Republic of Iraq
Ministry of Higher Education and Science Research
University of Babylon
College of Science for Women
Department of Computer Science

Speaker Identification Using Deep Learning Approaches

A Project

Submitted to the Council of the College of Science for Girls, University of Babylon, in partial fulfillment of the requirement to obtain a higher diploma in computer science

by:

Zahraa Adel Ali

Supervised by:

Dr. Ali Y. Yousif Al-Sultan

15 June 2022

قال تعالى :

"سورة الكهف، آية: ٦٦"

Supervisor Certificate

I certify that the project called "**Speaker Identification Using Deep Learning Approach**" was prepared in the Department of Computer Science/College of Science for women/Babylon University, by (**Zahraa Adel Ali)** as partial fulfillment of the requirements for the Higher Diploma in Computer Science.

Signature:

Name: **Dr. Ali Y. Yousif Al-Sultan**

Date: / /2022

Address: Computer Science Dept. College of Science for Women, University of Babylon, Babylon, Iraq.

The Head of the Department Certification

In view of the available recommendations, I forward the research entitled "**Speaker Identification Using Deep Learning Approach"** for debate by the examination committee.

Signature:

Name: Asst. prof. Dr Saif Mahmood Alalak

Date: / / 2022

Address: University of Babylon/College of Science for Women

Certification of the examination Committee

We are the chairman and members of the examination committee, certify that we have studied this project entitled (**Speaker Identification Using Deep Learning Approaches**) presented by the student (**Zahraa Adel Ali)** in the contents and is related with and that in our opinion it is adequate with (**Excellent**) standing as a project for the degree of higher diploma in Computer Science.

Signature:

Name: Dr. Suhad Ahmed Ali

Scientific Title: Professor

Address: University of Babylon /

College of Science for Women

Computer Science Department

Date: / /

(Chairman)

Signature:

Name: Dr. Mohammed Obeid Mahdi

Scientific Title: Assistants Professor

Address: University of Babylon /

College of Science for Women

Computer Science Department

Date: / /

(Member)

Signature:

Name: Dr. Ali Yakoob Yousif Al-Sultan

Scientific Title: Lecturer

Address: University of Babylon /College of Science for Women

Computer Science Department

Date: / /

(Supervisor)

Approved by the dean of college of Science for Women, University of Babylon

Signature:

Name: Prof. Dr. Abeer F. Al-Rubaei

Scientific Title: Professor

Date: / /

Dean of College of Science for Women-University of Babylon.

Dedication

I Dedicate This Work

To My Parents

And All People

Who Helped Me

During

My Years of Study.

Zahraa Adel Ali

2022

Acknowledgments

I would like to express my gratitude and thanks to my supervisor "Dr. Ali Y. Yousif Al-Sultan" Who help me throughout this thesis. I also would like to express my wholehearted thanks to my family for their generous support they provided me throughout my entire life.

Last, I would also like to thank the people who have helped me and inspired me during my High diploma study.

Zahraa Adel Ali

2022

ABSTRACT:

Humans can recognize a speaker by listening to their speech; nevertheless, deep learning algorithms have a significant problem in acquiring this essential human particular talent. Deep learning, like human listeners, employs aspects of a speaker's voice to determine the speaker's identity. The computational challenge of identifying speakers using features collected from their voices is known as speaker identification. In this project , a deep learning model for speaker identification based on a convolution neural network (CNN) is developed. The suggested CNN-based technique employs the traditional Mel-frequency Cepstral coefficients (MFCCs)-based feature extraction method, which is the most widely used for audio and speech signal feature extraction. This research article provides a quick overview of the speaker identification system before delving into the overall architecture of the speaker identification system utilizing the CNN model. the suggested CNN approach is compared with the LSTM method, which is performed 100 times on the same dataset as the tests. For comparison, the maximum and average categorization accuracy are shown. The suggested CNN-based technique have 99.83% average accuracy. It reveals the proposed CNN-based method gets the highest accuracy of 99.86%.and the average accuracy is 97.6%, When the LSTM method is used, and it exposes not a very good result, only 81% on an average accuracy. the proposed system compared with a previous study based on Gated Recurrent Unit(GRU) method and it exposes 98% high accuracy and 91% average accuracy result, the average accuracy of GRU method is less than the proposed method around 6%. And the average accuracy of LSTM method is less than the proposed method around 17% lower than that of the suggested method. As a result, the suggested CNN network model outperformed all others in terms of model training duration, recognition accuracy, and stability. The proposed approach for speaker identification is quite effective.

List of Contents

LIST OF FIGURES:

Symbol	Meaning
CNN	**Convolutional Neural Network**
LSTM	**Long short Term Memory**
MFCC	**Mel Frequency Cepstral Coefficients**
GRU	**Gated Recurrent Unit**
GMM_DNN	**Gaussian Mixture Model**
DNN	**Deep Neural Network**
CGAN	**Conditional Generative Adversarial Network**
ResNet	**Residual Network**
EER	**Equal Error Rate**
ML	**Machine Learning**
AI	**Artificial Intelligence**
ANN	**Artificial Neural Network**
PLP	**Perceptual Linear Predictive**
ReLU	**Rectifier Linear Unit**
wandb	**Weights And Biases**
FFT	**(Fast Fourier Transform)**
DCT	**Discrete cosine Transform**

Chapter One

Introduction

Chapter One
Introduction

1.1 Introduction

In the daily lives, speech is most likely the most important instrument for communication. As a result, developing a voice recognition system is always desired. Speech recognition is the process of converting an audio signal into a collection of words. As with instructions and control, data entry, and document creation software. Voice recognition technology may be separated into two sub-areas: speech recognition and speaker recognition [1] , Speech recognition is a method of assessing the contents of a speaker's speech. Many algorithms are used by voice recognition systems to transform sound waves into useful data for processing, which is subsequently interpreted by speech recognition techniques. [2]. The computational process of confirming a user's stated identification using extracted information from their speech is known as speaker recognition. Speaker identification and speaker verification are two subsets of speaker recognition. Speaker identification is the method of identifying the speaker from a particular statement by comparing the statement's voice biometrics to prior statements' models. The classical Mel-frequency cepstral coefficients (MFCCs) approach is a typical feature extraction method used for speech signal identification systems. MFCCs are utilized as feature extractors in speech signal identification systems to create speaker models for identification [3]. The fundamental goal of feature extraction is to portray a voice signal using a preset number of signal characteristics. This is due to the fact that all of the information in the acoustic signal is too vast to cope with, and part of it is unrelated to the identification job. [4].

1.2 Problem Statements

The problem of identification system nowadays is that the system can simply be fooled. There are still certain ways to trick the system despite the use of biometric identity that is unique from everyone else.

Speaker Identification systems are useful for confirming a person's identification and for automatically regulating voice-activated services like banking transactions and the delivery of sensitive data. It is possible to think of speech as a non-invasive biometric that may be recorded either consciously or unconsciously and sent across large distances by telephone.

1.3 Project Objective

The proposed work aims to build a Speaker Identification system using Mel Frequency Cepstral Coefficients (MFCC) and deep learning approach using Convolutional Neural Network (CNN) model.

1.4 Project Motivation

Speaker Identification is a useful biometric identification technique that has been used in several industries, including computers, voice dialing devices, banking, databases, and extremely guarded regions. Speaker identification has long attracted increasing interest from academics in several sectors of information security due to the distinctive features of voice signals.

1.5 Related Works

Speaker Identification systems have become area of intense research in recent years due to its wide range of applications in voice matching biometric identification, mobile access security, health care management and transportation below some of related works of Speaker Identification systems:

Shanin et, al 2020 [5], present an effective approach to enhance text-independent speaker identification performance in emotional talking environments based on novel classifier called cascaded Gaussian Mixture Model-Deep Neural Network (GMM-DNN). this work focuses on proposing, implementing and evaluating a new approach for speaker identification in emotional talking environments based on cascaded Gaussian Mixture Model-Deep Neural Network as a classifier. The results point out that the cascaded GMM-DNN classifier improves speaker identification performance at various emotions. This system has low the in angry and noisy talking environments with reduced computational complexity .

Chen,et, al, 2020 [6] ,present speaker identification with the conditional generative adversarial network (CGAN). It allows the adversarial networks for distinguishing real/fake samples and predicting class labels simultaneously. He configure the generator and the discriminator in SpeakerGAN with the gated convolutional neural network (CNN) and the modified residual network (ResNet) to obtain generated samples of high diversity as well as increase the network capacity. The multiple loss functions are combined and optimized to encourage the correct mapping and accelerate the convergence. Experimental results show that SpeakerGAN reduces the classification error rate by 87% and 16% compared with the traditional i-vector system and the state-of-the-art DNN based method. Under the scenario of limited training data, SpeakerGAN obtains significant improvement over the baselines. In the case of taking 1.6 s of each speaker for testing, SpeakerGAN achieves the identification accuracy of 98.20%, which suggests the promise for short-utterance speaker identification.

Jahangir, et, al, 2021 [7] presents hierarchical classification approach for speaker identification using robust time domain features. In hierarchical classification approach, the first level classifier identifies the gender voice (i.e. male or female voice). In addition, the second level classifier identifies the specific speaker voice. The experiments were conducted using the speech data

collected from ten subjects including male and female to ensure the generalized model evaluations. Several time domain features were extracted from the collected dataset that were proven highly discriminative for gender identification and specific speaker identification. The proposed method helps to reduce computational time with increase speaker identification rate. The experimental results obtained highest accuracy of 96.9% using random forest classifier for gender identification. Moreover, for specific speaker identification the highest accuracy of 78% was observed in male speaker identification and 88.7% accuracy was obtained in female speaker identification using random fores.

Nugroho, et,al, 2021 [8] proposes a Data Augmentation strategy using Adding White Noise techniques, Pitch Shifting, and Time Stretching, which are processed using a Deep Neural Network to produce a new model in speaker identification as an approach called as DA-DNN7L. The Data Augmentation approach is used as a solution to increase the limited data quantity of Indonesian ethnic speakers, while the seven layer DNN is an architecture that provides the best accuracy performance compared to other multilayer approach models, besides that the 7 layer approach used in several other studies achieves a high degree of accuracy. Research that has been carried out using the best performance seven-layer Deep Neural Network Data Augmentation strategy resulted in an accuracy rate of 99.76% and a loss of 0.05 in the 70%:30% split ratio and the addition of 400 augmentation data. After seeing the performance of this model, it can be concluded that Data

Augmentation Deep Neural Network can improve the speaker's recognition performance using the Indonesian ethnic dataset.

Feng Ye, et,al,2021 [9] propose a deep neural network (DNN) model based on gated recurrent unit (GRU) for speaker identification. In the network model design, the convolutional layer is used for voiceprint feature extraction and reduces dimensionality in both the time and frequency domains, allowing for faster GRU layer computation .The network models were evaluated on the Aishell-1 speech dataset. The experimental results showed that our proposed DNN model, which we call deep GRU, achieved a high recognition accuracy of 98.96%.

Chapter Two
Theoretical Background

Chapter Two
Theoretical Background

2.1 Introduction

Speaker identification is a task of identifying persons from their voices. The goal of speaker identification is to locate the speaker using distinguishing traits gleaned from voice signals. Speaker identification is the process of identifying anonymous speakers using one-to-many comparisons (1: n), in which the speaker's voice is compared to the voices of speakers who are listed in a database whose fundamental structure is the speaker's identification. The extraction and categorization of characteristics are both parts of a speaker identification technique. Features Extraction is the approach to data reduction with the best representative features. Mel-Frequency Cepstral Coefficients (MFCCs) are the most exact feature extraction approach for speaker identification [10].Classifications are another critical component of the speaker identification system. Patterns are classified into several classes at the categorization level. DWT, GMM, SVM, and Neural Networks are among the classifiers employed. Neural Networks is one of the most used methodologies for classification. The selection of this classifier is a significant challenge. However, there are no predefined criteria for selecting classifiers. Several pattern classifiers are being explored for the development of language processes such as emotion categorization, voice recognition, speech verification, and speech recognition [11].

2.2 Artificial Intelligence (AI)

Artificial intelligence (AI) is a discipline of computer science that focuses on creating intelligent machines that behave and act like humans. Artificially intelligent computers are designed to do a wide range of

functions such as speech recognition, learning, planning, and problem-solving. [12].. A subfield of computer science called artificial intelligence seeks to build intelligent machines. It is now a crucial component of the technological sector. Deep Learning has made it possible for machine learning to be used in many real-world contexts, and by extension, for the whole field of AI. All types of machine assistance appear plausible, if not realistic, thanks to the way deep learning breaks down jobs. Better preventative healthcare, driverless automobiles, and even better movie suggestions are all either now available or in the future. AI is both the here and now. AI may even reach the science fiction world we've long envisioned with the aid of deep learning. The essence of machine learning is the use of algorithms to analyze data, learn from it, and then decide or predict anything about the world. Instead of manually programming software routines with a specific set of instructions to perform a task. [12]., the machine is "trained" using massive quantities of data and algorithms to learn how to carry out the operation. Artificial neural networks (ANNs) and the more advanced deep learning technologies are the most effective AI tools for handling exceedingly tough problems, and both will be enhanced and employed in the future. figure (2-1) depicts a schematic of the artificial intelligence components.

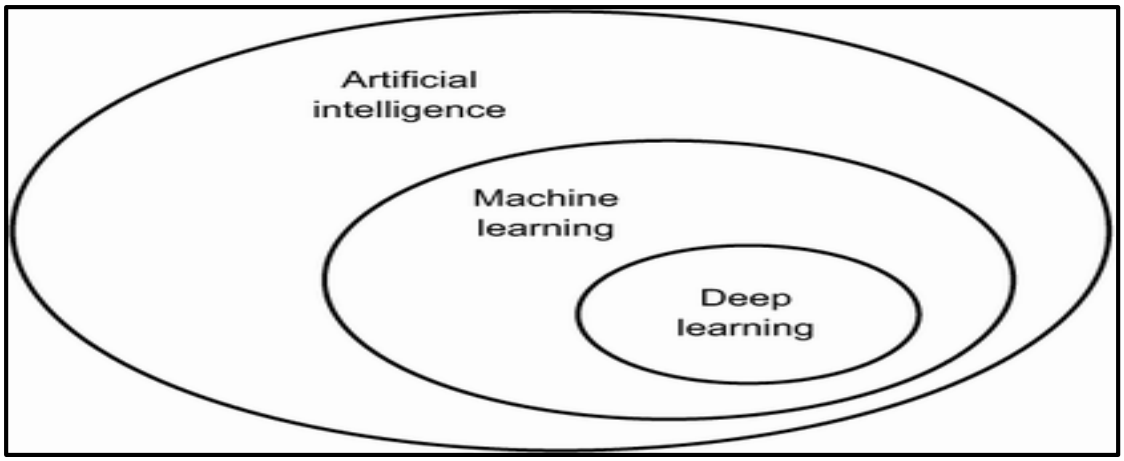


FIGURE 2-1 DIAGRAM OF THE COMPONENTS OF ARTIFICIAL INTELLIGENCE. [12].

2.3 Machine learning:

Machine learning is a broad term for computer algorithms that utilize experience to improve performance or make precise predictions. In this sense, experience refers to prior information that the learner has access to, which is typically in the form of electronically acquired data that is made available for analysis. This knowledge may take the form of digitalized training sets with human labels or some other kind of information gleaned through interacting with the environment. Its size and quality are always essential to the accuracy of the learner's predictions. Machine learning involves creating precise and effective prediction systems [13]. The time and space complexity of these algorithms, like that of other disciplines of computer science, are essential markers of their excellence. However, to estimate the sample size required for the algorithm to learn a family of concepts, we will also require a sample complexity concept in machine learning. More broadly, the size of the training sample and the complexity of the idea classes under consideration influence an algorithm's theoretical learning guarantees. Because the success of a learning system is dependent on the data it utilizes, data analysis and statistics are closely tied to machine learning. Learning techniques, in a wide sense, are data-driven approaches that integrate fundamental computer science principles with ideas from statistics, probability, and optimization [13].

2.4 Deep Learning

Deep learning is a subset of machine learning that focuses on the acquisition of successive layers of more meaningful representations. Deep learning is a unique way to learning data representations. The term "deep learning" refers to several layers of representations rather than any type of

deeper knowledge obtained by the process. The number of layers of a data model is referred to as its depth [14] Deep learning techniques of today commonly include tens or even hundreds of representational layers that are automatically learned by exposure to training data. Other machine learning approaches, known as shallow learning, are more likely to focus on learning only one or two layers of data representations (for example, collecting a pixel histogram and then applying a classification algorithm). These layered representations are taught in deep learning utilizing neural network models that are physically formed in layers on top of one another. Although some of the basic principles in deep learning were inspired in part by our knowledge of the brain [12].

2.5 Artificial Neural Network

Artificial neural networks (ANN) process input signals in the same way as the human brain do and convert them to output signals [15]. There are strong modeling techniques available, which allow for nonlinearity between feature variables and output signals. Without making any assumptions about a specific function, ANN may learn from data. Neural networks are made up of processing components that generate networks with weighted functions for each input. These components are often organized in a succession of layers with numerous connections. The structure of neural networks is made up of three main sorts of layers. an input layer that receives data from external sources, a hidden layer that does calculations according to the function given, and an output layer that generates output based on input [16].similar to the human brain Neuron cells, which number in the billions, make up the brain. Each neuron is composed of a cell body that carries and processes information to and from the brain (inputs and outputs) [17]. The fundamental notion of these

networks is based on how the biological brain system operates, which involves processing input and information to promote learning and knowledge production. The fundamental component of this approach is the design of new structures for the information processing system. [18]. Figure (2-2) depicts the architecture of an artificial neural network.the system is composed of multiple, intricately linked processing units known as neurons, which work together to solve problems and convey data via synapses, or electromagnetic connections. The neurons are organized into layers and are densely interconnected. The input layer accepts data while the output layer generates the result. One or more concealed layers are usually put between the two. Because of this setup, forecasting or knowing the precise flow of data is difficult.

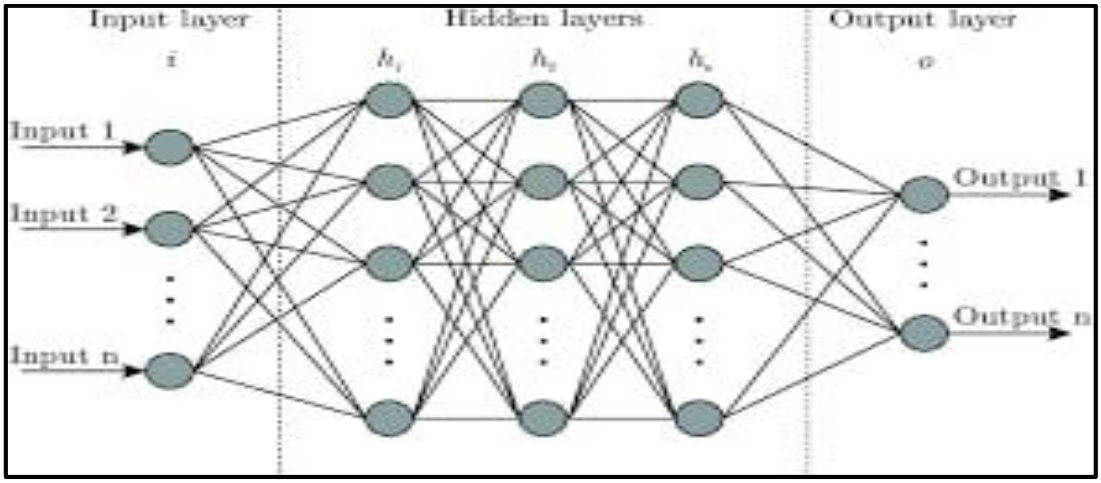


FIGURE 2-2 ARTIFICIAL NEURAL NETWORK ARCHITECTURE **[18]**

Each neuron has a threshold value and an activation function, as does each connection [19]. It is decided if each input has a positive or negative weight based on the weight sign and the weight has an effect on the signal intensity of a connection [20].Neurons have a threshold over which a signal is only transmitted when the total signal exceeds a certain threshold. The signal from the activation value, which is the weighted sum of the summing unit,

is used to generate the output. Figure (2-3) displays the Connection between the ANN system's input and output, as well as the weight of each element.

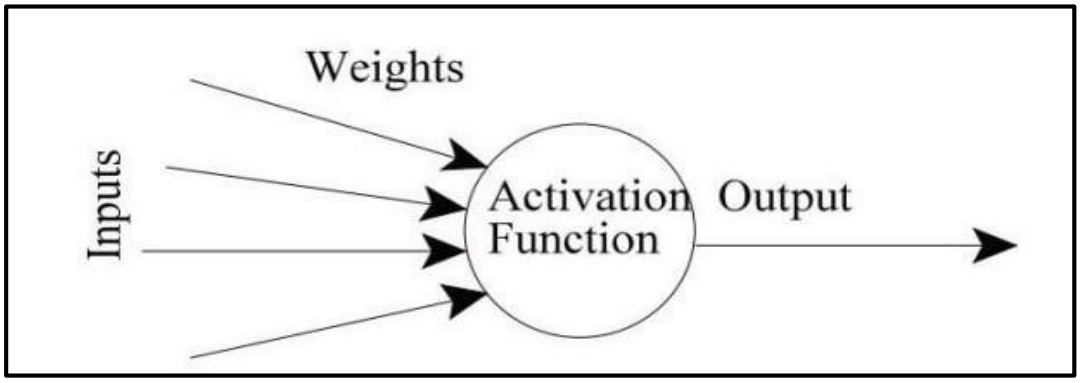


FIGURE 2-3 WEIGHT OF EACH ELEMENT AND INPUT AND OUTPUT OF THE ANN SYSTEM [20].

Other cells can use these networks to compensate for the loss of a damaged cell and help in cell regeneration. These networks are capable of learning. In essence, the ability of a system to learn is its most crucial component. Because it is more versatile and easier to program, a learning system can respond to new problems and equations more effectively. During the learning process, a neural network is programmed to perform certain tasks such as pattern recognition and information categorization. Artificial neural networks, like humans, learn via experience. For example, injecting touch nerve cells trains the cells to avoid the hot body, and this strategy teaches the system to correct its error [21].

2.5 Activation Functions IN Neural Networks

The activation function in a neural network model's responsible of translate input values into output values [22](see Figure (2-4)), where input values are determined by computing the weighted sum of neurons' input values. It is crucial for an activation function to be differentiable so that error backpropagation may be used to train the model.

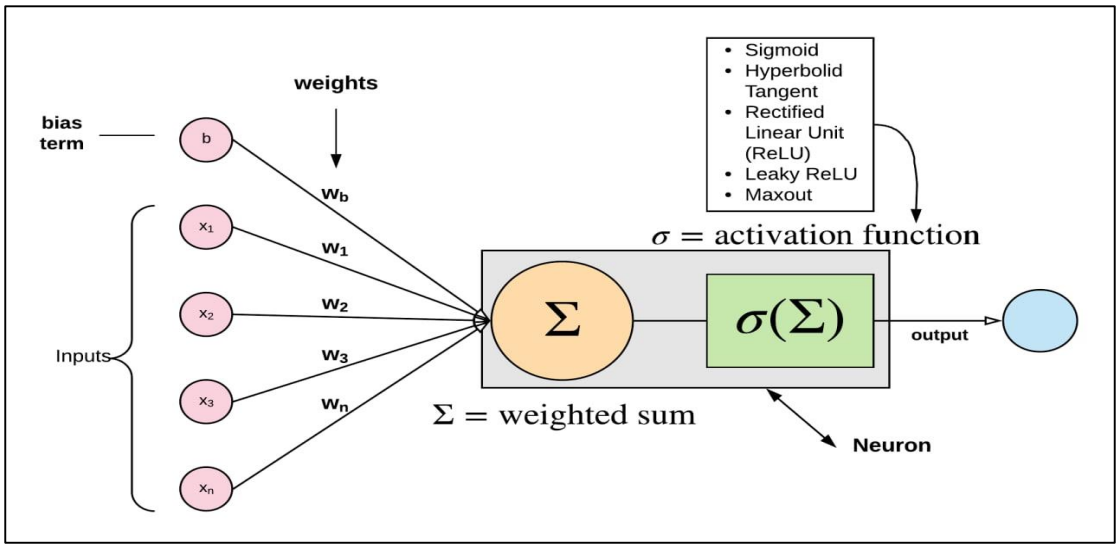


FIGURE (2-4) ACTIVATION FUNCTION IN NEURAL NETWORKS

The following is a description of the most popular activation functions in deep neural networks, including CNN :

A. Sigmoid

The input for the sigmoid activation function is a real number, and the output is bound to the range [0, 1].The sigmoid function's curve has a "S" form (see figure 2-5). The sigmoid is represented mathematically as follows: [23]

$$f(X)_{sigm} = \frac{1}{1+e^{-x}} \quad \text{_____(1)}$$

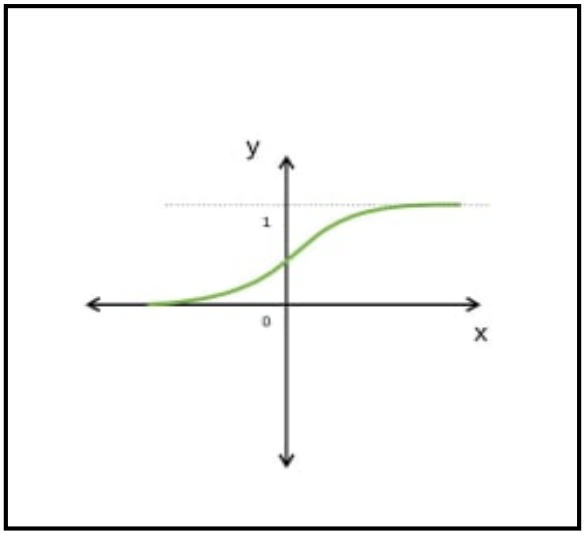


FIGURE (2-5) SIGMOID ACTIVATION FUNCTION **[22]**

B. ReLU:

The most often used activation function in convolutional neural networks is the Rectifier Linear Unit (ReLU). It is used to change every value from the input to a positive integer. [23] ReLU has the benefit of requiring a much lower computational burden than other methods (figure (2-6) shows the ReLU Activation function). ReLU is mathematically represented as follows:

$$f(x)_{ReLU}=\max(0,X\quad)\_\_\_\_\_(2)$$

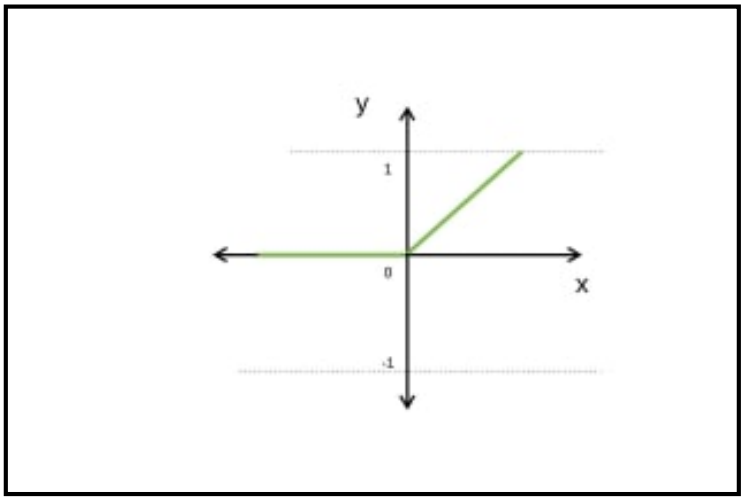


FIGURE (2-6) **RELU** ACTIVATION FUNCTION **[22]**

C. Softmax

Another form of activation function utilized in neural computing is the Softmax function. A vector of real values is utilized to compute the probability distribution. The output of the Softmax function is a range of values between 0 and 1(see figure(2-7)), with the probability total being equal to 1. The connection is used to construct the Softmax function [22].the formula of Softmax activation function explained follow:

$$f(x) = \frac{\exp(xi)}{\exp(xj)} \text{ - (1.12)} \quad \text{_____(3)}$$

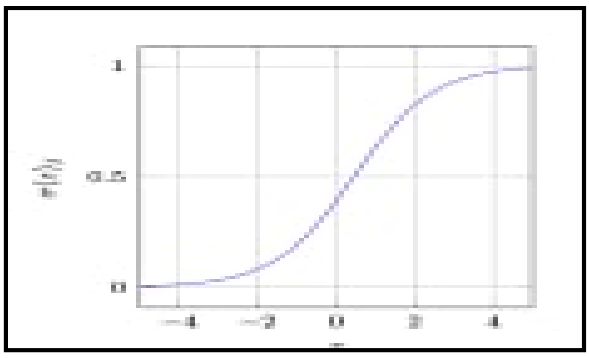


FIGURE (2-7) GRAPHIC REPRESENTATION OF SOFTMAX ACTIVATION FUNCTION **[22]**

The Softmax function yields probabilities for each class in multi-class models, with the target class having the highest probability. In practically all of the output layers of deep learning systems, the Softmax function may be seen.

The Sigmoid and Softmax AF vary primarily in that the former is used for binary classification jobs while the latter is utilized for multivariate classification tasks.

2.6 Understanding how deep learning neural network works

A layer's weights, which are just a collection of numbers, define what the layer performs with the incoming input. Technically speaking, the weights of a layer parameterize the change that layer makes (see figure 2-8) (Weights are also frequently referred to as a layer's parameters.) Finding a set of weight values for each layer in a network such that it can accurately translate example inputs to their corresponding goals is what is meant by

"learning" in this context. A deep neural network, however, has the capacity to store tens of millions of parameters. Finding the right settings for each of them could seem like an overwhelming effort, especially when changing the value of one parameter will change how the others behave [12].

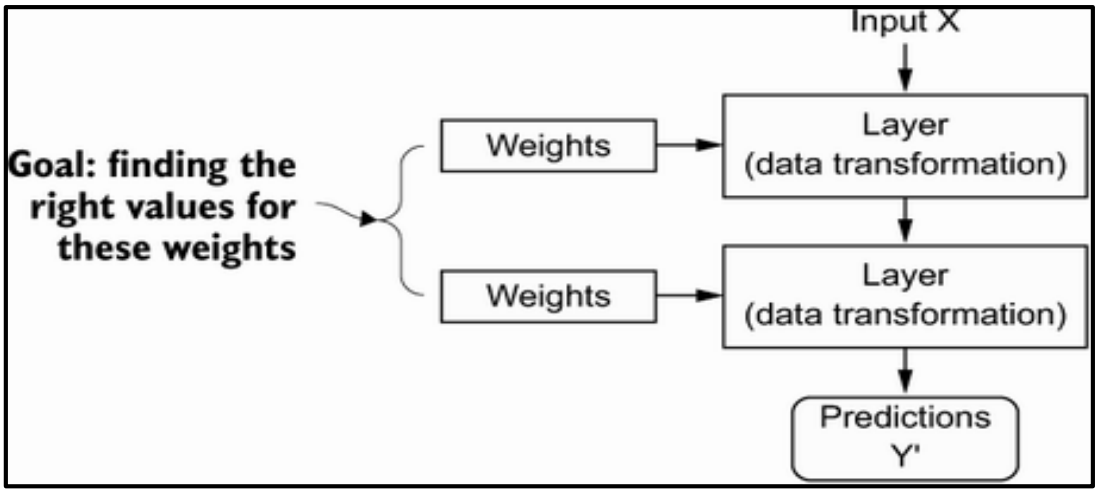


FIGURE (2-8) A NEURAL NETWORK IS PARAMETERIZED BY ITS WEIGHTS

To be able to govern a neural network, you must be able to determine how much its output deviates from your expectations. This is the responsibility of the network's loss function, Using the network predictions and the real goal (what you expected the network to output), the loss function computes a distance score, which measures how well the network performed in this specific case [12] (See figure (2-9)).

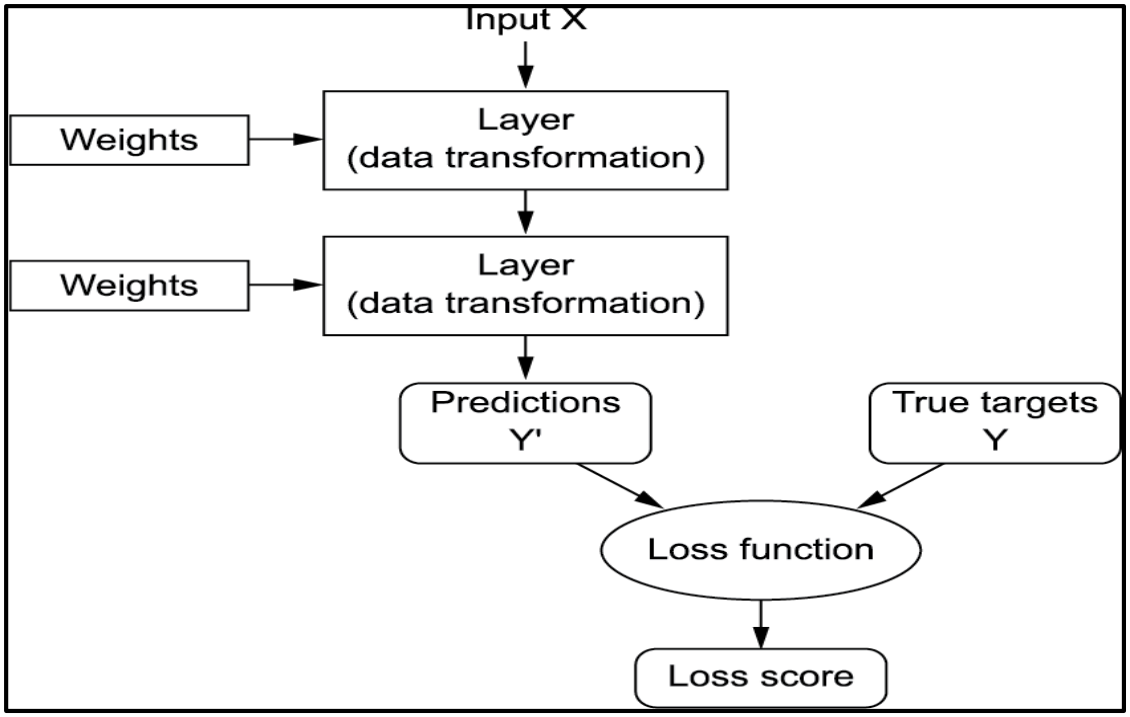


FIGURE (2-9) A LOSS FUNCTION MEASURES THE QUALITY OF THE NETWORK'S OUTPUT.

The key to deep learning is to use this score as a feedback signal to gently adjust the value of the weights such that the network's loss score is reduced (see figure2-10). This adjustment is the responsibility of the optimizer, which implements the Backpropagation algorithm, the main algorithm in deep learning [12].

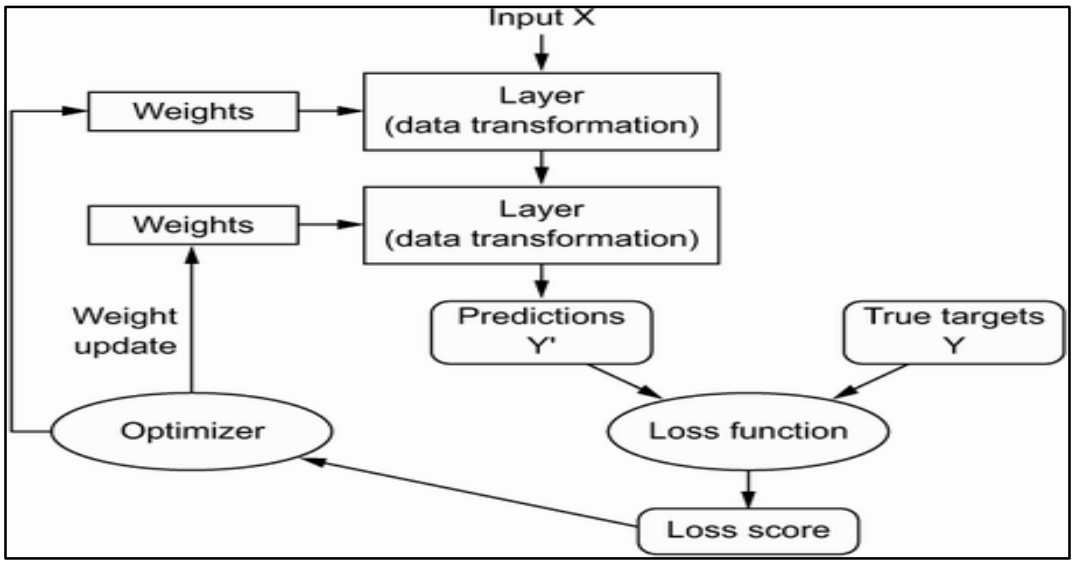


FIGURE (2-10) THE LOSS SCORE IS USED AS A FEEDBACK SIGNAL TO ADJUST THE WEIGHTS

Because its weights are originally assigned random values, the network employs a series of random adjustments. Its productivity is far lower than it should be, and as a result, the loss score is extraordinarily high. However, when the network analyzes more samples, the weights move slightly in the right direction, and the loss score decreases. When the training loop is sufficiently repeated, it generates weight values that minimize the loss function. A trained network has outputs that are as close to the objectives as feasible and has a low loss. Once again, it's a simple solution that looks to be magical when scaled [20].

2.7 Convolution Neural Network (CNN)

A. Convolution Neural Network (CNN)

Convolutional Neural Network (CNN) is a special kind of multi-layer neural network, CNN is designed to recognize visual patterns directly from images with minimal processing. figure (2-11) represent a graphical representation CNN network. The discipline of neural networks was originally inspired by the goal of modeling biological neural systems, but since then it has forked in different directions and has become a matter of engineering and performing good results in machine learning studies [24]

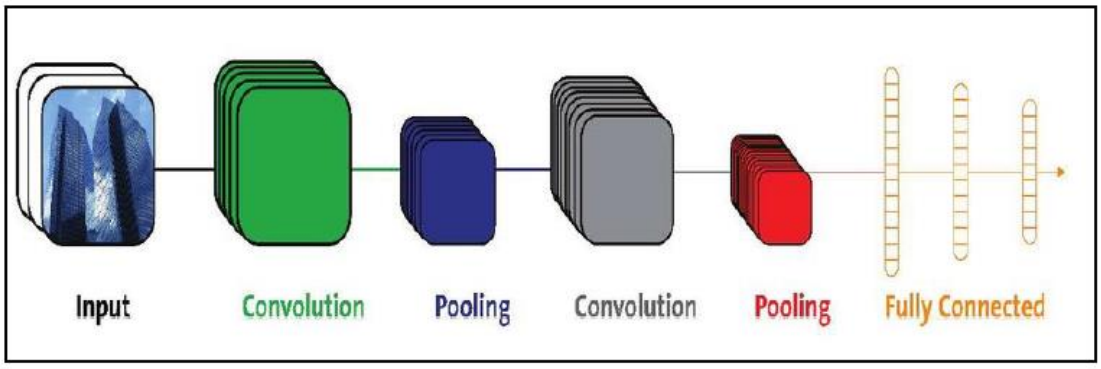


FIGURE (2-11)SIMPLE ARCHITECTURE OF A CNN MODEL

CNN Model Layers

The CNN Model contains 4 layers:

•Convolution layer:

The most crucial element of any CNN design is the convolutional layer. It has a number of convolutional kernels, also known as filters, which when combined with the input image's N-dimensional metrics yield a feature map as the result. The main objective of convolution in relation to Convolution Network is to extract features from the input data. [7] (see figure(2-12)

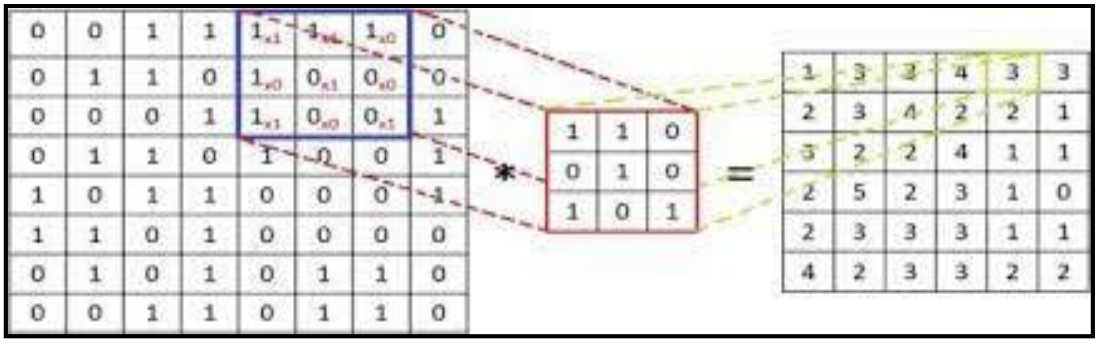


FIGURE (2-12)CNN CONVOLUTION LAYER

What is a kernel?

A kernel can be thought of as a grid of discrete values or integers, each of which represents the kernel's weight. All of a kernel's weights are randomly allocated at the

beginning of a CNN model's training phase (different approaches are also available there for initializing the weights). After that, the weights are adjusted and the kernel learnt to extract useful features with each training period. We have a 2D filter displayed in Figure (2-13)

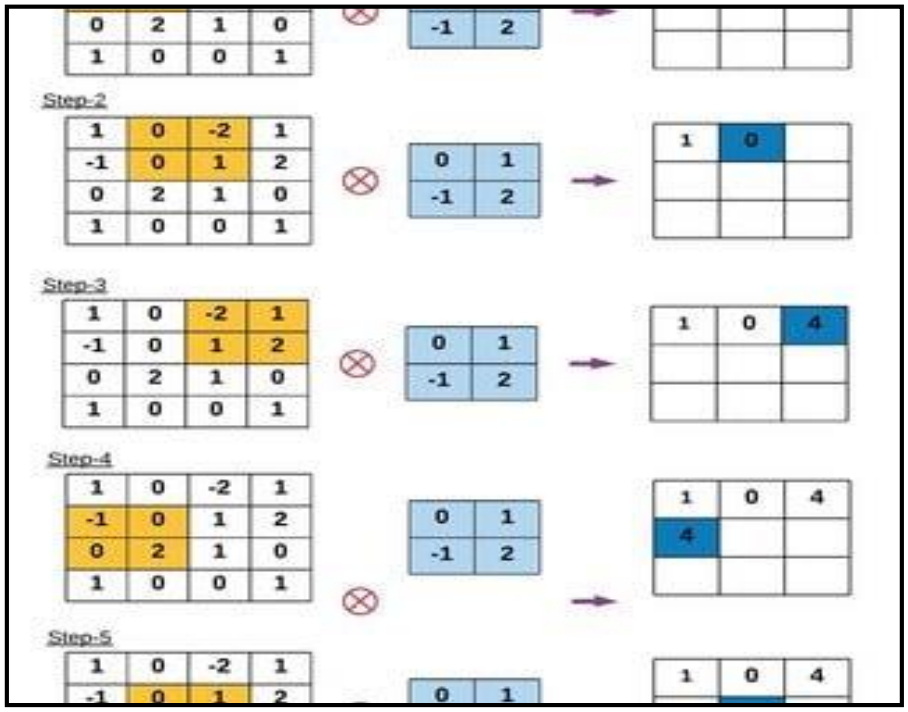


FIGURE (2-13) EXAMPLE OF A 2 × 2 KERNEL

Each convolutional layer in CNN will employ a variety of filters, allowing each filter to extract a variety of information.

•Pooling Layer:

The feature maps (generated after convolution operations) are sub-sampled using the pooling layers, i.e., the larger size feature maps are shrunk to smaller size feature maps.

The most significant features (or information) in each pool stage are always preserved when the feature maps are shrunk. Similar to the convolution operation, the pooling operation is carried out by defining the size of the pooled region and the operation stride. Different pooling approaches, such as max pooling, min pooling, average pooling, gated pooling, tree pooling, etc., are employed in various pooling layers. The most common and widely used method is max pooling. [23]

The pooling layer used to control the overfitting by progressively reducing the spatial size of the representation to reduce the number of parameters and computation as shown in figure (2-14). [24]

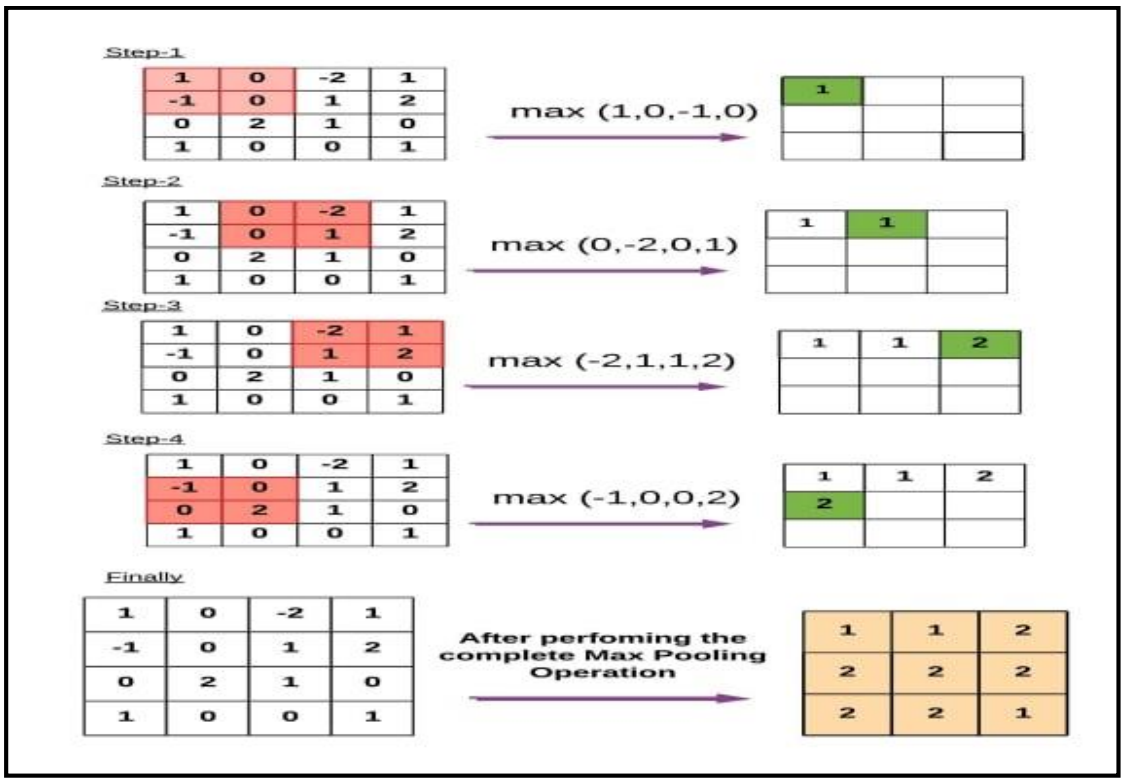


FIGURE (2-14)AN EXAMPLE THAT DEMONSTRATES BOTH THE END RESULT OF THE MAX-POOLING OPERATION AND SOME BEGINNING STEPS

- **Flatten layer:**

  This layer responsible of converting the data into a 1-dimensional array for inputting it to the next layer [24] as explained in the figure (2-15)

.

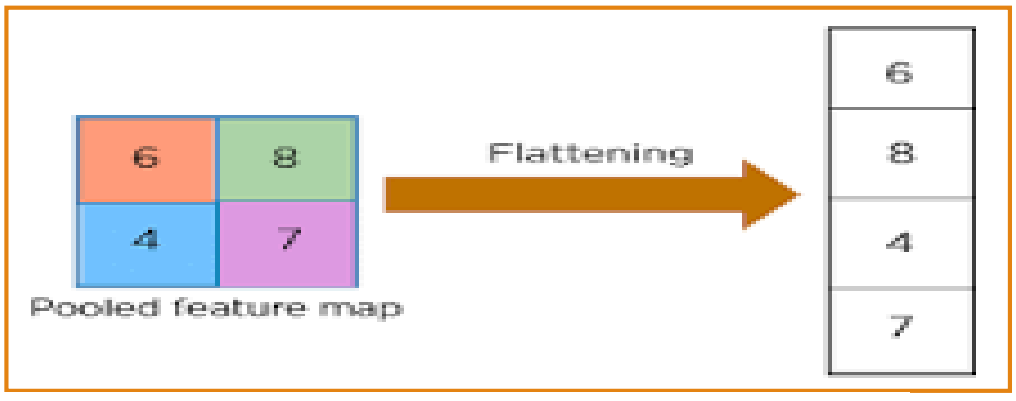


FIGURE (2-15) THE WORK OF FLATTEN LAYER

- **Fully connected layer:**

Despite being fully linked, this layer may be used to learn characteristics and Categorize data in addition to connecting all the neurons in one layer to all the neurons in another layer [24] .figure (2-16) shows the CNN model when not

fully

connected and when it fully connected

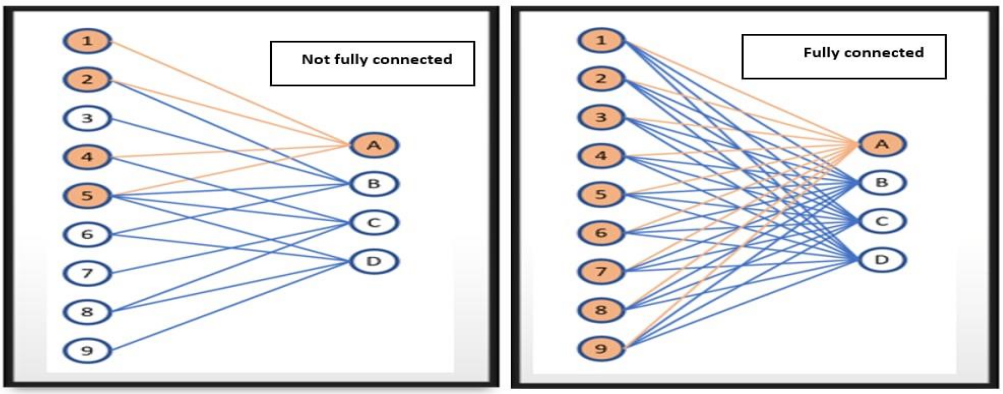


FIGURE (2-16) CNN MODEL WHEN NOT FULLY CONNECTED AND WHEN IT FULLY CONNECTED **[24] .**

2.8 Gated Recurrent Unit (GRU):

The gated recurrent unit(GRU) is a special type of optimized LSTM-based recurrent neural network. The GRU internal unit is similar to the LSTM internal unit , except that the GRU combines the incoming port and the forgetting port in LSTM into a single update port. The GRU considered simpler to calculate and implement. It retains the LSTM immunity to the vanishing gradient problem. Its internal structure is simpler and, therefore, it is also easier to train, as less calculation is required to upgrade the internal states. The update port controls the extent to which the state information from the previous moment is retained in the current state, while the reset port determines whether the current state should be combined with the previous information [25].

2.9 Feature Extraction

One of the most important steps in the voice recognition process is feature extraction. The objective of speech feature extraction is to transform voice signals into a coefficient vector feature that only includes the data required to identify the phrase. Speech includes distinct and varied features that may

be retrieved from several feature extraction proposals and utilized for speech recognition tasks since they are present in the uttered words. With feature extraction, the quantity of the data is reduced while maintaining the features of the speech signal in each frame that may be utilized as a characteristic. By transforming the shape of the speech signal into a parametric representation, features may be extracted.

Mel Frequency Cepstral Coefficients (MFCC), established in, and the perceptual linear predictive (PLP) feature, introduced in, are the two best-presented feature extraction algorithms. When comparing them, MFCC features are the most prevalent, popular, and reliable approach for feature extraction in speech recognition systems currently in use, especially in clean speech or clean environments [27]. Compressing a voice signal into streams of acoustic feature vectors, also known as speech feature vectors, is the first step in speech recognition. The recovered vectors are thought to contain enough data and be sufficiently small for effective recognition. The idea of feature extraction is really separated into two steps: first, the speech signal is transformed into feature vectors; second, the relevant characteristics that are impervious to changes in the surroundings and speech variation are selected[28]. The idea of feature extraction is really separated into two steps: first, the speech signal is transformed into feature vectors; second, the relevant characteristics that are impervious to changes in the surroundings and speech variation are selected[29]. In speech recognition systems, however, where accuracy has drastically declined in the case of their existence, changes in ambient variables and variances in speech are significant. Examples of environmental conditions that can alter include the transmission channel, microphone characteristics, cocktail effects, background noise, etc. Accent variations and variances between the male and female vocal tracts are two examples of speech variations.

Speech characteristics must be insensitive to these changes and fluctuations in order to create strong speech recognition. The Mel Frequency Cepstral Coefficients (MFCC) features, which are definitely the most commonly used speech characteristics, are the most popular and trustworthy owing to their exact estimate of the speech parameters and excellent computational model of speech [26].In figure (2-17), the method for extracting the MFCC features is described and summarized as follows:

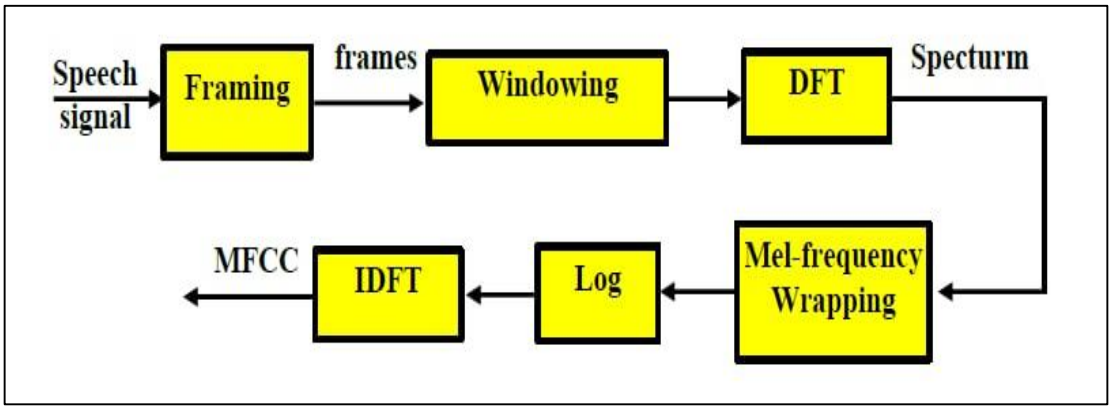


FIGURE (2-17) THE STANDARD PROCEDURES OF MFCC FEATURE EXTRACTION

Mel-frequency cepstral coefficients (MFCC) are now the most well-known and widely used features, both for speech recognition and for other purposes. The short-term evaluation serves as the foundation for the MFCC computation. The key distinction is the application of crucial bank filters in order to produce Mel-frequency warping How well humans can hear determines which frequencies have the smallest bandwidths. In this sense, a Mel is a measuring unit focused on the sound frequency perceived by the human ear in the usual methods of MFCC feature extraction. Mel-frequency calculation Calculating cepstrum is comparable to computing cepstral coefficients. The discrepancy is due to mel-frequency warping before performing logarithm and reverse DFT. The Mel-frequency scale is the true frequency scale that the warping converts to the apparent human

frequency scale. The linear spaces are less than 1 kHz, but the logarithmic scale surpasses 1 kHz [26].

The following steps explain the MFCC extraction process:

A.Framing and blocking

In this step the signal are blocked into small frames of N samples

B. **Windowing**

Windowing is done for minimizing the disruptions at the starting and at the end of the frame, the
frame and window function is being multiplied.

C. ***FFT (*Fast Fourier Transform)**

FFT is used for doing conversion from the spatial domain to the frequency domain.

D. ***Mel scale***

In this step, the above calculated spectrums are mapped on Mel scale to know the approximation about
the existing energy at each spot with the help of Triangular overlapping window also known as
triangular filter bank.

E. **Discrete cosine Transform (DCT)**

This process of carrying out DCT is done in order to convert the log Mel spectrum back into the spatial domain.

MFCC Advantages:

The following are some of the benefits of MFCCs [27]:

- Low coefficient correlation
- Not reliant on linear features; hence, comparable to the human auditory perception system

- Capture of important phonetic properties

Disadvantages of MFCC:

Among the MFCC's drawbacks are [27]:

• Low noise resistance

• In a continuous speech environment, a frame may contain information from two successive phonemes rather than just one phoneme

• Inflexible since the same basic wavelets must be used for all speech signals

2.10 Model Evaluation:

To finish the training of the system model. It is time to evaluate the model. Before evaluating the model started, there are some functions must be implemented plotting the model errors and printing the validation accuracy:

1. **Optimizer**: an optimizer is a method or algorithm used to adjust the properties of a neural network such as weights and learning rate to decrease loss. The optimizer used in the suggested model is named "adam," and it is the best optimizer that gives the most accuracy to the CNN model.
2. **Loss Function**: is the difference between the train and test data, or is prediction of error, the suggested model employs the' categorical_crossentropy' loss to compute the cross-entropy loss between labels and predictions

   The below equation illustrate how to compute the loss function for "Caterogical_CrossEntropy" loss:

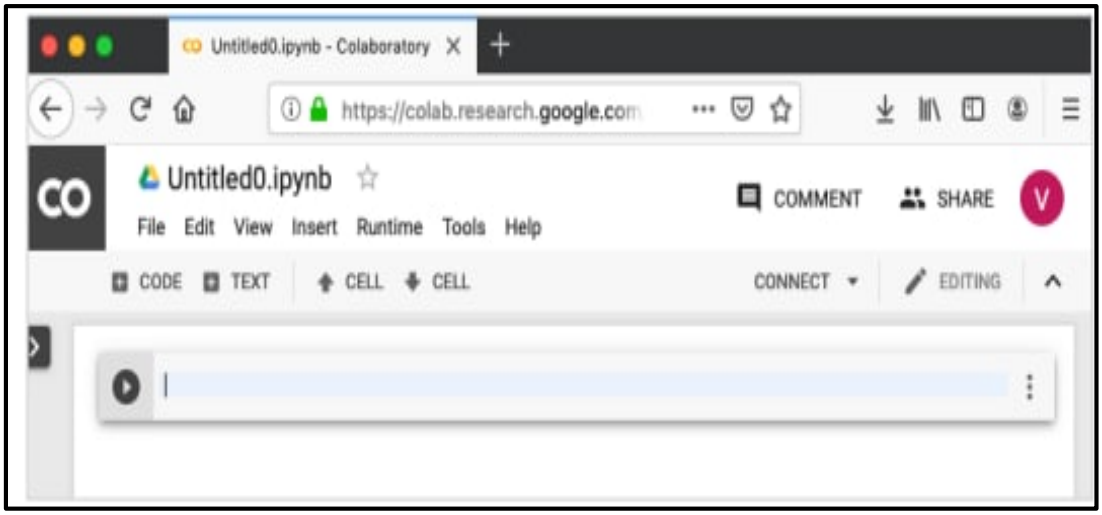
$$CE\ Loss = -\frac{1}{n}\sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij} \cdot \log(p_{ij}) \quad (3)$$

3. **Metrics:** a function used to determine model performance; in the proposed model, the accuracy metric is employed.
   The accuracy value computed through the below equation :

$$accuracy = \frac{true_{positive}+true_{negative}}{true_{positive}+true_{nagative}+false_{positive}+false_{nagative}} \quad \text{_____(4)}$$

2.11 Google Colaboratory

Colab is a cloud-based notebook environment that is completely free. It allows you and your team members to modify documents in the same manner that you do with Google Docs. Many common machine learning libraries are supported by Colab and can be quickly loaded into your notebook. Google Colab is a Google development tool. Since 2017, it has been freely available to everyone. Another interesting feature that Google offers developers is the use of GPU. Colab is fully free and works with GPUs. [28]If you have already used a Jupyter notebook, Google Colab is straightforward to use (see figure (2-18)). Colab is a Jupyter notebook environment that is available for free. Cloud-based, the notebooks you create may be edited concurrently by your team members, just way you edit documents in Google Docs, and it doesn't require any setup. Colab supports several well-known machine learning libraries, which are simple to load in your notebook.



FIGURE(2-18) GOOGLE COLAB NOTEBOOK

2.12 Wandb (Weights & Biases):

Wandb is a platform of the machine learning for developers to build better models faster. They can rapidly monitor experiments, use Wandb to update and iterate on datasets, assess model performance, reproduce models, display code outputs and discover regressions, and communicate findings with their peers [29]. Weights & Biases (Wandb) is a python package that allows developers to easily monitor their training in real-time (see figure 2-19). It can be easily integrated with popular deep learning frameworks like Tensor flow, Pytorch, or Keras

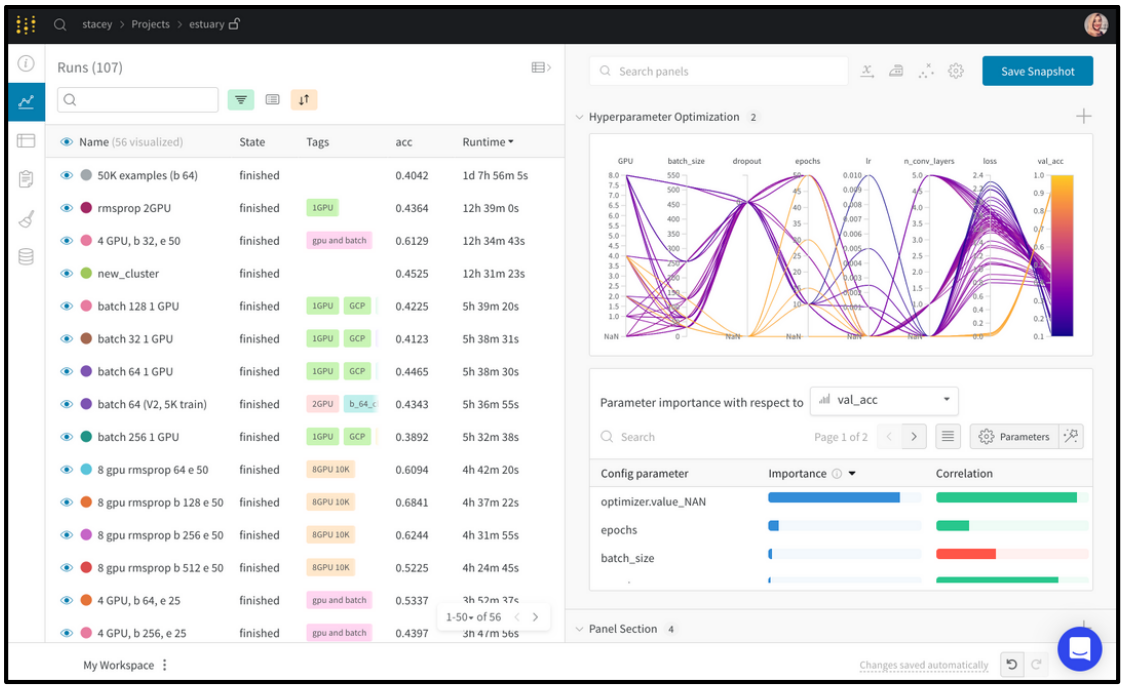


FIGURE (2-19) WANDB PLATFORM

Chapter Three

Proposed System

Chapter Three

Proposed System

3.1Introduction

This chapter will offer a presentation on the methodology utilized, the community and sample of the study, as well as the tools used for this study, and how to create this tool and assess its validity and reliability after dealing with the theoretical part of the subject of the current investigation. Figure (3-1) shows the proposed system to identify the speaker based on the CNN model.

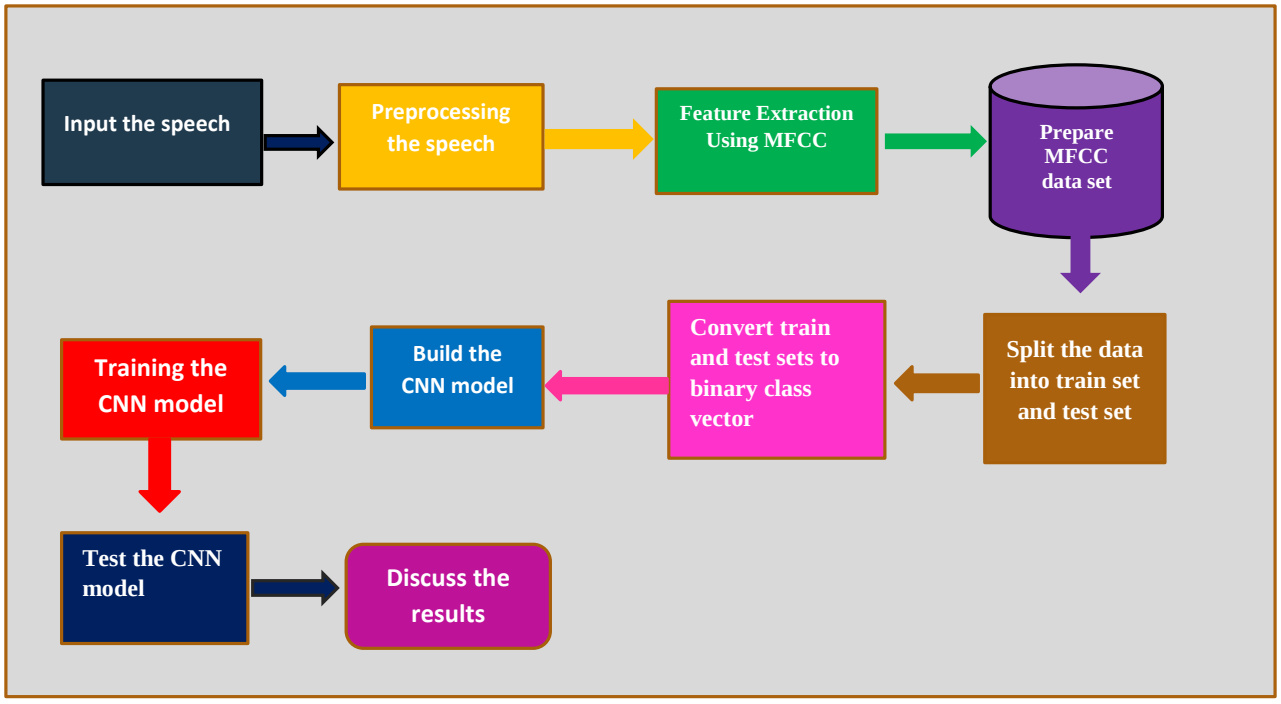


FIGURE (3-1) DIAGRAM OF THE PROPOSED SPEAKER IDENTIFICATION SYSTEM

3.2Dataset:

Speaker identification tests are done with the "16000 pcm speeches" dataset to examine the performance of the model in the proposed work. "16000 pcm speeches" is a publicly accessible speech dataset from "Kaggle" platform that comprises of English speech recorded by 5 speakers, each comprising 1700 recorded audio files of 2 seconds length. The suggested system's experimental configuration split the data set into "Jens Stoltenberg," "Benjamin Netanyau," "Magaret Tarcher," "Julia Gillard," and "elson Mandela." Table (3-1). shows the specific use of the "16000 pcm speeches" dataset

Table 3-1. The usage of the "16000_pcm_speeches" dataset in experiments.

Speaker	utterances
" Jens_Stoltenberg"	1500
"Benjamin_Netanyau"	1495
"Magaret_Tarcher"	1500
" Julia_Gillard"	1501
"Nelson_Mandela"	1502

In the proposed work the data was divided into train set and test set with the proportions of 70%, 30% respectively.

3.3Methodology:

Each speaker identification system has two phases: preprocessing and classification which identify the speaker identity using speaker identifier. the preprocessing applied on speaker signal. performing preprocessing for the input data before in order to prepared for feature extraction. During preprocessing phase, the speaker's voice is recorded and next the important

features are extracted to form a voice print, template, or model. In the identification phase, a speech sample or "statement" is compared against a previously created voice print

3.3.1 Load speech files.

The audio files in the "16000 pcm speeches" dataset are used as input data for the proposed system, which comprises of audio files of spoken speech in an audio format such as ".wav"

3.3.2 Preprocessing:

This process include Using sampling rate as the number of samples picked per second to convert wave (analog) to digital and passing to construct training, testing, and model. figure (3-2) explains the difference between analog and digital signals.

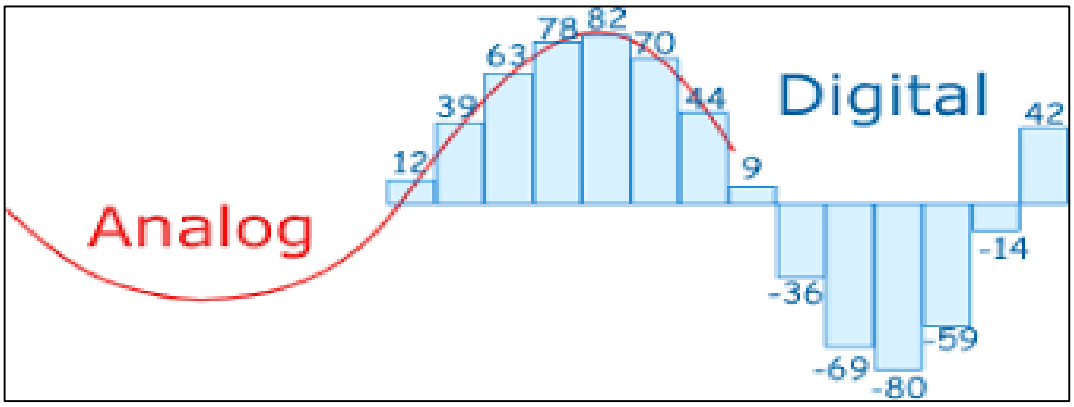


FIGURE (3-2) THE DIFFERENCE BETWEEN ANALOG AND DIGITAL SIGNALS

An analog signal is a time-varying continuous representation of a signal. An analog signal contains an endless number of samples between any two time periods. and a discrete representation of a signal across time is a digital signal. A limited number of samples exist between any two time periods in this case. The analog digital translation is required because analog signals consume a lot of memory because they have an endless number of samples and processing them is extremely computationally intensive. As a result, it appears that a technology to convert analog impulses to digital signals is required to

conveniently deal with voice files.in python sampling (convert the analog wave to digital) is performed using Librosa python library.

3.3.3 Feature Extraction

The voice waveform was transformed into a set of characteristics in this section. Signal processing front end is the name of this procedure. The Mel-Frequency Cepstrum Coefficients technique is used to extract features (MFCC). In order to extract the key elements of the speech stream, MFCC leverages the human ear's variance for critical bandwidths As a result, low-frequency filters perform linearly separated while high-frequency filters perform logarithmically spaced. The Mel-frequency scale is the name of this technique. The overall MFCC structure incorporated into the suggested system the recorded input signal has a sampling rate greater than 10000 Hz. The chosen sample frequency lessens the impact of aliasing during the conversion process from analog to digital. Additionally, all frequencies up to 5 kHz, which contain the majority of the energy in auditory human signals, might be captured.

figure (3-3)shows mfcc extracted features of speech signal of the proposed system.

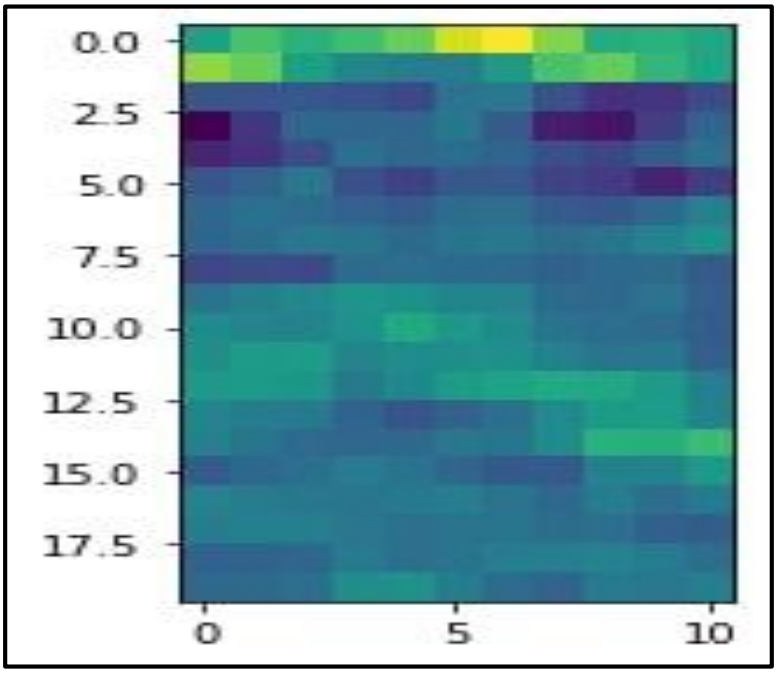


FIGURE (3-3) MFCC OF SPEECH SIGNAL IN THE PROPOSED SYSTEM

3.3.4 Train and Test split

The data in the proposed study was separated into train and test sets in proportions of 70% and 30%, respectively. The split procedure is carried out in Python using the sklearn python module. The model selection package's train test split function in the sklearn library is in charge of splitting data into training and test sets.

3.3.5 Sound as data

The dataset is made up of wave files that are all about two seconds long. To make use of the data, each file is sampled into a vector at a rate of 16000. The difference between two-dimensional array and audio characteristics is that an audio file must be transformed into an image (spectrogram) in order for the network to learn, do data analysis, and generate predictions.

The extraction of characteristics from the raw waveform is a typical method for voice recognition. Spectrograms, log-Mel filter banks, and Mel-frequency cepstral coefficients (MFCC) are common speech characteristics that transform the raw waveform into a time-frequency domain. These characteristics are then fed as image into a Convolutional neural network model.

3.3.6 Feature Dimension:

Set values to the essential parameters of the proposed model:

Epoch: is the number of times the model will iterate through the data.in the proposed method 100 epoch was used to train the network.

Batch Size: The batch size specifies the number of samples that will be sent across the network. Assume you have 1050 training samples and wish to set the batch size to 100. The technique trains the network using the first 100 samples (from 1st to 100th) from the training dataset. The network is then trained using the second 100 samples (from 101st to 200th). We can repeat this approach until all samples have been spread throughout the network. In this step set the batch size of the proposed system to 100.

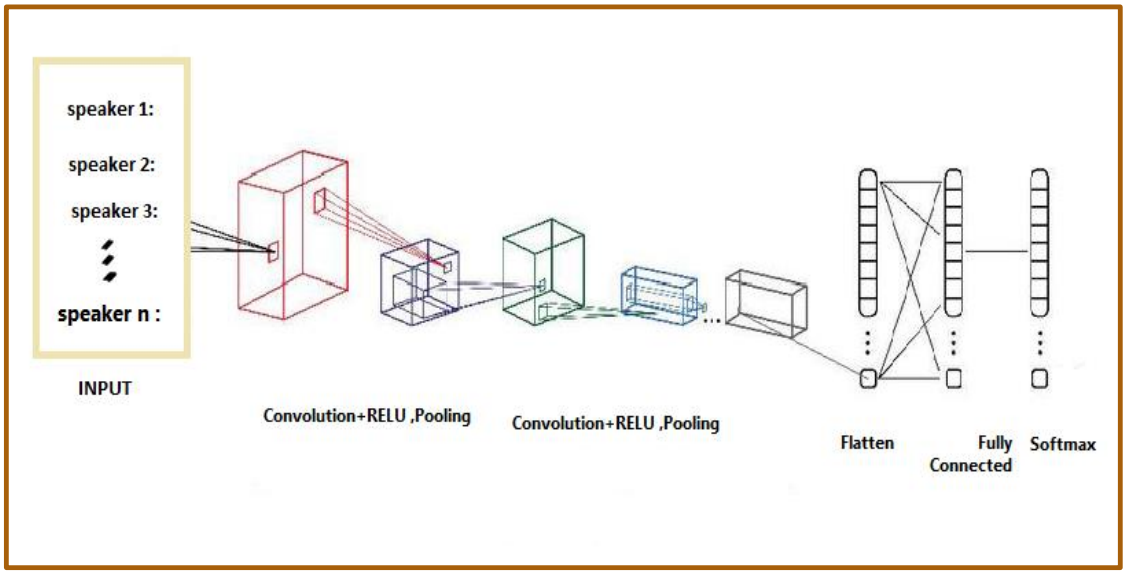
Channels: The depth of the matrices engaged in the convolutions is represented by the number of channels.

The channel value of the proposed system is 1.

3.3.7 Modeling:

Modeling includes build the convolution neural network of the system using Keras library in python, figure (3-4) shows the CNN model of the proposed system. **RELU** activation function used in the hidden layers of the proposed CNN model, the advantage of ReLU activation function over other activation functions is that it does not stimulate all neurons at the same time. The Softmax

activation function was employed at the last layer to learn and score the features in order to generate speaker identification results.



FIGURE(3-4)CNN MODEL OF THE PROPOSED SYSTEM

The following steps summerize the Modeling process in the proposed CNN model:

Step1:determine the type of the CNN ,the Sequential model is simple and efficient mode type.

Step2: add CNN model layers:

the suggested work employs the CNN model to aid in the identifying procedure. the model used by Google Collaboratory Notebook The suggested CNN model's structure for speaker identification , shown in figure (3-5) .

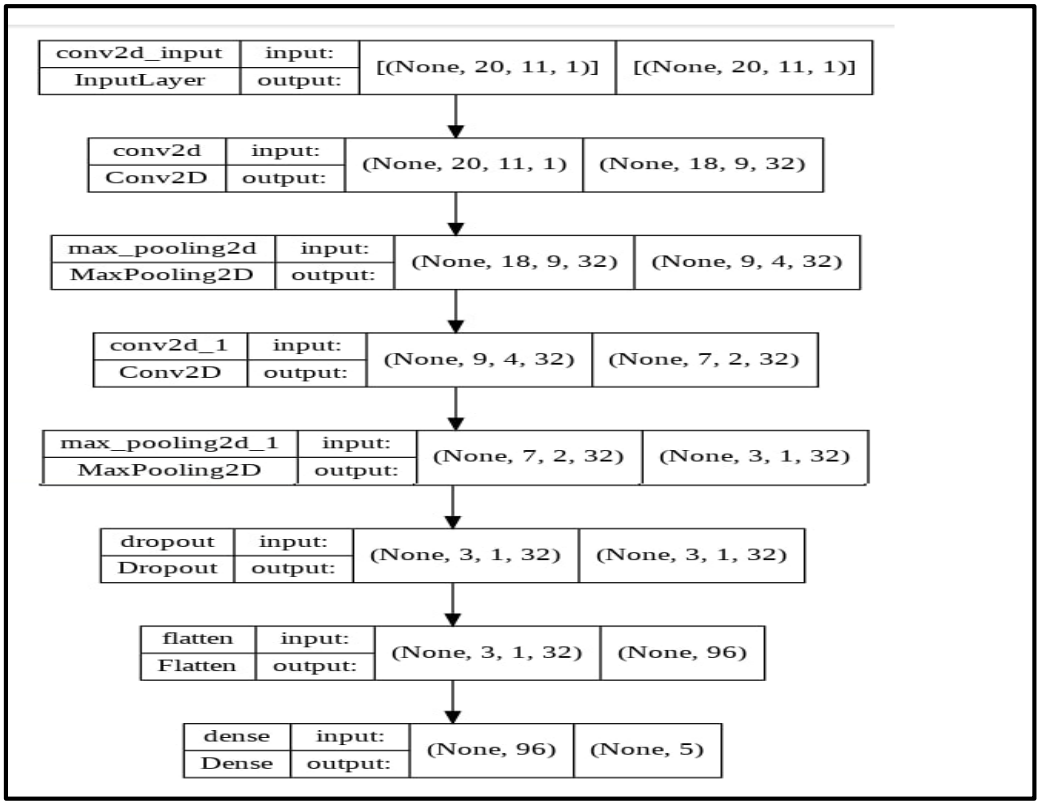


FIGURE (3-5) THE CNN STRUCTURE OF THE PROPOSED SYSTEM

The following steps explain the CNN layers of the proposed system:

1) The first layer in the proposed CNN model is the input layer , input layer with input dimensions matching our MFCC characteristics As a result, the feature matrix for an audio sample is large (20, 11). Keras will automatically deduce the form of every subsequent layer after the first. This implies that you just need to configure the input dimensions for the first layer. The first layer in the previous code snippet, sets the input dimension to (20, 11, 1), where 20 and 11 are the MFCC feature dimensions and 1 is

the channel size. The first layer's output is sent into the second layer. This sequence of sending output to the next layer continues until the last layer, which is the model's output.

2) The second layer is convolution layer which responsible of apply filters and Find the features .a filter with size (2*2)applied on the input data to extract the information from the input data .the result of this layer called feature map with values (18,9,32)

   The output of this layer used as input to Relu activation function becouse Many of the convolutional layers were interleaved with nonlinearity. In a nutshell, ReLU is used to filter information as it travels over the network from this layer to the next.

3) Layer three:The feature map used as input to the third layer which is the pooling layer at this work the max pooling filter used with size (2*2).this layer reduce the features to (9,4,32) .It aids in the preservation of crucial information or characteristics of the input data and shortens calculation time.

4) The forth layer: is convolution layer to apply more filters on the input data , a filter with size (2*2)applied on the input data again this layer produce a feature map with values (7,2,32) .and the Relu activation function also used for the same reason .

5) Layer five:pooling layer reduce the number of features to (3,1,32) while keeping the most relevant ones, and set the Max Pooling value to 2.

6) The sixth layer: droupout layer added with value (0.15) this layer used to prevent the overfitting (it randomly drops neurons from the neural network during training in each iteration).

7) The seven layer :in the proposed cnn model is flatten layer This transforms the data into a 1-dimensional array for input to the next Dence layer the result of this layer is single array.
8) Layer eight (output layer) :Dense (full connected) layer: neuron in the dense layer receives input from all neurons of its previous layer.

Step3: Compile the model: Now that the training data and model have been defined, it is necessary to configure the learning process. This is performed by invoking the Sequential model class's compile () function.

Step 4: Training the model: is a process of fitting the model on the train data and test data the 'fit ()' function used to train the model this function take the following parameters: training data (X_train), target data (y_train), validation data, and the number of epochs, Validation data represented with X_test and y_test.

Chapter Four
Experimental Results And Discussion

Chapter Four

Experimental Results and Discussion

4.1 Result:

The proposed network approaches are implemented in Python using the Keras deep learning framework, which can be run on a graphics processing unit (GPU). The GPU is typically more efficient than a central processing unit due to parallel computing (CPU). The "adam" optimizer and the "categorical crossentropy" loss function were chosen during the CNN network model training stage. In the CNN network model training stage, 100 epochs and a batch size of 100 are employed in the model's train stage. The training pairings are randomized after each training epoch. The validation set is utilized for early halting and changing hyper-parameters. To test the suggested CNN model's capacity to extract features, the model training results are depicted in the following image. figures (4-1)(4-2) show that the CNN network model employed in the experiment converged after around 90 rounds of training, and the validation set results are likewise very ideal:

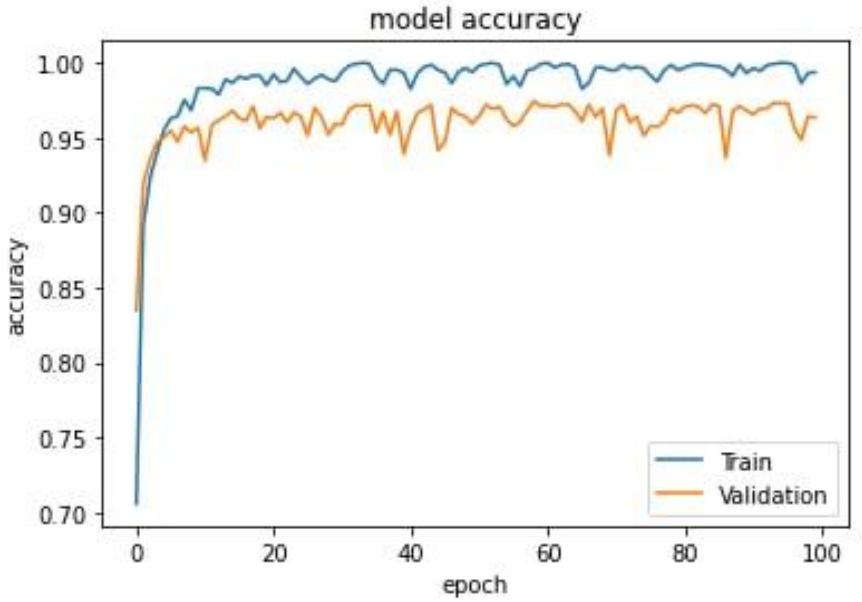


FIGURE (4-1) ACCURACY VALUE OF THE SYSTEM

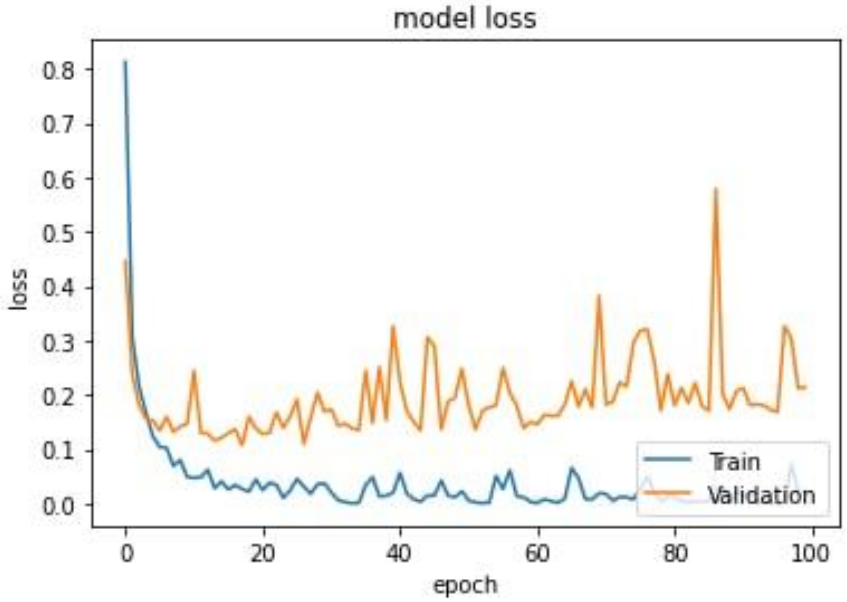


FIGURE (4-2) LOSS VALUE OF THE PROPOSED MODEL

The activation function has a significant effect on the prediction accuracy when replacing the Relu activation function with the sigmoid activation function in the hidden layers resulted in a significant effect on the accuracy of the system .the accuracy reduced to 92%.

TABLE (4-2) THE EFFECT OF ACTIVATION FUNCTION ON ACCURACY AND LODD VALUES

Activation function	**Accuracy**	**Validation accuracy**	**Loss**	**Validation loss**
ReLU	100%	99.82	0.45%	2%
Sigmoid	98%	93%	0.3%	2%

The suggested technique is compared to the LSTM method, which is run on the same dataset 100 times to the testing set for each method. For comparison, the maximum and average categorization accuracy are shown. The batch size for the LSTM network model is set to 100 during the training phase, and the optimization is carried out using the Adam optimizer. The "categorical

crossentropy" loss function is chosen, and the optimal configuration for each network that provides the greatest performance results is chosen. The dropout layer is used in the CNN and LSTM models to prevent the tough problem of overfitting in deep learning model training. To prevent the overfitting problem that is common in deep neural networks, a dropout of 0.1 at the CNN networks and a dropout of 0.25 at the LSTM model were used in this study. Convolutional and dense layers were used to avoid the overfitting problem that is common in deep neural networks. figures (4-3)(4-4) provide an example of the deep LSTM network model training process .

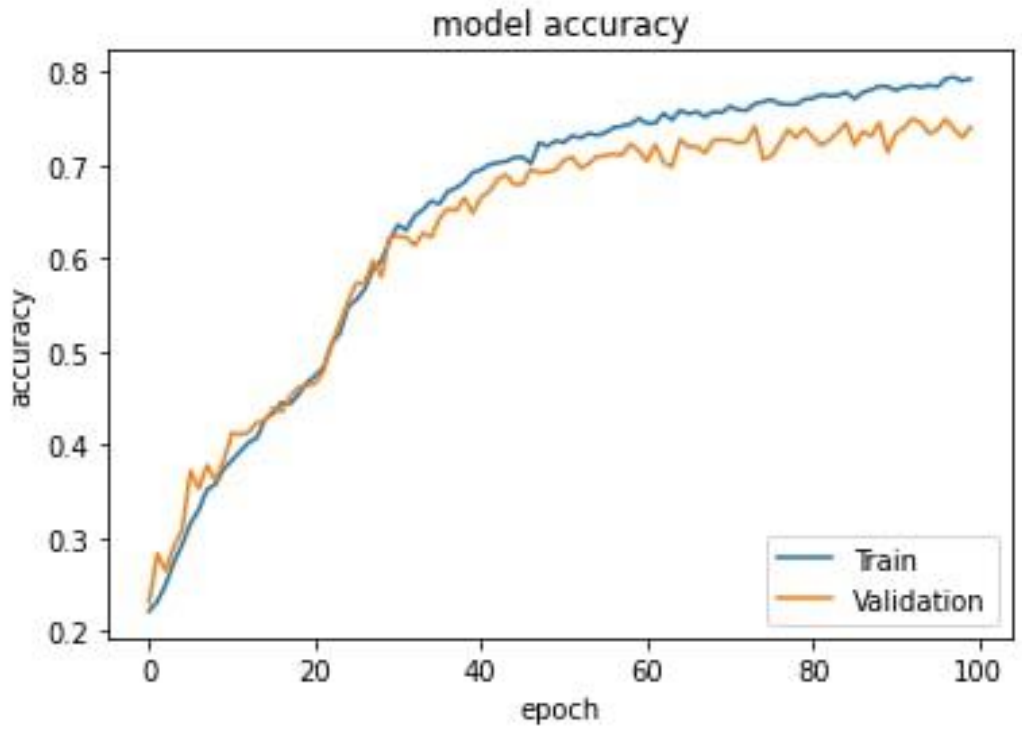


FIGURE (4-3) ACCURACY OF THE LSTM MODEL

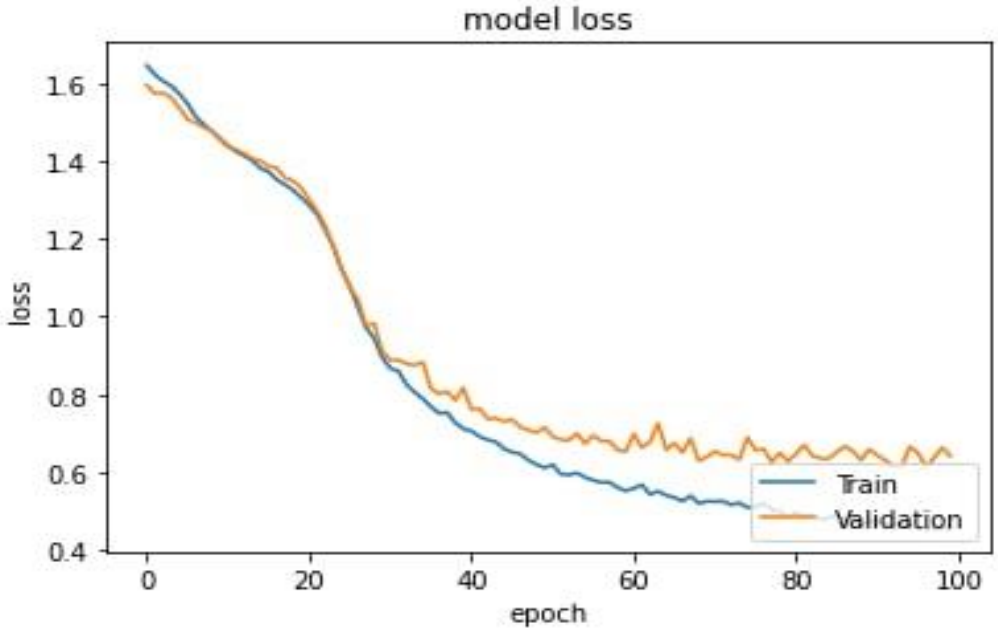


FIGURE (4-4) LOSS OF LSTM MODEL

the proposed system is also compared to a previous work using the GRU model that work on "aishell-1 dataset"[1] (Feng Ye and Jun Yang) which divided into 90% training set, 5% validation set, and 5% test set. The GRU method gives high accuracy 98% and the average accuracy 91% .GRU give results better than LSTM but the proposed methods has better results than the GRU method.

table (4-1):Shows the comparison of the accuracy and loss values of CNN and LSTM and GRU models

TABLE (4-1) EXPLAIN THE DEFFERENCE BETWEEN CNN AND LSTM AND GRU MODELS

Model	High Accuracy	Average accuracy	Validation Accuracy	Loss	Validation Loss
CNN Model	100%	99.80%	97.83	0.45%	2%
LSTM Model	81%	79%	74%	4%	6%
GRU Model	98%	91%	91%	.....	.....

It reveals the proposed CNN based method get the highest accuracy of 99.86%.and the average accuracy is 97.6%, When LSTM method is used, and it exposes not very good result, only 81% on an average accuracy. In case of the GRU method it exposes 98% high accuracy and 91% average accuracy result .the average accuracy of GRU method is less than the proposed method around 6%. And the average accuracy of LSTM method is less than the proposed method around 17%.

Chapter Five

Conclusion and Future Work

Chapter Five

Conclusion and Future Work

5.1 Introduction

This chapter discusses the project's conclusion as well as future work that can be added to the proposed method in the future.

5.2 Conclusion

Identifying a person through speech is a very effective technique that is suitable for the requirements of the current era .In order to verify the identity of a person without the need for direct contact between the person and the machine .The accuracy of the system in determining the identity of a person is affected by several factors, including:

- Activation function used in neural network layers.
- The number of layers in a neural network
- Does a neural network suffer from overfitting or not?

5.3 Future Work

Speaker identification as a sequence learning challenge will be further investigated in future studies employing representation learning techniques like recurrent neural networks. The use of discriminative training, such as when utilizing a Siamese architecture, to the pervasive unsupervised learning problem of speaker identification n is particularly intriguing. Additionally, more research should be done on convolutional filters like pooling and stride and how they might be used to comprehend how the auditory system processes speech over time.

References

[1] T. I. N. K. a. K. K. Supaporn Bunrit, Text-Independent Speaker Identification Using Deep, International Journal of Machine Learning and Computing,, 9, No. 2, April 2019.

[2] R. J. P. C. A. K. a. J. W. Shaun V. Ault, On Speech Recognition Algorithms, International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018.

[3] S. R. S. ,. S. G. W. Sreenivas Sremath Tirumalaa, Speaker identification features extraction methods: A systematic, Expert Systems With Applications, 2017.

[4] A. J. A. A. Nima Yousefian, Speech recognition with a competitive Probabilistic Radial Basis Neural Network, Mashhad, Iran: Proceeding of 3rd International Conference on Information and Knowledge (IKT07), 2007.

[5] 2. B. N. 3. H. 1Ismail Shahin, "1 Novel Cascaded Gaussian Mixture Model-Deep Neural Network Classifier for Speaker Identification in Emotional Talking Environments," *Neural Computing and Applications,* vol. 32, pp. 2575-2587, 2020.

[6] Y. L. ,. W. X. ,. Y. W. ,. H. X. Liyang Chen, "SpeakerGAN: Speaker Identification with Conditional Generative Adversarial Networ," *Neurocomputing,* pp. 211-220, 2020.

[7] R. J. Y. W. T. H. F. N. G. M. M. A. A.-G. I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications,* vol. 171, 2021.

[8] E. N. ,. M. ,. R. I. ,. S. Kristiawan Nugroho, "Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network," *Journal of King Saud University - Computer and Information Sciences,* vol. 32, no. 7, pp. 4375-4384, 2022.

[9] 2. a. J. Y. 1. Feng Ye 1, "A Deep Neural Network Model for Speaker Identification," *MDPI,* 2021.

[10] A. S. E.-F. A.-E. N. El-Sayed Mahmoud El-Rabaie, Speaker identification based on normalized pitch frequency and Mel Frequency Cepstral Coefficients, International Journal of Speech Technology, 2018.

[11] P. S. Jyoti B. Ramgire, A Survey on Speaker Recognition With Various Feature Extraction And Classification Techniques, International Research Journal of Engineering and Technology (IRJET), 2016.

[12] F. Chollet, deep-learning-with-python-second-edition, October 2021.

[13] M. I. J. A. T. M. MITCHELL, Machine learning: Trends, perspectives, and prospects, Science, 2019.

[14] F. Chollet, Deep Learning with Python 1st Edition.

[15] S. B. Wesolowski M, ArtiFcial neural networks theoretical background and pharmaceutical applications, 2012.

[16] K. I. Ahamed, A Study on Neural Network Architectures, Computer Engineering and Intelligent Systems , 2016.

[17] M. a. S. B. Van Gerven, "Artificial neural networks as models of neural, Frontiers in Computational Neuroscience, 2017.

[18] F. J. M. G. a. V. D. F. Bre, "Prediction of wind pressure coefficients, Energy and Buildings, . 2018.

[19] H. N. A. K. S. S. a. N. N. Balakrishnan, "ChaosNet: A chaos based artificial neural network architecture for classification.", Chaos: An Interdisciplinary Journal of Nonlinear Science, 2019.

[20] Y. W. J. L. Y. a. T. W. Li, On neural networks and learning systems for business computing, Neurocomputing, 2018.

[21] M. S. Roza Dastres, Artificial Neural Network Systems, March 2021.

[22] W. I. A. G. a. S. M. Chigozie Enyinna Nwankpa, Activation Functions: Comparison of Trends in Practice and Research for Deep Learning, arXiv:1811.03378v1, 2018.

[23] A. S. S. C. Anirudha Ghosh, Fundamental Concepts of Convolutional Neural Network, researchgate, 2020.

[24] M. R. K. P. Mohit Sewak, Practical Convolutional Neural Networks, Packt Publishing, 2018.

[25] M. M. ,. J. T. F. ,. R. A. Balduíno César Mateus, "Comparing LSTM and GRU Models to Predict the Condition of," *MDBI,* 22 October 2021.

[26] S. K. K. M. A. Anusuya, Front end analysis of speech recognition: a review, International Journal of Speech Technology,, 2011.

[27] H. B. J. Sanjay A. Valaki, A Survey on Feature Extraction and Classification, Computer Engineering Department, L.D. College of Engineering.

[28] T. Point, "Colab," 2019.

[29] W. &Biases, "Weights & Biases Documentation," WandB, 1/9/2022.

[30] T. A. Shaneh M, Voice command recognition system based on MFCC and VQ algorithms, World academy of science. Engineering and Technology, 2019.

الخلاصة:

يمكن للبشر التعرف على المتكلم من خلال الاستماع إلى كلامه ؛ ومع ذلك ، تواجه خوارزميات التعلم العميق مشكلة كبيرة في اكتساب هذه الموهبة البشرية الأساسية. يستخدم التعلم العميق ، مثل المستمعين من البشر ، جوانب من صوت المتكلم لتحديد هوية المتكلم. يُعرف التحدي الحسابي المتمثل في تحديد المتحدثين باستخدام الميزات التي تم جمعها من أصواتهم باسم تحديد المتكلم . في هذا المشروع ، تم تطوير نموذج التعلم العميق للتعرف على المتكلم بناءً على الشبكة العصبية الالتفافية (CNN).تستخدم التقنية المستندة إلى CNN المقترحة طريقة استخراج ميزات Cepstral التقليدية ذات التردد Mel (MFCCs) ، والتي تُستخدم على نطاق واسع لاستخراج ميزة الإشارات الصوتية والكلامية. تقدم هذه المقالة البحثية نظرة عامة سريعة على نظام تحديد المتكلم قبل الخوض في البنية العامة لنظام تحديد المتكلم باستخدام نموذج CNN .تتم مقارنة نهج CNN المقترح مع طريقة LSTM ، والتي يتم إجراؤها ١٠٠ مرة على نفس مجموعة البيانات. للمقارنة ، يتم عرض الحد الأقصى والمتوسط لدقة التصنيف. التقنية المقترحة المعتمدة على شبكة CNN لديها دقة متوسطة بنسبة ٩٩,٨٣٪. يكشف أن الطريقة المقترحة المعتمدة على CNN تحصل على أعلى دقة ٩٩,٨٦٪ ومتوسط الدقة ٩٧,٦٪ ، عند استخدام طريقة LSTM لا تكشف نتيجة جيدة جدًا فقط ٨١٪ على متوسط الدقة. ايضا تمت مقارنة النظام المقترح بدراسة سابقة تعتمد على طريقة الوحدة المتواترة (GRU) وتكشف دقة عالية بنسبة ٩٨٪ ومتوسط نتيجة دقة ٩١٪ ، ومتوسط دقة طريقة GRUأقل من الطريقة المقترحة حوالي ٦٪. ومتوسط دقة طريقة LSTM أقل من الطريقة المقترحة بحوالي ١٧٪ أقل من الطريقة المقترحة. نتيجة لذلك ، تفوق نموذج شبكة CNN المقترح على جميع النماذج الأخرى من حيث مدة تدريب النموذج ودقة التعرف والثبات.لذا النهج المقترح لتحديد المتكلم فعال جدا.

وزرة التعليم العالي والبحث العلمي
جامعة بابل
كلية العلوم للبنات
قسم علوم الحاسوب

التعرف على المتكلم باستخدام طرق التعلم العميق

بحث
مقدم إلى مجلس كلية العلوم للبنات - جامعة بابل
وهي جزء من متطلبات نيل درجة الدبلوم العالي
في علوم الحاسوب

من قبل الطالبة :زهراء عادل علي
اشراف : الدكتور علي يعقوب يوسف السلطان

2022 /9/1