Qaida Accuracy Teller Using Convolutinn Neural Networks

***Abstract*—The Quran is the Holy Book and a book of guidance. In Islam, Muslims are recommended to recite Quran to earn Sawab(Virtue). Quran is in Arabic Language and there is a set of rules for the correct pronunciation of arabic letters with all their qualities called Tajweed. To learn the Recitation of Quran with Tajweed, one starts learning from 'Qaida'. 'Qaida' is an Arabic Word which means the 'foundations'. Nearly all of the Tajweed and pronunciation guidelines for reciting the Quran are contained in the Qaida.There is a need of Qaari for learning Qaida who teaches and supevises wheather one is doing correct or wrong. A Qaari is a person who recites the Quran with proper Rules of Recitation. In Muslim countries like Pakistan, Qaaris are easily available, but in non-muslim countries where are very less or few Qaari availability most people doesn't get the privilege of recitation of Quran with Tajweed. The goal of this paper is to discover a machine learning algorithm that gives the Accuracy of Arabic words spoken on the basis of Tajweed so people who can't have a Qaari could learn and get accuracy of their recitation with Tajweed. The data we are using contains of audios of every arabic alphabet Recited with Tajweed by the students of Darul-ul-Uloom. This data set was based on Mel-Frequency Cepstral Coefficients images of audios.**

***Index Terms*—Quran, Arabic, Qaida, Qaari, Tajweed, machine learning, deep learning, Mel-Frequency Cepstral Coefficients**

I. Introduction

Accuracy calculation of an input data on the basis of similarity of the actual standardized set of data can be done through deep learning. Models can adapt and learn tasks to help better understand the similarity. As a result, we may employ these Deep Learning models for various tasks. Accuracy in voice is a field where deep learning models can be applied.

One of the major concerns in Recitation of Quran is reciting it with Tajweed. There had always been a need of Qaari for teaching Tajweed, speacially for the children and who doesn't know about it. Its a traditional background of teaching and learning Quran by person to person. That has now taken as a role by Qaaris. The Qaari listens the recitation of the student and tells how much he/she is correct and how they need to recite. The students again listens and tries to pronounce same as Qaari. this all shows the accuracy has been measured by Qaari in brain according to the standards of Tajweed. Here is the part where machine learning algorithms come in play.

In our research, we want to make a deep learning model that when given an Audio file of an alphabet, it can tell how much accurate that alphabet had been spoken, based on the feature of the audio: MFCCs. The data that has been used to train the model was not available as we needed the standard data to train the model accurately. We created a valid and authentic dataset from Darul-ul-Uloom Karachi and all voices were recorded under the supevision of a Qaari.

II. TAJWEED

The word "Tajweed" means to improve, make better. Tajweed is defined as articulating each letter from its point of articulation and providing it its appropriate rights and qualities. It protects the tongue from pronunciation errors made while reading the Glorious Qur'an. To learn a Tajweed, one must comprehend and diligently adhere to its many tenets. Tajweed rules are classified as Noon and Meem Mushaddad, Al –Qalqalah, Noon saakinah and Tanween Rules, Meem Saakinah Rules, Al-Madd Rules.
The rules gets divide wit alphabets, joining of alphabets, Al-Mad that means extending the length or pronouncing too long for that particular alphabet. [3]. A model for Tajweed rules can be seen as follows:

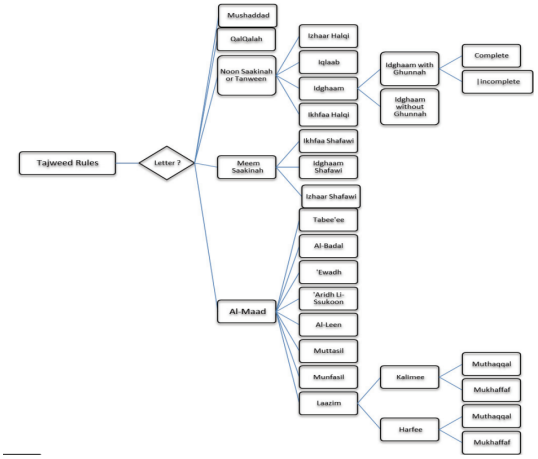


Fig. 1. A model for Tajweed Rules

III. LITERATURE-REVIEW

Convolution neural networks (CNNs) are used in this project. lets first dive into what is CNNs and why they are used. CNN's, also known as ConvNets, are a subclass of neural networks that excel at processing data with a grid-like architecture, like images. The moment we perceive an image, the human brain begins processing a massive amount of data. Every neuron has a distinct receptive field and is coupled to other neurons so that they collectively cover the whole visual field. Convolutional, pooling, and fully connected layers make

up the conventional architecture of a CNN.

Convolution Layer: The foundational component of the CNN is the convolution layer. It carries the majority of the computational load on the network.This layer creates a dot product between two matrices, one of which is the kernel—a collection of learnable parameters—and the other of which is the constrained area of the receptive field. Compared to a picture, the kernel is smaller in space but deeper.

Pooling Layer: By calculating an aggregate statistic from the surrounding outputs, the pooling layer substitutes for the network's output at specific locations. As a result, the representation's spatial size is reduced, which reduces the amount of computation and weights needed. Maximum output from the neighbourhood is reported by the most well-liked process, max pooling. By calculating an aggregate statistic from the surrounding outputs, the pooling layer substitutes for the network's output at specific locations.

Fully Connected Layer: As in a conventional FCNN, all of the neurons in the layer above and below have a complete connection with one another. Consequently, it can be calculated using a matrix multiplication followed by a bias effect. Between the input and the output, the FC layer aids in mapping the representation.

Non-Linearity Layers: Non-linearity layers are frequently included right after the convolutional layer to add non-linearity to the activation map because convolution is a linear operation and images are anything but linear. Non-linear operations come in a variety of form such as Sigmoid, Tanh, ReLU.

1) **Sigmoid:** The mathematical formula for the sigmoid nonlinearity is:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

2) **Tanh:** A real-valued number is condensed by Tanh to the range [-1, 1]. Similar to sigmoid neurons, the activation saturates, but unlike them, its output is zero-centred.
3) **ReLU:** Recently, there has been a huge increase in interest in the Rectified Linear Unit (ReLU). The formula f(x)=max(0,x) is calculated. To put it another way, threshold activation is all that is happening.

ReLU speeds up convergence by six times and is more dependable than sigmoid and tanh. ReLU may be vulnerable while being trained. It can be updated by a strong gradient that prevents the neuron from ever updating further. However, by choosing an appropriate learning rate, we can accomplish this.

Applications:

There are many applications of Convolution neural networks. Some of them are:

Decoding Facial Recognition: A convolutional neural network separates facial recognition into the following key elements. recognising each individual face in the image. despite other influences like light, angle, stance, etc., concentrating on each face. recognising distinctive qualities. The labelling of scenes is done in a manner akin to this.

Analyzing Documents: Document analysis can also make use of convolutional neural networks. This has a significant impact on recognizers in addition to being helpful for handwriting analysis. A machine needs to process approximately a million commands per minute in order to scan someone's writing and compare it to its extensive database. Though its full testing has not yet been widely observed, it is claimed that the error rate has been reduced to a minimum of 0.4 per cent at a character level with the use of CNNs and newer models and algorithms.

Understanding Climate: CNNs can be incredibly useful in the fight against climate change, particularly in figuring out why we're seeing such dramatic changes and what we might try to do to slow them down. It is claimed that the information contained in these natural history collections can also help researchers gain deeper social and scientific insights, however this would necessitate the availability of qualified human resources, such as researchers who can physically visit these kinds of archives. To conduct more in-depth experiments in this area, more personnel are required.

We applied CNNs on our data, and different models of CNNs for our project.

IV. DATASET

We are helping people to accurately pronounce the Arabic Qaida. One thing what we need to do most perfect in our whole project is our Data set collection. We need to standardize our data which means we can not simply extract data from the internet and use it, or give some potential source, if we fail to give the justification of our data to our user of this whole project that from where we are extracting the data or from which source we are judging your accuracy of Arabic Qaida then the purpose of this project comes to zero. Because of this data collection and it's validity was become our most difficult task. Thinking of the user's perspective we have to collect data manually from the authentic source, which lead us to another problem which is data deficiency. We decided that we will take five voices for the same Arabic letter from the same source, so from this we have 145 different voices from the same source. There are 29 letters in Arabic Qaida, our task to take 5 different voices of each letter from the same authentic source to train our model. The source which we consider as authentic is either a Hafiz-e-Quran or a person which has a very good Tajweed (good pronunciation of Arabic letters). We have to choose some schools where students are entertained with some good Qur'an Arabic teaching. We also added the negative voices for each letter of Arabic alphabets. Negative voice means the wrong pronunciation voice which doesn't follow the rules of tajweed. We had around 3190 voices then at all. The model should be trained with some voices that have noises in background. For which we add background voices for noise in each audio such as dishwasher working sound, noise of people standing near at the supermarket ambience, etc. We then got over all around 9570 audios in total.

For our data set we have selected Jamia Darul Uloom, Karachi which is a seminary for Muslims in Karachi, Pakistan. It carries on the Darul uloom tradition started by Darul Uloom Deoband, India. Boys and girls have separate faculties in the secondary schools. The school offers a mixed curriculum that includes traditional Islamic studies and modern academic disciplines and upholds the highest standards of Islamic education. It is included with Wafaq ul Madaris Al-Arabia as one of the Islamic schools. In 1951, Maulana Noor Ahmad and the late Mufti Muhammad Shafi founded the seminary [1]. He had belonged to Darul Uloom Deoband, where he served as Grand Mufti, but after Pakistan gained independence in 1947, he relocated to Pakistan. Grand Mufti of Pakistan Muhammad Rafi Usmani currently serves as president of Darul Uloom Karachi, and his vice president is Muhammad Taqi Usmani.

All the audios were the wave files. Once we generated the

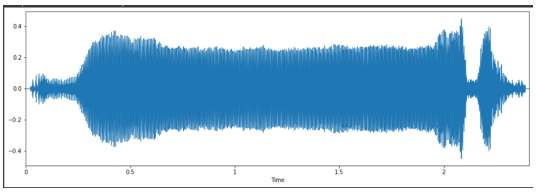


Fig. 2. Audio image using matplot.

enough amount of data from the authentic source then our task was to clean the voices and convert them into MFCC images. We added the silence at the end in some voices so all MFCCs would be of same size. We will generate the MFCC images from the python's librosa library.

Mel Frequency Cepstral Co-efficient Generation:

Mel frequency cepstral Co-efficient gives the 2D representation of our voices in terms of spectrogram with the frequency on the y-axis and time on the x-axis. The MFCC gives a discrete cosine transform (DCT) of a real logarithm of the short-term energy displayed on the Mel frequency scale [2].

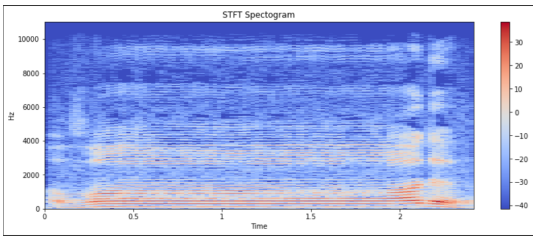


Fig. 3. STFF Spectogram.

The following is the way to calculate the MFCC.

- Frame the signal into short frames.
- For each frame calculate the periodogram estimate of the power spectrum.
- Apply the mel filterbank to the power spectra, sum the energy in each filter.
- Take the logarithm of all filterbank energies.
- Take the DCT of the log filterbank energies.
- Keep DCT coefficients 2-13, discard the rest.

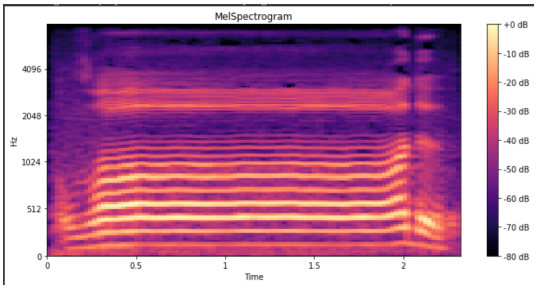


Fig. 4. MEL Spectogram.

The MFCC's then split into validation data set, and training set. In this process we have attained 3,190 original voices in which half of them are negative and half of them are positive. The data we have collected is the real-life voices in which standardization problems may occur, which means some voices are larger in duration as compare to other. So, as we all know that deep learning models cannot take unstandardized data set as input, we need to go through pre-processing procedure before using our dataset. We have used Python Pydup library to make our entire voices as one duration so from this we make our data in the standardized form. For gaining the correct accuracy the data we have collected was not enough so for gaining more data we start augment our dataset by adding three types of noises at the background. So the total number of voices in our dataset become 3 times of 3,190 which is 9,570. After generating this much amount of data we have used Python to read all the folders and generate the Mel spectrogram of each voices as we have to give an image as an input to our CNN model.

V. METHODOLOGY

The Major and Essential part of our project is gathering authentic data in order to train the model in a sufficient amount of epochs based on batch size. Since we chose the Convolution Neural Networks models to train on the data set. The process contains three main steps.

- Preprocessing and Augmentation of the data set.
- Generate Mel Frequency Cepstral Co-efficient from the data set.
- Training the data set on different Convolution Neural Networks.

A. Preprocessing on Data Set

1) Standardization of the data set: The gathered data set were not standardized in terms of speech length. The speaker's speech length varies according to the pronunciation of the different phonemes. We solved this issue by using the Sound modification software tool "Audacity" and "Pydub" Python Audio segmentation library. Why Standardization is important because standardization bounds the range of voice length which helps in the generation of MFCCs and ensures that the data is in the same sample size. We make and set each voice at 8 seconds by appending the silence at the end of each phoneme voice.

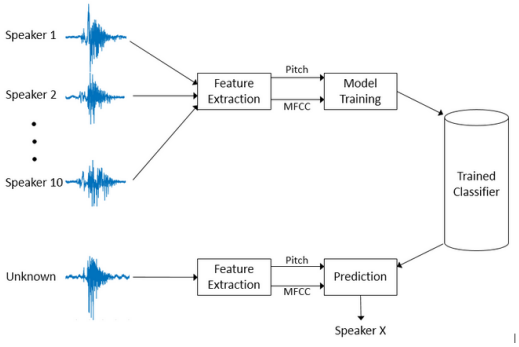


Fig. 5. Training diagram of MFFC model.

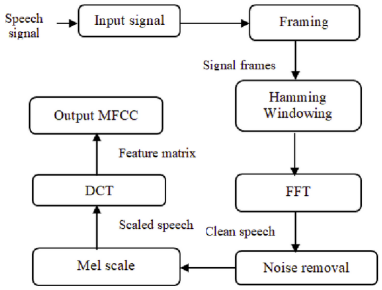


Fig. 6. MFCC extraction process diagram.

2) Augmentation of the data set: Augmentation is the set of techniques that are used to increase the amount of data by adding modified copies of already existing data. Sometimes, it creates new synthetic data from the existing data. Since the original data were not enough to get the model train properly, we augment the data by appending the real-time background noises on the original data set through the "Pydub" Python Audio segmentation library. Data augmentation acts as a regularizer and assists in managing the overfitting of data. Data augmentation increases the possibility of overfitting the model by generating additional training data and exposing the model to different versions of data.

B. Mel Frequency Cepstral Co-efficient Generation

The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. Mel Frequency Cepstral Coefficients (MFCC) give us a 2D representation of the audio signals with time on the x-axis and values on the y-axis. For our models, we used the Mel spectrogram with the following parameters.
The process to extract Mel Spectrogram.

- Extract STFT.
- Convert Amplitude to Dbs.
- Convert frequencies to Mel Scale

Set Parameters of Mel spectrogram

- Sampling rate (sr) = 22050
- Frequency Scale: MFCC
- Sampling time: 8s.
- Window Size: 180x180.

Once all MFCCs were generated, the dataset was split randomly into a training set (80%), and validation set (20%), and then our data is ready to go into the CNN model as input.

C. Convolution Neural Networks

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm that 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.
We researched a lot in this domain in order to solve our problem statement but we did not find any research techniques or methods. So, we decided to design and implement our own method. We made 29 classes one for each phoneme, each class containing 2 subclasses ["Negative" and "Positive"], and iterate through each class one by one on CNN models.

- **Pooling:**This method is essential as it not only ensures translation in-variance but greatly reduces the dimension of the feature map that is given as output by convolution.
- **Non-Linear Activation Function:**This adds enhanced feature extraction and retrieves more accurate predictions. ReLU function has been exercised in all models as it is quite reliable and avoids gradient vanishing.
- **Optimizers:**They minimize the loss function. For our models, we have used the Adam optimizer which is a default for most Neural Networks and gives optimum results.

VI. EVALUATION METRICS

The evaluation of the performance of the models was executed using the following two metrics based on Positive and Negative data which is split into train and validation set:
1. **Accuracy:** This metric evaluates how accurate the model's prediction is in comparison to the true values. This can also be taken as the percentage of validation data correctly classified.
2. **Loss:**It evaluates how well a model performs on the given data. In our models, we have used the Binary Cross Entropy loss function. Binary cross entropy compares each of the predicted probabilities to the actual class output which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value. That means how close or far from the actual value.

VII. EXPERIMENTS AND RESULTS

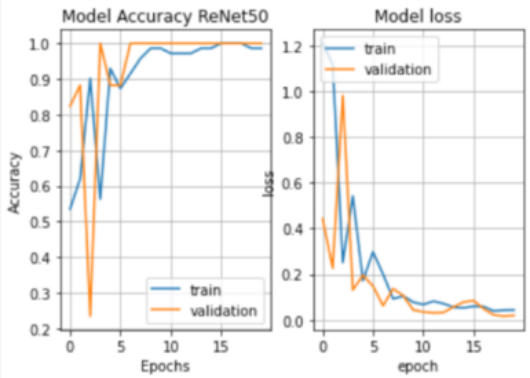
We trained and test six state of art CNN models on our data
1- ResNet-50
2- VGG-16
3- MobileNet V2

4- Xception
5- Inception V3
6- DenseNet 121

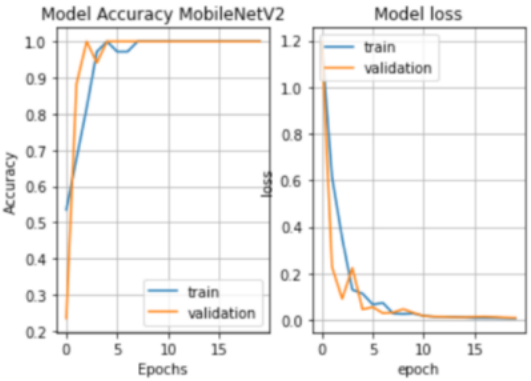
By narrowing our research down to the top three CNN models based on training accuracy. Then, these models were trained more rigorously with different parameters or included more noise and standardization in the data set and then evaluated again against the training accuracy. Initially, We tested our model on the phoneme ”Alif” explicitly and visualize the accuracy and loss graph between the training and validation data set.

A. MODELS EXPERIMENT ON WITH-OUT NOISE AND STANDARDIZATION

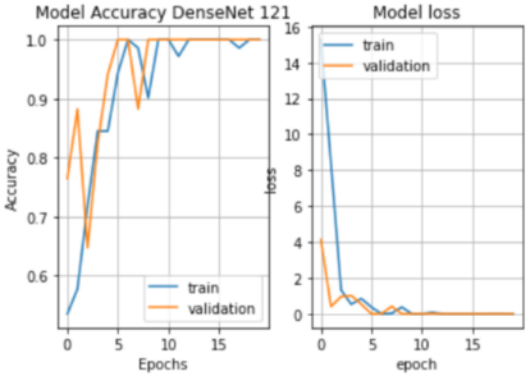
1) ResNet-50 Result:



2) MobileNet-V2 Result:

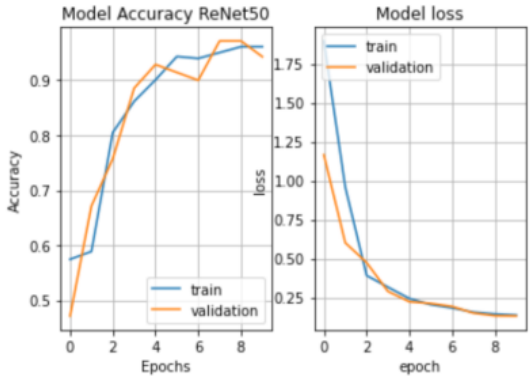


3) DenseNet-121 Result:

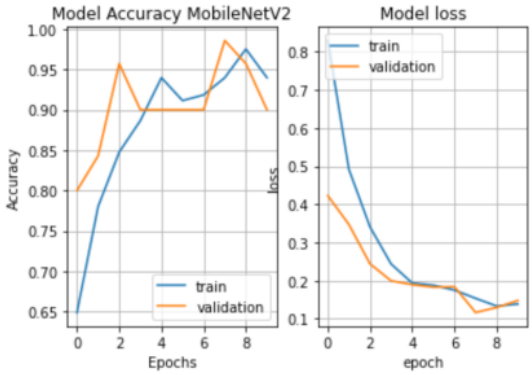


B. MODELS EXPERIMENT ON WITH NOISE AND STANDARDIZATION

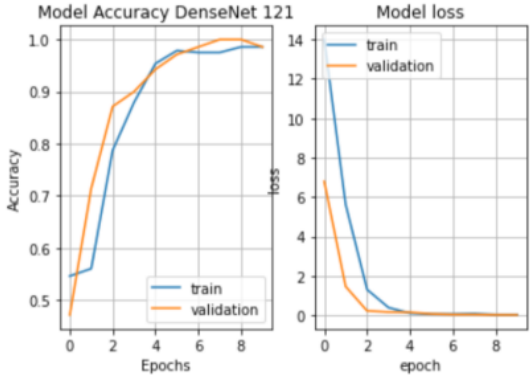
1) ResNet-50 Result:



2) MobileNet-V2 Result:



3) DenseNet-121 Result:



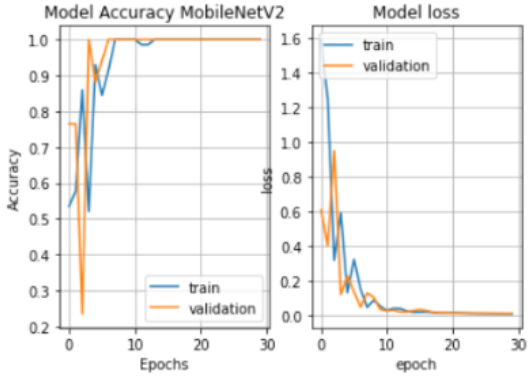
VIII. CONCLUSION

State of Art CNN Model Training Accuracy List

No.	*Models*	*Accuracy*
1	Resnet-50	72.33
2	MobileNet-V2	74.47
3	DenseNet-121	68.0
4	VGG-16	65.1
5	Inception-V3	67.3
6	Xception	52.34

We concluded that MobileNet-V2 came out with high accuracy on the given data set.

A. Improve More accuracy on MobileNet-V2 by increase number of epochs



B. Future Improvements

In the future we will expand our project to an application where a user comes, we will give him/her a sample voice taken from the positive dataset, user will listen to the voice and tries to speak like the given dataset voice. After that our application will give him the response of how accurately he learns. This response will be based on all the work we did on our model and the accuracy will be measured by our deep learning model. We have a limited dataset so the accuracy may vary.

REFERENCES

[1] Usmani, M., 2022. Mufti Muhammad Shafi‘ - The Grand Mufti of Pakistan. [online] Deoband.org. Available at: ¡https://www.deoband.org/2011/12/biographical-notes/shaykh-muhammad-shafi-the-mufti-of-pakistan/¿ [Accessed 17 October 2022].

[2] Ajibola Alim, S. and Khair Alang Rashid, N., 2022. Some Commonly Used Speech Feature Extraction Algorithms.

[3] M. J. Aqel and N. M. Zaitoun, “Tajweed: An Expert System for Holy Qur’an Recitation Proficiency,” in Procedia Computer Science, 2015, vol. 65, pp. 807–812. doi: 10.1016/j.procs.2015.09.029.

[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in ICLR, 2015.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition,” in ICASSP, 2016.

[6] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, “Deep Speech 2: End-to-End Speech Recogni- tion in English and Mandarin,” in ICML, 2016

[7] A. Graves, N. Jaitly, and A. Mohamed, “Hybrid Speech Recognition with Bidirectional LSTM,” in ASRU, 2013.

[8] O. Abdel-Hamid, L. Deng, and D. Yu, “Exploring Convo- lutional Neural Network Structures and Optimization Tech- niques for Speech Recognition,” in INTERSPEECH, 2013.

[9] Pankajray, “Convolutional Neural Network (CNN) and its Application- All u need to know,” Analytics Vidhya, Jan. 19, 2021. https://medium.com/analytics-vidhya/convolutional-neural-network-cnn-and-its-application-all-u-need-to-know-f29c1d51b3e5 (accessed Dec. 12, 2022).

[10] V. Choubey, “Text classification using CNN,” Medium, 22-Jul-2020. [Online]. Available: https://medium.com/voice-tech-podcast/text-classification-using-cnn-9ade8155dfb9. [Accessed: 12-Dec-2022].