AUTOMATED SIGN TO SPEECH CONVERTION

DEBABRATA SWAIN , DIVSEHAJ SINGH ANAND , ARYAN KUMAR GUPTA , SHREYAS KHARE , DHEERAJ SHARMA , PRACHI TAPADIYA

DEPARTMENT OF INFORMATION TECHNOLOGY

**Abstract** — Unable to communicate verbally is a disability. So in order to communicate there are may ways ,one of the most popular method is use of predefined sign languages . The purpose of this project is to contribute recognizing American sign languages(ASL) in the sector of automatic sign language recognition with maximum efficiency. This paper basically focuses on the recognition of ASL, the real time static gestures are collected from Laptop Webcam. The most challenging part in the design of an automatic sign language translator is the design of a good classifier that can classify the input static gestures with high accuracy. CNN architecture is used to design of classifier for sign languages recognition, in the proposed system . The system is trained by convolutional neural networks for the classification of five(5) alphabets using 150-200 images. With the help of different parameters configurations, system has trained the classifier with different parameter and tabulated the results. Compared to previous literature the proposed work attained an efficiency of 99.35% for our classifier .The result shows that the model accuracy improves as more and more data is collected from various subjects for training.

**Keywords** - gesture recognition, convolutional neural network,sign language recognition, deep learning.

# **1.INTRODUCTION**

Communication is a way of sharing our thoughts , ideas and feeling with others .Verbal and Non Verbal are the two modes of communication. Usually normal people communicate with each other verbally. But speech impaired people cannot communicate with each other verbally. So they communicate with normal people via sign language which is a Non Verbal mode of communication. Most prominently this method is used by mute and deaf people.

Deaf and mute people need to communicate with other people for daily needs and supplies.The deaf-mute people all over the world,use sign language as a medium between them and other people for communication .However,only people who have gone under special training to understand sign language can communicate with them. This leads to a big gap between normal people and deaf-dumb community .Usually, a sign language interpreters is used by deaf people to seek the help for translating their thoughts to normal people . But as a drawback the system is very costly and as a machine its lifetime is less than life period of a deaf person. So there’s a need for a system to automatically recognize the sign language gestures .which will eventually lead to minimization of the gap between deaf people and normal people in the society.There is a wide use of Sign language by the people who can’t speak and hear or people who can hear but can’t speak. Sign language is basically a composition of various gestures formed by making different hand shapes, their movements and orientations , or facial expressions.Sign language is the system which use hand positions to represent the letters of alphabet .Sign language also represent hand signs,convectional gestures,finger spelling .It can also represent a complete phrase not only words.

Sign language is a language which consist of body movements, especially of hand and arms, some facial expressions and some special symbols for alphabets and numbers. Normal people are not able to decode those sign gestures. There should be a system which acts as mediator between mute and normal people. So the proposed system aims at converting those sign gestures into speech which is understood by normal people. So this Automated Sign To Speech Conversion system helps in decoding those symbols without the need of any expert person who understands the sign language. The proposed system uses the Convolutional Neural Network (CNN)architecture. CNN network consist of different layers which processes the input alphabets and symbols and produces the desired output. We have successfully created a model which can help society in a broader way by bridging the communication gap between normal people and mute people.

There are various sign languages all over the globe, each having their own vocabulary. These include American Sign Language (ASL) in Northern America, British Sign Language (BSL) in Great Britain, Japanese Sign Language (JSL) in Japan, South African Sign Language (SASL), Indian Sign language (ISL) etc. We propose a system, acting as an interface between Deaf-dumb community and normal people based on American Sign Language (ASL) as shown in figure-1.

Our model identified the alphabet and gives the output in a speech format. Later these alphabets can be used to from sentences. The model helps the muted people to communicate with the normal people and express themselves. This not only makes their life emotionally better but also makes them more employable and independent. It becomes a lot easier for the normal people to understand the muted people who would in-turnbable to help them.

The feature detection is done using various contour analysis and feature extraction.[1]The feature is taken through the webcam which is then processed into a binary image upon which contour analysis is done and to optimize such actions OpenCV functions are used. The extracted feature is then passed into the neural network which passes the feature through various layers and predicts a single output which is then mapped to a its audio file and hence the audio is played. One of the most important objective of our model is to decrease the communication gap between speech and hearing impaired people with the normal people and use technology to its best in order to smooth integration of these differently abled people in our society.

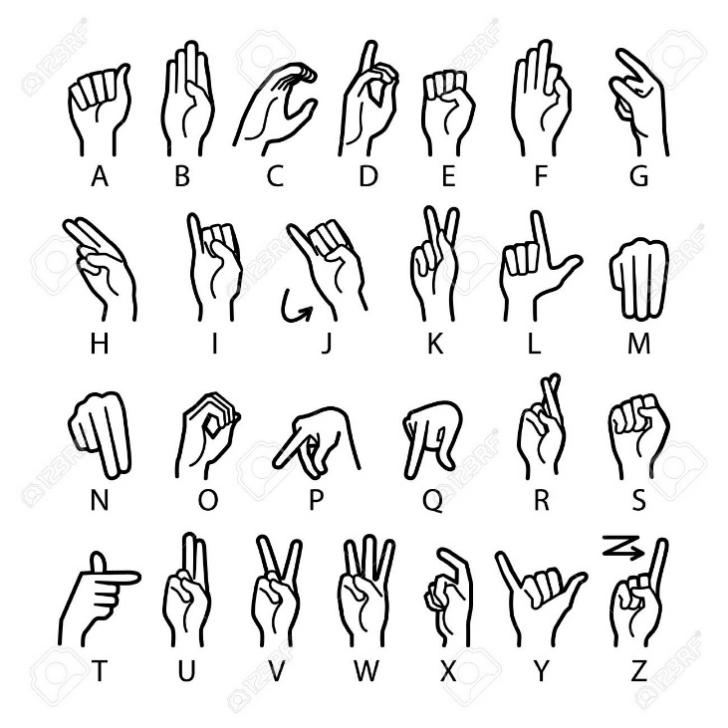


Fig.1 (ASL finger alphabets)

# **2.Literature Review**

Byeongkeun et al. using Convolutional Neural Networks [1] proposed a run time sign language finger. American Sign Language is specific and an important part of sign language recognition. To capture some additional information and to improve accuracy and processing time , the given system had used depth sensors. It is a system for converting sign language to text or speech by capturing the given gestures continuously and processing it without making use of hand-held gloves and sensors. In this method only few images were captured for recognition. The design of a system which bridges the communication gap between normal people and mute people has been created as a prototype.

The given system recognizes American Sign Language which consists of 30 words in its vocabulary. An appearance-based representations and hand tracking system was constructed that was classified by hidden Markov model (HMM). The RWTH-BOSTON-50 database has an error rate of 10.91%.

Another work related to this field was creating sign language recognition system by using pattern matching [2]. The main aim of this proposed work has been to create a system which will work on sign language recognition. Many researchers have already introduced about many various sign language recognition systems and have implemented using different techniques and methods. This proposed system is focusing on an approach which is to put the SLR system which will work on Signs as well as Text (which will be understandable by deaf and dumb persons and also by normal persons). The main task will be performed in two ways by the system. It will take input by the user in the form of text which will be then perform matching with the sign and vice-versa.

Sruthi Upendran [2] and *et.al* introduced “American Sign Language Interpreter System for Deaf and Dumb Individuals”. Out of 24 static alphabets 20 alphabets could be implemented in the procedure. Due to occlusion problem, the letters A,M,N and S couldn’t be recognized. The number of images used were limited.

The same can be implemented using an optimized approach by implementing the famous algorithm with LBP feature i.e. Viola Jones to recognize hand gestures in real time environment. Using this algorithm, we created Indian sign language interpreter with android implementation [3]. The advantage of this approach is that it takes less computational power to detect the gestures.

Chenyang Zhang, Yingli Tian [4] *et.al* “multi-modality American Sign Language recognition” . Two most important features of the system are: 1) The multiple signal modalities like sequence of depth images, different shapes of hand based on RCB colour scale, attributes of facial expressions and some attributes required to detect body joints and facial landmarks were considered. 2)Our system is capable of recognizing different components such as English alphabets , words and some special ASL grammar components which consist of head movements and facial expressions that gives some grammatical meaning to our sentences. For recognizing these ASL signs , we have performed signing sequences by gaining knowledge of ASl signs from professional linguistician.

Numerous algorithms and techniques have been produced with the help of image processing and artificial intelligence. Required patterns are obtained by some gesture based acknowledgment framework which learns and recognizes different signs and symbols. With the help of MATLAB Double handed Indian Sign Language are computed and caught as a progression of pictures. After that it is changed into text and speech.[5]

EMG is a signal based hand gesture identification. Two sensors have been used for its implementation. These sensors are placed on forearm and the results of movements of wrist and fingers are recognized experimentally with a great accuracy by using these sensors. One finger movements are accurately recognized using this model. Wrist and finger movements are recognized by collecting sufficient information from the measurement of muscle activity.[6]

A novel and real-time System for interacting with an application or video game via hand gestures is presented by N.H. Dardas and N.D. Georganas [7]. Using skin detection and hand posture contour comparision algorithm , the proposed system detects and tracks bare hand in cluttered background. Multiclass support vector machine (SVM) is used to recognize via bag – offfeatures. The system also builds a grammar which generates gesture commands to control an application.

A classification system for recognizing various Sign Languages of India has been proposed by J.Singha, K. Das [8] using Eigen value weighted Euclidean distance technique. Skin Filtering, Hand Cropping, Feature Extraction are used for classification.

A prototype system for hand tracking and hand posture recognition has been recognized by L.Bretzner, et.al [9]. Representation of Hand postures is done by using hierarchies of muli-scale image features at different scales which has qualitative interrelations in terms of scale, position and orientation.

**3. Convolutional Neural Network**

Covolutional Neural Network(CNN) is a deep learning algorithm which takes image as an input . It contains many layers viz. Convolutional layer, Pooling layer, Relu layer, Fully connected layer.

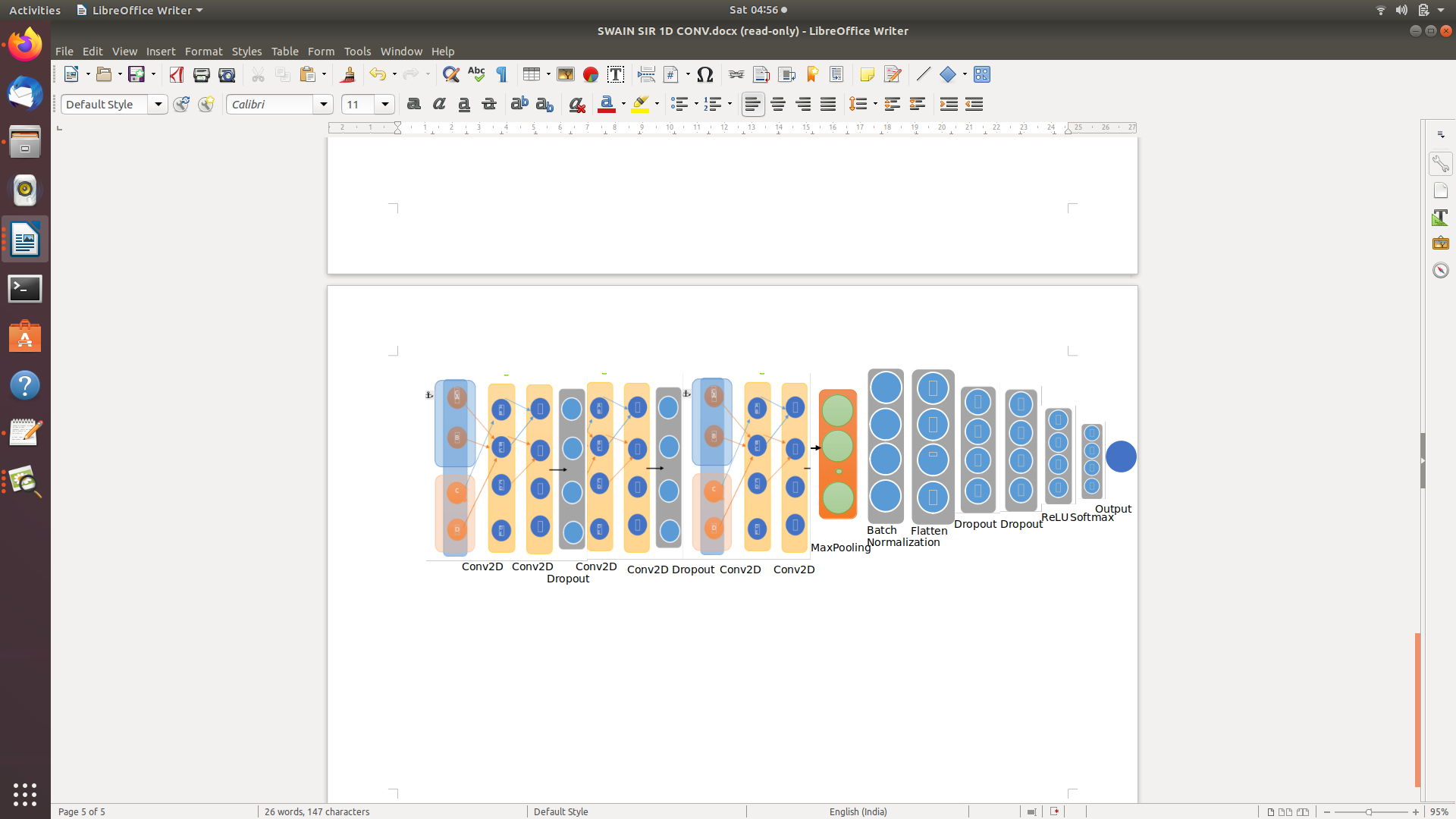


Fig 2 : (Proposed System Architecture)

## **Convolutional Layer**

The convolutional layer is also known as filtering layer. As the input image is taken, the patterns in sections of that image are checked. The filters, vectors which are stacks of weights, are multiplied by the output values by the convolutional layer hence this works. From the combinations of these various filters the content of image is predicted. A Convolution Layer is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

## **ReLu Layer**

The full form of ReLu is Rectified Linear Unit. It is the activation layer of CNN. For model to learn faster and perform better, the vanishing gradient problem is conquered by this ReLu layer. The rectified linear activation function is a piecewise linear function that will output the input directly if is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance. Relu layer is used to avoid vanishing gradient descent problem. In vanishing gradient descent problem the weight updation is very much negligible. But using Relu the weight updation is quite significant that impacts the training in an efficient manner.

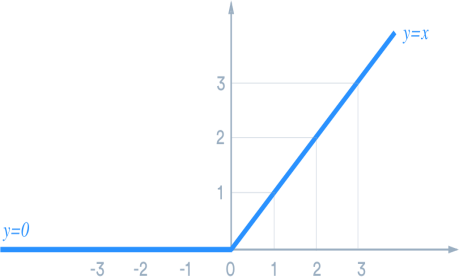


Fig 3.(Relu Activation Function)

## **Sigmoid Function**

The sigmoid function is a activation function which is also known as squashing function. This function limits the output value between 0 and 1, which is useful in the prediction of probabilities. The following fig 4 shows the graph of sigmoid function. Therefore, it is especially used for models where we have to predict the probability as an output.

A screenshot of a cell phone

Description automatically generated

Fig 4.(Sigmoid function)

## **Pooling Layer**

The spatial size of the convoled feature can be reduced by using the pooling layer.The computational power needed to analyse the data decreases via dimensionlity reduction. A pooling layer is a new layer added after the convolutional layer. Specifically, after a nonlinearity (e.g. ReLU) has been applied to the feature maps output by a convolutional layer. The addition of a pooling layer after the convolutional layer is a common pattern used for ordering layers within a convolutional neural network that may be repeated one or more times in a given model.The two types of pooling layers are Max pooling and the Average pooling.

## **Max pooling:**

## The kernel covers the value which is maximum from portion of image which is returned by this layer. In the example below out of the 4 values the maximum value will be chosen by the max pooling layer. The max pooling layer hence reduces the data size coming from the previous layers.

maxpooling =MAX(y1,y2,y3,y4,……, yn) (1)

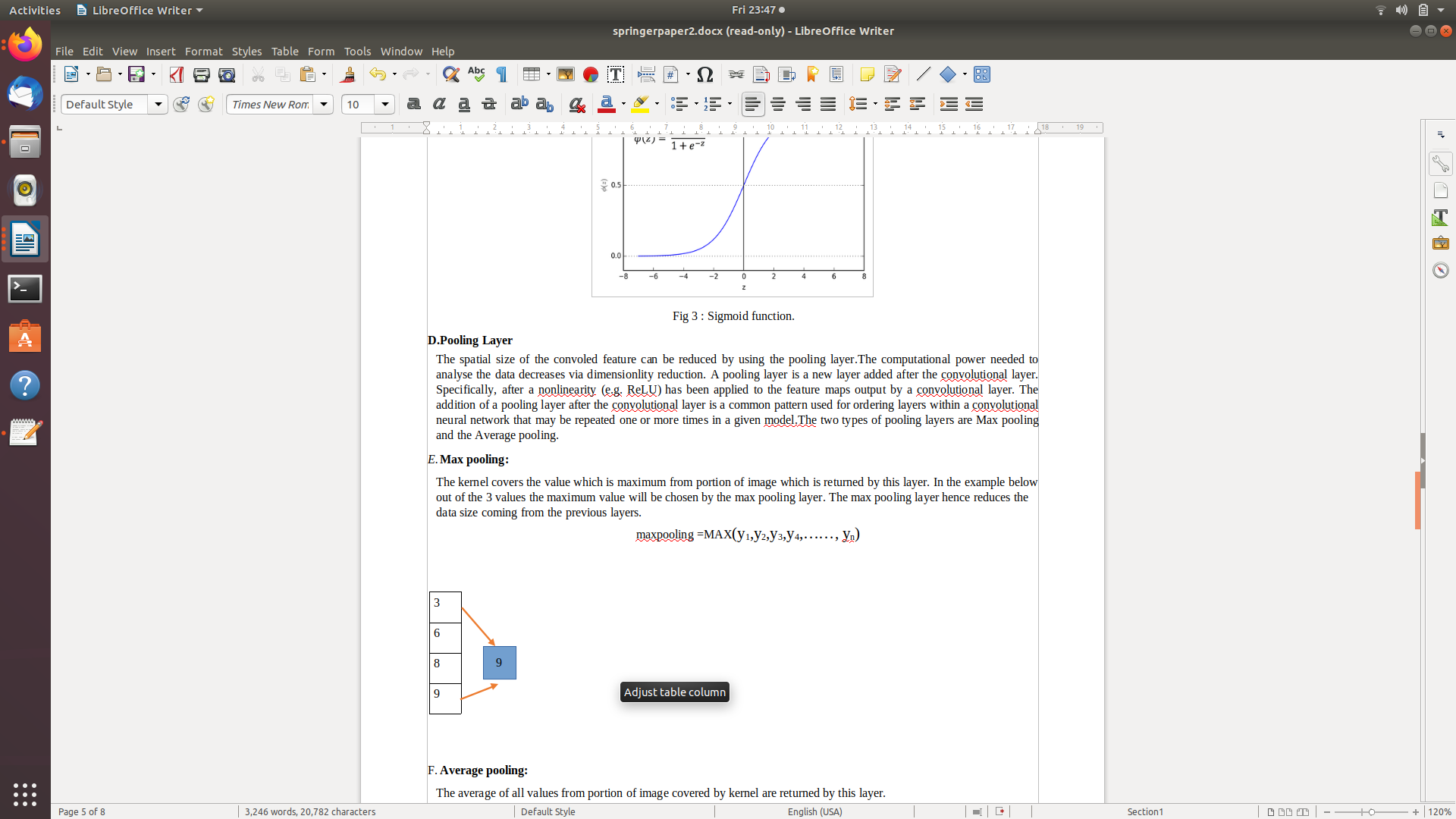


Fig 5.(MaxPooling Layer)

## **Average pooling:**

## The average of all values from portion of image covered by kernel are returned by this layer. In the example below the average of the 4 values will be calculated and replaced with the patch of 4 values. The input to this layer is obtained from the previous layer.

average pooling=1/n(∑ni=1 yi) (2)

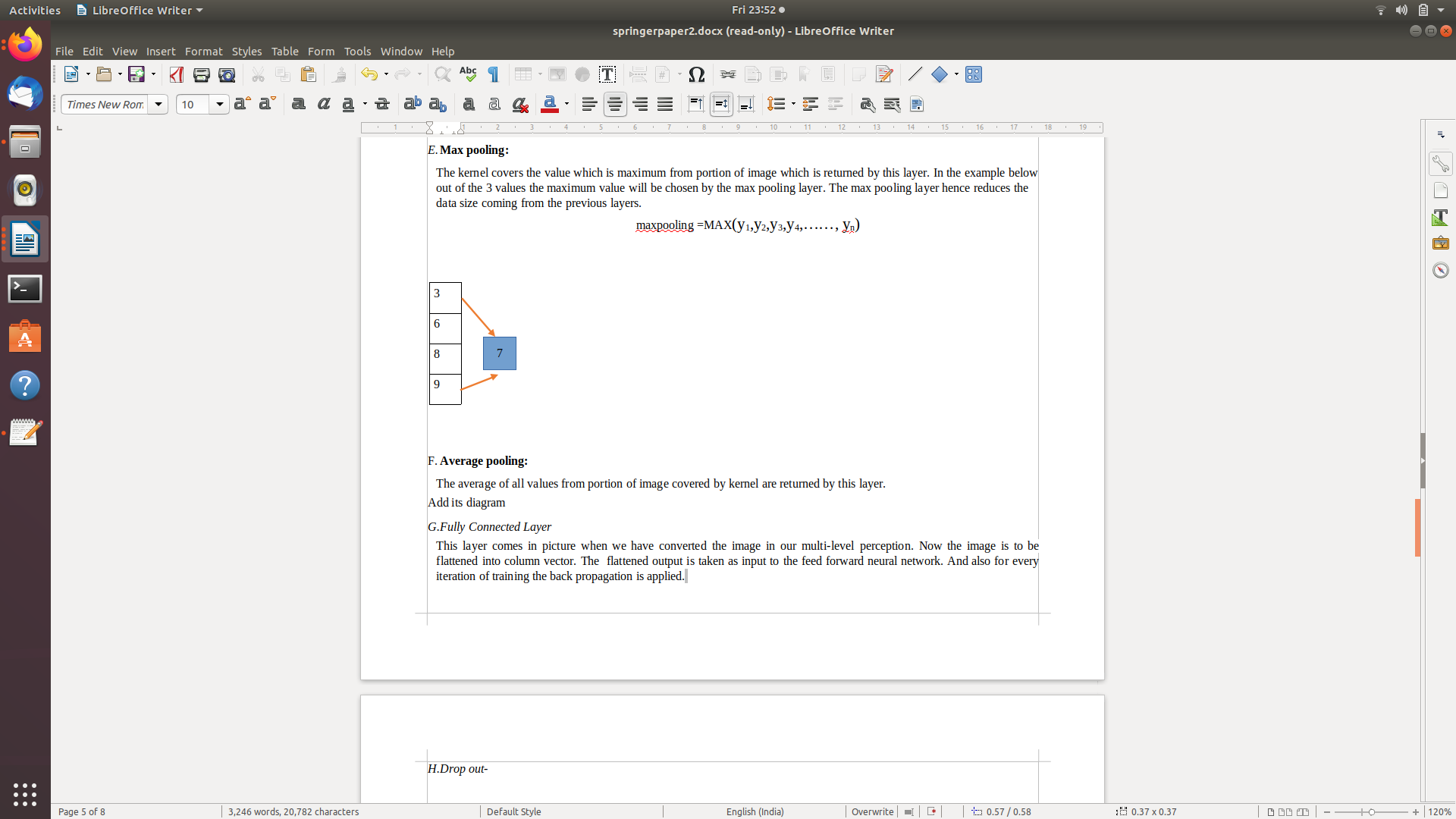


Fig 6.(Average Pooling Layer)

## **Fully Connected Layer**

This layer comes in picture when we have converted the image in our multi-level perception. Now the image is to be flattened into column vector. The flattened output is taken as input to the feed forward neural network. And also for every iteration of training the back propagation is applied.

## **Drop out-**

As a neural network learns , randomly seected neurons are ignored or droped out. This is done to avoid the model to get overfit.Dropout finds the probability of every input neuron for a particular class of an output neuron. In the example below the dropout value is specified as 0.1. This means that the neurons whose probability os less than 0.1 are removed from the output neuron.

P(xi | yi =1) < 0.1 (3)

## **Flatten**

This layer returns a single vector which can be the input for the next stage.The flattened layer is formed from the output of the previous layers.

## **Batch Normalization-**

Training neural networks with a lot of layers is difficult.Batch normalization is used to standardize the inputs to a layer for each mini batch.This process reduces the number of training epochs required to train the neural network and stabilizes the learning process.

## **Softmax Function-**

Softmax is an activation function.Softmax is used to normalize neural networks output between zero and one.The normalization is calculated by dividing the exp value of the computed output by the summation of the exp value of each possible output.

Softmax (xi)=exp(xi) / ∑jexp(xj) (4)

## **Loss Function**

Loss functions are optimizers.It is used for evaluating how well an algorithm models the given data. It basically error measures the deviation of predicted and actual outputs.We use optimizers to change the attributes of the neural network such as weights and learning rate in order to reduce the losses.In the proposed system Adam optimizer has been used to get a fine set of parameters.Here the prediction is a single label so categorical cross entropy has been used.The categorical cross entropy will compare the distribution of the predictions with the true prediction.Below is the equation of the categorical crossentropy ,where ŷ is the predicted value.

L(y,ŷ) = M∑j=0 N∑ i=0 (yij \* log(ŷij)) (5)

# **3.Methodology/Experimental**

## **3.1DATASET COLLECTION**

It is a very crucial stage in the research. This a sole stage on which the training phase and the testing phase all are interdependent. Here, as our model has the ability to sense and detect 5 alphabets namely A,B,C,D,E. So,we have proceeded by collecting 400 images of each alphabets. Thus in total we have 2000 images out of which we have 300 images to the testing part equaling to the 15% and the rest 1700 images have been sent to the training part equaling to the 85% of the total images. Hence, these training part again has actual training and the validation split training part. So, in the actual training part we have 1500 images and, in the validation, split training part we have 170 images in total for it.

## **A. Training**

We trained the model by varying the hidden layers from 1 to 4 and by varying the epochs 100 in each case. The number of nodes in each hidden layer also varies. Out of the 400 samples of each image we used 300 images for creating the training set, 60 for validation and 40 for testing. The data sets were pickled and were used for training. Given flowchart shows the training process. After the model was trained the net parameters were saved so that it can be used in the testing phase for testing the accuracy of the model and also for classification of the input symbols

A picture containing text, map

Description automatically generated

Fig 4.(Training Phase Workflow)

## **B. Testing**

In the testing phase the accuracy of our trained model is tested. The saved network parameter is loaded to test the dataset and determined the accuracy. The method used for all test cases and tabulated the obtained accuracies. The model allows the user to select the images of the sign language static gestures that need to be classified. The application used our trained CNN network to classify these symbols and produce their corresponding labels. The labels are then turned to their corresponding alphabets/numbers. The words/numbers are then spoken out using the python text to speech module.

# 

Fig 5.(Testing of the model)

**4.Result**

The model successfully recognizes the 26 English Alphabets and converts it into speech with a testing accuracy of 99.51%.

**Confusion Matrix:**

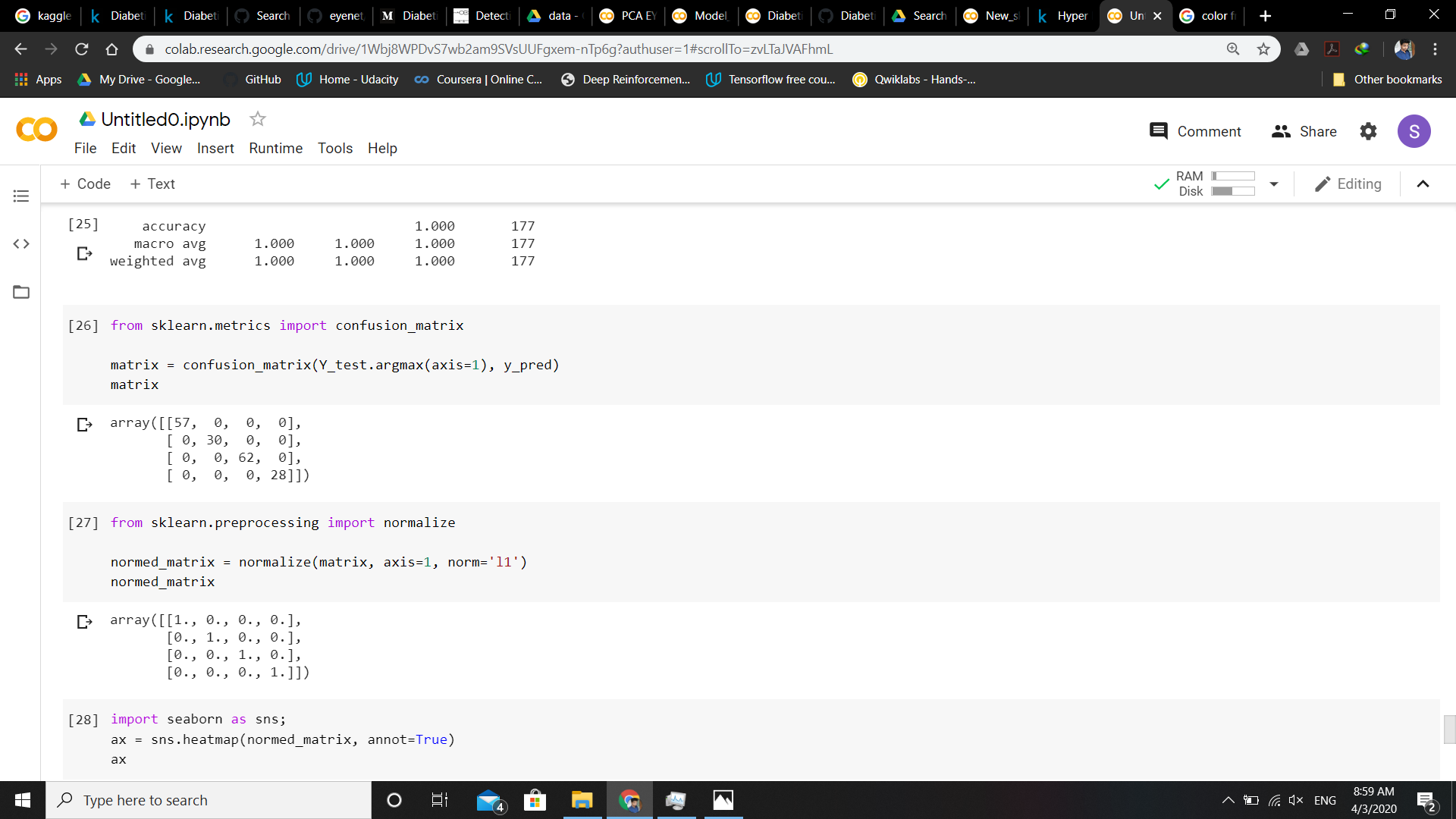


Fig 6.(Confusion Matrix)

**Normalized Matrix:**

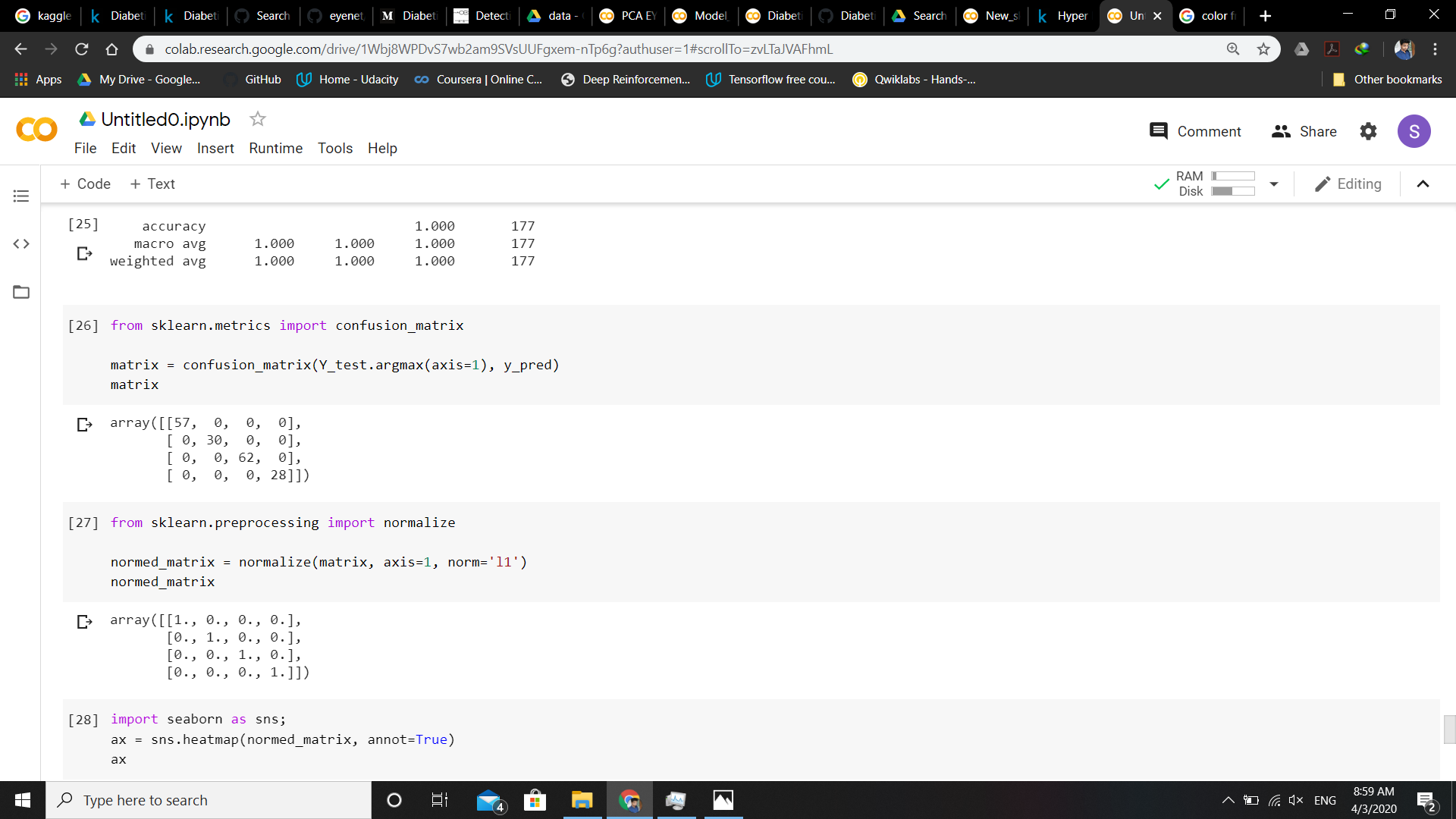


Fig 7.(Normalized Matrix)

# **5.Performance Analysis**

**A. Heatmap**

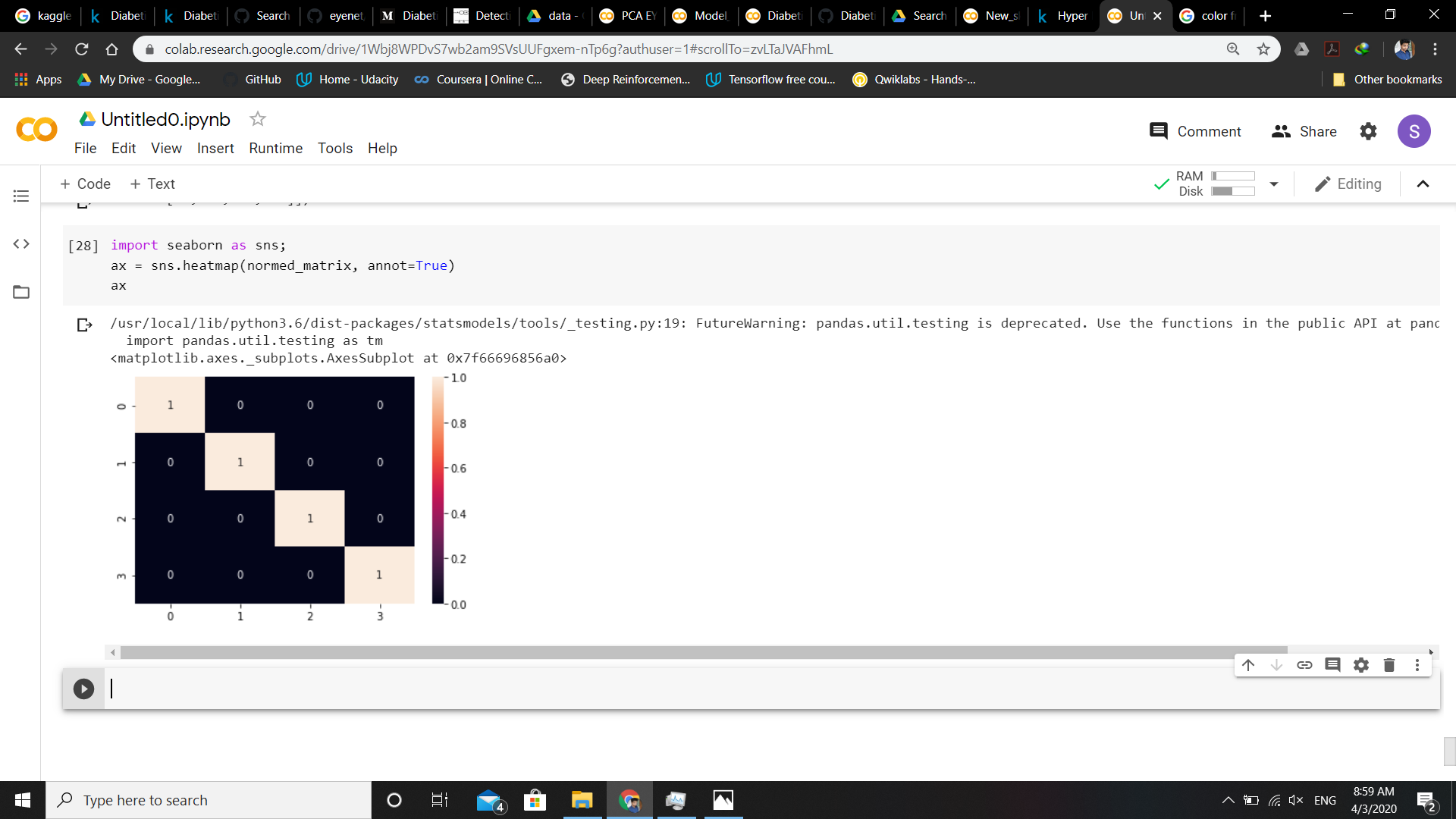


Fig 8.(Heatmap)

**B. Metrics Classification Report**

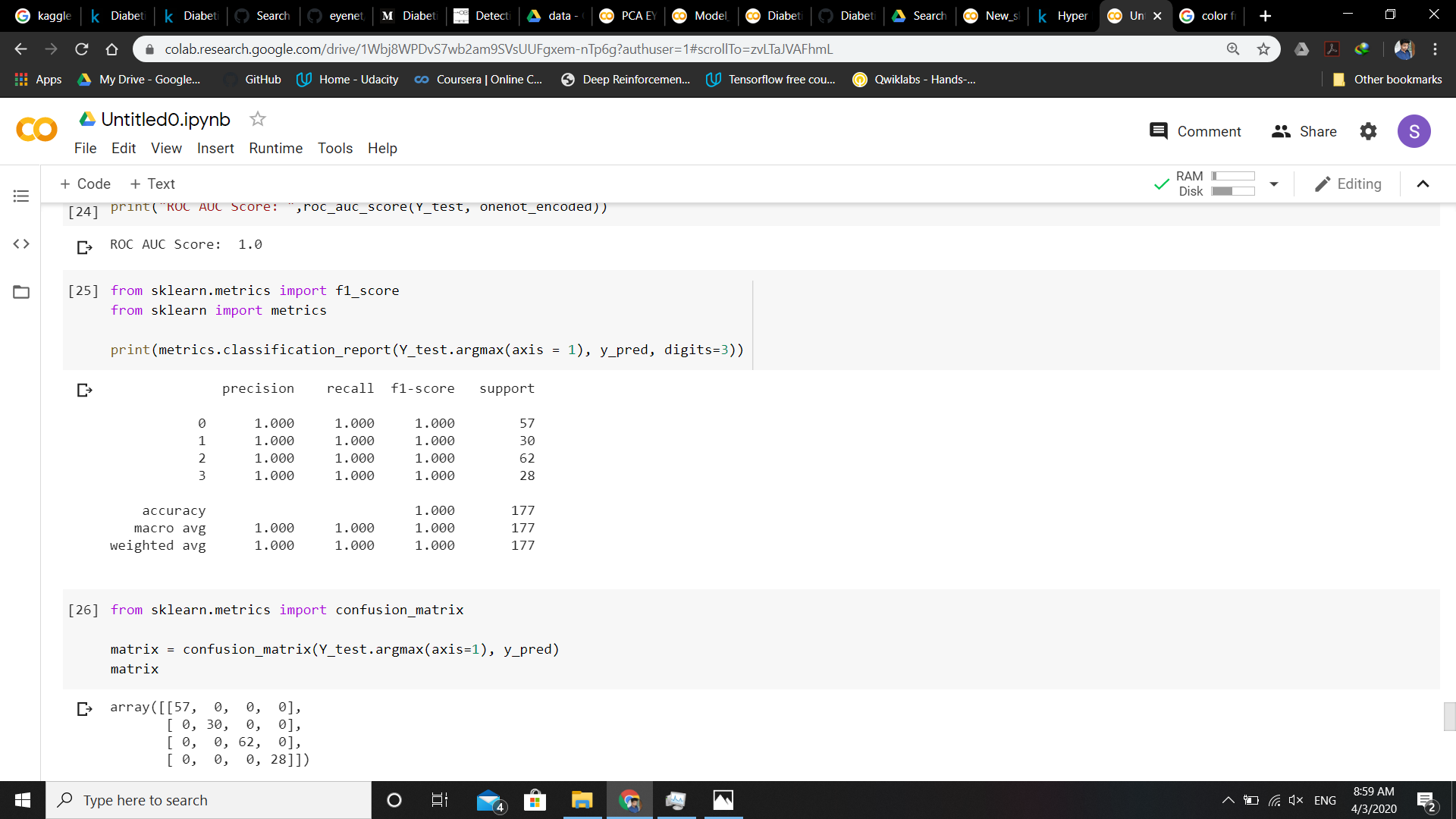


Fig 9.(Metrices Classification Report)

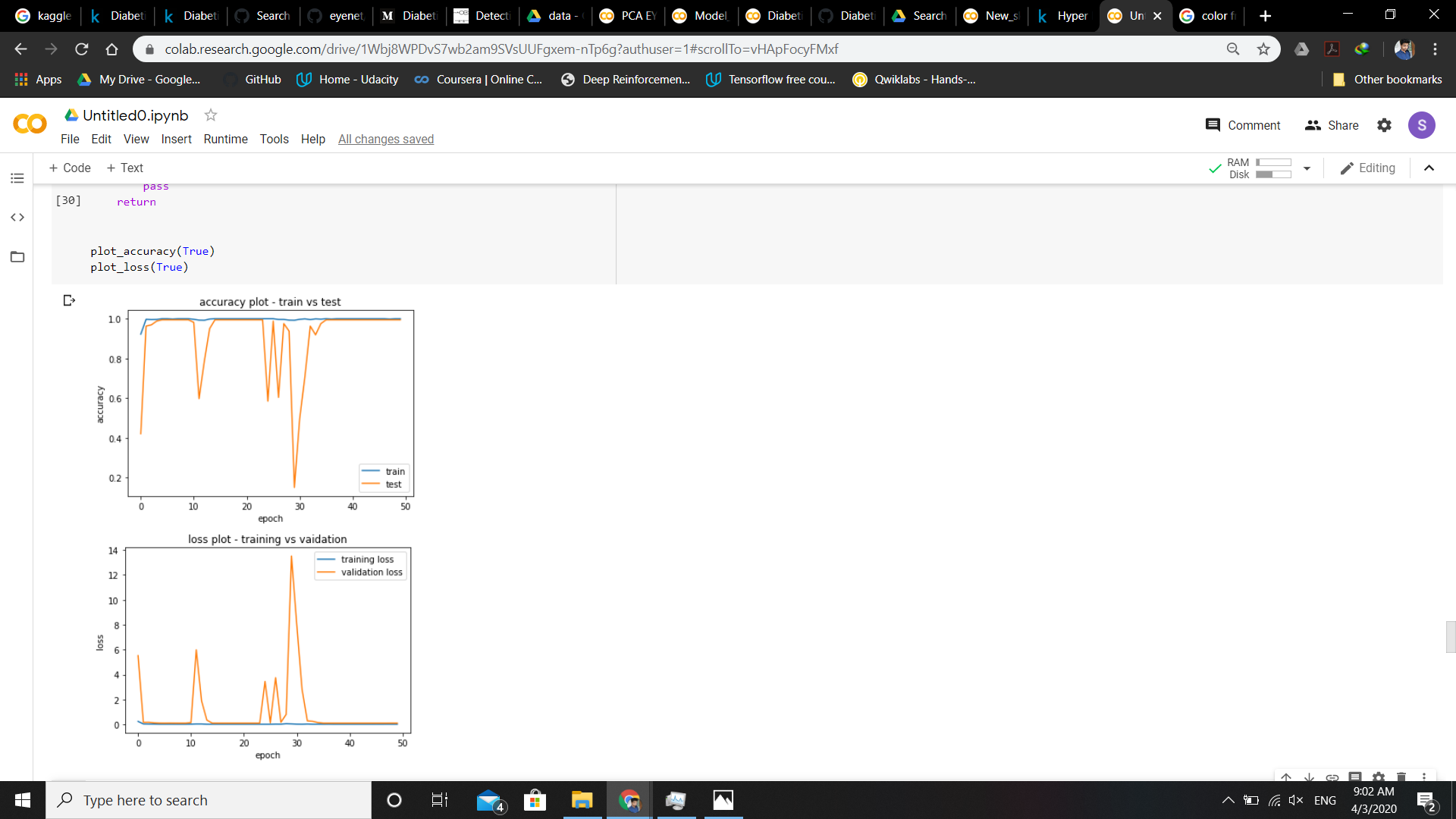
**C.Accuracy Plot**

Fig 10.(Accuracy Plot)

**D.Loss Plot**

# 

Fig 11.(Loss Plot)

# **6.Limitations**

In the project we have not predicted the numbers from 0 to 9 and A to E. Further the model fails in recognizing sentences and phrases. It is capable to take only sign at a time.The model also fails in detecting some expressions.The model is currently not detecting any feature or phrases.The model only recognizes static images and not dynamic images.

We need a large data set to get the best accuracy and to get better results in the shortest we might require a GPU.There are various types of sign languages like ASL(American Sign Language ), ISL (Indian Sign Language), BSL (British Sign Language) so it is difficult to make one model which can be globally accepted[6].

# **7.Future Scope**

In this project we have converted some alphabets and real time gestures into speech. Further by concatenating those alphabets we can convert phrases and sentences into speech.

By using high end processor we can have more number of data sets to increase the training and testing accuracy of the model.We also plan to use RNN and get the features of the hand using coordinates which would detect more gestures which in turn can be used to frame sentences rather than making sentences using each alphabet.The application can be integrated with mobile applications to improve the user interaction and experience.We also plan to make the model work for dynamic gestures as well and not restricting to only static images.

# **8.Conclusion**

The main idea behind this project is to overcome the barrier of communication between Deaf-mute people and hearing majority. Using CNN we recognized different alphabets of a sign language and some real time gestures and converted it into text. Further this text is converted into speech. The model allows users to select the static sign gestures as input and it will speak out the letters.[7]The whole point of making the solution is to make the lives of the muted people easier and comfortable by making a viable , user friendly and accurate product.

**9. References**

[9] L. Bretzner, et.al "Hand gesture recognition uses multi-scale color features, hierarchical models and particle \_ltering*", in Processing of* *International Conference. Au-tom. hand Gesture Recognition,* *Washington, DC, May 2002.*

[8] J. Singha and K. Das, “Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique*”,(IJACSA) International Journal of Advanced Computer* *Science and Applications, Volume 4, no. 2 , pp. 188-195,July 2013*.

[7] Nasser H. Dardas and Nicolas D. Georganas, “Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques,” *IEEE Transaction on Instrumentation and* *Measurement, volume 60, no. 11, pp.132-137 November 2011.*

[5] Kusurnika Krori Dutta, Satheesh Kumar Raju K, Anil Kumar G S, Sunny Arokia Swarny B, “Double handed Indian Sign Language to speech and text”, IEEE, 2015 Third International Conference on Image Information Processing.

[6] Andrea Ferrone; X. Jiang; L. Maiolo; A. Pecora; L. Colace; Carlo Menon ,”A fabric-based wearable band for hand gesture recognition based on filament strain sensors: A preliminary investigation”, IEEE, 2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT).

[4] <https://acadpubl.eu/jsi/2017-117-20-22/articles/20/2.pdf>

[1]<https://www.researchgate.net/publication/305233784_Real-time_sign_language_fingerspelling_recognition_using_convolutional_neural_networks_from_depth_map>

[3]<https://pdfs.semanticscholar.org/f63c/fcdd63bd34bcd6ec028169e6fe144e9cc83c.pdf>

[2]Convolutional Neural Networks for Speech Recognition Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yuhttps://www.microsoft.com › uploads › 2016/02 › CNN\_ASLPTrans2-14