INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND APPLICATIONS
VOLUME XVII, ISSUE VI, June 2023, WWW.IJCEA.COM, ISSN 2321-3469

DETECTING LIPS MOVEMENT AND PREDICTING PHRASES USING CNN

Shivanand Gadgi[1], Shubhangi Wartale[2] ,Manisha Mane[3],Prajakta Narole[4] and Vilas Ghonge[5]

Department of Information Technology, Savitribai Phule Pune University, India

ABSTRACT:

The audio-visual speech recognition system using lip movement uprooted from side- face images to attempt to increase noise- robustness in mobile surroundings. Although utmost former bimodal speech recognition styles use anterior face (lip) images, these styles aren't easy for druggies since they need to hold a device with a camera in front of their face when talking. Our proposed system landing lip movement using a small camera installed in a handset is more natural, easy and accessible. This system also effectively avoids a drop of signal- to- noise rate (SNR) of input speech. Visual features are uprooted by optic- inflow analysis and combined with audio features in the frame of CNN- grounded recognition.

Keywords: Convolutional Neural Network, Deep Learning, Image processing.

[1] INTRODUCTION

Speech plays an important parameter for communication, which is easy, simple, and everyone can speak without the help of any device and substantially the specialized skill set isn't demanded. The problem with the primitive interfacing bias is, some percent- age of introductory position of skill set is important necessary to use those interfaces. So it'll be delicate to interact with similar bias for people who are each not apprehensive of specialized skill set. As in this work, main attention is on speech recognition, any specialized skill set isn't needed so this will be helpful for the people to speak to the computers in given language rather than giving inputs from the other bias of the systems. Currently, common technological issues are with the computer

operation, similar as how effectively the commerce is there with the computers and how exactly stoner-friendly it's with lower conventional styles. It has come nearly mandatory of knowing the English literature to interact with the computers for penetrating the information technology. This restricts common people to stay out from the operation of the computers and other electronic bias. As there's a lot of enhancement in the information technology it's important necessary for common people to be in the lane of technological growth.

Besides this restriction, there will a most approachable system need to be constructed, similar as the bias which can read and take the input as the speech of the indigenous languages and respond to those indigenous effects for the stylish stoner-friendly system. This helps common people to make operation of similar technological growth. The aural noise in the terrain cannot lose the reciprocal features handed by the visual information. As the aural features are used for speech recognition are well understood. The major issues are the choice of visual features, emulsion model for the visual and audio data, along with a choice of the recognizer.

The most important conception behind the VSR (visual speech recognition) is the visual parameters. This won't be affected by any aural noise and disturbances in a noisy terrain. Visual speech is an intriguing content of exploration that has substantially used in intriguing fields like enhancing operations in mortal computer commerce, security, and digital entertainment. Therefore in proposed methodology, we're concentrating on only visual parameters to fete the speech. The mentioned data have motivated the experimenters carried out on particular VSR (visual speech recognition) that too with the AVSR (audio-visual speech recognition). This is known as automatic lip reading system for the visual speech recognition. In present days there are several automatic speech recognition styles proposed that combine both audio and visual features. For all similar type of systems, an important ideal of the visual speech recognizers is to ameliorate recognition delicacy, substantially under noisy environmental conditions. In this particular work main focus is on VSR (visual speech recognition) for Indian languages using lip parameters, the whole conception will be depending on the selection of input videotape with all light and environmental conditions by rooting the textbook affair. In order to achieve necessary parameters, numerous algorithms like canny edge discovery particularly for detecting the lips edge, GLCM(Gray Level Cooccurrence Matrix) and Gabor convolve for rooting the shape, texture features of lips. Eventually by applying CNN classifier according to point vector attained affair can be classified [7].

[2] LITERATURE REVIEW

In this paper [1], they proposed the Spatio-Temporal fusion module (STFM) and a convolutional sequence-to-sequence model based on the temporal focal block (TF-block) for lip

reading. Our STFM can be combined with most lip reading models to improve the utilization of local spatial information and the proposed TF-block can extract short-range temporal dependencies which are critical to lip reading. Our method achieves the state-of-the-art results on GRID and LRW datasets and comparable results with state-of-the-art approaches on LRS2-BBC and LRS3-TED datasets using much less training data and training time.

In this paper [2], in both datasets (sentence and word), it was observed that the ResNet-18 model achieved higher classification success than other models. According to table-10, in the word dataset, although ResNet-18′s training took 1 h and 15 min longer than GoogleNet's, ResNet-18 gave a better result by about 15%. The training time of the ResNet-18 model increased by 26% compared to the GoogleNet model. It is thought that the reason for such a difference in training times of ResNet-18 and GoogleNet models is the effect of 11.5 M parameters of ResNet-18. This number is 40% more than the number of parameters of GoogleNet. Similar results are also seen for the sentence dataset. ResNet-18′s training time is 7% longer than GoogleNet. However, the difference in SRR between ResNet-18 and GoogleNet is 16%. Due to the size of the dataset, processing is difficult and takes a lot of time. For this reason, increases in running times have been observed depending on the data size, especially during the training phase of the methods. The training time of the 40-words and 3-subjects (persons) dataset of the ResNet-18 model increased by 400%, from 64 min to 352 min when it was training with 111 words and 18 subjects. In addition to these, the runtime graphs gave similar results in proportion to the size of the data contained in both datasets. These results showed us that the running time of the methods mostly depends on the size of the data. A pre-trained CNN model is preferred. The reason for this is to check and compare the status of the dataset and analyze the success performance of other models and datasets.

In this paper[3], they presented an overview of deep-learning based approaches for audio-visual speech enhancement (AVSE) and audio-visual speech separation (AV-SS). As expected, 22 visual information provides a benefit for both speech enhancement and separation. In particular, AV-SE systems either outperform their audio-only counterpart for very low signalto-noise ratios (SNRs) or show similar performance at high SNRs. Performance improvements can be seen across all visemes, with better results for sounds easier to be distinguished visually [12]. Regarding speech separation, audiovisual systems not only outperform their audio-only counterpart, but, since vision is a strong guidance, they are also unaffected by the source permutation problem, occurring when the separated speech signals are assigned inconsistently to the sources. Throughout the paper, they surveyed a large number of approaches, deliberately avoiding to advocate a method over another based on their performance.

In this paper[4], The proposed structure and essential steps are addressed in depth in the four parts that follow. The dynamic lip videos must first be preprocessed, which includes separating audio and video signals, extracting keyframes, and positioning the mouth. Second, CNN is used to extract features from the preprocessed picture dataset. Then, to learn sequence information and attention weights, we combine LSTM with an attention mechanism. Finally, the ten-dimensional characteristics are mapped using two completely linked layers, with the SoftMax layer predicting the result of automatic lip-reading recognition. SoftMax normalizes and categorizes the output of fully linked layers based on likelihood.

In this paper[5], they present an end-to-end visual speech recognition system suitable for small-scale datasets which jointly learns to extract features directly from the pixels and perform classification using LSTM networks. Results on four datasets, OuluVS2, CUAVE, AVLetters and AVLetters2, demonstrate that the proposed model achieves state-of-the-art performance on all of them significantly outperforming all other approaches reported in the literature, even CNNs pre-trained on external databases.

In this paper[6], they have investigated the sensor technologies and the integration of machine learning and deep learning techniques in the field of spoken communication for voice recognition and production. They first introduced methods and techniques to acquire biosignals from muscle activity, brain activity, and articulatory activity as well as their applications in voice recognition and production. It is important for voice recognition technologies to be of high quality and to enable people to express themselves more accurately. In order to overcome the limitations of voice recognition, there have been invented various mouth interface technologies for voice recognition and production with various traditional sensors like EMG, EOG, EEG, EPG, gyro, image, and ultrasonic.

In this paper[7], they proposed an AVSR model based on the transformer with the DCM attention and a hybrid CTC/attention architecture. We constructed the DCM attention for proper alignment information between audio and visual modality even with noisy reverberant audio data, and applied a hybrid CTC/attention structure to enhance monotonic alignments. In general, our model provided better recognition performance than the compared models based on the transformer, even for out-of-sync data, and the hybrid CTC/attention loss further improved the performance. In the future, they will focus on more efficient fusion strategy of audio and video information and extend to audio–visual speech recognition including a speech enhancement model.

[3] PROPOSED METHODOLOGY

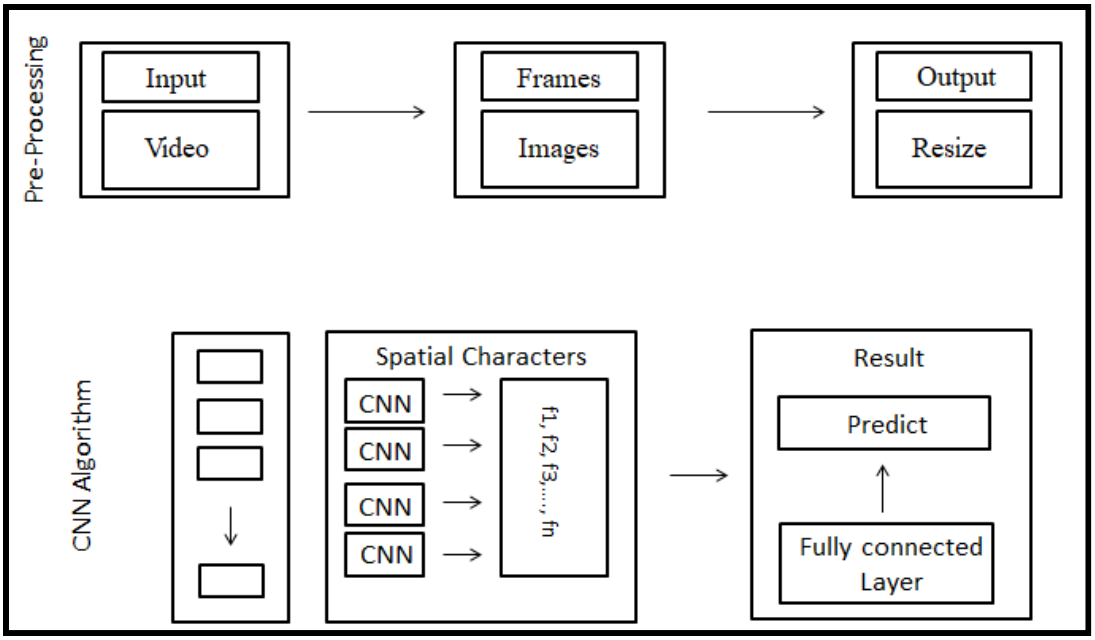

Figure: 1. System Architecture.

CNN or the convolutional neural network (CNN) could be a order of deep literacy neural networks. in brief consider CNN as a machine learning formula which will absorb AN input image, assign significance(learnable weights and impulses) to varied aspects objects within the image, and be able o separate one from the opposite. CNN works by rooting options from the film land.

LAYERS OF CNN:

By mounding multiple and fully different layers during a CNN, advanced infrastructures area unit designed for bracket issues. Different kinds of layers area unit most common complication layers, pooling subsampling layers, non-linear layers, and absolutely connected layers.

A. Convolutional Layer:

The general ideal of the complication operation is to prize high- position options from the image. We're suitable to always add over one complication subcaste once erecting the neural network, wherever the primary Convolution Layer is responsible for landing slants whereas the alternate subcaste captures the peripheries. The addition of layers de- pends on the quality of the image thus there aren't any magic figures on what number layers to point.

B. Pooling/Subsampling Layer:

The pooling subcaste applies anon-linear down- slice on the convolved point of- ten appertained to as the activation maps. This is substantially to reduce the computational complexity needed to reuse the huge volume of data linked to an image. Pooling isn't mandatory and is frequently avoided. Generally, there are two types of pooling, Max Pooling that returns the maximum value from the portion of the image covered by the Pooling Kernel and the Average Pooling that pars the values covered by a Pooling Kernel. Figure below provides a working illustration of how different pooling ways work.

C. Fully connected layers:

These layers mathematically total a weight of the former subcaste of options, indicating the precise admixture of "constituents" to see a named target affair result. Just in case of a completely connected subcaste, all the options of the former subcaste get employed in the computation of every element of every affair point.

[4] PROPOSED FRAMEWORK

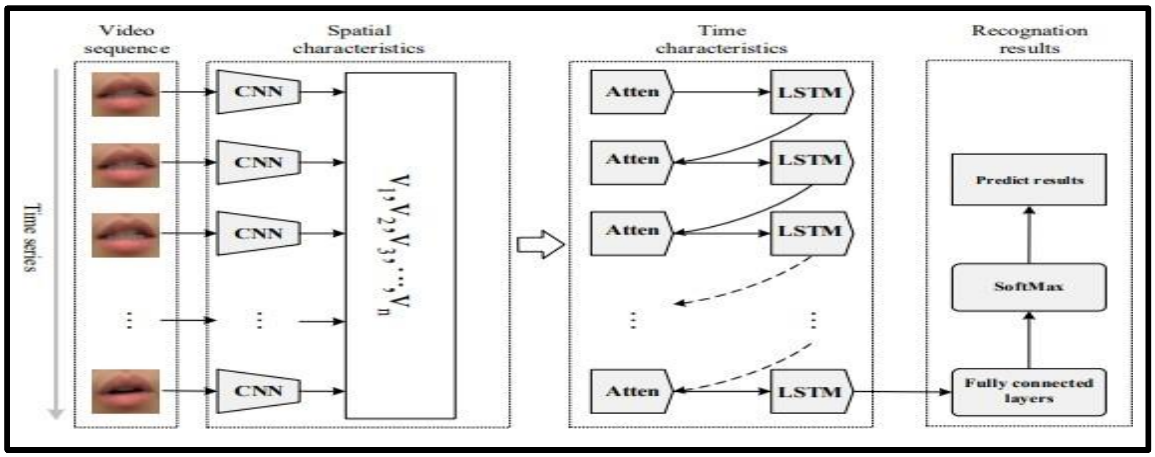

Figure: 2. Proposed Framework

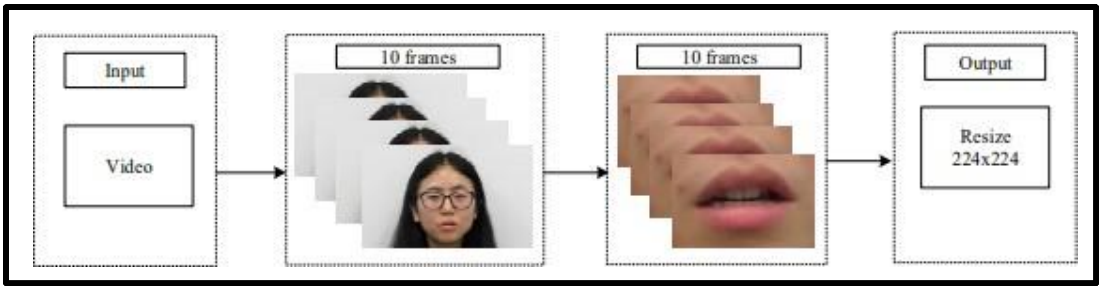

Figure: 3. System Steps

The proposed framework and main steps are discussed in detail according to the following four parts. Firstly, we need to preprocess the dynamic lip videos, including separating audio and video signals, extracting keyframes and positioning the mouth. Secondly, features are extracted from the preprocessed image dataset by using CNN. Then, we use LSTM with attention mechanism to learn sequence information and attention weights. Finally, the ten-dimensional features are mapped through two fully connected layers, and the result of automatic lip-reading recognition is predicted by SoftMax layer. SoftMax normalizes the output of the fully connected layers and classifies it according to probability.

[5] TEST STRATEGY

The test strategy consists of a series of different tests that will fully exercise the system. The primary purpose of the test is to uncover the system limitations. Following are the tests results:

Sr.No	Description	Test Case I/P	Actual Result	Expected	Test Criteria (P/F)
1	Install Python	Python Exe	Should get install properly	Proper Installed	P
2	Installing Libraries	Library command for install	Should Get installed	Libarary Installed Success-fully	P
3	Training Dataset	Dataset Training	Error in Training Model	Trained Model	F
4	Training Dataset	Dataset Training	Trained Model	Trained Model	P
5	Login Cre-dentials	User Name and Pass-word	Login Un-successful	Unsuccessful Login	F
6	Login Cre-dentials	User Name and Pass-word	Login Suc-cessful	Successful Login	P
7	Password	Current and New Pass-word	Password Updated	Update Password	P
8	Select Lim-ited Dataset	Number of rows	Should select and train the selected data	Trained Model	P
9	Prediction	Video as in-put	Should Pre-dict the re-sult	Result Pre-dicted	P

Figure: 4. Test Results

[6] PROJECT DESCRIPTION

Phase	Task	Description
Phase 1	Analysis	Analyse the information given in the IEEE paper.
Phase 2	Literature survey	Collect raw data and elaborate on literature surveys.
Phase 3	Design	Assign the module and design the process flow control.
Phase 4	Implementation	Implement the code for all the modules and integrate all the modules.
Phase 5	Testing	Test the code and overall process weather the process works properly.
Phase 6	Documentation	Prepare the document for this project with conclusion and future enhancement.

Fig.5. Task Description

[7] RESULTS AND DISCUSSIONS

A. Dataset and Description

We have created Lip Reading dataset of Indian English accent of short videos. It can be used in various fields of research including visual speech recognition, face detection, biometrics etc. Four speakers (1 male and 3 female) were recorded using Desktop Camera placed 1.5 feet away from the speaker and at exact height as if the speakers face where they were asked to utter a set of ten phrases. The dataset contains a total of 60 sequential images captured at 30fps. It is dataset created by the authors of the paper designed specifically for the use in visual speech recognition.

ID	Phrases
0	Could you repeat that please?
1	Thank you so much.
2	I don't understand.
3	What's your phone number?
4	Nice to meet you.
5	What do you do?
6	How can I help?
7	Excuse Me.
8	I am sorry.

Fig.6. Task Description

B. Pre-Processing:

Dataset contains videos for every author speaking every expression. Similar videos which are recorded at 30 fps are of large size and the pre-processing begins with conversion of the vids to frame s size. For every 3- 5 alternate video tape we get 90 150 frames saved in same successional image structure. Similar preprocessed datasets are now generally available for free to Use like the MIRACL- VC2 dataset. Once the frames are attained, every frame is cropped to prize the ROI(Region of interest) i.e. the mouth region which is done by the shape predictor 68 face corner which is a function available in the Dlib library.

C. Train/Test Split

In statistics and ML, we typically resolve our data into two subsets training data and testing data, and fit our model on the train data, so as to make vaticinations on the test data. At the point when we do that, one of two effects may do we over fit our model or we under fit our model. Overfitting implies that the model we prepared has trained" exorbitantly well" and is presently, well, fit also near the training dataset. As opposed to overfitting, when a model is Under fitting, it implies that the model does not fit the training data and, in this way, misses the patterns in the information.

D. Model building:

Now we continue with the creation of our CNN architecture. We've kept pool size and kernal_size = 2 and 3. We're setting time cycles as 5. We've also used different activation functions for different layers. For retired layers, we will be using the activation function named “RELU”. Also, for affair subcaste, we will be using the activation function named “SOFTMAX”. We use python import to keep a check val_acc during the time process. However, val_acc is dropping or dwindling its value, the we come to know that the literacy rate is getting changed, if for the two nonstop ages. Now we will fit our model and save the model as an". h5” extension.

[8] COMPARISONS

Models	Dataset	Accuracy
Existing Models		
LST M-5	LRW	66
LST M-5	Custom	25.76
D3D	LRW	78
D3D	Custom	34.76
3D+2D	LRW	83
3D+2D	Custom	38.19
Our Model		
VSR	Custom	73.0586836

Fig.7. Numerical Comparison

[9] CONCLUSION

Utmost recent workshop suggest that the optimal modeling of temporal sequences is still an open problem, which is presently been dived by means of intermittent neural net- workshop. Specifically, CNN have been extensively used for modeling sequences because of their capability to retain both short- and long- term environment information in their cell structures, although it isn't clear how to take full advantage of similar capability. For case, several authors have tried to model different scales of environment by adding multiple CNN layers, aiming to introduce constraints related to bigger speech structures similar as connected phonemes, syllables, words or rulings.

[10] REFERENCES

[1] Zhang, Xingxuan, Feng Cheng, and Shilin Wang. ”Spatio-temporal fusion based convolutional sequence learning for lip reading.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] Kurniawan, Adriana, and Suyanto Suyanto. ”Syllable-Based Indonesian Lip Reading Model.” 2020 8th International Conference on Information and Com- munication Technology (ICoICT). IEEE, 2020.

[3] Michelsanti, Daniel, et al. ”An overview of deep-learning-based audio-visual speech enhancement and separation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021).

[4] Desai, Dhairya, et al. ”Visual Speech Recognition.” International Journal of Engineering Research Technology (IJERT) 9.04 (2020).

[5] Petridis, Stavros, et al. ”End-to-end visual speech recognition for small-scale datasets.” Pattern Recognition Letters 131 (2020): 421-427.

[6] Lee, Wookey, et al. ”Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review.” Sensors 21.4 (2021): 1399.

[7] Lee, Yong-Hyeok, et al. ”Audio–visual speech recognition based on dual crossmodality attentions with the transformer model.” Applied Sciences 10.20 (2020): 7263.

[8] Alex M. Goh and Xiaoyu L. Yann, (2021), “Food-image Classification Using Neural Network Model” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 12-22, DOI 10.30696/IJEEA.IX.III.2021.12-22

[9] Jeevan Kumar, Rajesh Kumar Tiwari and Vijay Pandey, (2021), “Blood Sugar Detection Using Different Machine Learning Techniques” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 23-33, DOI 10.30696/IJEEA.IX.III.2021.23-33

[10] Nisarg Gupta, Prachi Deshpande, Jefferson Diaz, Siddharth Jangam, and Archana Shirke, (2021), “ F-alert: Early Fire Detection Using Machine Learning Techniques” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 34-43, DOI 10.30696/IJEEA.IX.III.2021.34-43

[11] Reeta Kumari, Dr. Ashish Kumar Sinha and Dr. Mahua Banerjee, (2021), “A Comparative Study Of Software Product Lines And Dynamic Software Product Lines” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 01-10, DOI 10.30696/IJEEA.IX.I.2021.01-10

[12] MING AI and HAIQING LIU, (2021), “Privacy-preserving Of Electricity Data Based On Group Signature And Homomorphic Encryption” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 11-20, DOI 10.30696/IJEEA.IX.I.2021.11-20

[13] Osman Goni, (2021), “Implementation of Local Area Network (lan) And Build A Secure Lan System For Atomic Energy Research Establishment (AERE)” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 21-33, DOI 10.30696/IJEEA.IX.I.2021.21-33.

[14] XIAOYU YANG, (2021), “Power Grid Fault Prediction Method Based On Feature Selection And Classification Algorithm” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 34-44, DOI 10.30696/IJEEA.IX.I.2021.34-44.

[15] Xiong LIU and Haiqing LIU, (2021), “Data Publication Based On Differential Privacy In V2G Network” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 34-44, DOI 10.30696/IJEEA.IX.I.2021.45-53.

[16] Mandava Siva Sai Vighnesh, MD Shakir Alam and Vinitha.S, (2021), “Leaf Diseases Detection and Medication” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 01-07, doi 10.30696/IJEEA.IX.I.2021.01-07

[17] Pradeep M, Ragul K and Varalakshmi K,(2021), “Voice and Gesture Based Home Automation System” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 08-18, doi 10.30696/IJEEA.IX.I.2021.08-18

[18] Jagan K, Parthiban E Manikandan B,(2021), “Engrossment of Streaming Data with Agglomeration of Data in Ant Colony” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 19-27, doi 10.30696/IJEEA.IX.I.2021.19-27

[19] M. Khadar, V. Ranjith, K Varalakshmi (2021), “Iot Integrated Forest Fire Detection and Prediction using NodeMCU” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 28—35, doi 10.30696/IJEEA.IX.I.2021.28-35

[20] Gayathri. M, Poorviga. A and Mr. Vasantha Raja S.S, (2021), “Prediction Of Breast Cancer Stages Using Machine Learning” Int. J. of Electronics Engineering and Applications, Vol. 7, No. 1, pp. 36-42, doi 10.30696/IJEEA.IX.I.2021.36-42