# Facial Expression Recognition using Deep CNN


Aksa Benny
*School of Computer Science*
*University of Windsor*
Windsor, Canada
bennya@uwindsor.ca

Puneet Jain
*School of Computer Science*
*University of Windsor*
Windsor, Canada
jain24@uwindsor.ca

Jaykumar Patel
*School of Computer Science*
*University of Windsor*
Windsor, Canada
patel8g3@uwindsor.ca

Jess Joseph Benny
*School of Computer Science*
*University of Windsor*
Windsor, Canada
bennyj@uwindsor.ca



***Abstract*— This project aims to deliver a facial expression recognition system using a deep learning technique. Facial expression recognition is pivotal in affective computing, mental illness prognosis, and rehabilitative services. Because of its potential applications in acknowledging people's mental health, facial expression recognition from images or videos has piqued the research community's interest. For that, we developed a 2-dimensional convolutional neural network to classify facial expressions from photos and video. The algorithm will classify emotions such as anger, happiness, neutrality, sadness, and surprise. An Object Detection Algorithm is utilized to detect faces in images or real-time video. After that, we have used the model to identify face landmarks and embeddings of the face in that image. We compared the performance of VGGNet as the base model with ResNet, and SE-Net, three standard image classification CNN architectures.**

# I. INTRODUCTION

Facial expression recognition is the process of distinguishing sensory perception, facial movement, and feature points from photographic images or videos and classifying them into abstract classes based solely on graphical and visual information. Humans' internal feelings are spontaneously reflected on their faces. The recognition of facial expression is plausible due to its resemblance. The expressions cause pseudo gestures of the face, consequently evolving the alignments of facial curvature. Humans' inner feelings are frequently reflected instinctively on their faces; thus, the face is known as the 'mirror of the mind. ' Thus, expression recognition aids in the interpretation of psychological operations and the differentiation of facial cues. Many physiological and mental health experts worldwide are working relentlessly to ensure the mental stability and health of society, which will lead to effective sustainability and reliability and the critical thinking ability of the workforce. So the project was primarily focused on implementing real-time facial expression recognition using the relatively modern technique in the field of deep learning using convolutional neural network

Our experiments leveraged data from the FER [2013] open data set. Rather than laboratory-controlled data, we required emotionally comprehensive data so that there would be an ample number of instances of expression utilization to investigate from in the wild conditions with sufficient training data and expression-unrelated alterations, such as luminance, head pose, and ethnicity bias. The open data set contained more than 25,000 images.

The successful consequence of this project can be assessing and surveilling the real-life emotions of individuals in multiple scenarios, including people in Covid confinement institutions, halfway houses, assisted living facilities, etc.

We evaluated the performance of 3 major CNN architectures for image classification, namely VGGNet, ResNet, and SE-Net. We considered VGGNet as the base model. We analyzed resource utilization matrices, including training and prediction speed, as well as accuracy and loss metrics, to develop a model that can be easily deployed in real-life circumstances.

# II. PROPOSED MODEL

## A. *Motivation*

The relevance of facial expression in nonverbal communication cannot be overstated. The hallmark of civilized culture is to read others' emotions through facial expressions. The fundamental goal of a man's innovations and discoveries is to enhance his society and himself. During the unprecedented Covid 19 outbreak, we realized how challenging it is to monitor the emotions of isolated victims, which damages their psychological health. We are inspired to develop a unanimous tool that helps organizations monitor the inmates' mental conditions.

It can be utilized by covid rehabilitative institutions to assess the emotional state of inmates under containment without having to actually meet or oversee them, which can be dangerous. It can also be adopted by people with Asperger's syndrome who cannot read others' sentiments whilst conversing. This aids them in developing an efficient communication model.

## B. *Description*

Even though few models allow video surveillance systems or emotions monitoring, there are few disadvantages. In traditional video surveillance, the surveilling person has to be in front of the camera and check what the person is doing and the emotions of the people who are under quarantine. He is able to see what is going on in the day-to-day life of that person, and it affects the privacy of the surveillant. But in computer-based surveillance systems, algorithms will extract required information(Emotions in our project).

The remaining of the surveillant's activities are entirely private. Unless the system recognizes someone as miserable or upset, they are given total privacy, and professional assistance is provided when only needed. Traditional machine learning algorithms were already utilized to classify facial expressions from photographs and videos. The manual retrieval of features from the image is among the main shortcomings of the approach. For example, for each mood, we must manually identify the face landmarks and their attributes. Consider the eyebrow as a facial feature; the shape of the eyebrow varies depending on the emotion. For example, in the event of surprise, the eyebrows will be lifted, and in the event of sadness, the eyebrows would be lowered. When using CNN-based models, we can input images directly instead of manually extracting features like in traditional ML Models. CNN can extract patterns from images and create models automatically.

## III. LITERATURE REVIEW

M. Munasinghe [1] proposed and developed a methodology to identify facial emotions using facial landmarks and a random forest classifier where the famous extended Cohn-Kanade database has been used to train random forest and test the system's accuracy.

Kaihao zhang,[2] in his paper, proposed a part-based hierarchical bidirectional recurrent neural network (phrnn) to analyze the facial expression information of temporal sequences where both recognition and verification signals as supervision are used to calculate different loss functions, which are helpful to increase the variations of other expressions and reduce the differences among identical expressions.

Ali mollahosseini[3] used deep neural networks whose architecture is comparable to or better than the state-of-the-art methods and better than traditional convolutional neural networks.

Huiyuan yang[4] has proposed facial expression recognition by extracting the expressive component through a de-expression learning procedure called De-expression Residue Learning (DeRL).

In his paper, T. Jabid[5] analyzes the performance of a new feature descriptor, Local Directional Pattern (LDP), to represent facial expressions.

## IV. CENTRAL IDEA

Our project's central notion relies upon contemporary convolutional neural networks' outstanding classificational capabilities and their recent results in image recognition. Artificial neural networks are based on the biological neural networks that make up animal brains, and they can perform complicated tasks when hundreds of these neurons are placed together.

Convolutional layers and pooling layers are two new building blocks introduced with the concept of CNN[6]. Convolutional layers can extract unique features from the input using filters, whereas pooling layers subsample the input. In this project, we will be making use of the latest three well-proven CNN architectures, VGGNet, ResNet, and SENet.

### A. *VGGNet*

Simonyan and Zisserman introduced the VGG network design in their study, "Very Deep Convolutional Networks for Large Scale Image Recognition," published in 2014[10]. It featured a straightforward and conventional design, with two or three convolutional layers, a pooling

layer, and so on (for a total of only 16 convolutional layers), plus a final dense network with two hidden layers and the output layer. It only used 3 x 3 filters, but there were a lot of them.

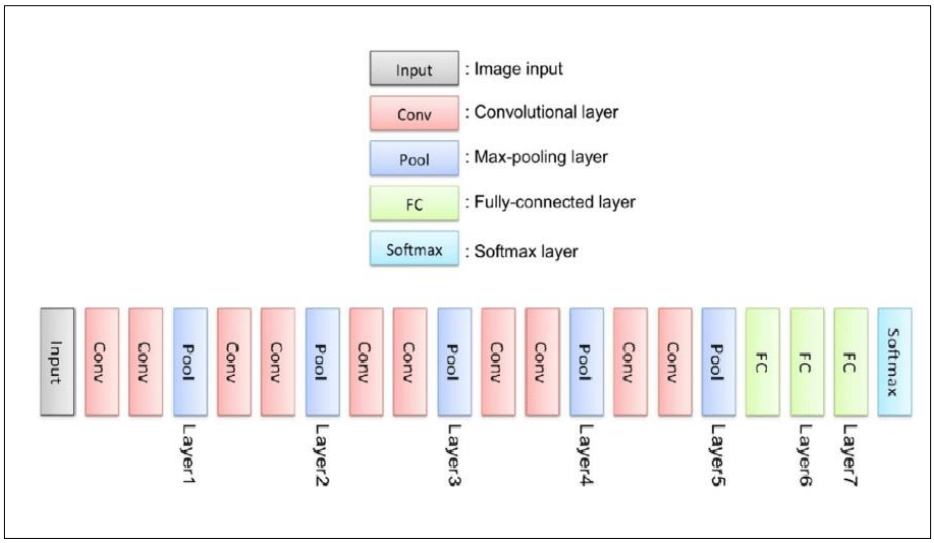


*Figure 1 Architecture of VGGNet*

## *B. ResNet*

ResNet, also known as Residual Network, developed by Kaiming He et al. ResNet[11], is inspired by the idea that "models are getting deeper and deeper, with fewer and fewer parameters." The use of skip connections (also known as shortcut connections) is vital to training such a deep network: the signal entering into one layer is also added to the output of a layer a little higher up the stack.

While training a neural network, the goal is to make it model a target function h(x). If we add the input to the output of the network(skip connection), then the network will be forced to model f(x) = h(x) – x rather than h(x). This is called residual learning, and This will speed up the training process.

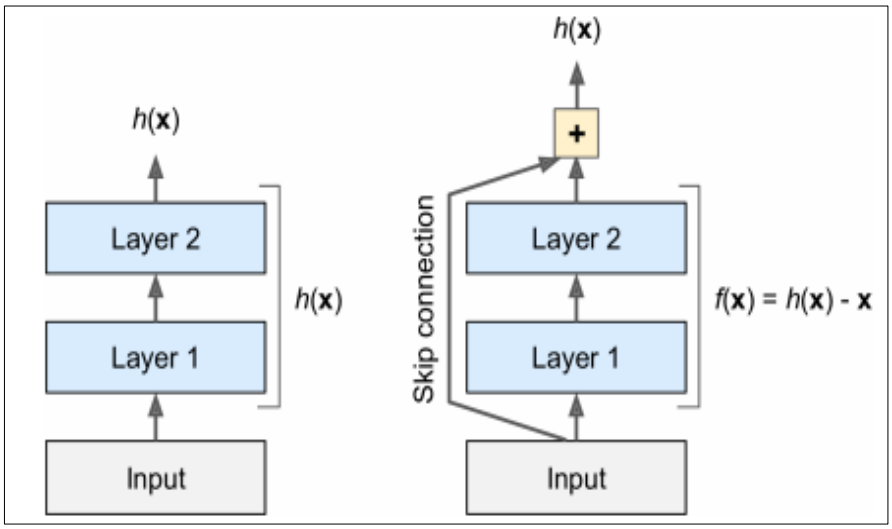


*Figure 2 Structure of residual unit*

ResNet architecture is similar to conventional CNN architectures, where convolutional layers are followed by pooling layers in the beginning and Fully connected dense layers at the end. In between, it will have a deep stack of residual units.

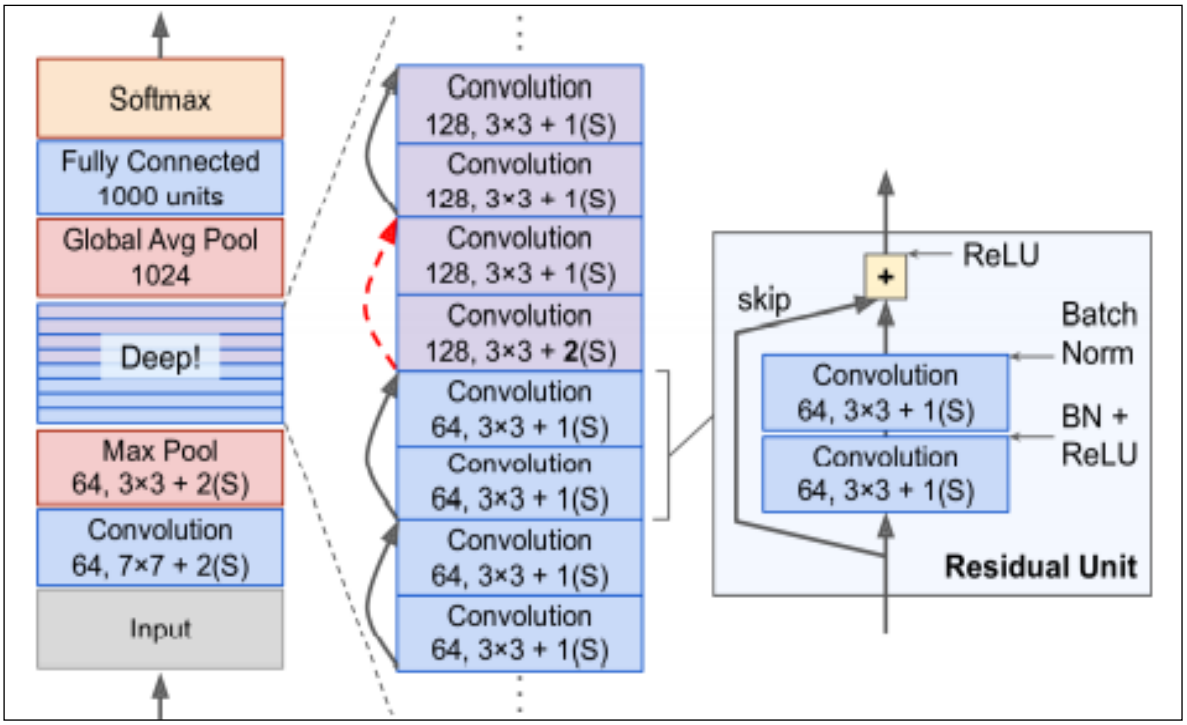


*Figure 3 Architecture of ResNet*

## C. *SENet*

SENet is also known as Squeeze-and-Excitation Network, is the winning architecture in ILSVRC 2017 challenge[12]. This architecture extends existing architectures such as ResNets and boosts their performance. A SENet boosts performance by adding a small neural network called a SE Block to each unit in the original architecture. A SE Block examines the unit's output to which it is connected, focusing solely on the depth dimension, and learns which characteristics are typically active together. An SE Block usually contains just three layers, a global average pooling layer, a dense layer using ReLu activation, and another dense layer using sigmoid activation. We will be experimenting with SENet extended on ResNet, known as the SE-ResNet model (Figure 4).

## D. *Tensorflow and Keras*

We will be using two famous APIs for developing deep learning models, namely Tensorflow and Keras. Google developed TensorFlow. It is now available as an open-source library under the Apache license. Keras is a high-level deep-learning library written in python. The latest builds of Keras are running on top of Tensorflow.

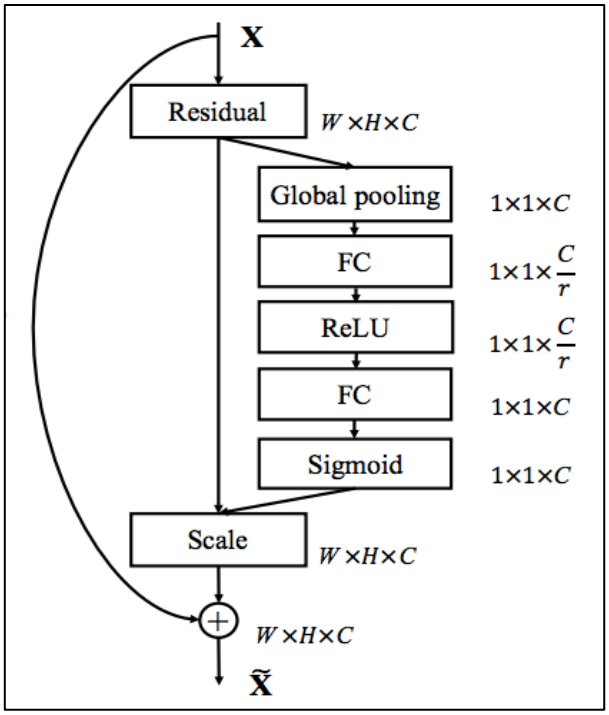


*Figure 4 Architecture of SE-ResNet*

## V. METHODOLOGY

Deep-learning algorithms used in the field of computer vision, such as CNN, have made significant progress in recent decades. For feature extraction, classification, and recognition tasks, several deep-learning-based methods have been applied. The fundamental benefit of a CNN is that it allows "end-to-end" learning directly from input images, which eliminates the need for physics-based models and/or other pre-processing approaches. CNN has produced cutting-edge achievements in a variety of disciplines, including object identification, face recognition, scene understanding, and FER, because of these factors. Facial expression is recognized and classified using a few simple steps. Firstly, we need to determine if there exists a face in the image or not. To complete this task, we take the help of Haar Cascade [13]. Haar Cascade is a machine learning-based approach that involves training the classifier using many positive and negative images. Positive images - These photos contain the images that our classifier is supposed to recognize. Negative Images - Images of everything else that is not the object we are looking for. In other words, it is an Object Detection Algorithm used to identify faces in an image or a real-time video. Models are stored in XML files and can be read with the OpenCV methods. Include models for face detection, eye detection, upper body and lower body detection, license plate detection, etc. Once it is determined that the image does contain a face, our next task is to process that image and identify face landmarks and embeddings of the face in that image. To complete these tasks, the image is passed through various Convolution Neural Network Layers (CNN) (Figure 5), and each layer performs functions like Relu - rectified linear activation function [17], the piecewise linear function that will output the input directly if it is positive; otherwise, it will output zero, and Max pooling - pooling operation that calculates the maximum, or largest, value in each patch of each feature map. Once the image is passed through various CNN layers, the SoftMax function will be performed as the final part of image processing; SoftMax is a function that turns a vector of K real values into a vector of K real values that sum to 1. After processing the image, the last step is to classify the image into categories such as happy, angry, sad, surprised, and neutral.

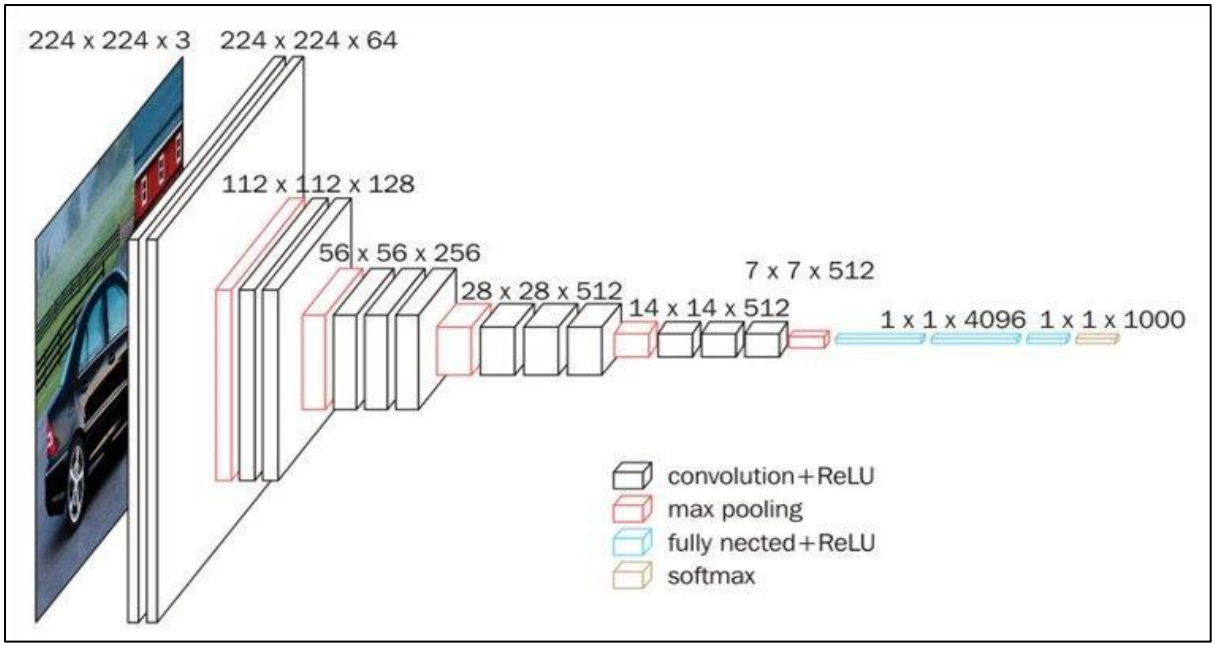


*Figure 5 Model Architecture*

## VI. EXPERIMENTS

Many databases have been used for comparative and extended experimentation in the field of facial expression recognition. Human face expressions have traditionally been researched with either 2D static photographs or 2D video sequences. Large position fluctuations and delicate face behaviors are challenging to handle with a 2D analysis. The exploration of the fine structural changes inherent in spontaneous expressions will be aided by the analysis of 3D facial emotions. We have used FER – 2013 [14] dataset. The data consists of grayscale images of faces at a resolution of 48x48 pixels. The faces have been automatically registered so that they are centered in each image and take up roughly the same amount of space. The aim is to categorize each face into one of five groups depending on the emotion expressed in the facial expression (Angry, Happy, Sad, Surprise, Neutral). The training dataset contains approximately 25 thousand cases, while the validation dataset contains approximately 6 thousand examples.

Contents of the training dataset (Figure 6) :

- Happy: 7164
- Angry: 3993
- Sad: 4938
- Neutral: 4982
- Surprised: 3205

Similarly, the contents of the validation dataset are :

- Happy: 1825
- Angry: 960
- Sad: 1139
- Neutral: 1216
- Surprised: 797

*Figure 6 Images from the dataset showing different facial expressions, top – happy, middle left – angry, middle right – surprised, bottom left – sad, bottom right – neutral*

In the hope of achieving high accuracy, we tried to expand our dataset. We made various modifications to each image and used the added augmented images to train our model. To expand our dataset, each image was rescaled, rotated a little bit, sheared, zoomed, width and height are changed.

```
train_datagen = ImageDataGenerator(
                    rescale=1./255,
                    rotation_range=30,
                    shear_range=0.3,
                    zoom_range=0.3,
                    width_shift_range=0.4,
                    height_shift_range=0.4,
                    fill_mode='nearest')
```

*Figure 7 Code Snippet That Performs Augmentation of each image.*

While training the model, each image is passed through various CNN Layers, Relu, and Max pooling is performed at each layer (Figure 8).

```
model.add(Conv2D(32,(3,3),
                 padding='same',
                 kernel_initializer='he_normal',
                 input_shape=(img_rows,img_cols,1)))
model.add(Activation('elu'))
model.add(BatchNormalization())
model.add(Conv2D(32,(3,3),
                 padding='same',
                 kernel_initializer='he_normal',
                 input_shape=(img_rows,img_cols,1)))
model.add(Activation('elu'))
model.add(BatchNormalization())
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.2))
```

*Figure 8 Code Snippet of Model Design*

After passing through multiple CNN layers, the final step is SoftMax (Figure 9).

```
model.add(Dense(num_classes,
                kernel_initializer='he_normal'))
model.add(Activation('softmax'))
```

*Figure 9 Code Snippet - Final Dense and Softmax Layers*

Once the training is complete, it is time to test our model. To test our model, we tried to capture Facial Expressions in real-time with the help of a webcam on the laptop, and the results are quite satisfactory (Figure 10).

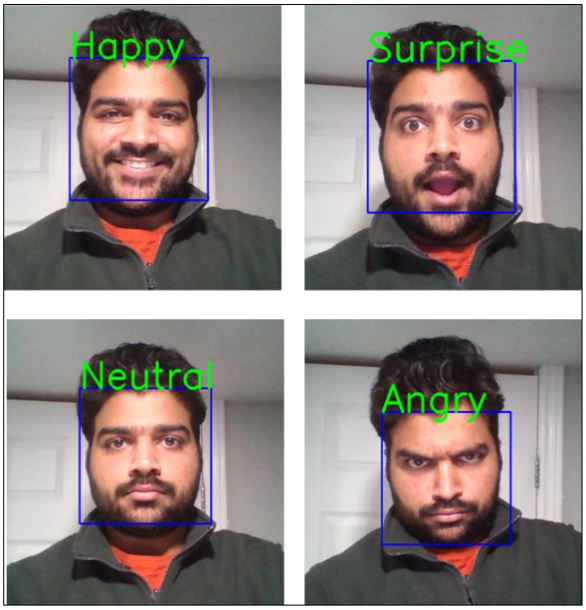


*Figure 10 Real-time classification images from camera input*

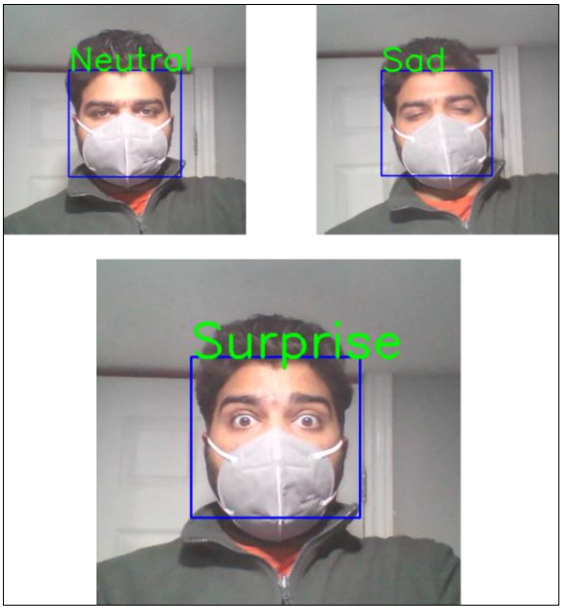


*Figure 11 Real-time classification of images from camera input with a mask on*

We also tried to put our model to the test while a person is wearing various accessories, such as a face mask (Figure 11).

## VII. RESULTS AND DISCUSSIONS

In this study, we evaluated the performance of VGGNet, the base model, with SeNet and ResNet, two alternative models. To determine the correctness of the models, we have checked the validation loss and validation accuracy of the models and precision and recall of the facial expressions using the confusion metrics. Each model's average training time, average prediction time, memory usage, and other resource utilization have been used as performance measures.

To improve our classification of CNN models, VGGNet, SeNet, and Resnet, we use categorical_crossentropy as a loss function and optimizer as adam. We avoided overfitting by using data augmentation and drop-out layers.

The Drop-out layer randomly sets input units to 0 with a frequency of rate at each step during training time. Inputs that aren't set to 0 are scaled up by 1/(1 - rate) so that the total sum remains the same. Note that the Drop-out layer is only active when training is set to True, which means that no values are dropped during prediction.

```
model.compile(optimizer = 'adam',
              loss='categorical_crossentropy',
              metrics=['accuracy'])
```

*Figure 12 Code Snippet - loss function & optimizer*

We have a large validation dataset that has been matched to predicted classes; thus, we achieved the validation accuracy of 73.8% with training speed and prediction speed of 26ms per second and 1 second per 5000 inputs, respectively, in our base model VGGNet (Figure 13).

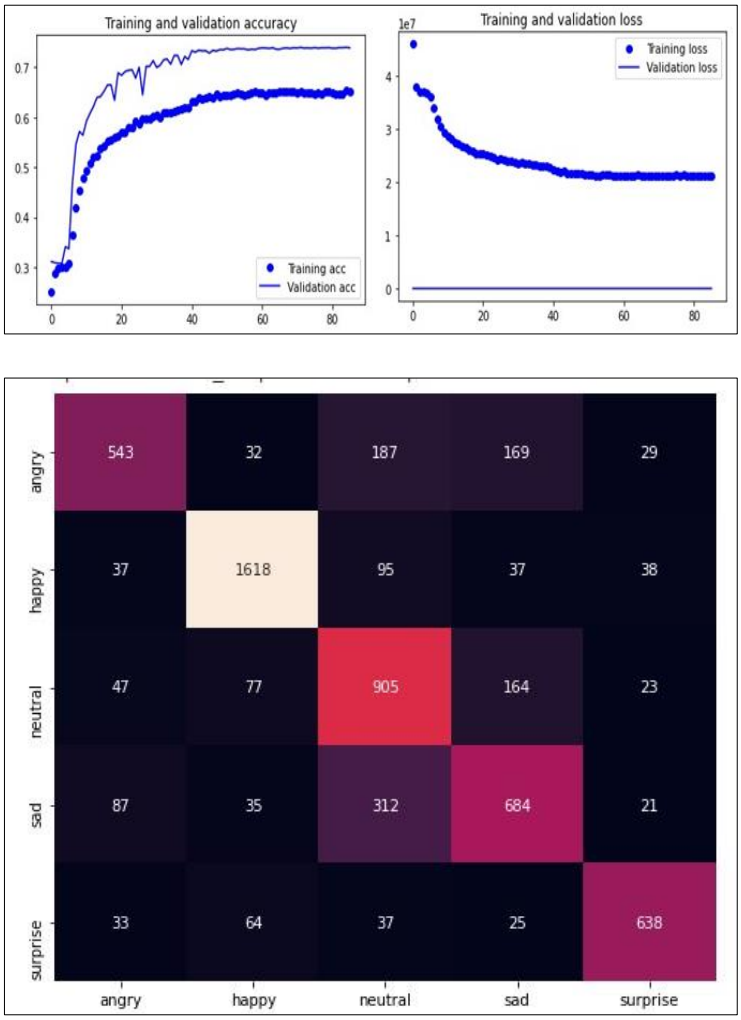


*Figure 13 VGGNet Accuracy, Loss & Confusion Matrix*

For the ResNet model, training speed and prediction speed were almost the same, but the accuracy was 68.61% (Figure 14).

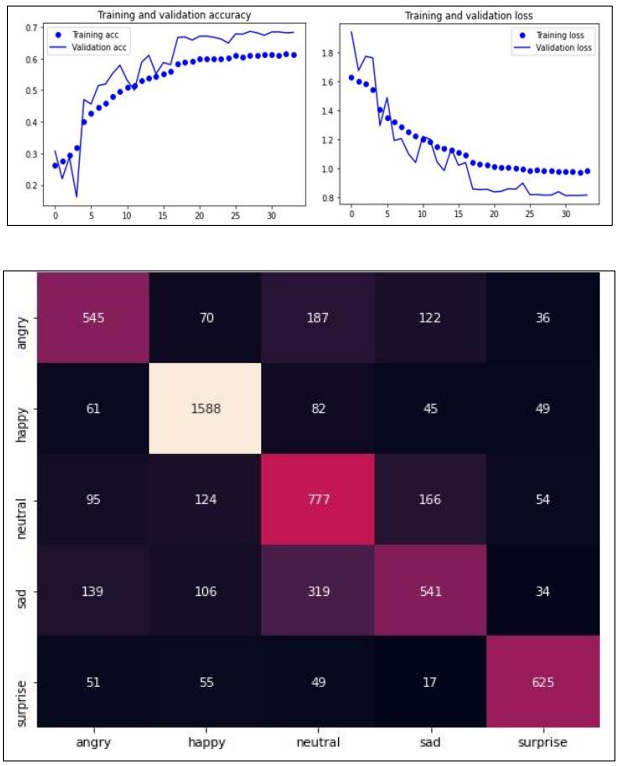


*Figure 14 ResNet Accuracy, Loss & Confusion Matrix*

SeNet had the highest validation accuracy of 75.89%, with training speed and prediction speed of 217 ms per second and 12 seconds per 5000 inputs, respectively.

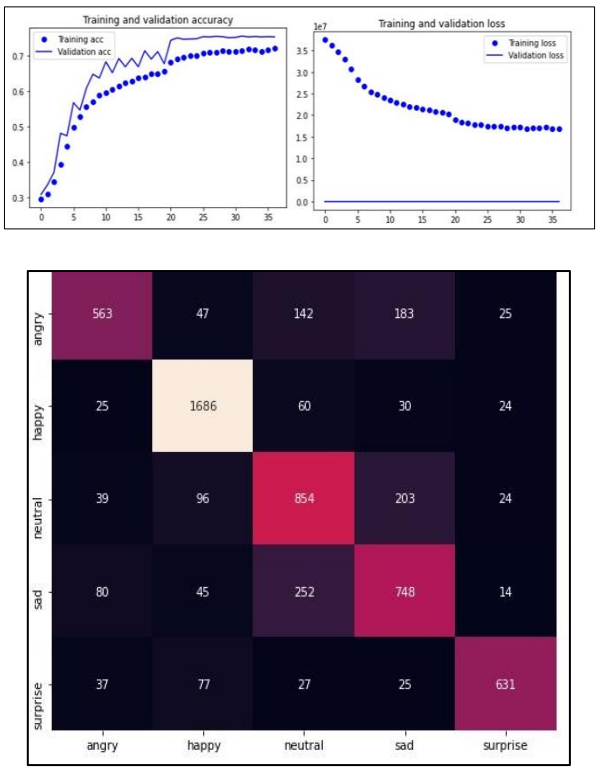


*Figure 15 SeNet Accuracy, Loss & Confusion Matrix*

For models, VGGNet, ResNet, and SeNet, however, the lowest recall and precision were seen for sad and neutral facial expressions, among other facial expressions. This pattern could be interpreted from the confusion matrix of the three models. As seen in the dataset, sad and neutral expressions express very little. For a normal individual, the expansion of lips and eyes, among other things, are nearly identical in both expressions.

| *Algorithm* | Recall | Precision | Accuracy |
|---|---|---|---|
| *VGGNet* | 72.734% | 74.38% | 73.8% |
| *ResNet* | 66.094% | 66.726% | 68.61% |
| *SeNet* | 73.22% | 75.37% | 75.89% |

*Table 1 Performance Metrics of all three models*

Moreover, for all the models, we have the training accuracy more than the validation accuracy and training loss less than the validation loss because of the data augmentation and drop-out layers. Imbalance in training data for the classes could model bias towards for majority class. Hence, weights are assigned to every class so that every class will be classified equally during the validation phase.

## VIII. CONCLUSION

In conclusion, the SeNet model has the highest validation accuracy, precision, and recall, whereas the ResNet model has the lowest. Not only will we compare accuracy matrices, but we will also compare resource utilization matrices like testing and prediction time to determine the most suitable model. VGGNet has a resource consumption advantage over SeNet due to its reduced resource utilization and even though accuracy is a little less. Because of its low processing power requirements, VGGNet could be employed in various small devices, such as IoT devices.

## IX. FUTURE SCOPE AND OPEN PROBLEMS

A better dataset could help the model predict sad and neutral emotions more accurately as FER-2013 is not a laboratory-created dataset. In the future, using datasets such as CK+[15] and AffectNet[16] could provide better results. By using autoencoders or better pre-processing, we will ResNet and SeNet, as they are well-proven face recognition models and tuning them for facial expression identification could improve accuracy. Moreover, we try to deploy this so that we can deal with real-life scenarios and check the accuracy of the model.

## X. REFERENCES


[1] M. I. N. P. Munasinghe, "Facial Expression Recognition Using Facial Landmarks and Random Forest Classifier," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018, pp. 423-427, DOI: 10.1109/ICIS.2018.8466510.

[2] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," in IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4193-4203, Sept. 2017, DOI: 10.1109/TIP.2017.2689999.

[3] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-10, DOI: 10.1109/WACV.2016.7477450.

[4] Huiyuan Yang, Umur Ciftci, Lijun Yin " Facial Expression Recognition by De-Expression Residue Learning" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168-2177.

[5] T. Jabid, M. H. Kabir and O. Chae, "Facial expression recognition using Local Directional Pattern (LDP)," 2010 IEEE International Conference on Image Processing, 2010, pp. 1605-1608, DOI: 10.1109/ICIP.2010.5652374.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Aurélien Géron "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow", 2nd Edition June 2019, ISBN-13 : 978-1492032649

[7] Yann LeCun et al. "Gradient-based learning applied to document recognition, "Proceedings of the IEEE, 1998, pp. 2278-2324, DOI: 10.1109/5.726791.

[8] Yann LeCun et al. "Gradient-based learning applied to document recognition, "Proceedings of the IEEE, 1998, pp. 2278-2324, DOI: 10.1109/5.726791.

[9] Christian Szegedy et al., "Going Deeper With Convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

[10] Karen Simonyan and Andrew Zisserman "Very Deep Convolutional Networks for Large-Scale Image Recognition " conference paper at ICLR,2015

[11] Kaiming He et al., "Deep Residual Learning for Image Recognition", https://arxiv.org/abs/1512.03385

[12] Jie Hu et al. "Squeeze-and-Excitation Networks", CVPR 2018, https://arxiv.org/abs/1709.01507

[13] Padilla, R., Costa Filho, C. F. F., & Costa, M. G. F. (2012). Evaluation of haar cascade classifiers designed for face detection. World Academy of Science, Engineering and Technology, 64, 362-365.

[14] Wolfram Research, "FER-2013" from the Wolfram Data Repository (2018).

[15] The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression

[16] AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild

[17] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.