International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

# Facial Emotion Recognition System using Deep Learning and Convolutional Neural Networks

Hanusha T[1], Dr. Varalatchoumy[2]

[1] PG-Scholar, Department of Computer Science & Engineering, Cambridge Institute of Technology, Bangalore
[2] Associate Professor, Department of Computer Science & Engineering, Cambridge Institute of Technology, Bangalore

***Abstract*-- In the field of education online learning plays a vital role. For effective and quality learning engagement of Listener is the key point. The fundamental problem facing in the online learning environment is the low engagement of Listener to the Preceptor. The educational institutions and Preceptors are responsible to guarantee best learning environment with maximum engagement in educational activities for online learners. An efficient system for detecting Listener's engagement helps Preceptor and educational institutions to monitor their classes and Listener's behavior in real time. There is lot of different methods to detect engagement in online class. In this paper conducting a survey on various automatic engagement detection methods. Classify the existing methods into major five categories based on their techniques used for engagement detection, Through - Mouse Click, ECG & Body & Head movement, Eye movement, Facial Expression. To overcome the flow the existing system, the proposed System based on CNN deep learning model for the engagement detection in Listener's based on the survey proceedings.**

## INTRODUCTION

In the Digital era, digitization in educational field makes lot of advancement in the education system. In this 21st century digitization focus on delivering high quality education to all Listeners via online. They can study from anywhere in the world without any fail. But there are lot of challenges facing by instructors and Preceptors during online class. The relationship between Listener and Preceptor is the main component of meaningful education. One of the challenges is that the tutor/instructor is not knowing about how well the Listeners receiving information from the lecture. When a Listener properly involves in a learning environment, the content delivered from Preceptor is meaningfully received by the Listener and this is called as Listener engagement in learning.

A psychological investment of Listener in understanding some skill sets or some academic works can be described as Listener engagement. When the engagement of Listener increases at the same time Listeners achievement also increases, they we can say that engagement is directly proportional to Listener's achievements. For developing better interactive applications the measurement of Listener (user) engagement is required.

Importance of engagement of Listeners in education is first discussed by Newmann, he said about the effort taken by Listener to understand the skills and academic works through online. When we go deeper in to Listener engagement then we realized it is a complicated phenomenon and it has different dimensions. Because of this complexity currently there are some different engagement measurement techniques are available. Those methods can be grouped in to two as manual and automatic measurement systems. Manual method includes self-reports and Preceptor checklists, this is a primitive method and because of popularity still it is used as one way to measure Listener engagement. It is a cost effective and easily implemented method for engagement measurement. There is lot of limitations for manual methods; one of the limitations is the time consumption. The Preceptor need to ask some questions and wait for replay from each Listener. In some cases, Listeners will not answer honestly and chances of malpractices may increase. This will affect proper measurement of engagement in real time. Automatic methods overcome the limitations of manual methods. It includes different techniques like video based and sensor based. It will increase the accuracy of measuring data and decrease the time consumption for data collection. This will help to improve the quality of personalized learning environment.

In the starting of 1990s the virtual classroom concept introduced to the world. This idea was widely accepted by many educational institutions for offering better infrastructure for their education system. The World Wide Web (www) was also become popular in the same time and with the collaboration, it offers huge contents for learning . Then there was a rapid growth in the field of online learning. One of the main problems arise in the online learning is the lack of direct supervision by the Preceptor/instructor. Another problem is the interest of Listener to certain topics, they will have low engagement for not interesting topics and it will affect Listeners academic performance and drop-out rate increases. Therefore, it is needed to maintain engagement during learning process and an automatic efficient system required for Listener engagement detection.

## ONLINE LEARNING:

While countries are at different points in their COVID-19 infection rates, worldwide there are currently more than 1.2 billion children in 186 countries affected by school closures due to the pandemic. In Denmark, children up to the age of 11 are returning to nurseries and schools after initially closing, but in South Korea students are responding to roll calls from their teachers online. With this sudden shift away from the classroom in many parts of the globe, some are wondering whether the adoption of online learning will continue to persist post-pandemic, and how such a shift would impact the worldwide education market.

Even before COVID-19, there was already high growth and adoption in education technology, with global EdTech investments reaching US$18.66 billion in 2019 and the overall market for online education projected to reach $350 Billion by 2025. Whether it is language apps, virtual tutoring, video conferencing tools, or online learning software, there has been a significant surge in usage since the COVID-19 pandemic.

## CHALLENGES FACED:

There are, however, challenges to overcome. Some students without reliable internet access and/or technology struggle to participate in digital learning; this gap is seen across countries and between income brackets within countries. For example, whilst 95% of students in Switzerland, Norway, and Austria have a computer to use for their schoolwork, only 34% in Indonesia do, according to OECD data.

In the US, there is a significant gap between those from privileged and disadvantaged backgrounds: whilst virtually all 15-year-olds from a privileged background said they had a computer to work on, nearly 25% of those from disadvantaged backgrounds did not. While some schools and governments have been providing digital equipment to students in need, such as in Australia, many are still concerned that the pandemic will widen the digital divide.

## RISK FACTORS:

There are specific listener retention strategies around all of these risk factors, and this one in particular requires much-needed support for students who are trying to integrate coursework with family time, full-time jobs and other commitments. These listeners were trying to carve out time for many aspects of their already hectic lives before they added online courses to the mix. One of the major obstacles for online listener as they navigate new learning management systems, wikis and software they are either unfamiliar with or that malfunction. Helping listeners through their technological challenges will allow them to take advantage of all the ways they can engage with their classmates and online community. This is especially important for distance learners and has shown to aid in student retention.

Often times, it’s been a while since these listeners have been in a physical learning environment. They may have struggled academically in the past, these swapped multiple times without graduating or be nervous to try a new mode of education. These are just a few of the academic challenges for adult learners. It’s important for online listeners to be held accountable as online classes offer flexibility, but include deadlines and other requirements that can creep up on them.

## EARLY INVENTIONS:

In 1924, the first testing machine was invented. This device allowed students to test themselves. Then, in 1954, BF Skinner, a Harvard Professor, invented the “teaching machine”, which enabled schools to administer programmed instruction to their students. It wasn’t until 1960 however that the first computer-based training program was introduced to the world. This computer-based training program (or CBT program) was known as PLATO-Programmed Logic for Automated Teaching Operations. It was originally designed for students attending the University of Illinois, but ended up being used in schools throughout the area.

The first online learning systems were really only set up to deliver information to students but as we entered the 70s online learning started to become more interactive. In Britain, the Open University was keen to take advantage of e-learning. Their system of education has always been primarily focused on learning at a distance. In the past, course materials were delivered by post and correspondence with tutors was via mail. With the internet, the Open University began to offer a wider range of interactive educational experiences as well as faster correspondence with students via email etc.

## CONVOLUTIONAL NEURAL NETWORKS:

The convolutional neural network, or CNN for short, is a specialized type of neural network model designed for working with two-dimensional image data, although they can be used with one-dimensional and three-dimensional data. Central to the convolutional neural network is the convolutional layer that gives the network its name. This layer performs an operation called a “convolution “. In the context of a convolutional neural network, a convolution is a linear operation that involves the multiplication of a set of weights with the input, much like a traditional neural network. Given that the technique was designed for two-dimensional input, the multiplication is performed between an array of input data and a two-dimensional array of weights, called a filter or a kernel.

The filter is smaller than the input data and the type of multiplication applied between a filter- sized patch of the input and the filter is a dot product. A dot product is the element-wise multiplication between the filter-sized patch of the input and filter, which is then summed, always resulting in a single value. Because it results in a single value, the operation is often referred to as the “scalar product “. Using a filter smaller than the input is intentional as it allows the same filter (set of weights) to be multiplied by the input array multiple times at different points on the input. Specifically, the filter is applied systematically to each overlapping part or filter-sized patch of the input data, left to right, top to bottom.

This systematic application of the same filter across an image is a powerful idea. If the filter is designed to detect a specific type of feature in the input, then the application of that filter systematically across the entire input image allows the filter an opportunity to discover that feature anywhere in the image. This capability is commonly referred to as translation invariance.

## MULTIPLE CHANNELS:

Color images have multiple channels, typically one for each color channel, such as red, green, and blue. From a data perspective, that means that a single image provided as input to the model is, in fact, three images. A filter must always have the same number of channels as the input, often referred to as "depth ". If an input image has 3 channels (e.g., a depth of 3), then a filter applied to that image must also have 3 channels (e.g., a depth of 3). In this case, a 3×3 filter would in fact be 3x3x3 or [3, 3, 3] for rows, columns, and depth. Regardless of the depth of the input and depth of the filter, the filter is applied to the input using a dot product operation which results in a single value.

This means that if a convolutional layer has 32 filters, these 32 filters are not just two- dimensional for the two-dimensional image input, but are also three-dimensional, having specific filter weights for each of the three channels. Yet, each filter results in a single feature map. Which means that the depth of the output of applying the convolutional layer with 32 filters is 32 for the 32 feature maps created.

## MULTIPLE LAYERS:

Convolutional layers are not only applied to input data, e.g., raw pixel values, but they can also be applied to the output of other layers. The stacking of convolutional layers allows a hierarchical decomposition of the input. Consider that the filters that operate directly on the raw pixel values will learn to extract low-level features, such as lines. The filters that operate on the output of the first line layers may extract features that are combinations of lower-level features, such as features that comprise multiple lines to express shapes. This process continues until very deep layers are extracting faces, animals, houses, and so on.

## LITERATURE SURVEY

Mushtaq Hussain et al [1], proposed an engagement detection system for virtual learning environment. They developed the system for identifying low engaged students in a course conducted by open university. The uses number of clicks as input from the Listener on the virtual learning environment application. The time spent by a Listener on each activity is calculated by the number of clicks per activity. Zhaoli Zhang et al [2], uses mouse cursor location and time-stamp in the learning web page to collect data of Listener engagement. According to mouse movement by Listeners captured by the system and labeled with the time- stamp for increasing accuracy of collected data. Feature extraction is done using the collected data and it predicts weather the Listener is engaged or not engaged. Jia Yue et al [3], proposed a method which mouse cursor position, wheel movement, click event as the data for detecting engagement. The x and y axis of mouse cursor coordinates and time-stamped records of mouse click events are also considered for better accuracy.

Michail N et al [4], proposed a wristband model system which has an EEG cap. ENOBIO EOG correction mechanism is used for calibrating data. The user wore the EEG cap and the concentration and attention level while learning is measured. Mohamed El Kerdawy et al [5], uses 14 channel EEG headset to record EEG signals. Two more channels are there in the headset as reference point. Each channel points are fixed around different locations of head. Engagement scores according to user's interest classifies the signal as engaged or not engaged model. Hamed Monkaresi et al [6], introduces a video-based heart rate detection for detect user's engagement. High-accurate heart rate sensing mechanism is used for this system and the data is saved for different conditions of user. Finally, machine learning technique used to validate model.

Lakshmi Priya et al [7], uses head movement for detecting concentration level of leaner. It also has a feedback mechanism to the enhance e-learning and for better content delivery. Head movement is decoded from the video captured using web cam, it avoids physical contact with leaner for data collection. There system can decode learners' interest and involvement in topics which helps the instructor to improve teaching methods and update to interesting topics. Timothy Patterson et al [8], proposed a sensor-based system for prediction. This system detects human activity using accelerometer sensor. They use smart phone and android application for data collection. Athanasios Psaltis et al [9], conducted an experiment with 72 students to play serious game applications to detect engagement level. They stored different engagement levels of players for machine learning model creation. Prabin Sharma et al [10], system also had head movement tracker. From the captured video the Listener head movement is detected and further used for engagement detection. K. Keerthana et al [11], also used head movement detection with facial emotion detection for increasing accuracy of their system.

Jia Yue et al [3], used eye movement for predicting learner's behavior. According to the study, eye movement pattern for different cases like watching video, reading and writing notes are differ from each other. The horizontal and vertical movement of eye tracked from the video stream using camera. Using this the duration of Listener focused on a particular screen is recorded. They used random forest classifier for model creation and prediction. Prabin Sharma et al [10], proposed a method to utilize built-in web cam in laptop for capturing video of Listener and decode eye movement from it. Form the acquired data concentration level of Listeners identified and it helps to find weather the Listeners engaged in the class or not. Work by Meredith Carrol et al [12], proposes a experimental approach of with eye tracking and without eye tracking. With eye tracking method gives more accuracy when they tested their model in real time.

Mohamed El Kerdawy et al [5], records the face while conducting online classes and tests. Then they convert the recorded video to image frames and face feature are extracted. Several features like nose to jaw length, nose to chin length, mouth and eye aspect ratio are used for detecting emotions of student. They developed different machine learning models for comparing it for better model. K. Keerthana [11], proposed a system with extract facial features and detect emotions. They used CNN algorithm in machine learning for engagement detection. They predicted different emotions like sad, happy, bored, angry etc. using facial data and it helps to predict involvement of Listener in the class. Yifeng Zhao et al [13], uses traditional deep learning method for engagement detection. They predict emotions like fear, sad, happy, neutral etc with the huge JAFFE dataset. They introduced a capsule network model for prediction. Awais Mahmood et al [14], divides face image into different blocks for better representation and feature extraction. It decreases time complexity of real-time implementation.

Veena Mayya et al [15], proposed a method to detect facial expression from single image. They detected six major expressions like anger, sadness, surprise, fear and disgust. They also determined age and gender with the model, because deep convolution neural network is used for model creation. Vladimir Soloviev et al [16], proposed a facial expression based engagement detection. They used a boosted decision tree classifier algorithm. Their system is capable to measure engagement of group of students not only individual. They also measure engagement level of each individual. In this section, the shortcomings of the approaches considered in this survey of engagement detection are deliberated. Mushtaq Hussain et al [1], says that performance of their system decreases according to the quality of acquired data. They are not considering total number of clicks for each test/class. Zhaoli Zhang et al [2], describes about the issue with short content pages. For those page clicks, scroll and mouse movements are very less and prediction of engagement is more difficult. Jia Yue et al [3], research not considering online learning with mobile phone and eye their method cannot be implemented easily on all students' computers. Prabin Sharma et al [10], says that the body movement of Listener will affect the eye and head movement tracking because when Listener moves forward or backward changes the focus of camera.

Michail N. Giannakos et al [4], proposed a costly system. Cost of EEG module is very high and real time implementation is not technically feasible. They system shows more than 20% error rate. Hamed Monkaresi [6], says that the data collection is a difficult task. Due to head movement, they could not extract features properly and it affects the performance of their system. Another problem faced by them is the subject movement, the signal will distract when the subject moves. Krithika.L.Ba et al [7], says that some of students sit more close to their camera and the video taken will be unfocused or blurred. Because of this Listener detection is not possible and it will decrease the accuracy of the system. According to Timothy Patterson et al [8], approach the data can be collected only using smart phones. Because it is not feasible in real-time scenario. Athanasios Psaltis et al [9], says they can only achieve less accuracy and need to include other technologies to improve accuracy. Meredith Carrol [12], while testing with different situations they noticed that while body moves eye tracking is not possible. And they say that combination of eye movement and body movement detection is required for better accuracy.

## PROPOSED METHODOLOGY

The idea was to build a single consolidated system that is able to effectively recognise the emotions of learners during online form of education with the help of convolutional neural networks and plot emotion metrics based on the results. Below Fig 1, shows the overall system design.

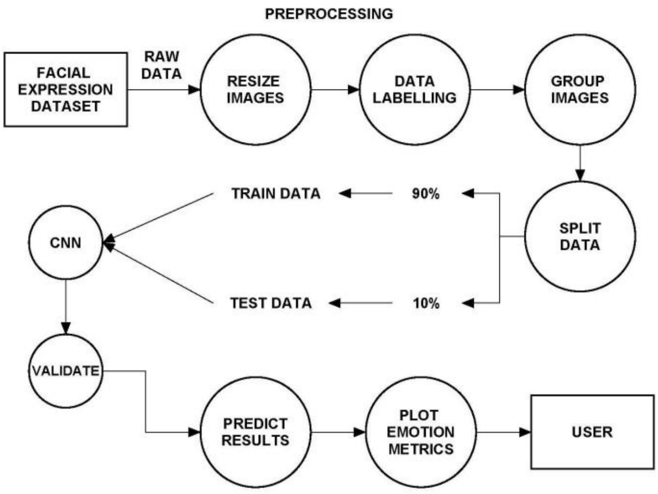

Figure 1: System Architecture

### Image Acquisition and Data Pre-processing:

Acquiring image data is where deep learning plays an essential that needs to work on the image being acquired and to work on the models as well. While acquiring datasets, some aspects need to be followed that can help for prediction, to train correctly on learning the convolutional neural network model. Generally, more data in the dataset would achieve a better result. The data collected from Kaggle website is totally un-processed and raw. In Fig 2, an example has been shown for the angry class in the dataset.

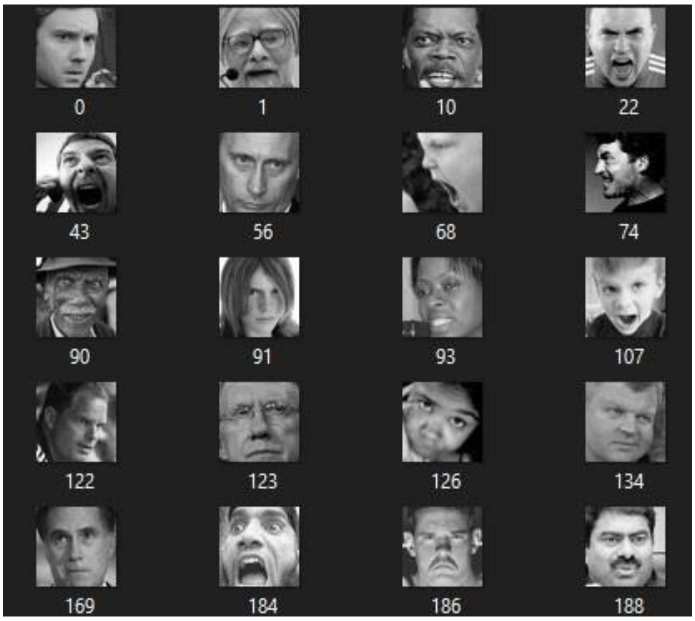

Figure 2: Dataset (class: angry)

The following are the steps involved in processing the raw data. The huge amount of image data collected from the facial expressing dataset is then passed through data pre-processing, where we convert this raw data, and clean it to build a dataset that is trainable. Here, images are then grouped based on their expression names, such as happy, sad, angry, surprise and neutral. All these emotions are different individual classes in the facial expression dataset.

**Feature Selection and Preparation of Data:**

Feature Engineering is defined as the process of acquiring and using the domain knowledge of data to create some features which the machine learning uses to work with the algorithms. If the feature work engineering is done accurately and correctly, then it tends to increase the predictive power CNN in deep learning to facilitate the DL process. Also, feature engineering is what which creates a vast difference between good DL model and bad one. It is the processing of raw data transformation to essential features which could better represent the data and build the predictive models that could result in improved accuracy model on the unseen data. In this module, the most prominent features of the images are chosen and confined. These features are then used to represent a particular class in the dataset. The process of feature selection needs to be done meticulously as the quality of the features are directly dependent on the prediction accuracy of the CNN.

**CNN Construction and Training:**

Training set is the subset of the training a model and the Test set is the subset to test on the trained model. CNN construction is the process of custom designing a neural network architecture that can facilitate this particular use-case. In this way we can give the CNN the training data as the input which should contain the correct answer, known as the target attribute. Fig 3 shows the construction of MobileNet.

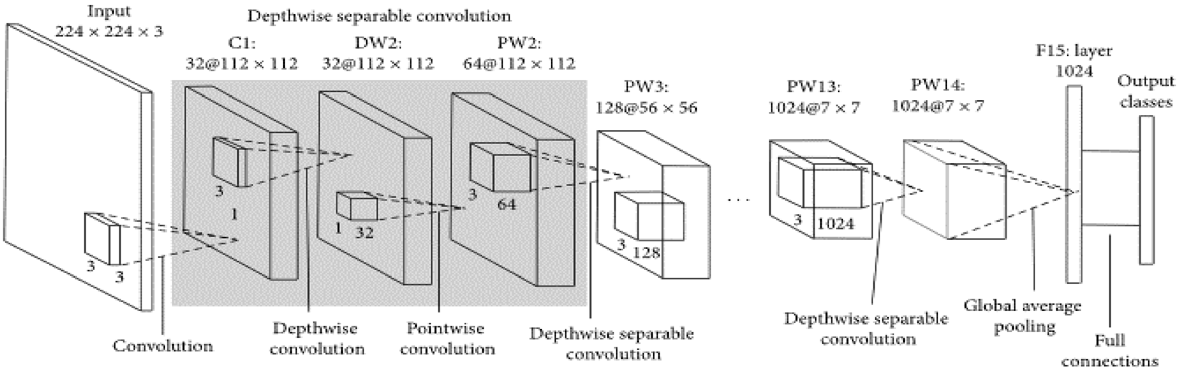

Figure 3: MobileNet Construction

This project takes advantage of MobileNet CNN implementation from Google. MobileNet uses depth wise separable convolutions. It significantly reduces the number of parameters when compared to the network with regular convolutions with the same depth in the nets. This results in lightweight deep neural networks. CNN architecture of MobileNet is represented below in Fig 4.

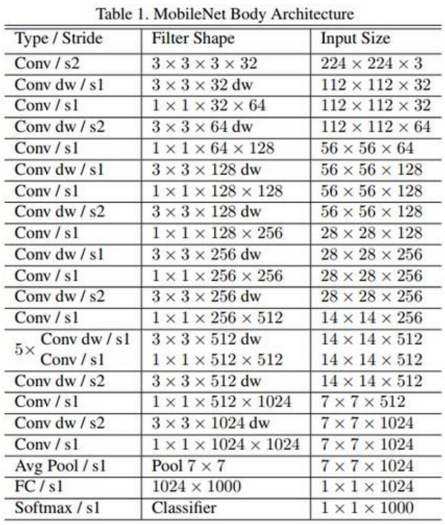

Table 1. MobileNet Body Architecture

| Type / Stride | Filter Shape | Input Size |
|---|---|---|
| Conv / s2 | 3 × 3 × 3 × 32 | 224 × 224 × 3 |
| Conv dw / s1 | 3 × 3 × 32 dw | 112 × 112 × 32 |
| Conv / s1 | 1 × 1 × 32 × 64 | 112 × 112 × 32 |
| Conv dw / s2 | 3 × 3 × 64 dw | 112 × 112 × 64 |
| Conv / s1 | 1 × 1 × 64 × 128 | 56 × 56 × 64 |
| Conv dw / s1 | 3 × 3 × 128 dw | 56 × 56 × 128 |
| Conv / s1 | 1 × 1 × 128 × 128 | 56 × 56 × 128 |
| Conv dw / s2 | 3 × 3 × 128 dw | 56 × 56 × 128 |
| Conv / s1 | 1 × 1 × 128 × 256 | 28 × 28 × 128 |
| Conv dw / s1 | 3 × 3 × 256 dw | 28 × 28 × 256 |
| Conv / s1 | 1 × 1 × 256 × 256 | 28 × 28 × 256 |
| Conv dw / s2 | 3 × 3 × 256 dw | 28 × 28 × 256 |
| Conv / s1 | 1 × 1 × 256 × 512 | 14 × 14 × 256 |
| 5× Conv dw / s1<br>Conv / s1 | 3 × 3 × 512 dw<br>1 × 1 × 512 × 512 | 14 × 14 × 512<br>14 × 14 × 512 |
| Conv dw / s2 | 3 × 3 × 512 dw | 14 × 14 × 512 |
| Conv / s1 | 1 × 1 × 512 × 1024 | 7 × 7 × 512 |
| Conv dw / s2 | 3 × 3 × 1024 dw | 7 × 7 × 1024 |
| Conv / s1 | 1 × 1 × 1024 × 1024 | 7 × 7 × 1024 |
| Avg Pool / s1 | Pool 7 × 7 | 7 × 7 × 1024 |
| FC / s1 | 1024 × 1000 | 1 × 1 × 1024 |
| Softmax / s1 | Classifier | 1 × 1 × 1000 |

Figure 4: MobileNet Architecture

Training was done with about 90% of the facial expression dataset, and the rest 10% was used for testing purpose. This gives the CNN more amount of data to train upon, which leads to good prediction accuracy. Predicted value's weightedaverage is calculated iteratively to get thefinal predicted value.

**CNN Validation and Analysis of Results & Plotting of Metrics:**

In this phase, testing the model with the input set of data takes place. After the training process, the test data is used to compare the ground truth with the predicted outcome, from the CNN's predicted output prediction. These results are basically metrics that can facilitate with the analysis of prediction results. Performance of the proposed CNN must be estimated to measure the effectiveness of our proposed algorithms. The performance metrics to be calculated are precision, recall, f1- score and accuracy.

Precision gives the True Positive Rate which is the quality of the model and is calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (1)$$

Recall is also known as sensitivity. Recall gives output quality of how many true relevant results are obtained. It is calculated as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (2)$$

F1-score is calculated as the weighted average of both precision and recall. The f1- score is calculated as

$$\text{F1 score} = 2 * \frac{\text{precision} \; * \; \text{recall}}{\text{precision} \; + \; \text{recall}}$$

(3)

In performance metrics, accuracy is the complete overall validation of the classifier and formulated as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP}+\text{TN} + \text{FP}+\text{FN}} \tag{4}$$

In this project, these final prediction metrics are referred to as “emotion metrics” and “emotion metric percentage”. These values are mostly helpful for the plotting of the metrics once the prediction is done. Graphs are a common method to visually illustrate relationships in the data. The sole purpose of a graph is to present data that are too numerous or complicated to be described adequately in the text and in less space. However, use graphs for small amounts of data that could be conveyed succinctly in a sentence. Likewise, do not reiterate the data in the text since it defeats the purpose of using a graph. The user can set a specific time for which the system remains working and also actively keeps calculation emotion metrics in the background. After the said time is elapsed, it can effectively plot the required bar-graph to represent the metrics. The system learns by getting knowledge from the training data provided to it during the initial phase of training (learning). For any given input of a video with a person face portraying their emotions, the accurate algorithm analyses and predicts the emotion shown on the person’s face as a result, these emotions are tallied in the background, an emotion metrics percentage graph is plotted to determine how long the person was happy, sad, neutral, angry or surprised for a given amount of time ‘t’.

## EXPERIMENTAL RESULTS

Proposed system can be tested with a video file given to it as an input, and let it identify and recognize emotions on the person’s face. In Fig 5, we can see the system in action. Furthermore, in Fig 6 and Fig 7, we can see the other emotions, detected in the video.

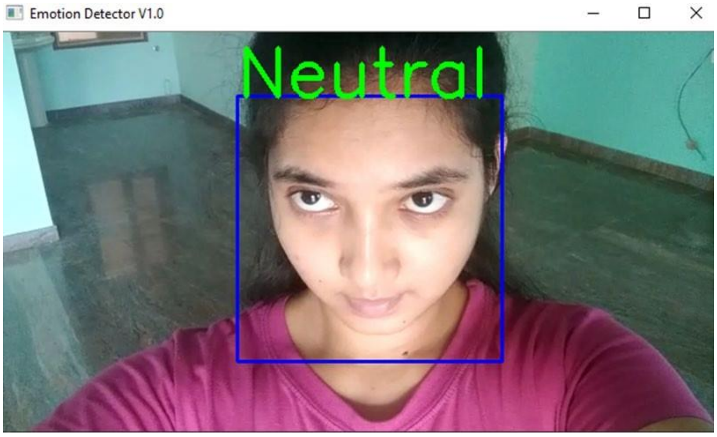

Figure 5: System during run-time (Neutral)

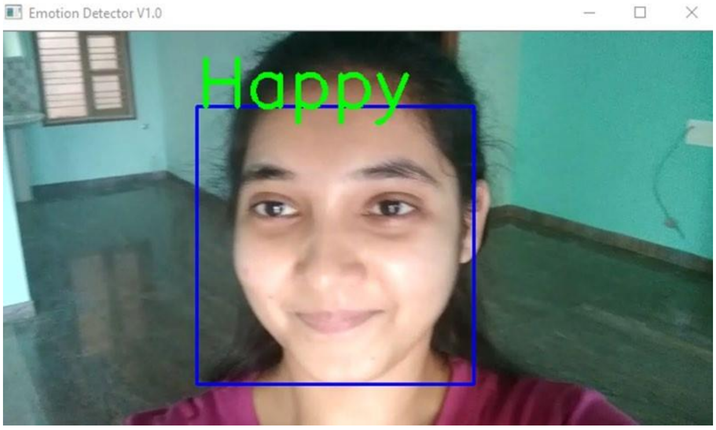

Figure 6: System during run-time (Happy)

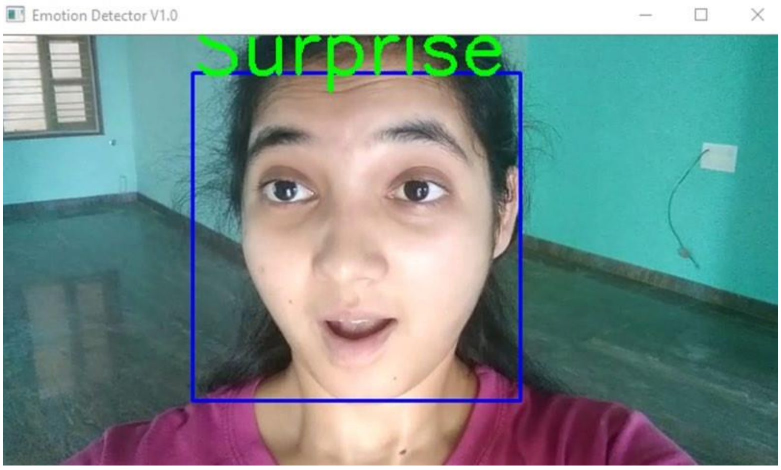

Figure 7: System during run-time (Surprise)

Based on the emotions recognized in the input video fed into the system, as the part of the final output, it plots an emotions metrics graph which compares each emotion with its respective emotion percentage. This is represented by Fig 8.

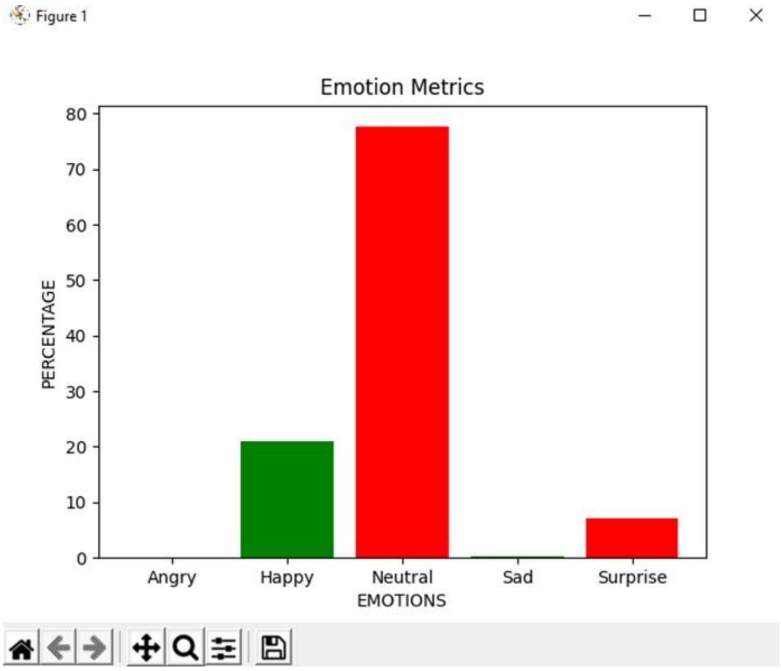

Figure 8: Emotion Metric Graph

The Performance metrics are obtained using the accuracy function in Python. In Table 1, we can see the accuracy percentage of the system. Here, the comparison has been done solely based on the current application, which is emotion recognition. With respect to these metrics, the graph in Fig 9 shows performance metrics, visually.

Table 1: Performance Metrics

| Implementation | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| **MobileNet** | 87.75 | 0.88 | 0.87 | 0.87 |
| **Fast R-CNN** | 86.23 | 0.87 | 0.86 | 0.86 |
| **ShuffleNet** | 84.34 | 0.85 | 0.84 | 0.84 |
| **ResNet50** | 83.52 | 0.84 | 0.83 | 0.83 |

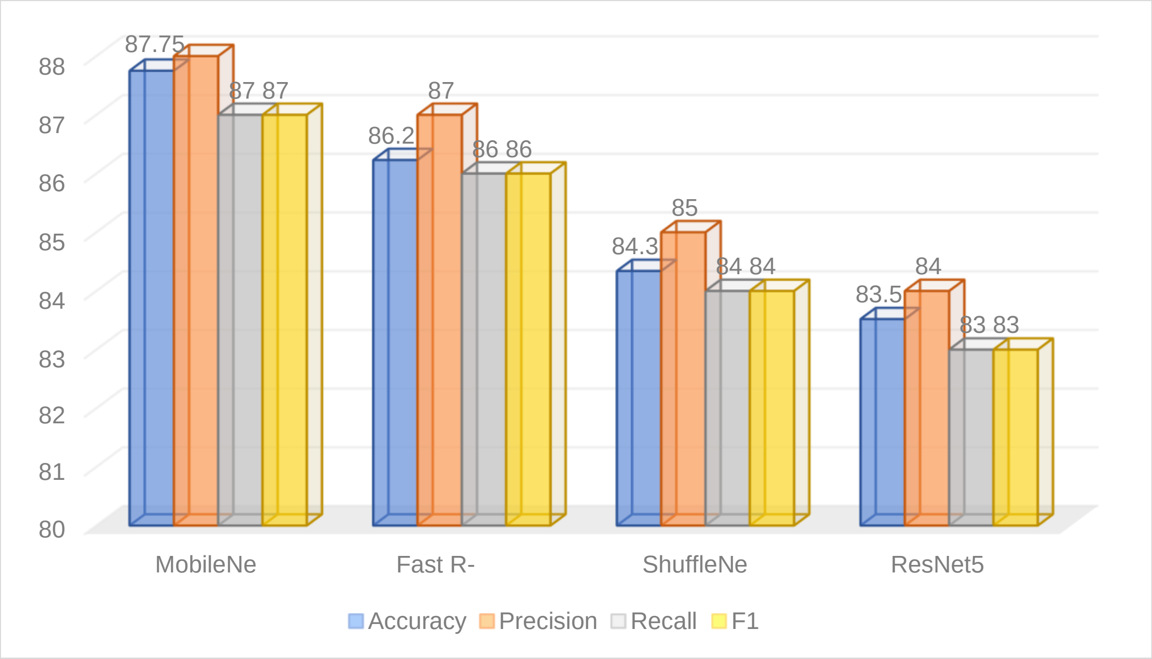

Figure 9: System performance metrics

## CONCLUSIONS

Online learning is increasing day by day at the current scenario the importance of student's engagement detection is to be accepted. Focusing on automated engagement detection the limitations of primitive methods are eliminated. In this project Deep Neural Networks have been used for engagement detection. After conducting a numerous tests and experiments with the system, it has been concluded that facial expression data is one of the best for detecting Listener behavior from their emotions. The addition of CNN for deep learning model creation with facial data shows an excellent result in the prediction of engagement and it provides about 88% accuracy, using Mobile Net architecture. CNN becomes more suitable method for Listener engagement detection in real-time.

## REFERENCES

[1] Mushtaq Hussain , Wenhao Zhu , Wu Zhang , and Syed Muhammad Raza Abidi, "ListenerEngagement Predictions in an e-Learning System and Their Impact on ListenerCourse Assessment Scores", Computational Intelligence and Neuroscience, vol. 2018, Oct 2018.

[2] Zhaoli Zhang, Zhenhua Li, Hai Liu,Taihe Cao, and Sannyuya Liu, "Data-drived Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology", Journal of Educational Computing Research, vol. 58, 2020.

[3] Jia Yue, Feng Tian, Kuo-Min Chao, Nazaraf Shah, Longzhuang, Yan Chen , And Qinghua Zheng, "Recognizing Multidimensional Engagement of E-Learners Based onMulti-Channel Data in E-Learning Environment", IEEE Access, vol. 7, pp. 149554 - 149567, Oct 2019.

[4] Michail N. Giannakos, Kshitij Sharma, Ilias O. Pappas, Vassilis Kostakos, Eduardo Velloso, "Multimodal data as a means to understand the learning experience", International Journal of Information Management, vol. 48, pp. 108-119, 2019.

[5] Mohamed El Kerdawy ,Mohamed El Halaby, Afnan Hassan, Mohamed Maher, Hatem Fayed, Doaa Shawky and Ashraf Badawi, "The Automatic Detection of Cognition Using EEG and Facial Expressions", Sensors, vol. 20, Jun 2020.

[6] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, Sidney K. D'Mello, "Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate", IEEE Transactions on Affective Computing, vol. 8, pp. 15-28, 2017.

[7] Krithika.L.Ba,Lakshmi Priya GG, "ListenerEmotion Recognition System (SERS) for e- learning improvement based on learner concentration metric", Procedia Computer Science, vol. 85, pp. 767-776, 2016.

[8] Timothy Patterson, Naveed Khan, Sally McClean, Chris Nugent, Shuai Zhang, Ian Cleland, Qin Ni, "Sensor-Based Change Detection for Timely Solicitation of User Engagement", IEEE Transactions on Mobile Computing, vol. 16, pp. 2889 - 2900, 2016.

[9] Athanasios Psaltis, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, and Petros Daras, 'Multimodal ListenerEngagement Recognition in Prosocial Games", IEEE Transactions on Computational Intelligence and AI in Games, 2017.

[10] Prabin Sharma, Shubham Joshi, Subash Gautam, Sneha Maharjan, Vitor Filipe, Manuel Cabral Reis, "ListenerEngagement Detection Using Emotion Analysis, Eye Tracking and Head Movement With Machine Learning", Computer Vision and Pattern Recognition, vol. 1, 2020.

[11] K. Keerthana, D.Pradeep, Dr. B. Vanathi, "Learner's Engagement Analysis for E-Learning Platform", International Journal of Scientific Development and Research, vol. 5, 2020.

[12] Meredith Carroll, Mitchell Ruble, Mark Dranias, Summer Rebensky, Maria Chaparro, Joanna Chiang, Brent Winslow, "Automatic Detection of Learner Engagement Using Machine Learning and Wearable Sensors", Journal of Behavioral and Brain Science, vol. 10, pp. 165-178, 2020.

[12] Yifeng Zhao, Deyun Chen, "A Facial Expression Recognition Method Using Improved Capsule Network Model", Scientific Programming, vol. 2020, Oct 2020.

[13] Awais Mahmood, Shariq Hussain, Khalid Iqbal, and Wail S. Elkilani, "Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion", Mathematical Problems in Engineering, vol. 2019, Aug 2019.

[14] Veena Mayya, Radhika M. Pai, Manohara Pai M, "Automatic Facial Expression Recognition Using DCNN", Procedia Computer Science, vol. 93, pp. 453-461, 2016.

[15] Vladimir Soloviev, "Machine Learning Approach for ListenerEngagement Automatic Recognition from Facial Expressions", Scientific Publications of the State University of Novi Pazar Series A Applied Mathematics Informatics and mechanics, vol. 10, pp. 79-86, 2018.