Models allow hidden events like parts of speech tags to be incorporated into a probabilistic model. They are used to assign labels, i.e., words, syllables, phrases, etc., in the sequence, as sequence model in speech recognition [14]. These labels map the supplied input so that the best suitable label sequence may be determined.

N-grams: This is the most elementary kind of linguistic model (LM) with probability for phrases or sentences. An N-gram is the N-word sequence. For instance, "pizza order" is either a 3-g or a trigram and "pizza order" is a 4-g. In order to increase recognition and precision, grammar and the likelihood of particular word sequences are employed.

Neural networks: primarily utilized by neural networks for deeper learning algorithms, training data is processed by imitating the human brain's interconnections through node layers. The inputs, weights, bias and output of each node are included. If this value is greater than the stated threshold, it "fires" or turns the node on, transmitting data in the network to the next tier. Neural networks learn to map through supervised learning and adapt via gradient downgrade depending on the loss function. Although neural networks tend to be more precise and can absorb more data, they tend to be slower to train as compared to tradition methods.

Diarization of the speaker (SD): Algorithms of the speaker diarization identify and talk by speaker identity. This helps to identify more people in a discussion in programmes and is often employed at customers and sales workers in contact centres.

III Various Classifiers

AdaBoost Classifier

Adaptive Boosting is the AdaBoost algorithm, which is used for machine learning as an ensemble method. Adaptive Boosting is so named because the weights are reassigned to each instance, with larger weights applied to mistakenly categorized instances. Boosting is used in supervised learning to minimize bias as well as variation. It is based on the progressive development of learners. With the exception of the first, each succeeding student is developed from previously developed learners. In other words, weak learners are transformed into strong ones. The adaboost

 DOI : 10.35291/2454-9150.2021.0325

algorithm operates on the same principles as boosting, however there is a subtle variation in how it works. Stacking approach used in Machine learning is shown in Figure 1.

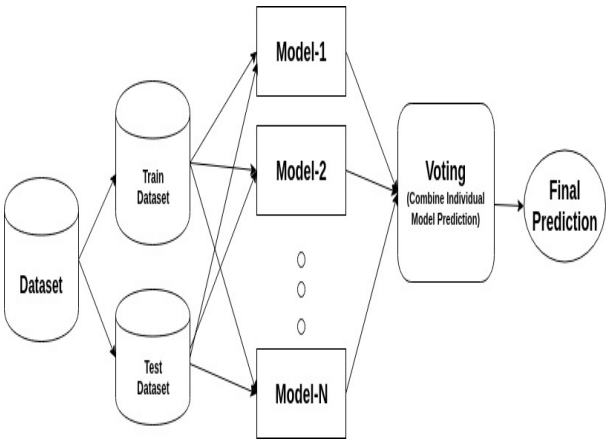


Figure 1 stacking approach used in Machine learning

The classification report and confusion matrix obtained using adaboost classifier is shown below on IEMOCAP database.

Classification Report_____

	precision	recall	f1-score	support
Negative	0.99	0.95	0.97	6163
Positive	0.36	0.71	0.48	230
accuracy			0.94	6393
macro avg	0.67	0.83	0.72	6393
weighted avg	0.97	0.94	0.95	6393

Evaluation Metrics_____

Accuracy: 0.944001

Weighted Precision :0.966245

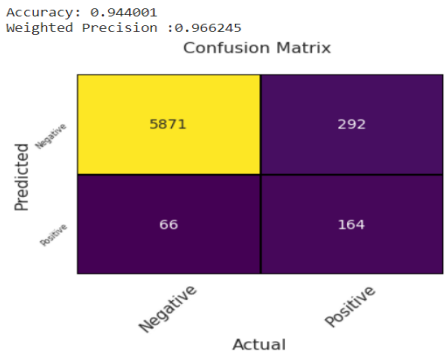


Figure 2 Confusion Matrix obtained using adaboost classifier

Neural Net:

Inspired by the process of learning in the human brains, neural networks. It consists of an artificial system of functions, termed parameters, enabling the computer to learn and adapt by processing fresh data. Each parameter is a function that creates an output after receiving one input or more inputs, sometimes also called neurons. Those outputs are subsequently transmitted to the next neural layer, where they are used as inputs of their own function. These results then are sent to the next neuronal layer and thus proceed until each neuronal layer is evaluated and the neurons input is received. The simple neural net is shown in Figure 3.

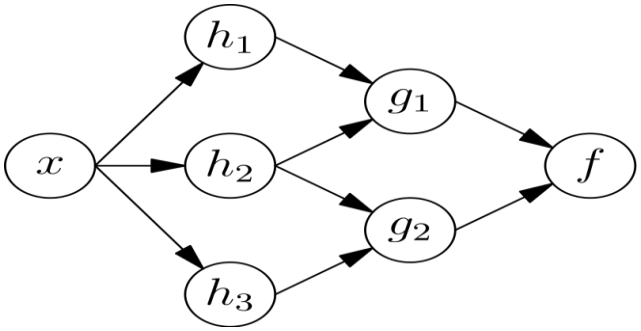


Figure 3 Simple Neural Net

The classification report and confusion matrix obtained using neural net classifier is shown below on IEMOCAP database.

Classification Report_____

	precision	recall	f1-score	support
Negative	1.00	0.93	0.96	6393
Positive	0.00	0.00	0.00	0
accuracy			0.93	6393
macro avg	0.50	0.46	0.48	6393
weighted avg	1.00	0.93	0.96	6393

Evaluation Metrics_____

Accuracy: 0.928672

Weighted Precision :1.000000

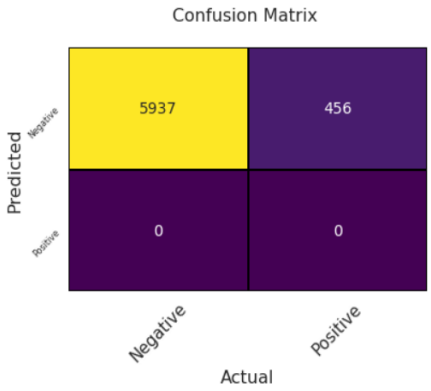


Figure 4 Confusion Matrix obtained using neural net classifier

Decision Tree Classifier

Decision Tree method belongs to the guided study algorithm family. The decision tree method may be used to solve issues like regression and classification, in contrast to other monitored learning techniques.

The objective of the Choice Tree is to construct an education model which can be used by learning basic decision rules based on predictions of the class or value of the destination variable (training data).

 DOI : 10.35291/2454-9150.2021.0325

In Decision Trees, we start from the tree's root to forecast a class label for a record. We compare the root attribute values with the attribute of the record. We follow the branch corresponding to this value based on comparison and jump to the next n.

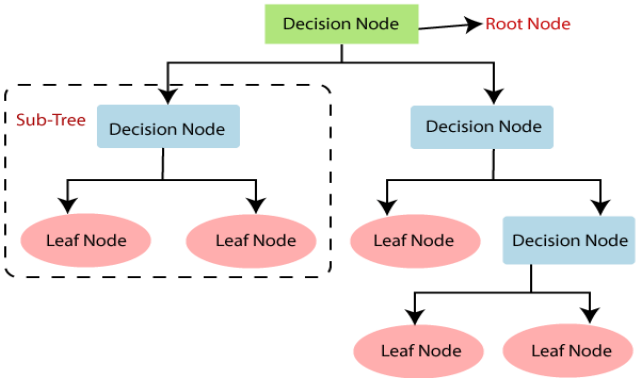


Figure 5 Working of DT

_Classification Report_____

	precision	recall	f1-score	support
Negative	0.97	0.97	0.97	5965
Positive	0.58	0.61	0.60	428
accuracy			0.94	6393
macro avg	0.77	0.79	0.78	6393
weighted avg	0.95	0.94	0.94	6393

Evaluation Metrics_____

Accuracy: 0.944001
Weighted Precision :0.945733

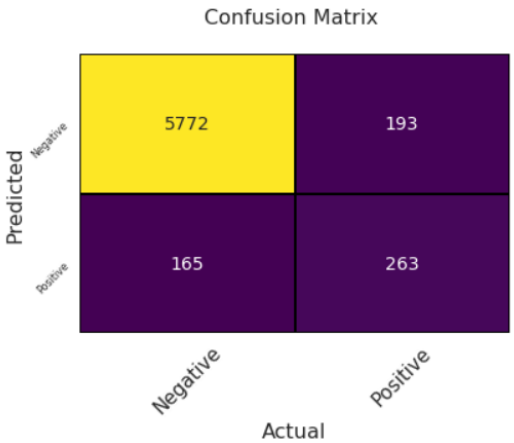


Figure 6 Confusion Matrix obtained using DT classifier

RBF SVM

The RBF kernel SVM decision region is a linear decision area as well. In fact, what the RBF Kernel SVM accomplishes is to construct nonlinear combinations of your samples in a higher-dimensional space where you may utilize a linear border of decision to split your classes:
If the dataset is linear or otherwise inseparable, the data set is nonlinear, kernel functions such as RBF are advised. One might use the linear kernel function (kernel="linear") for a linearly separable dataset (linear dataset). A clear grasp of when kernel functions should be used will assist you train the best model using the SVM technique.

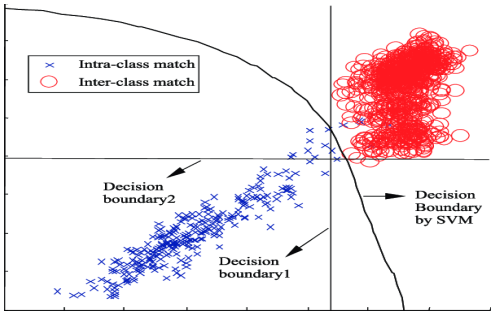


Figure 7 classification by RBF SVM

Classification Report_____

	precision	recall	f1-score	support
Negative	1.00	0.96	0.98	6171
Positive	0.45	0.92	0.60	222
accuracy			0.96	6393
macro avg	0.72	0.94	0.79	6393
weighted avg	0.98	0.96	0.96	6393

Evaluation Metrics_____

Accuracy: 0.958079
Weighted Precision :0.978122

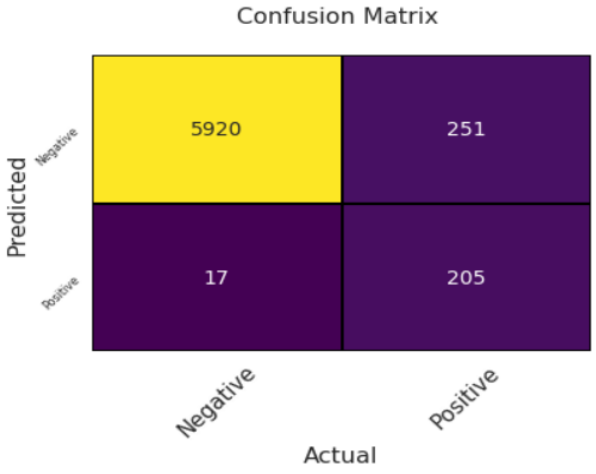


Figure 8 Confusion Matrix obtained using RBS SVM classifier

Nearest Neighbors Classifier

K-Nearest Neighbors is a simple yet important classification technique in Machine Learning. It is a supervised learning algorithm that is widely used in pattern recognition, data mining, and intrusion detection.
It is extensively applicable in real-world circumstances since it is non-parametric, which means it makes no underlying assumptions regarding data distribution.
Assume there are two categories, A and B, and we get a new data point x1, and we want to know which of these two categories this data point belongs to. A K-NN method is

 DOI : 10.35291/2454-9150.2021.0325

required to address this sort of problem. The category or class of a certain dataset may simply be identified by using K-NN.

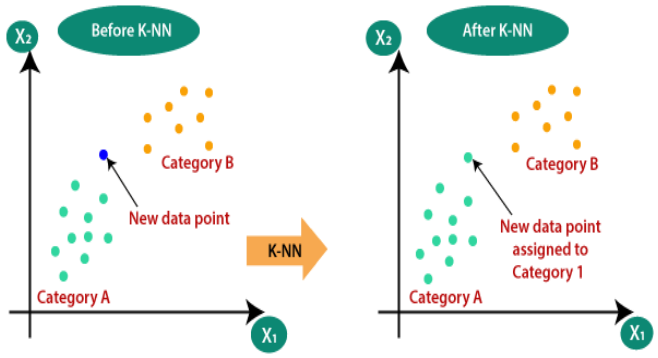


Figure 9 identify the category or class of a particular dataset using knn

Classification Report_____

	precision	recall	f1-score	support
Negative	1.00	0.94	0.97	6321
Positive	0.16	1.00	0.27	72
accuracy			0.94	6393
macro avg	0.58	0.97	0.62	6393
weighted avg	0.99	0.94	0.96	6393

Evaluation Metrics_____

Accuracy: 0.939934
Weighted Precision :0.990516

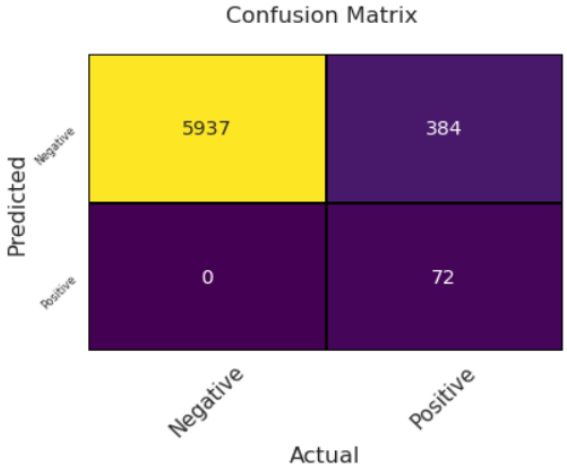


Figure 10 Confusion Matrix obtained using KNN classifier

Random Forest Classifier

Random forest is a versatile method that is easy to apply, which delivers a wonderful result, even without hyper parameters adjustment. It is also one of the most often utilised algorithms due to its simple and diverse nature. Random forests generate and blend several decision trees to provide a more precise and consistent forecast.

The random forests have a huge benefit in terms of their application in classification and regression issues, which make up the majority of today's machine learning systems.

Let's look at the classification of random forests because classification is sometimes viewed as the machine's building component. Underneath you can see how two trees seem like a random forest in Figure 11:

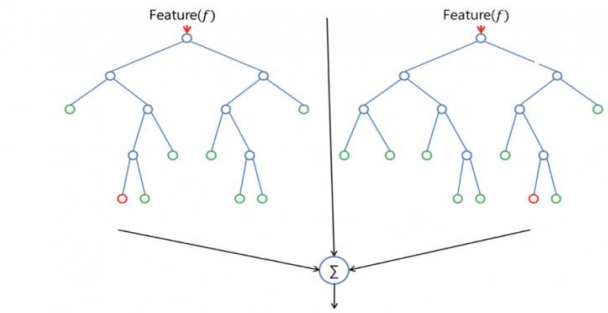


Figure 11 random forest look like with two trees

Random forest adds unpredictability to the model as trees grow. It looks at a random selection of characteristics instead of examining the main feature during splitting a node. This leads to a broad range, which often leads to a better model.

In the random forest, therefore, the technique is only used for dividing a node in a random subset of characteristics.

Classification Report_____

	precision	recall	f1-score	support
Negative	0.99	0.96	0.98	6125
Positive	0.52	0.88	0.65	268
accuracy			0.96	6393
macro avg	0.75	0.92	0.81	6393
weighted avg	0.97	0.96	0.97	6393

Evaluation Metrics_____

Accuracy: 0.960269
Weighted Precision :0.974358

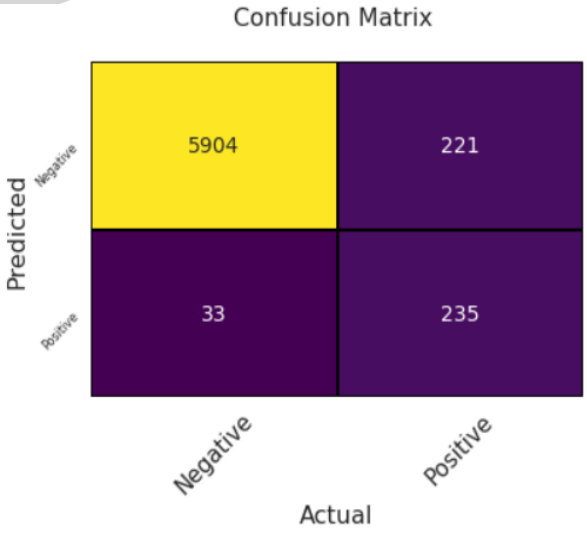


Figure 12 Confusion Matrix obtained using Random forest classifier

 DOI : 10.35291/2454-9150.2021.0325

Ludwig Classifier

Ludwig is a profound learning toolbox that allows models to be predicted and used without writing a word of code. It is built on top of TensorFlow and leverages a data-type-based abstraction to construct very many applications. Ludwig offers highly quick prototype and model iteration thanks to the declarative structure of the files. It is both suited for novices to train profound learning models without knowing all TensorFlow specifics and profound learning in general, and also enables experienced users to be far more productive, reducing jobs that would take months or minutes.

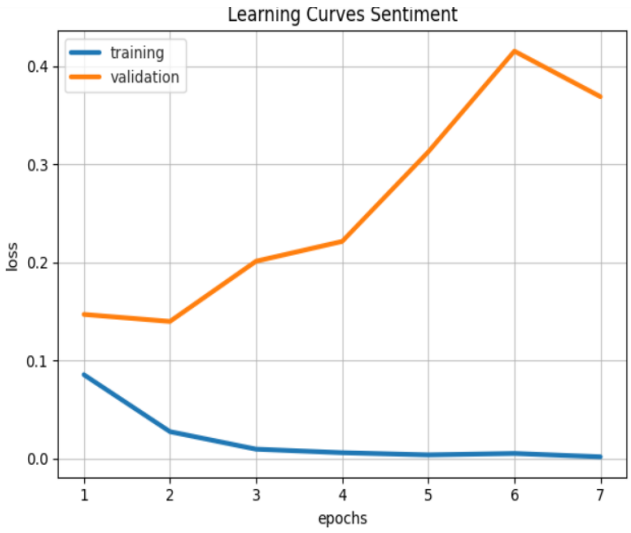


Figure 13 Learning Curve Sentiment

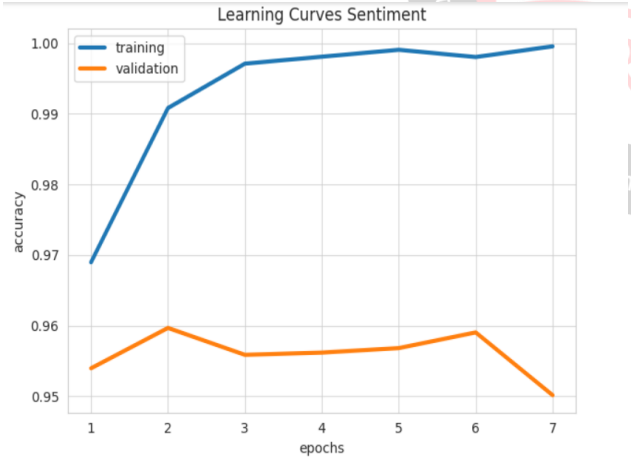


Figure 14 Learning Curve using Ludwig Classifier

The learning curve obtain through deep learning promise nearly 99.9% percent accuracy which is very good indicator at 7 epoch. The lose Vs epoch curve shown in Figure 13 shows lose in the training curve reduced to almost zero with increase in epoch, this itself describe the accuracy of the system. This clearly shown in Figure 14 that the accuracy increases when Ludwig classifier is used with number of epochs.

RESULT AND DISCUSSION

Several Classification methods, including Boosting, Logistic Regression, SVM, and Naive Bayes, were built in the Python programming language on Google Colab based on the pre-processed data sets. The findings of the top 8 algorithms reveal optimistic results in the scenario. The default functionality was used to train our classifier on the clean data set. Four measurements were utilised for evaluation: precision, F1-score, recall, accuracy, and ROC area. A 2-dimensional matrix that provides information on actual classes and expected classes of a classifier is an essential confusion matrix.

Accuracy is defined as the fraction of properly categorized predictions; it is calculated using the formula:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives.

Recall is an assessment of a classifier's order to consistently anticipate instances of a certain class; it is also known as the TPR (true positive rate):

Recall = TP/(TP+FN)

Precision is the fraction of positive predictions given by the classifiers that are truly positive:

Precision = TP/(TP+FP)

The harmonic average of accuracy and recall is the F1-score:

F1-Score = 2*Precision*Recall/ (Precision+Recall)

The various results are combined and shown in the Table 1

Table 1 Comparison among different classifier

Algorithm used	Precision	Recall	f1-score	Accuracy
AdaBoost	0.97	0.94	0.95	0.944001
Neural Net	1	0.93	0.96	0.928672
Decision Tree	0.95	0.94	0.94	0.944001
RBF SVM	0.98	0.96	0.96	0.958079
Linear SVM	1	0.93	0.96	0.931175
Nearest Neighbors	0.99	0.94	0.96	0.939934
Random Forest	0.97	0.96	0.97	0.960269
Linear SVC	0.97	0.96	0.97	0.96371

 DOI : 10.35291/2454-9150.2021.0325

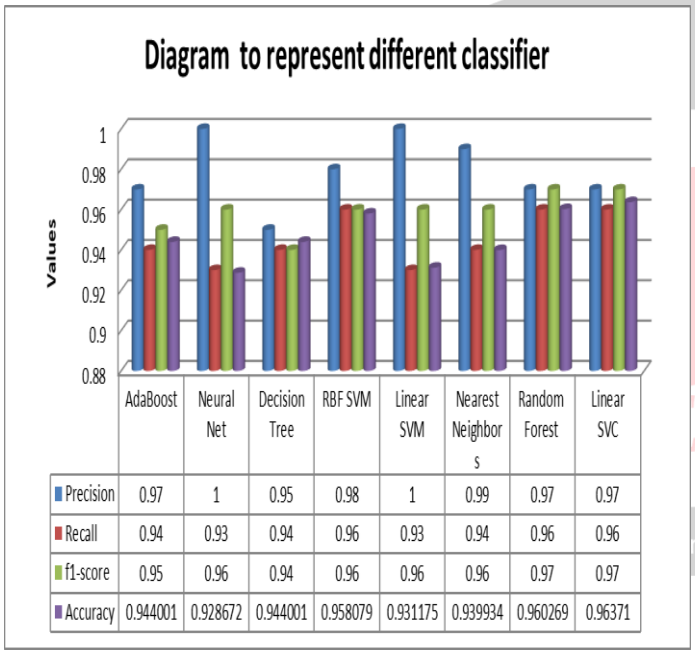
Inference from table 1

a) As far precision is concern it is found that Neural tet and Linear SVM have highest value which is unity . And for other classifiers it is above 0.95 except the Decision Tree which have 0.95.

b) Recall measures the ability of the classifier to predict correctly and it is max in RBF SVM, Random Forest and Linear SVC and in all others it is more than 0.93 which shows how accurate the prediction is.

c) F1 Score which is harmonic average of precision and recall has almost equal value and it is max in Random Forest and Linear SVC.

d) Accuracy which is proportion of correctly classified predictions is one of the important parameter for any machine learning application in this research study max accuracy is found to be about 97% in Random Forest and Linear SVC. Also, It is found that Accuracy nearly about 99.9% is achieved when a deep learning model using Ludwig Classifier is used as shown in Figure 14.



	AdaBoost	Neural Net	Decision Tree	RBF SVM	Linear SVM	Nearest Neighbors	Random Forest	Linear SVC
Precision	0.97	1	0.95	0.98	1	0.99	0.97	0.97
Recall	0.94	0.93	0.94	0.96	0.93	0.94	0.96	0.96
f1-score	0.95	0.96	0.94	0.96	0.96	0.96	0.97	0.97
Accuracy	0.944001	0.928672	0.944001	0.958079	0.931175	0.939934	0.960269	0.96371

Figure 15 Summary of different classifiers

CONCLUSION AND FUTURE SCOPE

In this paper, we have successfully used the various classification algorithms along with deep learning for emotion prediction. The analysis was implemented in the python programming language, and performance metrics like accuracy, recall, precision, f1-score were calculated. From the analysis, It is found that max accuracy is about 97% in Random Forest and Linear SVC. Also, It is found that Accuracy nearly about 99.9% is achieved when a deep learning model using Ludwig Classifier is used as shown in Figure 14 . This research study is very useful for the researchers working in the same Field.

REFERENCES

[1] Akansha Madan, Divya Gupta, ".Speech Feature Extraction and Classification: A Comparative Review", International Journal of computer applications, (0975-8887) Volume 90 – No 9, March 2014

[2] Y. Yuan, P. Zhao, and Q. Zhou, "Research of speaker recognition based on combination of LPCC and MFCC," in Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on, 2010, pp. 765-767

[3] Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson, IEEE Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Us-ing a Novel Distortion Model IEEE Transactions On Audio, Speech, And Language Processing, Vol. 16,No. 8, November 2008.

[4] IEEE Trans. Audio Speech Lang. Process., vol. 16, no. 1, pp. 65–73, Jan. 2008.

[5] L. S. Chee, Ooi Chia Ai, and S. Yaacob, "Overview of Automatic Stuttering Recognition System," in International Conference on ManMachine Systems (ICoMMS 2009) Penang, Malaysia, 2009.

[6] K. Hu and D. Wang, ''Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction,'' IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 6, pp. 1600–1609, Aug. 2011.

[7] Luengo and E. Navas, "Feature analysis and evaluation for automatic emotion identification in speech", IEEE Trans.on Multimedia, vol. 12,no. 6, pp. 267-270, Oct 2010.

[8] M. G. Sumithra and A. K. Devika, "A study on feature extraction techniques for text independent speaker identification," in Computer Communication and Informatics (ICCCI), 2012 International Conference on, 2012, pp. 1-5.

[9] Ajay Rupani, Deepa, Gajender and PawanWhig,"A Review of Technology Paradigm for IOT on FPGA", International Journal of Innovative Research in Computer and Communication Engineering, 2016, Vol. 5, Issue 9 pp.61-64. ISSN (Online): 2320-9801/ ISSN (Print): 2320-9798

[10] Pawan Whig and S. N Ahmad," Simulation and performance analysis of Multiple PCS sensor system, ", Electronics, 2016, Vol. 20,Issue 2 pp. 85-89(Scopus). ISSN: 1450-5843

[11] Biswas, S. Ahmad, and M. K. Islam Mollat, "Speaker Identification Using Cepstral Based Features and Discrete Hidden Markov Model," information and Communication Technology, 2007. ICICT International Conference on, 2007, pp. 303-306.

[12] Y. Yuan, P. Zhao, and Q. Zhou, "Research of speaker recognition based on combination of LPCC and MFCC," in Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference 2010, pp. 765-767.

[13] Pawan Whig and Ajay Rupani ," The development of big data science to the world", Engineering Reports, 2019,vol 2 no2, pp1-7

[14] PawanWhig ," Artificial intelligence and machine learning in business", Engineering Reports, 2019,vol 2 no2, pp8-13.

 DOI : 10.35291/2454-9150.2021.0325