SPEECH BASED EMOTION RECOGNITION USING MACHINE LEARNING

A Project Report

Submitted to the Faculty of Engineering of
JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY
In
COMPUTER SCIENCE AND ENGINEERING

By

G. HRITHIK
(18481A0583)

K. DEVI KIRAN
(18481A05A9)

K. N. V. NARESH
(18481A05B5)

Under the guidance of
Mrs. G. Bharathi, M.Tech, (Ph.D)
Sr. Gr. Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESHADRIRAO GUDLAVALLERU ENGINEERING COLLEGE
(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRIRAO KNOWLEDGE VILLAGE
GUDLAVALLERU – 521356
ANDHRA PRADESH

2021-2022

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project report entitled "SPEECH BASED EMOTION RECOGNITION USING MACHINE LEARNING" is a bonafide record of work carried out by G. Hrithik (18481A0583), K. Devi Kiran (18481A05A9), K. N. V. Naresh (18481A05B5) under the guidance and super vision of Mrs. G. Bharathi in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2021-22.

Project Guide
(Mrs. G. Bharathi)

Head of the Department
(Dr. M. BABU RAO)

External Examiner

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Mrs. G. Bharathi**, Sr. Gr. Assistant Professor, Department of Computer Science and Engineering for her constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. M. Babu Rao**, Head of the Department, Computer Science and Engineering for his encouragements all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. G.V.S.N.R.V Prasad** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project in time.

Team members

G. Hrithik (18481A0583)

K. Devi Kiran (18481A05A9)

K. N. V. Naresh (18481A05B5)

INDEX

LIST OF ABBREVIATIONS

Abbreviation	Explanation
ADLINE	Adaptive Linear Element
AI	Artificial Intelligence
AP	Acoustic Prosodic
ANN	Artificial Neural Networks
CNN	Convolutional Neural Networks
DL	Deep Learning
DT	Decision Tree
EAR	Emotion Association Rules
EEG	Electronic Encephalogram
FNN	Feed-Forward Neural Network
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Models
HCI	Human-Computer Interaction
HMM	Hidden Markov Models
HNR	Harmonics to Noise Ratio
KNN	K-Nearest Neighbours
LFCC	Log Frequency Power Coefficients
LPCC	Linear Prediction Cepstral Coefficients
MaxEnt	Maximum Entropy Model
MDT	Meta Decision Tree
ML	Machine Learning
MLP	Multi-Layer Perceptron

MFCC	Mel-Frequency Cepstral Coefficients
MNN	Modular Neural Network
NN	Neural Networks
PCA	Principal Component Analysis
PDP	Parallel Distributed Processing
RAVDESS	Ryerson Audio Visual Database of Emotional Speech and Song
RBF	Radial Basis Function
RBFNN	Radial Basis Function Neural Network
RNN	Recurrent Neural Networks
SL	Semantic Labels
SER	Speech Emotion Recognition
SVM	Support Vector Machine
TEO	Teager Energy Operator

LIST OF FIGURES

ABSTRACT

Speech Emotion Recognition (SER) plays an indispensable role in intelligent speech application. SER, is the act of attempting to recognize human emotion and affective states from speech. There are several modalities for expressing human emotions like body-posture, facial expression & voice. The human voice can be characterized by several attributes such as pitch, timbre, loudness, and vocal tone. It has often been observed that humans express their emotions by varying different vocal attributes during speech generation. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. SER is tough because emotions are subjective and annotating audio is challenging.

The proposed approach is based upon python modules like PyAudio, Librosa for audio input and analysing it and the MLP Classifier for Feature Extraction and Classification. For evaluated models of different experiment settings reports accuracy, f-score, precision and recall. However, the performance of previous work has been restricted by neglecting the interaction of different frequencies, since the converged communication of frequency is also critical for us to generate accurate emotion feature representations. The proposed method includes the previously neglected frequencies which can lead to accurate prediction of emotions.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION :

Emotion Recognition has been the subject of exploration for quite a long time. Emotion Recognition from speech signals has been studied to a great extent during recent times. Humans have a unique ability to convey themselves through speech. These days alternative communication methods like text messages and emails are available. Further, instant messages are aided by emojis that have paved the way for visual communication in this digital world. However, Speech is still the most significant part of human culture and is data rich. In recent times, speech emotion recognition (SER), which expects to investigate the emotion states through speech signals, has been drawing increasing consideration. Nevertheless, SER remains a challenging task, with the question of how to extract effective emotional features.

Emotion Recognition from Speech has evolved from being an experimental study to an important component for Human-Computer Interaction (HCI). In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications. At present, speech emotion recognition is an emerging crossing field of artificial intelligence and artificial psychology. Besides, it is a popular research topic of signal processing and pattern recognition. The research is widely applied in human-computer interaction, interactive teaching, entertainment, security fields, and so on. These systems aim to facilitate the natural interaction with machines by direct voice interaction instead of using traditional devices as input to understand verbal content and make it easy for human listeners to react. Some applications include dialogue systems for spoken languages such as call center conversations, onboard vehicle driving system and utilization of emotion patterns from the speech in medical applications.

Both paralinguistic and linguistic information is contained in the speech. Classical Automatic Speech Recognition systems focused less on some of the essential paralinguistic information passed on by speech like gender, personality, emotion, aim, and state of mind. The human mind utilizes all phonetic and paralinguistic data to comprehend the utterances' that has hidden importance and efficacious correspondence. The superiority of communication gets badly affected if there is any meagerness in the cognizance of paralinguistic features. Therefore, creating coherent and human-like communication machines that comprehend paralinguistic data.

Determining the emotional state of humans is an idiosyncratic task and may be used as a standard for any emotion recognition model. Amongst the numerous models used for categorization of these emotions, a discrete

emotional approach is considered as one of the fundamental approaches. It uses various emotions such as anger, boredom, disgust, surprise, fear, joy, happiness, neutral and sadness. The approach for speech emotion recognition (SER) primarily comprises two phases known as feature extraction and features classification phase. In the field of speech processing, researchers have derived several features such as source-based excitation features, prosodic features, vocal traction factors, and other hybrid features. The second phase includes feature classification using linear and nonlinear classifiers.

The most commonly used linear classifiers for emotion recognition include Bayesian Networks (BN) or the Multi Layer Perceptron (MLP) and Support Vector Machine (SVM). The most propitious technique for speech recognition is the neural network based approach. Artificial Neural Networks, (ANN) are biologically inspired tools for information processing. Speech recognition modelling by artificial neural networks (ANN) doesn't require any prior knowledge of speech process and this technique quickly became an attractive substitute to HMM. RNN can learn the sublunary relationship of Speech – Data and is capable of modelling time dependent phonemes. The conventional neural networks of Multi- Layer Perceptron (MLP) type have been increasingly in use for speech recognition and also for various other speech processing applications. Speech recognition is the process of converting an acoustic signal, captured by microphone or a telephone, to a set of characters. They can also serve as the input to further linguistic processing to achieve speech understanding, a subject covered in section. As we know, speech recognition performs tasks that similar with human brain.

1.2 OBJECTIVES OF THE PROJECT :

The main objectives of the project is :

- To consider an efficient dataset for model training that includes various modulations of emotions, so that the predictions made from the trained model can be accurate.
- To observe and find the best audio features for the feature extraction that could lead to more accurate emotion predictions.
- To test and train the dataset over different ratios to obtain the accurate model and find the Accuracy, Support, F-score and Recall for the proposed trained model.
- To improve the accuracy of the existing systems of Speech Emotion Recognition Systems, that has neglected over the features of some emotions which has lead to the inaccurate predictions.

1.3 PROBLEM STATEMENT :

In Artificial Intelligence, the automatic speech recognition field has been actively involved in generating the machines that communicate with human beings via speech. Human speech is the most common and expedient way of communication, and understanding speech is one of the complex mechanisms that the human brain performs. From a speech signal, numerous amounts of information can be gathered like gender, words, dialect, emotion, and age that could be utilized for various applications. In speech processing, one of the most arduous tasks for the researchers is speech emotion recognition. Various tasks such as speech to text conversion, feature extraction, feature selection, and classification of those features to identify the emotions must be performed by a well-developed framework that includes all these modules. The task of classification of features is yet another challenging work, and it involves the training of various emotional models to perform the classification appropriately. Now comes the second aspect of emotional speech recognition, the database used for training models. It involves selecting only the features that happen to be salient to depict the emotions accurately.

An SER system needs a classifier, a supervised learning construct, programmed to perceive any emotions in new speech signals. A supervised system like that introduces the need for labeled data with emotions embedded in it. Before any processing can be done on the data to extract the features, it needs preprocessing. For this reason, the sampling rate across all the databases should be consistent. The classification process essentially requires features. They help reduce raw data into the most critical characteristics only, regardless of whether it suffices to utilize acoustic features for displaying emotions. Classifiers' performance can be said to depend mainly on the techniques of feature extraction and those features that are viewed as salient for a particular emotion. Merging all the above modules in the desired way provides us with an application that can recognize a user's emotions and further provide it as an input to the system to respond appropriately. Various algorithms have been deployed over the years for the two important processes involved in SER (Speech Emotion Recognition) namely feature extraction and the decision making. To date, numerous acoustic features and classifiers have been put through experimentation to test their credibility, but the accuracy still needs to be improved.

1.4 EXISTING SYSTEM :

There are three classes of features in a speech namely, the lexical features (the vocabulary used), the visual features (the expressions the speaker makes) and the acoustic features (sound properties like pitch, tone, jitter, etc.).

The problem of speech emotion recognition can be solved by analyzing one or more of these features. Choosing to follow the lexical features would require a transcript of the speech which would further require an additional step of text extraction from speech if one wants to predict emotions from real-time audio. Similarly, going forward with analyzing visual features would require the excess to the video of the conversations which might not be feasible in every case while the analysis on the acoustic features can be done in real-time while the conversation is taking place as we'd just need the audio data for accomplishing our task. Hence, the existing systems have chosen to analyze the acoustic features in their work.

Furthermore, the representation of emotions can be done in two ways:

- **Discrete Classification :** Classifying emotions in discrete labels like anger, happiness, boredom, etc.
- **Dimensional Representation :** Representing emotions with dimensions such as Valence (on a negative to positive scale), Activation or Energy (on a low to high scale) and Dominance (on an active to passive scale)

Both these approaches have their pros and cons. The dimensional approach is more elaborate and gives more context to prediction but it is harder to implement and there is a lack of annotated audio data in a dimensional format. The discrete classification is more straightforward and easier to implement but it lacks the context of the prediction that dimensional representation provides. The existing systems have used the discrete classification approach for lack of dimensionally annotated data in the public domain.

The existing systems also worked on different classifiers, which are able to predict the emotion in most cases. But, the proposed system has been built to reduce the flaws of the existing systems, that has lead to inaccurate predictions. Linear discriminant classifiers, Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), k-nearest neighborhood (kNN) classifiers, Support Vector Machines (SVM), Decision Tree, and Artificial Neural Networks (ANN) are a few models that have been generally used to classify emotions dependent on their acoustic features of intrigue. The feature extraction techniques used by these classifiers are the determining factor for the performance of these classifiers. Proposed a system by observing that the MLP Classifier performs consistently better in both the accuracy and f-scores and also predicts the model in less time compared to other classifiers.

CHAPTER 2

LITERATURE REVIEW

Navya Damodar et al., proposed the use of decision tree and CNN as classifier to classify the emotions from the English and Kannada audio data. The performance of CNN and DT are potential for various emotions. Comparative study of the classifiers using various parameters is presented. The performance of CNN has been identified as the best classifier for emotion recognition. Emotions are recognized with 72% and 63% accuracy using CNN and Decision Tree algorithms respectively. MFCC features are extracted from the audio signals and Model is trained, tested and evaluated accordingly by changing the parameters. Speech Emotion Recognition system is useful in psychiatric diagnosis, lie detection, call centre conversations, customer voice review, voice messages.

Jianfeng Zhao et al., proposed about targets at studying deep elements from distinct information to recognize speech emotion. The authors designed a merged convolutional neural community (CNN), which had two branches, one being one-dimensional (1D) CNN department and every other 2D CNN branch, to research the high-level points from uncooked audio clips and log-mel spectrograms. The constructing of the merged deep CNN consists of two steps. First, one 1D CNN and one 2D CNN architectures had been designed and evaluated; then, after the deletion of the 2nd dense layers, the two CNN architectures have been merged together. To velocity up the education of the merged CNN, switch studying used to be added in the training. The 1D CNN and 2D CNN had been skilled first. Then, the realized facets of the 1D CNN and 2D CNN had been repurposed and transferred to the merged CNN. Finally, the merged deep CNN initialised with transferred facets used to be fine-tuned. Two hyperparameters of the designed architectures had been chosen via Bayesian optimisation in the training. The experiments performed on two benchmark datasets exhibit that the merged deep CNN can enhance emotion classification overall performance significantly.

S. Koelstra et al., presented a multimodal data set for the analysis of human affective states. The electroencephalogram (EEG) and peripheral physiological signals of 32 participants were recorded as each watched 40 one minute long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity. For 22 of the 32 participants, frontal face video was also recorded. A novel method for stimuli selection is proposed using retrieval by affective tags from the last.fm website, video highlight detection, and an online assessment tool. An extensive analysis of the participants' ratings during the experiment is presented. Correlates between the EEG signal frequencies and the participants' ratings are investigated. Methods and results are presented for single-trial classification of arousal, valence, and

like/dislike ratings using the modalities of EEG, peripheral physiological signals, and multimedia content analysis. Finally, decision fusion of the classification results from different modalities is performed. The data set is made publicly available and we encourage other researchers to use it for testing their own affective state estimation methods.

Babak Basharirad et al., presented that the attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech. Selection of suitable feature sets, design of a proper classifications methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems. This paper critically analyzed the current available approaches of speech emotion recognition methods based on the three evaluating parameters (feature set, classification of features, accurately usage). In addition, this paper also evaluates the performance and limitations of available methods. Furthermore, it highlights the current promising direction for improvement of speech emotion recognition systems.

Chen et al., aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method. This method classify different emotions from coarse to fine then select appropriate feature by using Fisher rate. The output of Fisher rate is an input parameters for multi- level SVM based classifier. Furthermore principal component analysis (PCA) and artificial neural network (ANN) are employed to reduce the dimensionality and classification of four comparative experiments, respectively. Four comparative experiments include Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. Consequence indicates in dimension reduction Fisher is better than PCA and for classification, SVM is more expansible than ANN for emotion recognition in speaker independent is. The recognition rates for three level are 86.5%, 68.5% and 50.2% separately in Beihang university database of emotional speech (BHUDES).

Nwe et al., proposed a new system for emotion classification of utterance signals. The system employed a short time log frequency power coefficients (LFPC) and discrete HMM to characterize the speech signals and classifier respectively. This method classified the emotion into six different categories then used the private dataset to train and test the new system. In order to evaluate the performance of the proposed method, LFPC is compared with the mel-frequency Cepstral coefficients (MFCC) and linear prediction Cepstral coefficients (LPCC). Result demonstrate the average and best classification accuracy achieved 78% and 96% respectively. Furthermore, results expose that LFPC is a better option as feature for emotion classification than the standard features.

Wu et al., proposed a fusion-based method for speech emotion recognition by employing multiple classifier and acoustic-prosodic (AP) features and semantic labels (SLs). In this fusion method, first AP features are extracted

then three different types of base-level classifier include GMMs, SVMs, MLP and Meta decision tree (MDT) are used. The maximum entropy model (MaxEnt) in the semantic labels method are applied. MaxEnt modeled the association between emotion association rules (EARs) and emotion states in emotion recognition. In the final state to define the emotion recognition outcome, the integrated consequence from the SL-based and AS-based are utilized. The experimental result on private dataset shows the performance based on MDT archives 80%, SL-based recognition archives 80.92, and mixture of AP and SL archives 83.55%.

Narayanan et al., proposed domain-specific emotion recognition by utilizing speech signals from call center application. Detecting negative and non-negative emotion (ex : anger and happy) are the main focus of this research. Different types of information include acoustic, lexical, and discourse are used for emotion recognition. In addition, information-theoretic contents of emotional salience is presented to obtain data at emotion information at the language level. Both k-NN and linear discriminant classifier are used to work with different types of features. Experimental result confirms that the best results are achieved by combination of acoustic and language data. Outcomes demonstrates by combining three information source instead of one source, classification accuracy increases by 40.7% for males and 36.4% for females. Compare to pervious work improvement range in accuracy is from 1.4% to 6.75% for male and 0.75% to 3.96% for female.

CHAPTER 3

PROPOSED METHOD

3.1 METHODOLOGY :

The proposed methodology for the speech emotion recognition system is by using MLP(Multi-layer Perceptron) classifier and RAVDESS (Ryerson audio visual database of emotional speech and song) dataset. At first, the audio is collected and pre-processed using the python libraries “Librosa” and “PyAudio” for removing the silence and noise reduction from the audio files. Then, the extracted features from the RAVDESS dataset that undergoes training and testing process are stored based on their classification and an machine learning model is created. Now the input audio clip gets the feature extraction done and the extracted features are compared with the model that had undergo testing and training from the RAVDESS dataset. Finally, the output emotion is predicted and displayed on the screen.

On doing the literature survey of various methods for accurate predictions of emotions from speech, we come to the conclusion that to predict the emotion from speech there are multiple approaches for classification of speech signals. But the best possible way to get accurate prediction are through MLP Classifier (Multi Layer Perceptron) which is a part of an ANN (Artificial Neural Network).

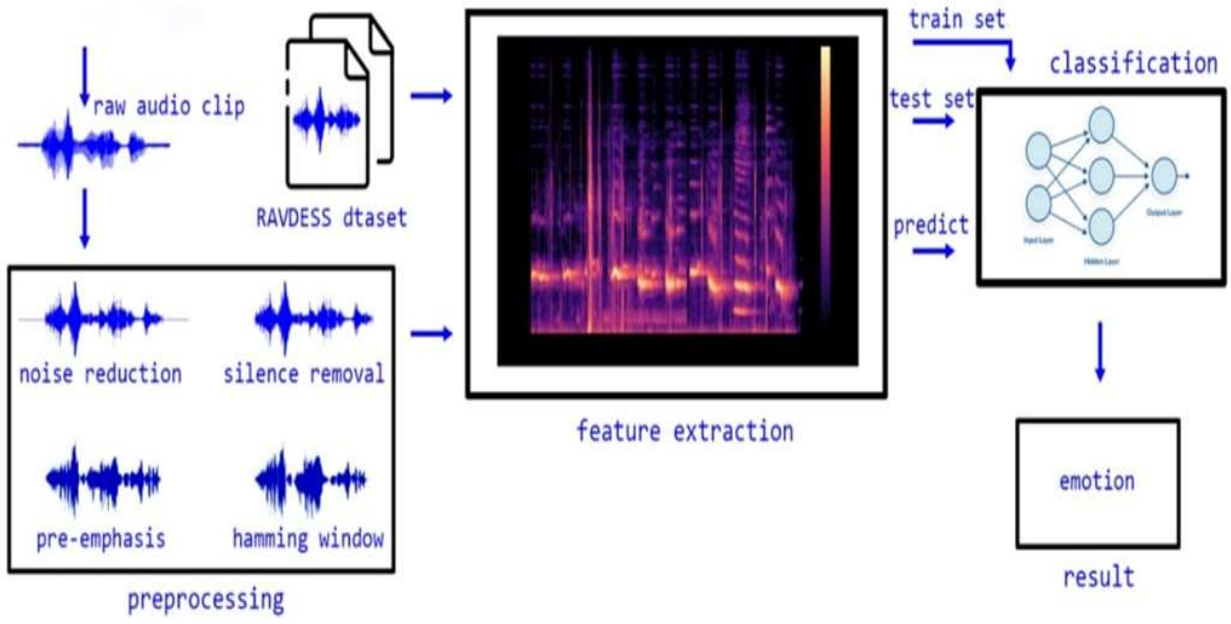


Fig : 3.1 : Architecture of the proposed system

3.1.1 MACHINE LEARNING :

Machine Learning (ML) is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans "The ability to learn". Machine learning is actively being used today. Machine learning is the branch of **Artificial Intelligence (AI)** which provide the ability to learning automatically learn and improve from experience. It was first introduced in **1959** by **Arthur Samuel**.

The primary aim is to allow the computer to learn automatically without human involvement or assistance and adjust actions accordingly. Many problems are historical very easy for humans, and very difficult for networks, Machine Learning (**Deep Learning in particular**) is currently our best solution for many of those problems.

Types of Machine Learning **:**

- **Supervised Learning** - "Train me!"
- **Unsupervised Learning -** "I am self-sufficient in learning!"
- **Reinforcement Learning -** "My life my rules (Hit and Trial)!"

3.1.1.1 Supervised Learning :

Supervised learning is the type of machine learning in which machines are trained using "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

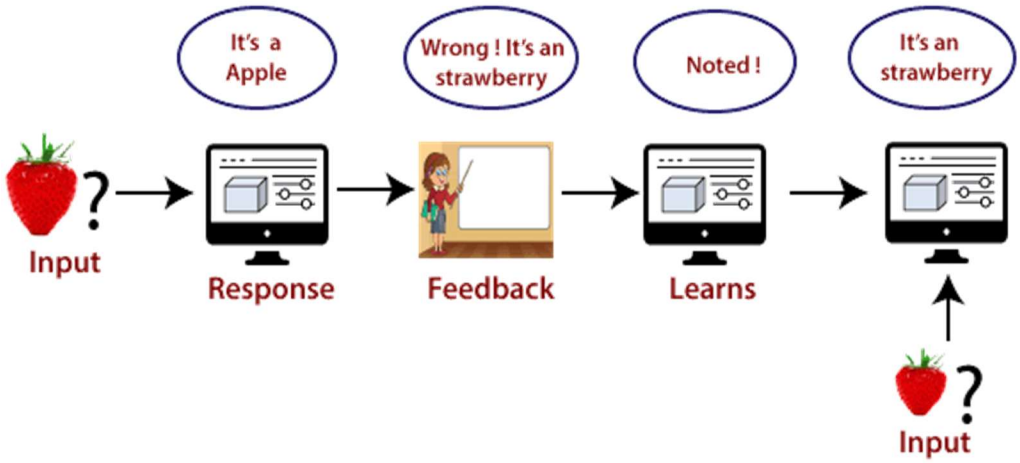


Fig : 3.1.1.1.1 : Example for Supervised Learning.

It can be grouped into two types:

- **Classification** is a technique that aims to reproduce class assignments. It produces the response value, and the data separated into "**classes**". Like recognizing a type of car in a photo.
- **Regression** is a technique which aims to produce the output value. Like predicting the price of a different product.

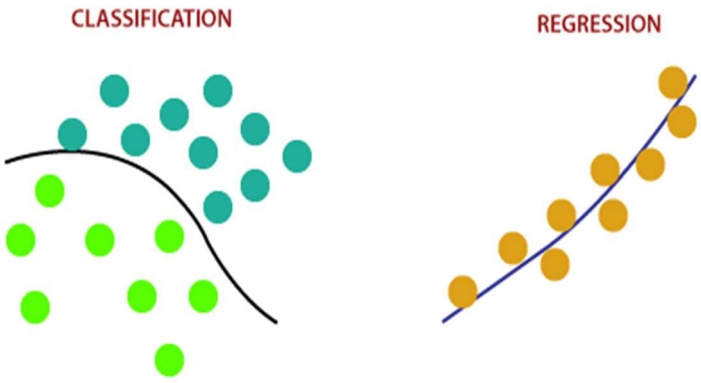


Fig : 3.1.1.1.2 : Example to depict the difference between Classification and Regression.

3.1.1.2 Unsupervised Learning :

The unsupervised machine learning algorithm is used when the train information is neither classified nor labelled. If the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it. Supposed we presented images of apples, bananas, and mangoes to the model, based on some patterns and relationships it creates cluster and divides the dataset into clusters. Now if a new data is delivered to the model, it adds it to one of the generated groups.

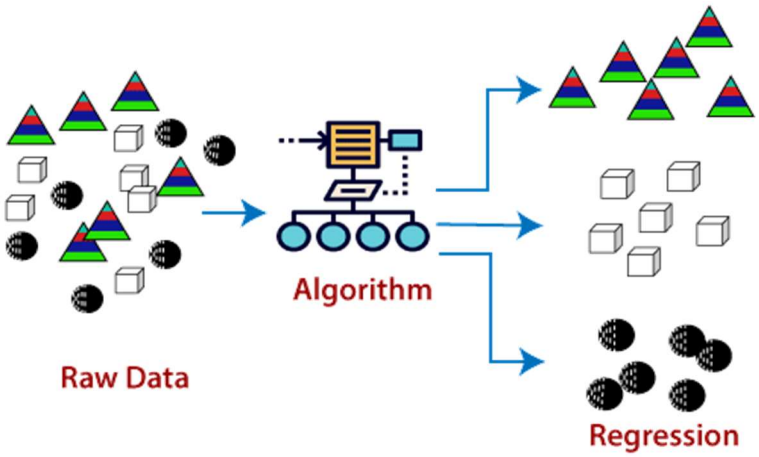


Fig : 3.1.1.2 : Basic Architecture of Unsupervised Learning.

It also has two types :

- **Clustering** is used to find likeness and differences in a particular thing. It groups similar things. This algorithm can help us to solve many obstacles. Like creating clusters of similar tweets based on their content, find a group of photos with similar cars, or identify different type of news.
- **Association** rules mining is another key of unsupervised data mining method, after clustering, which finds interesting associations (relationships, dependencies) in a large set of data items.

3.1.1.3 Semi-Supervised Learning :

Semi-Supervised Learning falls somewhere in between the supervised and unsupervised learning. So, they use both labelled and unlabelled data for training where a small amount of labelled data and a big amount of unlabelled data is used. Commonly, semi-supervised learning is chosen when the acquired labelled data requires skilled and significant resources to train it.

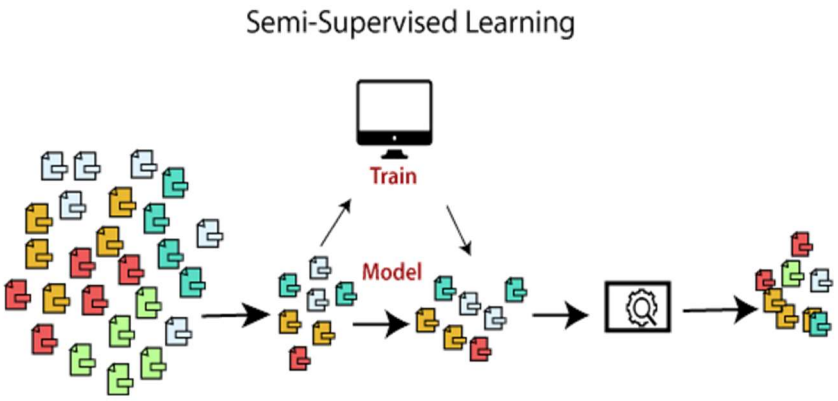


Fig : 3.1.1.3 : Basic Architecture of Semi-Supervised Learning.

3.1.1.4 Reinforcement Learning :

Reinforcement learning is the ability of an agent to interact with the environment and find out the best outcome. It chases the concept of **hit and trial** method. The agent is rewarded or condemned with a point for a correct or a wrong answer, and based on the positive rewards points gained the model train itself. And once again it is trained to predict the new data presented to it. The goal of the agent is to get the most reward points and to improve its performance.

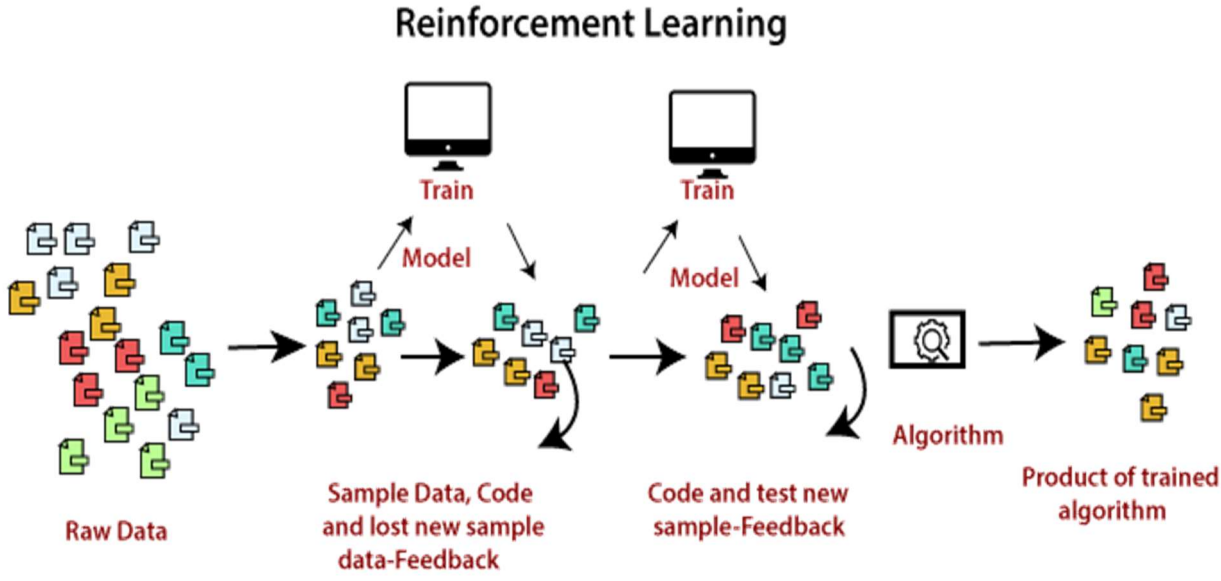


Fig : 3.1.1.4 : Basic Architecture of Reinforcement Learning.

3.1.2 DEEP LEARNING :

Deep learning is also a subset of Machine Learning in Artificial Intelligence. Learning can be supervised, semi-supervised, or unsupervised. Deep learning is a machine learning approach that prepares computers to achieve what comes naturally to humans. Deep learning is a crucial technology behind driverless cars, enabling them to recognize a stop sign.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers. While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

- Deep learning requires large amounts of labeled data. For example, driverless car development requires millions of images and thousands of hours of video.

- Deep learning requires substantial computing power. High-performance GPUs have a parallel architecture that is efficient for deep learning. When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

Deep learning applications are used in industries from automated driving to medical devices.

3.1.2.1 Working of Deep Learning :

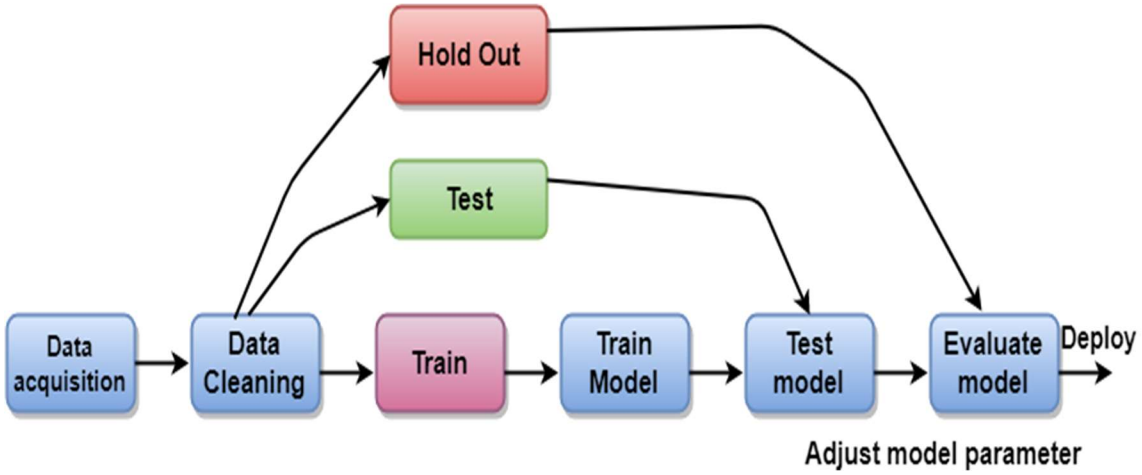


Fig : 3.1.2.1.1 : Process Flow of Deep Learning.

Deep learning is getting lots of attention lately and for a good reason. A computer model learns to perform the classification tasks directly from any images, text, and sound in deep learning. The term "**deep**" commonly refers to the number of hidden layer in the neural network. Conventional neural networks only contain 2-3 hidden layers, while deep networks can have 150.

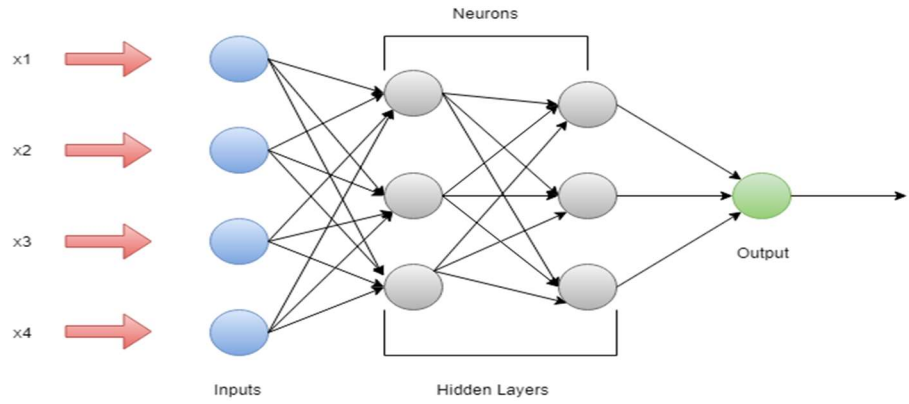


Fig : 3.1.2.1.2 : Layers in Deep Learning.

3.1.3 NEURAL NETWORKS :

A Neural Network is a computing system made of several simple, highly interconnected processing elements, which process information by its dynamic state response to external inputs.

A neural network can be made with multiple perceptrons. Where there are three layers, they are :

- **Input layer :** Input layers are the real value from the data.
- **Hidden layer :** Hidden layers are between input and output layers where three or more layers are deep network.
- **Output layer :** It is the final estimate of the output.

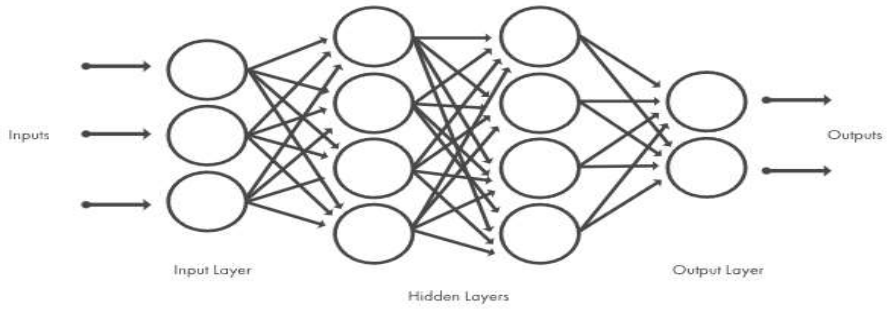


Fig : 3.1.3 : Layers in Neural Networks.

3.1.3.1 Artificial Neural Network :

The term "Artificial Neural Network" refers to a biologically inspired sub-field of artificial intelligence modeled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes.

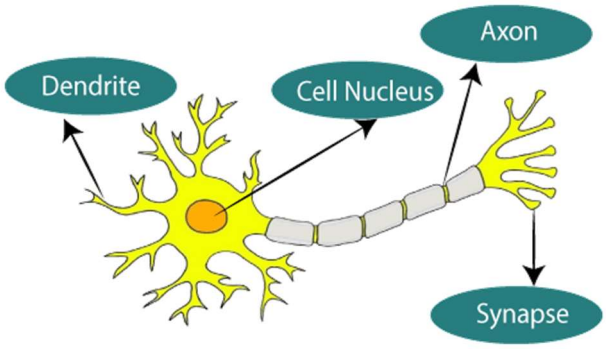


Fig : 3.1.3.1.1 : Illustrating the Biological Neural Network.

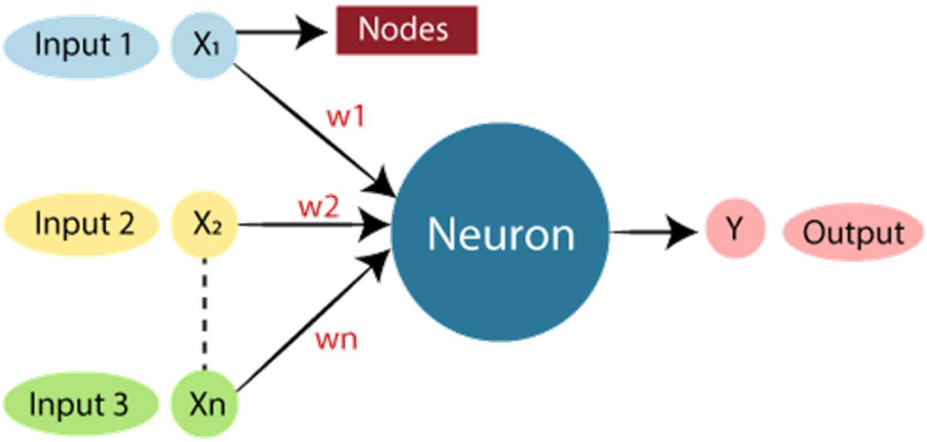


Fig : 3.1.3.1.2 : Typical Artificial Neural Network.

Dendrites from Biological Neural Network represent inputs in Artificial Neural Networks, cell nucleus represents Nodes, synapse represents Weights, and Axon represents Output. An **Artificial Neural Network** in the field of **Artificial intelligence** where it attempts to mimic the network of neurons makes up a human brain so that computers will have an option to understand things and make decisions in a human-like manner. The artificial

neural network is designed by programming computers to behave simply like interconnected brain cells.

There are around 1000 billion neurons in the human brain. Each neuron has an association point somewhere in the range of 1,000 and 100,000. In the human brain, data is stored in such a manner as to be distributed, and we can extract more than one piece of this data when necessary from our memory parallelly. We can say that the human brain is made up of incredibly amazing parallel processors.

We can understand the artificial neural network with an example, consider an example of a digital logic gate that takes an input and gives an output. "OR" gate, which takes two inputs. If one or both the inputs are "On," then we get "On" in output. If both the inputs are "Off," then we get "Off" in output. Here the output depends upon input. Our brain does not perform the same task. The outputs to inputs relationship keep changing because of the neurons in our brain, which are "learning."

3.1.3.1.2 Architecture of an Artificial Neural Network :

To understand the concept of the architecture of an artificial neural network, we have to understand what a neural network consists of. In order to define a neural network that consists of a large number of artificial neurons, which are termed units arranged in a sequence of layers. look at various types of layers available in an artificial neural network. Artificial Neural Network primarily consists of three layers:

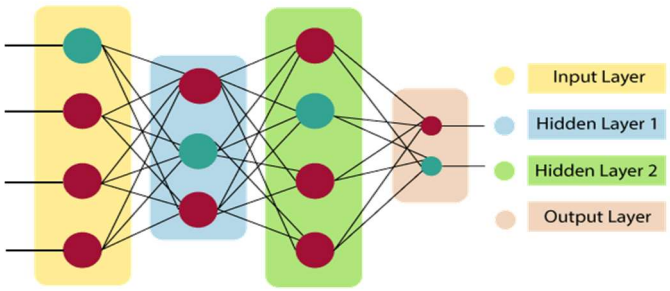


Fig : 3.1.3.1.2.1 : Layers in Artificial Neural Networks.

- **Input Layer** accepts inputs in several different formats provided by the programmer.
- **Hidden Layer** presents in-between input and output layers. It performs all the calculations to find hidden features and patterns.
- In **Output Layer,** the input goes through a series of transformations using the hidden layer, which finally results in output that is conveyed using this layer.

The artificial neural network takes input and computes the weighted sum of the inputs and includes a bias. This computation is represented in the form of a transfer function.

$$\sum_{i=1}^{n} Wi * Xi + b$$

It determines weighted total is passed as an input to an activation function to produce the output. Activation functions choose whether a node should fire or not. Only those who are fired make it to the output layer. There are distinctive activation functions available that can be applied upon the sort of task we are performing.

3.1.3.1.3 Working of Artificial Neural Networks (ANN) :

Artificial Neural Network can be best represented as a weighted directed graph, where the artificial neurons form the nodes. The association between the neurons outputs and neuron inputs can be viewed as the directed edges with weights. The Artificial Neural Network receives the input signal from the external source in the form of a pattern and image in the form of a vector. These inputs are then mathematically assigned by the notations x(n) for every n number of inputs.

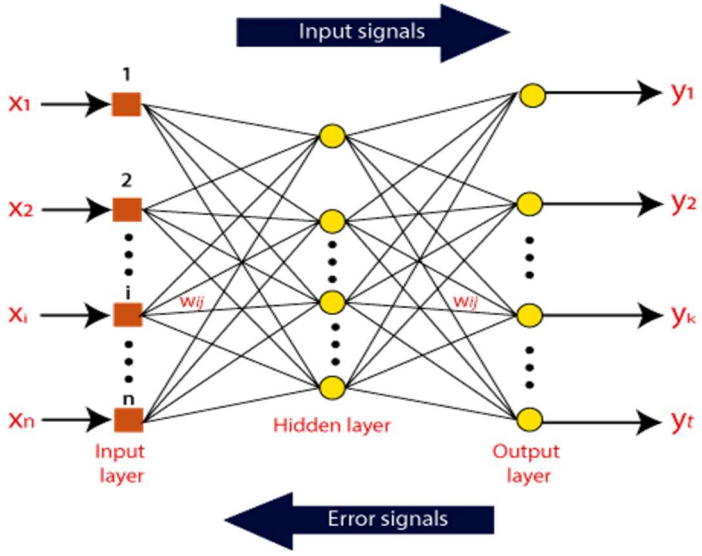


Fig : 3.1.3.1.3 : Working of an ANN.

Afterward, each of the input is multiplied by its corresponding weights (these weights are the details utilized by the artificial neural networks to solve a specific problem). In general terms, these weights normally represent the strength of the interconnection between neurons inside the artificial neural network. All the weighted inputs are summarized inside the computing unit.

If the weighted sum is equal to zero, then bias is added to make the output non-zero or something else to scale up to the system's response. Bias has the same input, and weight equals to 1. Here the total of weighted inputs can be in the range of 0 to positive infinity. Here, to keep the response in the limits of the desired value, a certain maximum value is benchmarked, and the total of weighted inputs is passed through the activation function.

The activation function refers to the set of transfer functions used to achieve the desired output. There is a different kind of the activation function, but primarily either linear or non-linear sets of functions. Some of the commonly used sets of activation functions are the Binary, linear, and Tan hyperbolic sigmoidal activation functions.

Artificial Neural Network (ANN) is entirely inspired by the way the biological nervous system work. For Example, the human brain works. The most powerful attribute of the human brain is to adapt, and ANN acquires similar characteristics. We should understand that how exactly our brain does? It is still very primitive, although we have a fundamental understanding of the procedure. It is accepted that during the learning procedure, the brain's neural structure is altered, increasing or decreasing the capacity of its synaptic connections relying on their activity. This is the reason why more relevant information is simpler to review than information that has not been reviewed for a long time. More significant information will have powerful synaptic connections, and less applicable information will gradually have its synaptic connections weaken, making it harder to review.

ANN can model this learning process by changing the weighted associations found between neurons in the network. It effectively mimics the strengthening and weakening of the synaptic associations found in our brains. The strengthening and weakening of the associations are what empowers the network to adapt. Face recognition would be an example of an issue extremely difficult for a human to precisely convert into code. An issue that could not be resolved better by a learning algorithm would be a loan granting institution that could use the previous credit score to classify future loan probabilities.

The learning rule is a technique or a mathematical logic which encourages a neural network to gain from the existing condition and uplift its performance. It is an iterative procedure. In this tutorial, we will talk about the learning rules in Neural Network. Here, we will discuss what is Hebbian learning rule, perception learning rule, Delta learning rule, correlation learning rule, out star learning rule? All these Neural Network Learning Rules are discussed in details given below with their mathematical formulas.

A learning rule or Learning process is a technique or a mathematical logic. It boosts the Artificial Neural Network's performance and implements this rule over the network.

3.1.3.1.4 Components of an Artificial Neural Network :

- **Neurons :**

  Neurons are similar to the biological neurons. Neurons are nothing but the activation function. Artificial neurons or Activation function has a "**switch on**" characteristic when it performs the classification task. We can say when the input is higher than a specific value; the output should change state, i.e., 0 to 1, -1 to 1, etc. The sigmoid function is commonly used activation function in **Artificial Neural Network**.

  F (Z) = 1/1+EXP (-Z)

- **Nodes :**

  The biological neuron is connected in hierarchical networks, with the output of some neurons being the input to others. These networks are represented as a connected layer of nodes. Each node takes multiple weighted inputs and applies to the neuron to the summation of these inputs and generates an output.

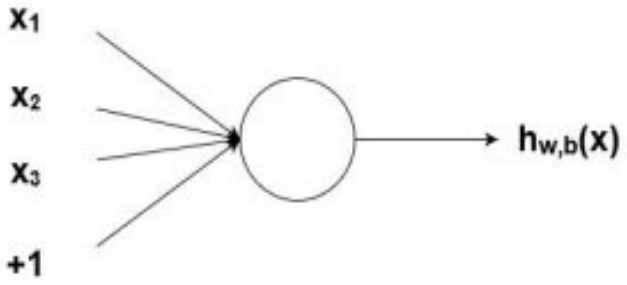


Fig : 3.1.3.1.4.1 :Representing Nodes in ANN.

- **Bias :**

In the neural network, we predict the output (y) based on the given input (x). We create a model, i.e. (mx + c), which help us to predict the output. When we train the model, it finds the appropriate value of the constants m and c itself. The constant c is the bias. Bias helps a model in such a manner that it can fit best for the given data. We can say bias gives freedom to perform best.

- **Algorithm :**

Algorithms are required in the neural network. Biological neurons have self-understanding and working capability, but how an artificial neuron will work in the same way? For this, it is necessary to train our artificial neuron network. For this purpose, there are lots of algorithms used. Each algorithm has a different way of working.

There are five algorithms which are used in training of our ANN

- Gradient Descent
- Newton's Method
- Conjugate Gradient
- Quasi Newton's
- Levenberg Marquardt

3.1.3.1.5 Advantages of Artificial Neural Network (ANN) :

- **Parallel processing capability.** Artificial neural networks have a numerical value that can perform more than one task simultaneously.
- **Storing data on the entire network.** Data that is used in traditional programming is stored on the whole network, not on a database. The disappearance of a couple of pieces of data in one place doesn't prevent the network from working.
- **Capability to work with incomplete knowledge.** After ANN training, the information may produce output even with inadequate data. The loss of performance here relies upon the significance of missing data.
- **Having a memory distribution.** For ANN is to be able to adapt, it is important to determine the examples and to encourage the network according to the desired output by demonstrating these examples to the network. The succession of the network is directly proportional to the chosen instances, and if the event can't appear to the network in all its aspects, it can produce false output.

- **Having fault tolerance.** Extortion of one or more cells of ANN does not prohibit it from generating output, and this feature makes the network fault-tolerance.

3.1.3.1.6 Disadvantages of Artificial Neural Network (ANN) :

- **Assurance of proper network structure.** There is no particular guideline for determining the structure of artificial neural networks. The appropriate network structure is accomplished through experience, trial, and error.
- **Unrecognized behavior of the network.** It is the most significant issue of ANN. When ANN produces a testing solution, it does not provide insight concerning why and how. It decreases trust in the network.
- **Hardware dependence.** Artificial neural networks need processors with parallel processing power, as per their structure. Therefore, the realization of the equipment is dependent.
- **Difficulty of showing the issue to the network.** ANNs can work with numerical data. Problems must be converted into numerical values before being introduced to ANN. The presentation mechanism to be resolved here will directly impact the performance of the network. It relies on the user's abilities.
- **The duration of the network is unknown.** The network is reduced to a specific value of the error, and this value does not give us optimum results.

3.1.3.1.7 Types of Artificial Neural Network :

There are various types of Artificial Neural Networks (ANN) depending upon the human brain neuron and network functions, an artificial neural network similarly performs tasks. The majority of the artificial neural networks will have some similarities with a more complex biological partner and are very effective at their expected tasks. For example, segmentation or classification.

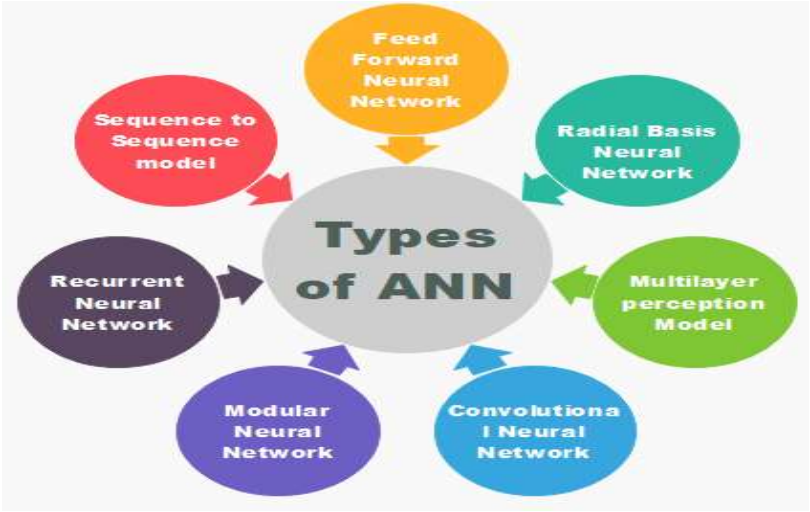
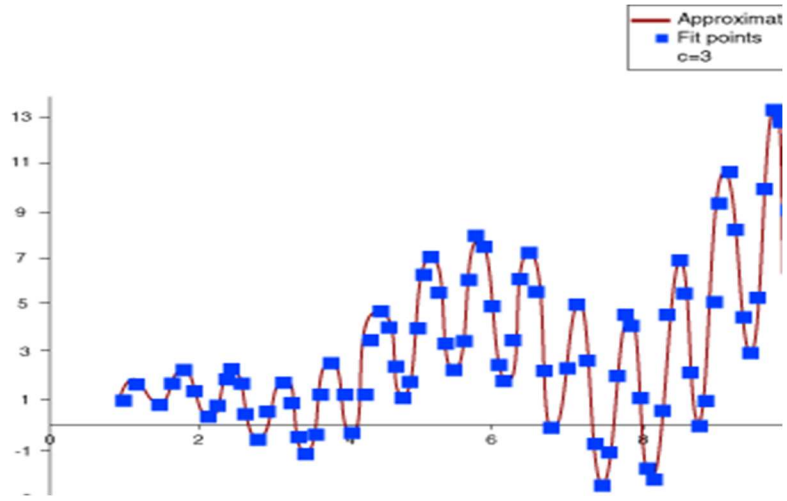


Fig : 3.1.3.1.7 : Types of ANN.

3.1.3.1.7.1 Radial Basis function Neural Network :

RBFNN find the distance of a point to the centre and considered it to work smoothly. There are two layers in the RBF Neural Network. In the inner layer, the features are combined with the radial basis function. Features provide an output that is used in consideration. Other measures can also be used rather than Euclidean.



Redial Basis Function defines the

- A receptor t.
- Confronted maps are drawn around the receptor.
- For RBF Gaussian Functions are generally used. So we can define the radial distance **r=||X-t||**.

Redial Function=$\Phi(r) = \exp(-r^2/2\sigma^2)$, where $\sigma > 0$

This Neural Network is used in power restoration system. In the present era power system have increased in size and complexity. It's both factors increase the risk of major power outages. Power needs to be restored as quickly and reliably as possible after a blackout.

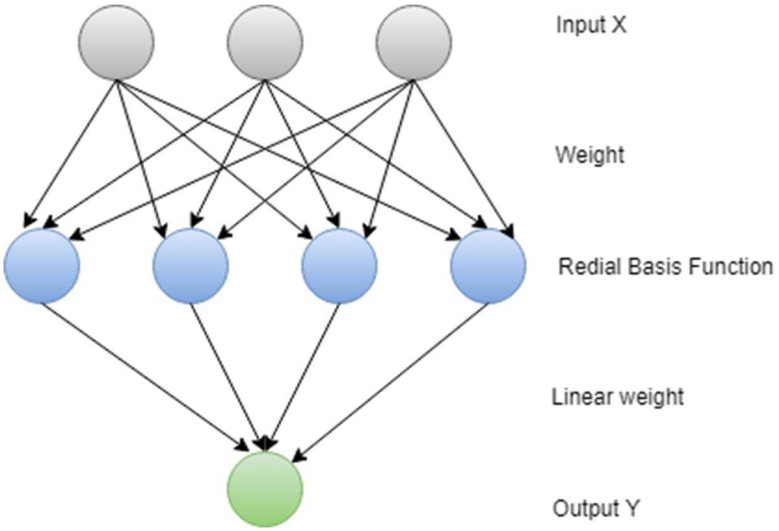


Fig : 3.1.3.1.7.1 : Layers in Radial Basis Neural Network .

3.1.3.1.7.2 Convolutional Neural Network :

In image classification and image recognition, a **Convolutional Neural Network** plays a vital role, or we can say it is the main category for those. Face recognition, object detection, etc., are some areas where CNN are widely used. It is similar to FNN, learn-able weights and biases are available in neurons.

CNN takes an image as input that is classified and process under a certain category such as dog, cat, lion, tiger, etc. As we know, the computer sees an image as pixels and depends on the resolution of the picture. Based on image resolution, it will see h * w * d, where h= height w= width and d= dimension. For example, An RGB image is 6 * 6 * 3 array of the matrix, and the grayscale image is 4 * 4 * 3 array of the pattern.

In CNN, each input image will pass through a sequence of convolution layers along with pooling, fully connected layers, filters (Also known as kernels). And apply Soft-max function to classify an object with probabilistic values 0 and 1.

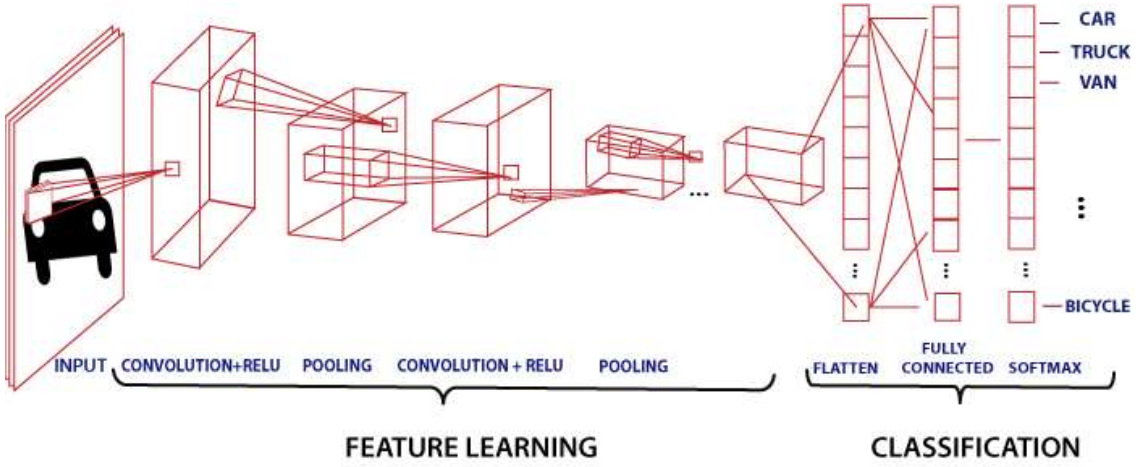


Fig : 3.1.3.1.7.2 : Basic Architecture of Convolutional Neural Network.

3.1.3.1.7.3 Recurrent Neural Network :

Recurrent Neural Network is based on prediction. In this neural network, the output of a particular layer is saved and fed back to the input. It will help to predict the outcome of the layer. In Recurrent Neural Network, the first layer is formed in the same way as FNN's layer, and in the subsequent layer, the recurrent neural network process begins. Both inputs and outputs are independent of each other, but in some cases, it required to predict the next word of the sentence.

Then it will depend on the previous word of the sentence. RNN is famous for its primary and most important feature, i.e., **Hidden State**. Hidden State remembers the information about a sequence.

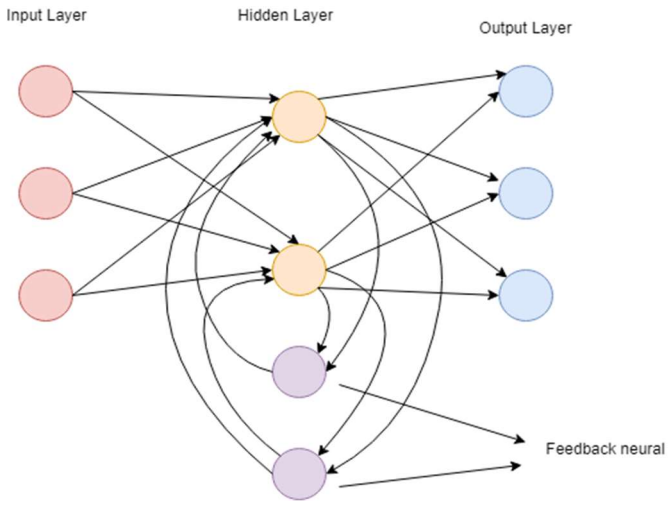


Fig : 3.1.3.1.7.3 : Layers in Recurrent Neural Network.

RNN has a memory to store **the result after calculation. RNN** uses the same parameters on each input to perform the same task on all the hidden layers or data to produce the output. Unlike other neural networks, RNN parameter complexity is less.

3.1.3.1.7.4 Modular Neural Network :

In **Modular Neural Network**, several different networks are functionally independent. In MNN the task is divided into sub-task and perform by several systems. During the computational process, networks don't communicate directly with each other. All the interfaces are work independently towards achieving the output. Combined networks are more powerful than flat and unrestricted. Intermediary takes the production of each system, process them to produce the final output.

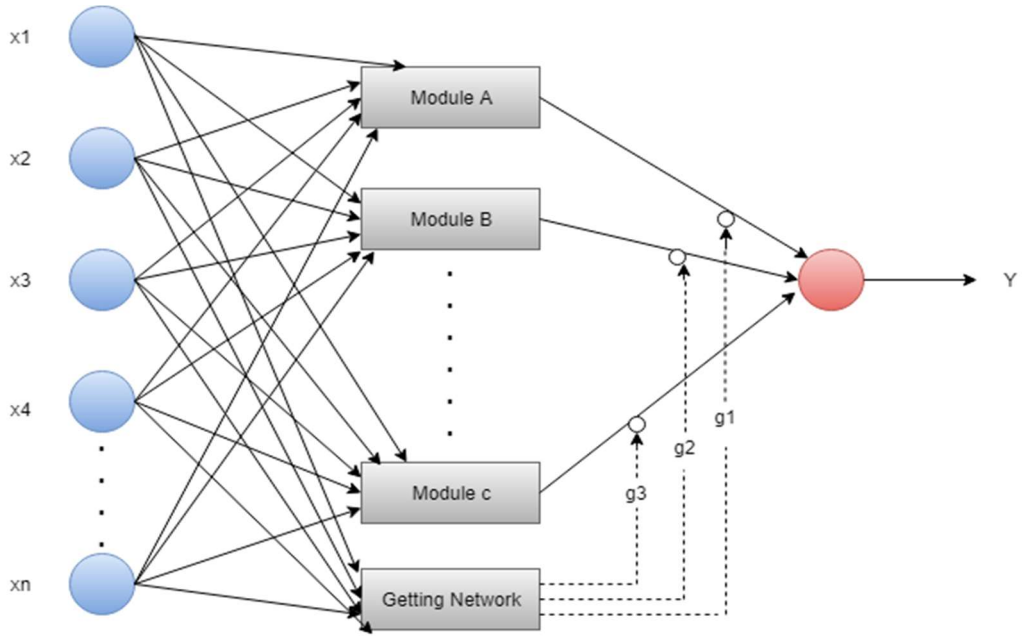


Fig : 3.1.3.1.7.4 : Layers in Modular Neural Network.

3.1.3.1.7.5 Sequence to Sequence Network :

It is consist of two recurrent neural networks. Here, encoder processes the input and decoder processes the output. The encoder and decoder can either use for same or different parameter. Sequence-to-sequence models are applied in chatbots, machine translation, and question answering systems.

3.1.3.1.7.6 Feed-Forward ANN :

A feed-forward network is a basic neural network comprising of an input layer, an output layer, and at least one layer of a neuron. Through assessment of its output by reviewing its input, the intensity of the network can be noticed based on group behavior of the associated neurons, and the output is decided.

The advancement of layered feed-forward networks initiated in the late **1950s**, given by **Rosenblatt's** perceptron and **Widrow's** Adaptive linear Element (ADLINE).The perceptron and ADLINE can be defined as a single layer networks and are usually referred to as single-layer perceptron's. Single-layer perceptron's can only solve linearly separable problems. The limitations of the single-layer network have prompted the advancement of multi-layer feed-forward networks with at least one hidden layer, called multi-layer perceptron (MLP) networks. MLP networks overcome various limitations of single-layer perceptron's and can be prepared to utilize the backpropagation algorithm. The backpropagation method was invented autonomously several times.

In 1974, Werbos created a backpropagation training algorithm. However, Werbos work remained unknown in the scientific community, and in 1985, parker rediscovers the technique. Soon after Parker published his discoveries, Rumelhart, Hinton, and Williams also rediscovered the method. It is the endeavors of Rumelhart and the other individual if the Parallel Distributed Processing (PDP) group, that makes the backpropagation method a pillar of neurocomputing.

FNN is the purest form of ANN in which input and data travel in only one direction. Data flows in an only forward direction; that's why it is known as **the Feedforward Neural Network**. The data passes through input nodes and exit from the output nodes. The nodes are not connected cyclically. It doesn't need to have a hidden layer. In FNN, there doesn't need to be multiple layers. It may have a single layer also.

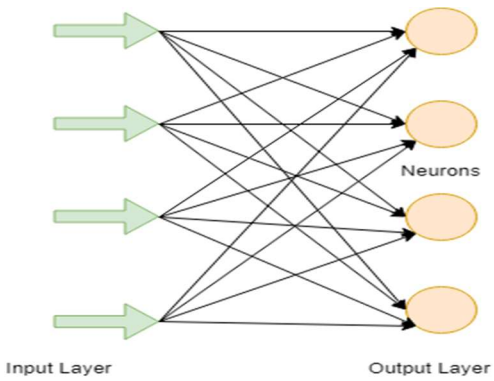


Fig : 3.1.3.1.7.6 : Basic Architecture of Semi-Supervised Learning.

It has a front propagate wave that is achieved by using a classifying activation function. All other types of neural network use backpropagation, but FNN can't. In FNN, the sum of the product's input and weight are calculated, and then it is fed to the output. Technologies such as **face recognition** and **computer vision** are used FNN.

3.1.3.1.7.6.1 Single-layer feedforward network:

Rosenblatt first constructed the single-layer feedforward network in the late 1950s and early 1990s. The concept of feedforward artificial neural network having just one weighted layer. In other words, we can say that the input layer is completely associated with the outer layer.

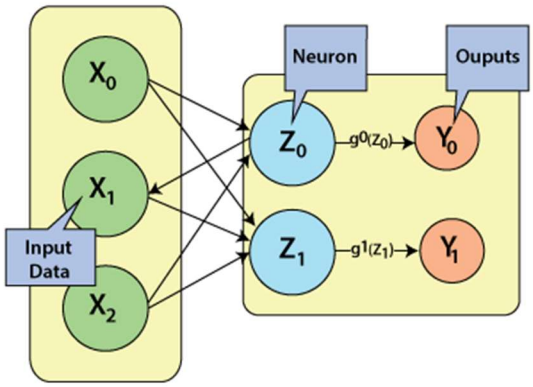


Fig : 3.1.3.1.7.6.1 : Single-layer feedforward network.

3.1.3.1.7.6.2 Multilayer feedforward network:

A multilayer feedforward neural network is a linkage of perceptrons in which information and calculations flow are unidirectional, from the input data to the outputs. The total number of layers in a neural network is the same as the total number of layers of perceptrons. The easiest neural network is one with a single input layer and an output layer of perceptrons. The concept of feedforward artificial neural network having more than one weighted layer. As the system has at least one layer between the input and the output layer, it is called the hidden layer.

Perceptron is a single layer neural network. It is a binary classifier and part of supervised learning. A simple model of the biological neuron in an artificial neural network is known as the perceptron. The artificial neuron has input and output.

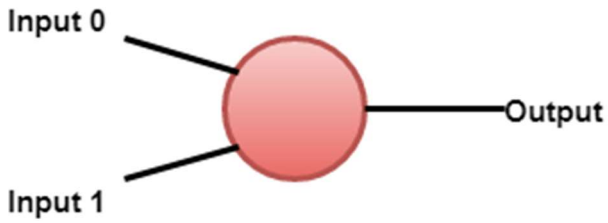


Fig : 3.1.3.1.7.6.2 : Representation of perceptron model.

3.1.3.1.7.6.2.1 Multi-Layer Perceptrons :

MLP networks are used for supervised learning format. A typical learning algorithm for MLP networks is also called **back propagation's algorithm**. A multilayer perceptron (MLP) is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input nodes connected as a directed graph between the input and output layers. MLP uses backpropagation for training the network. MLP is a deep learning method.

A **Multilayer Perceptron** has three or more layer. The data that cannot be separated linearly is classified with the help of this network. This network is a fully connected network that means every single node is connected with all other nodes that are in the next layer. A **Nonlinear Activation Function** is used **in Multilayer Perceptron**. It's input and output layer nodes are connected as a directed graph. It is a deep learning method so that for training the network it uses **backpropagation**. It is extensively applied in speech recognition and machine translation technologies. Multi-Layer perceptron defines the most complex architecture of artificial neural networks. It is substantially formed from multiple layers of the perceptron. TensorFlow is a very popular deep learning framework released by, and this notebook will guide to build a neural network with this library. If we want to understand what is a Multi-layer perceptron, we have to develop a multi-layer perceptron from scratch using Numpy.

The field of artificial neural networks is often just called neural networks or multi-layer perceptrons after perhaps the most useful type of neural network. A perceptron is a single neuron model that was a precursor to larger neural networks.

It is a field that investigates how simple models of biological brains can be used to solve difficult computational tasks like the predictive modeling tasks we see in machine learning. The goal is not to create

realistic models of the brain, but instead to develop robust algorithms and data structures that we can use to model difficult problems.

The power of neural networks comes from their ability to learn the representation in your training data and how to best relate it to the output variable that you want to predict. In this sense neural networks learn a mapping. Mathematically, they are capable of learning any mapping function and have been proven to be a universal approximation algorithm.

The predictive capability of neural networks comes from the hierarchical or multi-layered structure of the networks. The data structure can pick out (learn to represent) features at different scales or resolutions and combine them into higher-order features. For example from lines, to collections of lines to shapes.

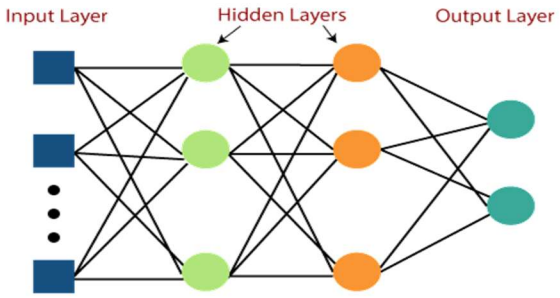


Fig : 3.1.3.1.7.6.2.1 : Layers in Multi-Layer Perceptron.

3.2 IMPLEMENTATION :

In order build the proposed system the following steps should be followed to train and build an efficient model that could lead to accurate emotion predictions from speech.

- Importing Modules
- DataSet Collection
- Feature Extraction
- Testing and Training
- Building the model
- User Audio Input
- Process the Audio
- Predict the Emotion
- Display the Emotion

3.2.1 IMPORTING MODULES :

The proposed system includes the modules Librosa and PyAudio for processing and classification of the audio. It also includes various other machine learning modules like Scikit-learn, Pandas, Numpy, Keras .etc. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Librosa is basically used to work with audio data like in Music Generation, Automatic Speech Recognition. It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals.

```
pip install librosa
```

PyAudio provides python bindings for port Audio, the cross-platform audio I/O library. With PyAudio, you can easily use Python to play and record audio on a variety of platforms.

```
pip install pyaudio
```

3.2.2 DATASET COLLECTION :

In order to build an efficient system, training on an efficient dataset is important that includes various modulations of emotions, so that the predictions made from the trained model can be accurate. For the proposed system includes RAVDESS Dataset, which includes voice of many actors, different modulations, etc.

3.2.2.1 RAVDESS DATASET :

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Speech folder contains 1440 files: 60 trials per actor x 24 actors = 1440. Each of the 7356 RAVDESS files has a unique filename (ex : 02-01-06-01-02-01-12.mp4).

Identifier	Coding description of factor levels
Modality	01 = Audio-video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06 = Fearful, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
Repetition	01 = First repetition, 02 = Second repetition
Actor	01 = First actor, . . ., 24 = Twenty-fourth actor

Fig : 3.2.2.1 : Explaining RAVDESS Dataset File Structure.

Filename example: 03-01-06-01-02-01-12.wav

- Audio-only (03)
- Speech (01)
- Fearful (06)
- Normal intensity (01)
- Statement "dogs" (02)
- 1st Repetition (01)
- 12th Actor (12) - Female, as the actor ID number is even.

3.2.3 FEATURE EXTRACTION :

The feature extraction process extracts the features from the given dataset. However, the features extracted will be different from each system based on the file type (i.e the features extracted from an image file will be different from the features extracted from the audio file). The following are the different features that can be extracted from audio.

- Prosodic Features

  Several features like rhythm and intonation that human beings can recognize are known as prosodic features, or para-linguistic features as these features manage the components of speech that are properties of massive units as in sentences, words, syllables, and expressions and sentences. Prosodic features are extricated from massive units and, thus, are long-term features. These features are the ones passing on unique properties of emotional substance for speech emotion recognition. Energy, duration,

and fundamental frequency are some characteristics on which broadly utilized prosodic features are based.

- Spectral Features

The vocal tract filters a sound when produced by an individual. The shape of the vocal tract controls the produced sound. An exact portrayal of the sound delivered and the vocal tract is resulted by precisely simulated shape. The vocal tract features are competently depicted in the frequency domain. Fourier transform is utilized for obtaining the spectral features transforming the time domain signal into the frequency domain signal.

- Mel Frequency Cepstral Coefficients

The most widely used spectral feature in automatic speech recognition is Mel Frequency Cepstral Coefficient (MFCC). MFCCs represent the envelope of the short-time power spectrum, which represents the shape of the vocal tract. The utterances are split into various segments before converting into the frequency domain using short-time discrete Fourier transform to obtain MFCC. Mel filter bank is utilized to calculate several sub-band energies. After that, the logarithm of respective sub-bands is computed. Lastly, MFCC is determined by applying the inverse Fourier transform.

- Linear Prediction Cepstral Coefficients

Linear prediction cepstral coefficients (LPCC) captures the emotion-specific information expressed through vocal tract characteristics. There are differences between the characteristics and emotions. Linear Prediction Coeficient (LPC) is primarily equivalent to the even envelope of the log spectrum of the speech, and the coefficients of all the pole-filters are used for obtaining the LPCC by a recursive method. The speech signal is flattened before processing to avoid additive noise error as LPCCs are more exposed to noise than MFCCs.

- Gammatone Frequency Cepstral Coefficients

Gammatone frequency cepstral coefficients (GFCC) is computed by a method similar to that of MFCC, except that Gammatone filter-bank is applied in place of Mel filter bank to the power spectrum.

- Voice Quality Features

Irrespective of other spectral features, the voice quality features define the qualities of the glottal source. The impact of the vocal tract is expiated to a large extent by inverse filtering. Some of the

automatic changes might deliver a speech signal that may distinguish between various emotions utilizing the features like harmonics to noise ratio (HNR), shimmer, and jitter. The emotional content and voice quality of the speech have a compelling correlation between them.

- Teager Energy Operator Based Features

  Teager energy operator (TEO) was introduced by Teager and Kaiser. TEO was framed on the confirmation that the hearing process is responsible for energy detection. It has been perceived that under stressful conditions, there is a change in fundamental frequency and critical bands because of the distribution of harmonics. A distressing circumstance affects the speaker's muscle pressure, resulting in modifying the airflow during the sound creation.

The proposed system is based on audio and there are many features to determine the audio classification. But for the proposed system these are the features considered to classify the audio.

Audio Features used for the Proposed System :

Mfcc : Mel Frequency Cepstral Coefficient, represents the short-term power spectrum of a sound. It basically includes windowing the signal, taking the log of the magnitude and then wrapping the features on a mel scale, followed by the inverse DCT.

Chroma : Pertains to the 12 different pitch classes. High-level features intend to build user related concepts out of low and mid level features.

Mel : Mel Spectrogram Frequency. It is a scale of pitches judged by listeners to be equal in distance one from another.

3.2.4 TESTING AND TRAINING :

Training Data is the observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables. **Testing Data** is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

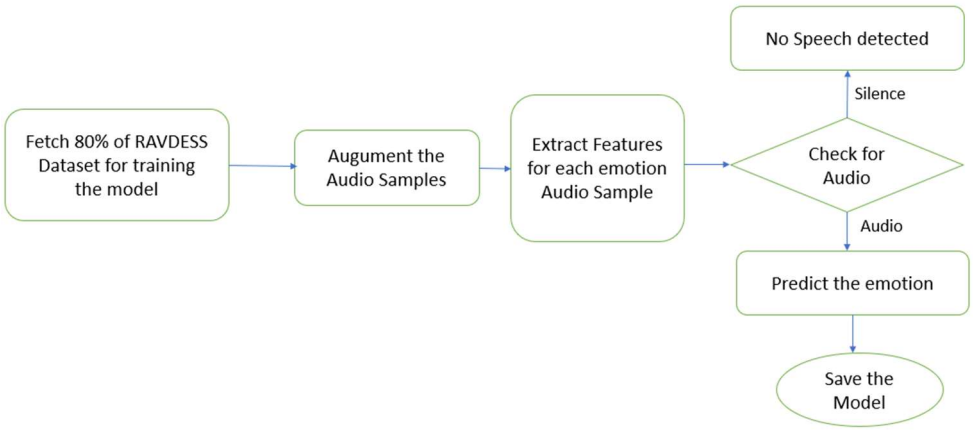


Fig : 3.2.4.1 : Training for the Proposed System.

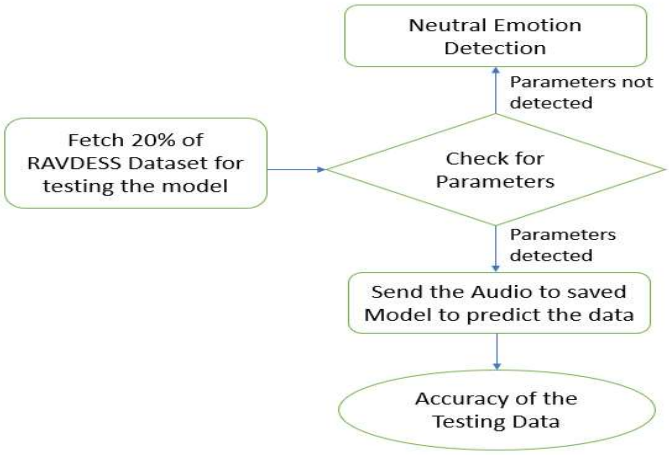


Fig : 3.2.4.2 : Testing for the Proposed System.

3.2.5 BUILDING THE MODEL :

After completing the training on the dataset, model is created with '.h5' extension. Saving the model and with the help of the built model, the emotion can be predicted from the speech.

The five steps to building a machine learning model include:

- Explore the data and choose the type of algorithm.
- Prepare and clean the dataset.
- Split the prepared dataset and perform cross validation.
- Perform machine learning optimization.
- Deploy the model.

3.2.6 USER AUDIO INPUT :

The user should give the audio input to the model, that he/she wants to predict the emotion, gender and the words spoken to be predicted and displayed on the screen.

3.2.7 PROCESS THE AUDIO :

Audio Processing means changing the characteristics of an audio signal in some way. Processing can be used to enhance audio, fix problems, separate sources, create new sounds, as well as to compress, store and transmit data.

There are two types of audio processing. Analog processing was the first and involves converting a sound wave into an electrical signal. Once it is in electrical form, the signal can be manipulated. The electrical signal used in analog devices closely resembles a sound wave, which allows the sound to be processed with the least amount of distortion. In digital audio processing, an audio signal is converted into digital information, often binary code, which can be interpreted by a computer. A digital signal changes the nature of sound from a continuous wave into discrete packages of information. For the proposed system includes digital processing for the silence removal, noise reduction, pre-emphasis .etc.

3.2.8 PREDICT AND DISPLAY THE EMOTION :

The output involves emotion of the person from the audio input, gender of the person and the words spoken by the person in the audio input are identified and printed on the screen. In order to predict and print the emotion the user has to specify the path of the audio file that he/she wants to know the emotion from the speech.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 FEATURES EXTRACTED :

The following figure depicts the total number of features extracted from the dataset used for training the model. A total of 180 features are extracted from the data given for training the model.

```
print((X_train.shape[0], X_test.shape[0]))
print(f'Features extracted: {X_train.shape[1]}')
```

```
(1152, 288)
Features extracted: 180
```

Fig : 4.1 : Features Extracted from the training of the RAVDESS dataset.

4.2 CLASSIFICATION REPORT & CONFUSION MATRIX :

Classification report is often defined as the performance metric for the machine learning model. It is used to determine the precision, recall, f1-score and support of the machine learning model. The following depicts the precision, recall, f1-score and support of the emotions (angry, calm, disgust, fearful, happy, neutral, sad, surprised) of the built machine learning model.

	precision	recall	f1-score	support
angry	0.71	0.69	0.70	52
calm	0.67	0.15	0.24	40
disgust	0.25	0.71	0.37	48
fearful	0.49	0.69	0.57	36
happy	0.55	0.36	0.44	47
neutral	0.12	0.03	0.05	29
sad	0.36	0.33	0.34	52
surprised	0.75	0.38	0.50	56
accuracy			0.44	360
macro avg	0.49	0.42	0.40	360
weighted avg	0.51	0.44	0.42	360

Fig : 4.2.1 : Classification Report for the Proposed System.

The following is the confusion matrix of the model of the given 8 (i.e 8*8 matrix) emotions.

```
[[36  0 10  1  2  0  2  1]
 [ 0  6 19  5  0  3  7  0]
 [ 3  0 34  3  4  1  2  1]
 [ 4  0  3 25  0  1  3  0]
 [ 6  0 13  2 17  2  4  3]
 [ 0  0 13  4  0  1 10  1]
 [ 1  3 19 10  1  0 17  1]
 [ 1  0 24  1  7  0  2 21]]
```

Fig : 4.2.2 : Confusion Matrix of the Proposed System.

4.3 ACCURACY:

The accuracy of the model using MLP (Multi Layer Perceptron) Classifier along with the LGB (Light Gradient Boost) combined by using the Voting Classifier resulted in higher accuracy compared to the previous systems.

```
for key in models.keys():
    fmodel=model(key,save=True,print_stat=False,cv=True)
```

```
Voting Classifier: MLP2, LGB CV Accuracy: 0.7013739883305101
```

Fig : 4.3 : Accuracy of the Proposed System.

4.4 OUPUT :

The user input audio file path should be specified in the predict function cell to predict the emotion of the speech and gender of the audio, along with the text spoken in the audio and they are displayed on the screen.

```
predict("/content/drive/MyDrive/Main Project/speech/Actor_14/03-01-01-01-01-01-14.wav")
```

```
['neutral_female', 'kids are talking by the door']
```

```
predict("/content/drive/MyDrive/Main Project/test.wav")
```

```
['surprised_female', 'dogs are sitting by the door']
```

```
predict("/content/drive/MyDrive/Main Project/Monday at 3-07 pm.wav")
```

```
['fearful_male', 'I am very afraid']
```

Fig : 4.4 : Output Emotion.

The output includes the gender of the person from the input audio file, along with the emotion of the person from the audio file and also displays the spoken text by the user in the audio file.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION :

The proposed system indicates that MLPs are very effective in classifying speech signals. Even with simplified models, a restricted set of characters can be identified without problems. The proposed system acquired greater accuracies as in contrast to different techniques for character emotions. The accuracy of the proposed model is 70%. The overall performance of a module is pretty structured on the great of pre-processing. Every human emotion has been completely studied, analyzed and the accuracy has been checked. The consequences received in this learning about identifying the speech emotion is feasible, and that MLPs can be used for any assignment regarding recognizing of speech and demonstrating the accuracy of every emotion existing in the speech.

5.2 FUTURE SCOPE :

Though the proposed system, leveraged Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots etc. A few possible steps that can be implemented to make the models more robust and accurate. Like an accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model. And also figuring out a way to clear random silence from the audio clip. Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum. And also Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

REFERENCES

[1] Tuomas Eerola and Jonna K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music", Psychology of Music, pp. 1–32, The Author(s) 2010.

[2] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell, "On the Importance of Both Dimensional and Discrete Models of Emotion" School of Psychology, The University of New South Wales, Australia.

[3] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey & Marc Schröder, 'FEELTRACE' Schools of Psychology and English, Queen's University Belfast.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–30, 2012.

[5] "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset," Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications(IAAI-17)).

[6] Chung, Seong Youb, and Hyun Joong Yoon. "Affective classification using Bayesian classifier and supervised learning."Control,Automation and Systems (ICCAS), 2012 12th International Conference on. IEEE, (2012).

[7] Jaebok Kim, Khiet P. Truong, Gwenn Englebienne, and Vanessa Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition",Human Media Interaction, University of Twente, Enschede, The Netherlands, 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).

[8] Babak Basharirad, and Mohammadreza Moradhaseli,"Speech emotion recognition methods: Literature review",A IP Conference Proceedings 1891, 020105 (2017).

[9] Kim, Jaebok and Englebienne, Gwen and Truong, Khiet P and Evers, Vanessa, "Towards Speech Emotion Recognition ``in the wild'' using Aggregated Corpora and Deep Multi-Task Learning", 'Proceedings of the INTERSPEECH', year-2017.

[10] B. Yang, M. Lugger, "Psychological motivated multistage emotion classification exploiting voice quality feature." F. Mihelic, J. Zibert, Speech Recognition, InTech, 2008, chapter 22.

[11] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 5–20, Jan.

2011.

[12] Steven R. Livingstone, Frank A. Russo,"The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English",journal.pone.0196391,May 16, 2018.

[13] Monorama Swain, Aurobinda Routray, Prithviraj Kabisatpathy,"Databases, features and classifiers for speech emotion recognition: a review", I. J. Speech Technology 2018.

[14] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, vol. 15, no. 2, pp. 84–115, 2012.

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems.

PSO2 : Design and develop web sites, web apps and mobile apps.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

Project Outcomes	
Course Outcome (CO1)	Identify and analyze the problem statement using prior technical knowledge in the domain of interest.
Course Outcome (CO2)	Design and develop engineering solutions to complex problems by employing systematic approach.
Course Outcome (CO3)	Examine ethical, environmental, legal and security issues during project implementation.
Course Outcome (CO4)	Prepare and present technical reports by utilizing different visualization tools and evaluation metrics.

Mapping Table

CS1537 : MAIN PROJECT															
Course Outcomes	Program Outcomes and Program Specific Outcome														
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12		PSO 1	PSO 2
CO1	3	3	1					2	2	2				1	1
CO2	3	3	3	3	3			2	2	2		1		3	3
CO3	2	2	3	2	2	3	3	3	2	2	2			3	
CO4	2		1		3				3	3	2	2		2	2

Note: Map each project outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped

SPEECH BASED EMOTION RECOGNITION USING MACHINE LEARNING

[1.] G. Hrithik, [2.] K. Devi Kiran, [3.] K. N. V. Naresh,
[4.]G. Bharathi

Student, Department of CSE, SeshadriRaoGudlavalleruEngineeringCollege, Seshadri Rao Knowledge Village, Gudlavalleru, Andhra Pradesh, India
Sr. Gr. AssistantProfessor, Department of CSE, SeshadriRaoGudlavalleruEngineeringCollege, Seshadri Rao Knowledge Village, Gudlavalleru, Andhra Pradesh, India

hrithikgunduboyina@gmail.com
kiranreddy1738@gmail.com
mahinaresh412@gmail.com
gbharthi@gmail.com

.

ABSTRACT-**In intelligent speech applications, Speech Emotion Recognition (SER) is critical. The act of understanding human emotions and emotional states from speech is known as speech recognition. Human emotions can be expressed in a variety of ways, including bodily posture, face expression, and speech. Pitch, timbre, loudness, and vocal tone are some of the characteristics of the human voice. Humans have been found to express their emotions by changing voice characteristics during speech production. The suggested method uses Python modules such as PyAudio and Librosa for audio input and analysis of the audio, as well as the MLP Classifier for feature extraction and classification. The suggested method reports accuracy, f-score, precision, and recall for tested models in various experiment scenarios. However, earlier efforts have been hampered by a lack of attention to detail in the emotion prediction.**

1. INTRODUCTION

Speech emotion recognition (SER) is becoming more significant in a variety of applications like naturalistic human-computer interaction (HCI). Speech emotion recognition is now a popular study in the fields of signal processing and pattern recognition. The first part of a voice emotion processing and recognition system is signal collection, followed by feature extraction, and finally emotion recognition. The neural network-based approach is the most advantageous technology for speech recognition. Artificial Neural Networks (ANN) are information-processing technologies inspired by biological neural networks. Artificial neural networks (ANN) require no prior knowledge to simulate speech recognition.

2. LITERATURE SURVEY

Jianfeng Zhao et al.,proposedto understand speech emotion by analysing deep aspects from several data sources. To investigate the high-level points from unprocessed audio recordings and log-mel spectrograms, the

authors created a merged convolutional neural community (CNN) with two branches, one of which was a one-dimensional (1D) CNN branch and the other was a 2D CNN branch. There are two phases involved in creating the merged deep CNN. After designing and evaluating one 1D CNN and one 2D CNN architecture, the two CNN designs were fused together after removing the second dense layer. Switch learning was included in the training to help speed up the instruction of the combined CNN. The first to be trained were the 1D and 2D CNNs.Finally, the fine-tuning of the combined deep CNN initialised with transferred facets was performed.

S. Koelstra et al ,proposed a speech emotion detection model i.e a multimodal data set for the investigation of human affective states, was provided. While watching 40 one-minute long samples of music videos, the electroencephalogram (EEG) and peripheral physiological data of 32 subjects were monitored. Each film was scored on arousal, valence, like/dislike, dominance, and familiarity by the participants. The use of emotive tags offered as a unique way for stimuli selection. The results of a thorough examination of the participants' ratings throughout the experiment are presented. The researchers look into the correlations between EEG signal frequencies and participant ratings.

Babak Basharirad et al.,proposed a model due to the availability of high computation capability, the attention of the emotional speech signals research has been enhanced in human machine interfaces. Many approaches for detecting emotional states through speech have been proposed in the literature. The fundamental problems in speech emotion detection systems are the selection of adequate feature sets, the creation of proper classification methods, and the preparation of an appropriate dataset. Based on the three evaluation parameters, this research is critically examined the existing accessible techniques to speech emotion recognition algorithms (feature set, classification of features, accurate prediction). In addition, the performance and limits of available approaches are evaluated in this research. It also emphasises the present promising approach for voice emotion recognition system advancement.

3.PROPOSED WORK

Using the librosa, PyAudio modules and the Multi Layer Perception Classifier (MLP Classifier) and the RAVDESS Dataset, to create a model to discern emotion from speech. From sound files, the suggested system will be able to discern emotion. The data will be loaded, features extracted, and the dataset divided into training and testing sets. The model will then be trained using an MLP Classifier that has been initialised. Finally, we'll display the emotion detected from the user input audio. The suggested work achieves more accuracy and precision than existing systems, as well as the detection of more emotions than the prior system.

3.1 IMPLEMENTATION

3.1.1 MLP CLASSIFIER

A feedforward artificial neural network model, the Multi Layer Perceptron (MLP), maps input data sets to a collection of relevant outputs. An MLP is made up of numerous layers, out of which each is fully connected to the one before it.Multilayer Perceptrons are frequently trained on a set of input-output pairs and they are learned to model the correlation (or dependencies) between the input-output pairs.

.

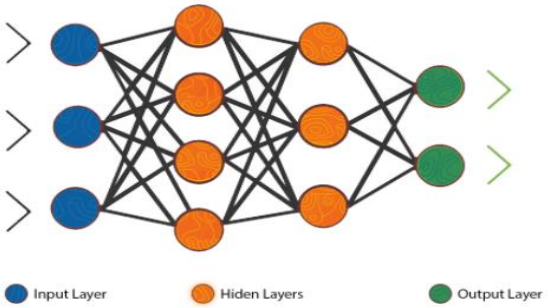


Fig 1:MLP Structure

3.1.2 ARCHITECTURE

The following diagram depicts the workflow of the proposed speech based emotion recognition system. At first, the audio is collected and pre-processed using the python library Librosa for removing the silence and noise reduction from the audio files. The extracted features from the RAVDESS dataset that undergoes training and testing process are stored based on their classification and a machine learning model is created. Now the input audio clip gets the feature extraction done and the features extracted are compared with the model that had undergo testing and training from the RAVDESS dataset. Finally, the output emotion is predicted and displayed on the screen.

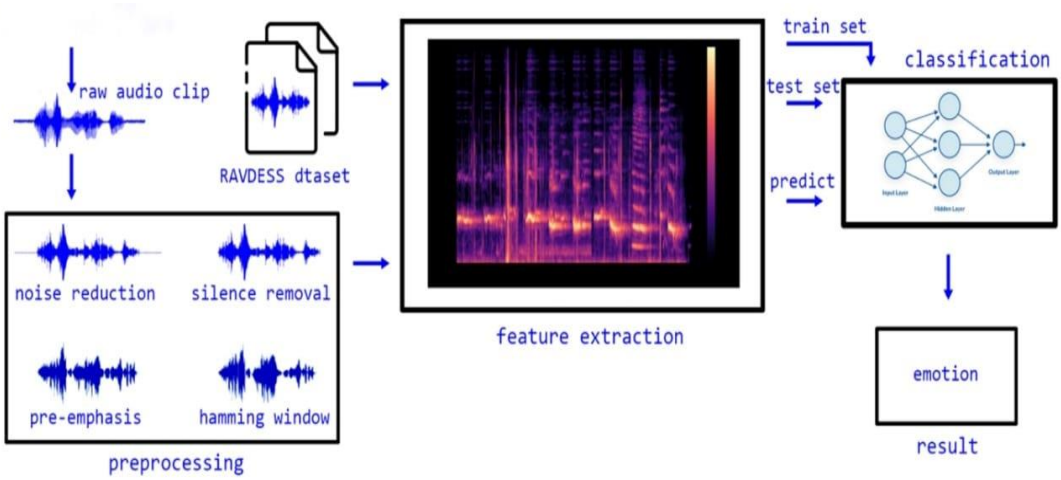


Fig 2:Work Flow of the proposed system

3.1.4. RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) includes 7356 files, which is nearly 25GB of size. The database is said to contain 24 professional actors out of which 12 are female and 12 are male, vocalizing two lexically-matched statements.

The speech in the datasetincludes 8 emotions such as angry, fearful, sad, calm, neutral, disgust, happy and surpriseexpressions. Even though the database contains the songs and vidoes, since the proposed system is speech emotion detection we are taking only speech files as the dataset for testing and training of the model.

Speech folder contains 1440 files: 60 trials per actor x 24 actors = 1440. Each of the 7356 RAVDESS files has filename that is unique.(ex : 02-01-06-01-02-01-12.mp4).

Identifier	Coding description of factor levels
Modality	01 = Audio-video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06 = Fearful, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
Repetition	01 = First repetition, 02 = Second repetition
Actor	01 = First actor, . . ., 24 = Twenty-fourth actor

We have get this dataset from thekaggle

https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio#:~:text=Ryerson%20Audio%2DVisual,number%20is%20even

3.1.5 EMOTIONS DETECTING :

The following emotions will be predicted in the proposed speech based emotion recognition system. The proposed system also returns the gender of the user input audio file that is sent for detecting the emotion.

The following figure indicates the name of the file for predicting the audio for classification in the model building(i.e if the file name contains '01' indicates the audio file is neutral emotion, which are required during the testing and training process of model building).

```
emotions={
  '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}

def gender(g):
    if int(g[0:2]) % 2 == 0:
        return 'female'
    else:
        return 'male'
```

4 FEATURE EXTRACTION

The feature extraction process extracts the features from the given dataset. However, the features extracted will be different from each other based on the file type(i.e the feature extraction for an image file will be different from the feature extraction for the audio file).

The proposed system is based on audio and there are many features to determine the audio classification. But for the proposed system these are the features used to classify the audio.

Audio Features :

Mfcc : Mel Frequency Cepstral Coefficient, defines the sound's short-term power spectrum. It entails windowing the signal, calculating the log of the magnitude, wrapping the features on a mel scale, and also performing the inverse DCT.

Chroma : Pertains to the 12 different pitch classes. High-level features intend to build user related concepts out of low and mid level features.

Mel : Mel Spectrogram Frequency. It defines the pitches judged by listeners to check the distance to be equal from each another.

5.RESULTS AND DISCUSSION

5.1 Features Extracted :

The following depicts the total number of features extracted from the dataset used for training the model. The feature extraction resulted by extracting 180 features from the dataset given for training the model.

```
print((X_train.shape[0], X_test.shape[0]))
print(f'Features extracted: {X_train.shape[1]}')
```

```
(1152, 288)
Features extracted: 180
```

5.2 Classification Report & Confusion Matrix :

Classification report is often defined as the performance metric for the machine learning model. It is used to determine the precision, support, f1-score and recall of the machine learning model. The following depicts the support, precision, f1-score and recall of the emotions (angry, fearful, sad, calm, neutral, disgust, happy and surprise expressions) of the built machine learning model.

	precision	recall	f1-score	support
angry	0.71	0.69	0.70	52
calm	0.67	0.15	0.24	40
disgust	0.25	0.71	0.37	48
fearful	0.49	0.69	0.57	36
happy	0.55	0.36	0.44	47
neutral	0.12	0.03	0.05	29
sad	0.36	0.33	0.34	52
surprised	0.75	0.38	0.50	56
accuracy			0.44	360
macro avg	0.49	0.42	0.40	360
weighted avg	0.51	0.44	0.42	360

The confusion matrix is also one of the performance metric for evaluating the efficiency of the model. The following is the confusion matrix of the model of the given 8 (i.e 8*8 matrix) emotions.

```
[[36  0 10  1  2  0  2  1]
 [ 0  6 19  5  0  3  7  0]
 [ 3  0 34  3  4  1  2  1]
 [ 4  0  3 25  0  1  3  0]
 [ 6  0 13  2 17  2  4  3]
 [ 0  0 13  4  0  1 10  1]
 [ 1  3 19 10  1  0 17  1]
 [ 1  0 24  1  7  0  2 21]]
```

5.3 Accuracy:

The accuracy of the model is the ultimate metric in evaluating the performance of the machine learning model. The accuracy of the proposedmodel using MLP(Multi Layer Perceptron) Classifier along with the LGB(Light Gradient Boost) combined by using the voting classifier resulted in higher accuracy compared to the previous systems.

```
for key in models.keys():
    fmodel=model(key,save=True,print_stat=False,cv=True)
```

```
Voting Classifier: MLP2, LGB CV Accuracy: 0.7013739883305101
```

5.4 Output:

The user input audio file should be specified in the predict function cell to predict the emotion of the speech and gender of the audio, along with the text spoken in the audio and they are displayed on the screen.

```
predict("/content/drive/MyDrive/Main Project/speech/Actor_14/03-01-01-01-01-01-14.wav")
```

```
['neutral_female', 'kids are talking by the door']
```

```
predict("/content/drive/MyDrive/Main Project/test.wav")
```

```
['surprised_female', 'dogs are sitting by the door']
```

```
predict("/content/drive/MyDrive/Main Project/Monday at 3-07 pm.wav")
```

```
['fearful_male', 'I am very afraid']
```

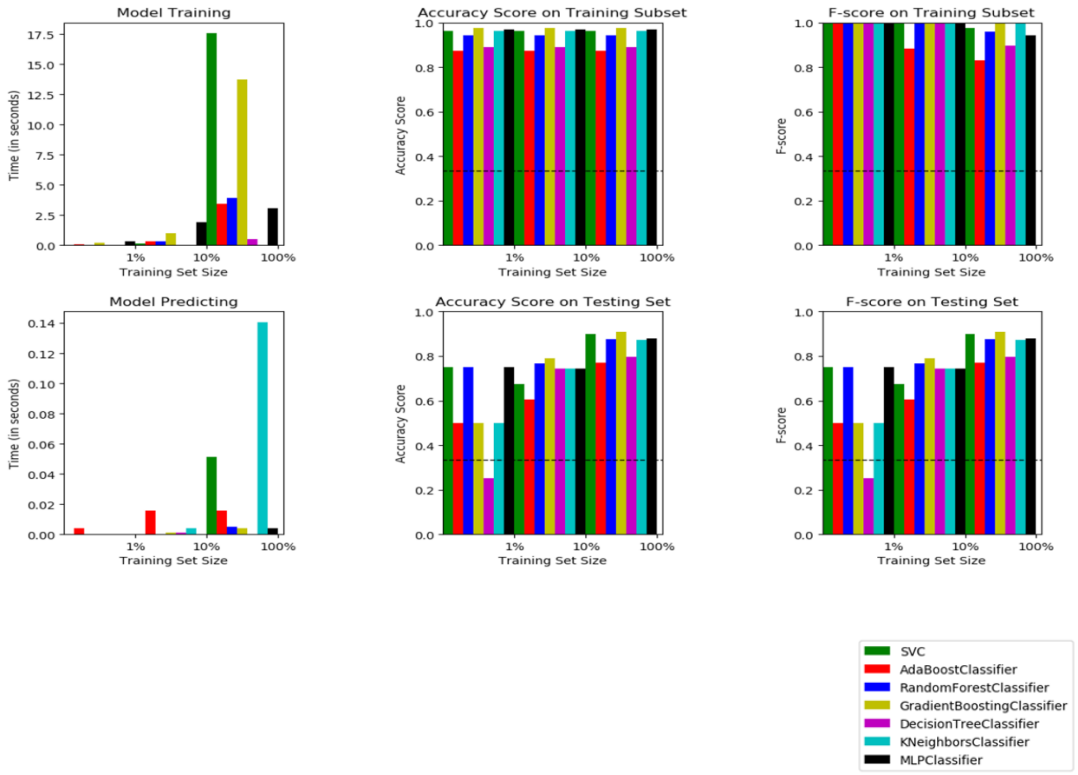


Fig 3 : Study on different Classifiers and their accuracies in Detecting emotion from speech.

The above graph depicts the time taken for predicting the emotion, accuracy scores and f-scores of both the testing set and training set of different classifiers in predicting the emotion from speech. The lesser the time taken for prediction indicates the better efficiency of the model. Some classifiers produced better results of accuracy scores and some classifiers produced better results of f-scores. However, the classifiers that produced greater results in accuracy and f-scores are taking more time in the model prediction. We observed that the MLP Classifier performs consistently better in both the accuracy and f-scores and also predicts the model in less time compared to other classifiers.

6.CONCLUSION

The paper indicates based on the proposed work that the MLPs are very accurate and effective in classifying and predicting emotion from the speech signals. Even with limited models, a restricted set of features can be extracted without problems identified. We have acquired greater accuracies as in contrast to different techniques for character emotions. The accuracy of the proposed model is 70%. The overall performance of the model is pretty structured on the great of pre-processing. Almost, every human emotion required for the proposed system, has been completely studied, analyzed and the accuracy has been checked.The consequences received in this learn about display that speech focus is feasible, and that MLPs can be used in recognizing the speech and depicting the accuracy of every emotion existing in the speech.

REFERENCES

[1] TuomasEerola and Jonna K. Vuoskoski, “A comparison of the discrete and dimensional models of emotion in music”, Psychology of Music, pp. 1–32, The Author(s) 2010.

[2] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell, “On the Importance of Both Dimensional and Discrete Models of Emotion” School of Psychology, The University of New South Wales, Australia.

[3] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey& Marc Schröder, ‘FEELTRACE’ Schools of Psychology and English, Queen's University Belfast.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis using physiological signals,” IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–30, 2012.

[5] “Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset,” Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications(IAAI-17)).

[6] Chung, SeongYoub, and Hyun Joong Yoon. ”Affective classification using Bayesian classifier and supervised learning.”Control,Automation and Systems (ICCAS), 2012 12th International Conference on. IEEE, (2012).

[7] Jaebok Kim, Khiet P. Truong, Gwenn Englebienne, and Vanessa Evers, “Learning spectro-temporal features with 3D CNNs for speech emotion recognition“,Human Media Interaction, University of Twente, Enschede, The Netherlands, 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).

[8] Babak Basharirad, and Mohammadreza Moradhaseli,“Speech emotion recognition methods: Literature review”,A IP Conference Proceedings 1891, 020105 (2017).

[9] Kim, Jaebok and Englebienne, Gwen and Truong, Khiet P and Evers, Vanessa, ”Towards Speech Emotion Recognition ``in the wild" using Aggregated Corpora and Deep Multi-Task Learning”, ’Proceedings of the INTERSPEECH’, year-2017.

[10] B. Yang, M. Lugger, “Psychological motivated multistage emotion classification exploiting voice quality feature.” F. Mihelic, J. Zibert, Speech Recognition, InTech, 2008, chapter 22.

[11] C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels,” IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 5–20, Jan. 2011.

[12] Steven R. Livingstone, Frank A. Russo,”The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English”,journal.pone.0196391,May 16, 2018.

[13] Monorama Swain, AurobindaRoutray, PrithvirajKabisatpathy,"Databases, features and classifiers for speech emotion recognition: a review", I. J. Speech Technology 2018.

[14] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, vol. 15, no. 2, pp. 84–115, 2012.