International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7s
DOI: https://doi.org/10.17762/ijritcc.v11i7s.7001
***Article Received**: 02 April 2023 **Revised**: 22 May 2023 **Accepted**: 01 June 2023*

An Automated System for Depression Detection Based on Facial and Vocal Features

Mohit Patil[1], Vijayshri Khedkar[2]
[1]Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
e-mail: mohit.patil.mtech2021@sitpune.edu.in
[2]Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
e-mail: vijayshri.khedkar@sitpune.edu.in

Abstract— Diagnosing depression is a challenge due to the subjective nature of traditional tools like questionnaires and interviews. Researchers are exploring alternative methods for detecting depression, such as using facial and vocal features. This study investigated the potential of facial and vocal features for depression detection using two datasets: images of facial expressions with emotion labels, and a vocal expression dataset with positive and negative words. Four deep-learning models were evaluated for depression detection from facial expressions, and two traditional machine-learning models were trained for sentiment analysis on the vocal expression dataset. The CNN model performed best for facial expression analysis, while the Naive Bayes model performed best for vocal expression analysis. The models were integrated into a web application for depression analysis, allowing users to upload a video and receive an analysis of their facial and vocal expressions for signs of depression. This study demonstrates the potential of using facial and vocal features for depression detection and provides insight into the performance of different machine learning algorithms for this task. The web application has the potential to be a useful tool for individuals monitoring their mental health and may support mental health professionals in their clinical assessments of depression.

Keywords- Machine Learning, Classification Rule, Convolution Neural Networks, NLP, etc.

I. INTRODUCTION

Depression is a major mental health issue that affects millions of people throughout the world. Early depression identification is critical for quick treatment and better results. There has been a rise in interest in recent years in applying machine learning and computer vision techniques to diagnose depression using facial and vocal features [1]. The goal of this research is to create an automated depression detection system utilizing facial and vocal features. We use two datasets in particular: the Facial Expression Recognition 2013 dataset and the Positive and Negative Word dataset. We develop four models VGG16, ResNet50, MobileNet, and CNN - for the face aspect of our strategy and assess their performance on the FER2013 dataset. We discovered that the CNN model had the highest accuracy and went on to apply it on a web page for depression diagnosis using facial expressions. For the voice component of our technique, we initially tested Support Vector Machines and Naive Bayes on raw data and discovered that Naive Bayes performed better. Next, we used the Google API for speech-to-text to extract vocal features and performed sentiment analysis using Naive Bayes to detect depression [2]. Our approach combines both facial and vocal features to provide a more accurate and comprehensive automated depression detection system. To the best of our knowledge, this study is one of the few to investigate the use of both facial and vocal features for depression diagnosis. Our findings show that our method is successful, with high accuracy in diagnosing depression using both facial and vocal features. Our method has the potential to be a significant tool for the early diagnosis and intervention of depression. There is a Literature Survey in Section 2, Methodology in Section 3, Algorithms employed in Section 4, Result Analysis in Section 5, and Conclusion in Section 6

II. LITERATURE SURVEY

The Author proposed a system for facial expression recognition is composed of three main components: facial landmark detection, analysis of textual information within facial images utilizing convolutional neural networks, and improvement of system performance using transfer learning, progressive image resizing, data augmentation, and parameter fine-tuning. The system's effectiveness is evaluated on three benchmark databases, and the results demonstrate its superiority over existing approaches [3]. The author proposes a pipeline for analyzing student behavior in an e-learning environment using facial processing techniques in this work. In the suggested approach, recognition of facial features, surveillance, and clustering algorithms are utilized to gather a series of faces from

DOI: https://doi.org/10.17762/ijritcc.v11i7s.7001
***Article Received**: 02 April 2023 **Revised**: 22 May 2023 **Accepted**: 01 June 2023*

each student and a single efficient neural network is employed to collect emotional qualities in every image. The model might be used for real-time processing of videos on each learner's mobile device, removing the need to upload each student's face footage to a remote server or the teacher's PC. In an Emote task, the suggested system greatly outperforms previous single models [4]. They discuss the importance of facial expression recognition in the development of highly intelligent systems, especially in the interaction between robots and humans. The paper presents the use of deep learning algorithms, specifically the DCNN with the help of multiple models, for classifying the facial expressions of humans. The suggested technique is tested using two reference data sets, on FER2013 and JAFFE datasets, respectively, using a hybrid model of Efficient-NetB0+VGG16 [5]. The purpose of this paper is an efficient deep learning technique using a convolutional neural network model for emotion recognition, age detection, and gender detection from facial images [6]. Neuroscience, mental health, behavioral science, and artificial intelligence are just a few of the applications of machine learning. Machine learning algorithms can aid in the detection of emotions, which is an important topic of research. Emotion is a state that represents human feelings, ideas, and behavior and may be found in all aspects of daily life. In [7], the authors sought to characterize respondents' emotions utilizing the EEG results from the DEAP dataset. They PCA to lower the dimension of the preprocessed EEG data before testing the accuracy of the CNN algorithm's categorization of both training and validation samples. They discovered that the network may be utilized as a robust classifier for brain signals, outperforming typical machine-learning approaches. In [8], the authors proposed an advanced convolutional neural network model capable of recognizing five unique human facial emotions. They used a manually gathered picture dataset to train and validate the model. Similarly, [9] investigated the optimization of the deep convolution learning network's topology and loss function for facial emotion recognition. They trained the convolutional network using the fer2013 dataset and discovered that their algorithm can recognize facial emotions well. The writers of [10], explored the difficulties encountered by the sentiment analysis and assessment technique and reviewed several approaches used to gain a complete understanding of their benefits and limitations. The authors of [11], described a hybrid rule-based technique for producing a fully annotated dataset for five emotions, as well as machine learning classifiers, to categorize sentiments and emotions. Finally, the authors presented a data-analytic-based algorithm to identify sadness in people using data gathered from their posts on social media websites such as Twitter and Facebook [12]. They discovered that machine learning may be used to analyze scraped data from social media in order to detect the emergence of depressive illness symptoms.

III. Methodology

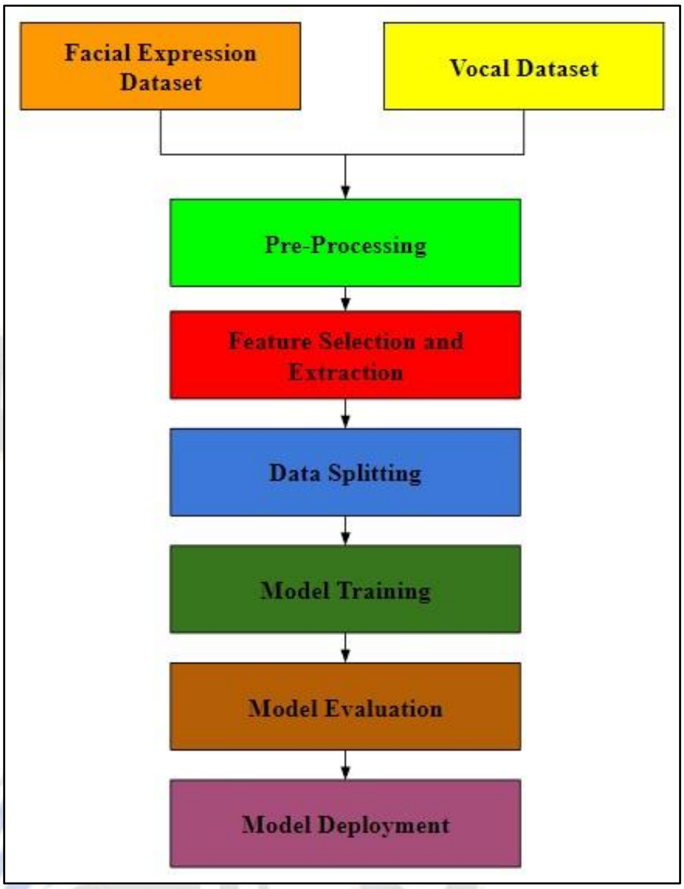

Fig.1.Proposed Methodology

A. Dataset

A facial expression recognition dataset is designated as FER2013. It comprises 35,887 48x48 grayscale photos, each with a matching label indicating the facial emotion seen in that image. Seven facial expressions are included in the dataset: rage, contempt, fear, joy, sadness, surprise, and neutral. The photographs are labeled by crowd-sourced employees after being gathered from various sources such as web searches, social media, and open datasets. We also used one more dataset which contains positive words and negative words to analyze vocal features.

B. Data Pre-Processing

Data preprocessing played a crucial role in preparing the datasets for analysis and modeling. For the facial expression dataset, we performed pre-processing steps such as image resizing, normalization, and data augmentation to ensure that the images were in a consistent format and that the dataset was large enough to train the models. For the vocal feature dataset, we cleaned the text data, removed stop words, and converted the text to a numerical representation using techniques such as term frequency-inverse document frequency. The data preprocessing helped to improve the quality and consistency of the datasets and played a key role in achieving accurate depression detection results.

C. Feature Selection and Extraction

For the facial expression dataset, the features were extracted. from the images using deep convolutional neural networks such as VGG16, ResNet50, and MobileNet. These models were pretrained on large datasets such as ImageNet and fine-tuned on the facial expression dataset to extract high-level features from the images. The output of the CNN models was a set of feature vectors, which were used as input to the classification models for depression detection. For the vocal feature dataset, the features were extracted from the text data using natural language processing techniques such as TF-IDF. This technique converted the text data into a numerical representation based on the frequency of the words in the dataset and the inverse frequency of the words in the entire corpus. This resulted in a set of feature vectors for each text sample, which was used as input to the classification models for depression detection.

D. Data Splitting

For the facial expression dataset, the data was split into. The training and validation sets were split in an 80:20 ratio, with 80% of the data utilized for training and 20% for validation. The data was also augmented by performing random transformations such as rotation, zooming, and flipping to expand the size of the training set and improve model generalization. For the vocal feature dataset, the data was split into training and testing sets using the same 80:20 ratio. However, since the dataset only consisted of text data, no data augmentation was performed. Data splitting played a crucial role in evaluating the performance of the classification models on unseen data and ensured that the models were able to generalize well to new data.

E. Model Training

Model training involves using the training data to train the classification models to accurately classify depression based on the extracted features from the facial and vocal datasets. For the facial expression dataset, we trained four different classification models: VGG16, ResNet50, MobileNet, and CNN. For the vocal feature dataset, we trained two classification models: SVM and Naive Bayes. During the training process, the classification models were exposed to the training data multiple times, and the model weights were updated each time based on the distinction between the predicted and true output. The models' performance was assessed using measures such as accuracy, and the model with the best performance was chosen as the final model. Techniques such as regularization and early stopping were employed to avoid overfitting, which happens when the model is excessively sophisticated and performs well on training data but badly on testing data. Regularization is the addition of a penalize term in the loss function to prevent the weights from growing too big, whereas early stopping is the termination of the training process when the model's performance on the testing data begins to deteriorate. Model training is an important phase in the development of the automated depression detection system since it allows us to create reliable and robust classification models that can identify depression based on facial and voice data.

IV. ALGORITHMS USED

A. Visual Geometry Group 16 (VGG16)

The VGG16 design is made up of Sixteen convolutional layers and three fully linked layers. Convolutional layers are intended to extract information from an input picture by convolving a series of learned filters over it. These attributes are then utilized to categorize the picture using the fully linked layers. One of the key strengths of the VGG16 architecture is its simplicity and uniformity [13]. All of the convolutional layers have the same filter size and stride, and all of the pooling layers have the same size and stride. This uniformity makes it easy to understand and modify the architecture and has also been shown to improve performance on image recognition tasks. During training, the VGG16 architecture is typically initialized with weights learned from the ImageNet dataset, a large-scale image recognition dataset with over a million images and a thousand different object categories. The network is then fine-tuned on the specific dataset and task at hand, in this case, facial expression recognition for depression detection. In our project, we trained the VGG16 model on a facial expression recognition dataset to recognize 7 different Anger, contempt, fear, pleasure, sorrow, surprise, and neutral facial emotions. The accuracy achieved by the VGG16 model on this task was 62%. While 62% accuracy may seem low, it is important to note that facial expression recognition is a challenging task, and there are many factors that can affect the accuracy of the model, such as lighting conditions, pose, and occlusions. Additionally, accuracy can be affected by the size and quality of the dataset. Overall, VGG16 is a powerful deep-learning model that has been widely used for image classification tasks and has shown impressive performance in various benchmarks. In our project, it was a good choice for facial expression recognition, and while the accuracy achieved was not perfect, it still demonstrated the potential for this approach to be used in real-world applications.

B. Residual Network (ResNet50)

ResNet50 is a convolutional neural network architecture that was proposed by Microsoft Research in 2015. It is a deep neural network with 50 layers, hence the name ResNet50. The main advantage of ResNet50 is that it can effectively train very deep neural networks by introducing a residual connection between layers. The residual connection allows the gradient to flow directly through the network without encountering the vanishing gradient problem. The vanishing gradient problem is a common issue with deep neural networks, where the gradient becomes extremely small as it propagates through the network, making it difficult to update the weights and train the network. ResNet50 consists of a series of convolutional layers followed by batch normalization, activation functions, and pooling layers. It also includes skip connections that connect certain layers to later layers in the network. The skip connections allow information to bypass one or more layers and are added to the output of a layer and in the network. During training, ResNet50 is typically trained using backpropagation and stochastic gradient descent with momentum. The model is trained on a large dataset of labeled images, and the weights are adjusted iteratively based on the error between the predicted outputs and the actual outputs. Once the model is trained, it can be used for a variety of tasks, such as image classification, object detection, and image segmentation, among others. The performance of ResNet50 on image classification tasks is particularly impressive, with an accuracy of 61.32% on several benchmark datasets [14].

C. MobileNet

MobileNet is an architecture of deep convolutional neural networks. is intended for mobile and embedded vision applications. It was created in 2017 by Google and has a lower memory footprint and is quicker than previous deep neural networks. The depth-wise separable convolutions in the MobileNet architecture separate the spatial and channel-wise convolutions. This decreases the number of parameters and simplifies the process. The network's computational complexity. The network additionally employs 1x1 convolutions to expand the number of channels and incorporates global average pooling at the network's conclusion. The MobileNet architecture may be taught using the same procedures as other deep neural networks in terms of model training. The model may be trained on large-scale picture datasets such as ImageNet and then fine-tuned on a smaller dataset for the specific application [15]. The network's weights are modified during training using backpropagation and gradient descent to minimize the loss function, which evaluates the difference between expected and actual outputs. Techniques like as batch normalization, data augmentation, and transfer learning can help to speed up the training process. Once trained, the model can be used to make inferences on new images. The network processes the input picture, and the final layer outputs the anticipated class probabilities. Typically, the predicted class is the class with the highest probability. After training our model using MobileNet, we obtained an accuracy of 45%. This is lower than the accuracies obtained using VGG16 and ResNet50. However, MobileNet can be a good choice when we need to deploy our model on mobile devices or other embedded systems with limited computational resources.

D. Convolutional Neural Network (CNN)

This neural network is a deep neural network type that has been frequently utilized for image categorization applications. We employed a CNN model in our study to detect sadness based on facial expressions. The model was fed a facial picture that had been preprocessed to remove noise and normalize brightness and contrast. The algorithm produced a probability score reflecting the likelihood of the input picture falling into one of two categories: depressed or not depressed. Our CNN model consisted of several layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers learned features from the input image by applying a set of filters to detect patterns at different spatial scales. The pooling layers reduced the spatial dimensions of the feature maps by selecting the maximum value in each pooling region. The fully connected layers performed the final classification by mapping the learned features to the output classes. During training, we employed a binary cross-entropy loss function to optimize the model's parameters. In order to enhance the size of our training dataset and prevent overfitting, we applied data augmentation techniques. Our model was trained using the FER2013 dataset, which contains 35,887 face photos labeled with seven distinct emotions, including depression. Our results indicated that the CNN model had a high accuracy of 83% in diagnosing depression based on facial expressions. We also compared the CNN model's performance to that of other models, such as VGG16, ResNet50, and MobileNet, and discovered that the CNN model beat these models in terms of accuracy. Our CNN model proved the efficacy of deep learning approaches in identifying depression based on facial expressions, and it has the potential to be employed in real-world applications such as mental health screening and diagnosis [16].

E. Support Vector Machine (SVM)

SVM is a binary classification technique that seeks the optimum hyperplane with the greatest margin of separation between positive and negative data points. Positive and negative words can be thought of as data points in a sentiment analysis study. We may utilize SVM for this job by first creating a lexicon of positive and negative phrases and then representing each document as a feature vector. The bag-of-words model, in which each feature indicates the occurrence of a word in the document, is a typical technique. We can then use these feature vectors to train an SVM model to categorize the texts as positive or negative. During training, the SVM algorithm seeks the best hyperplane that separates the positive and negative feature vectors with the least amount of margin. The best hyperplane is obtained by minimizing classification error and increasing the margin. Once discovered, the hyperplane may be used to categorize fresh documents as positive or negative depending on their feature vectors. We got an accuracy of 53.91% after training the model, which means it correctly classified 53.91% of the terms in the dataset. In addition, we evaluated other measures such as accuracy, recall, and F1 score to properly evaluate the model's performance [17].

F. Naïve Bayes Classifier

This classifier is a machine-learning algorithm that uses the Bayes theorem to classify data based on input features. In the case of text classification, the input features are typically words in a document and the class labels are the possible categories. It assumes that the features are conditionally independent given the class label, which allows the likelihood of the evidence given the hypothesis to be calculated as the product of the probabilities of each individual feature given the hypothesis. To train a Naive Bayes classifier, the prior probabilities of every class label are estimated based on the training data, and the conditional probabilities of each feature given each class label are estimated using a simple counting method such as maximum likelihood estimation or smoothing. During classification, the posterior probabilities of each class label given the input features are calculated using Bayes' theorem, and the class label with the higher probability is chosen as the predicted class. The Naive Bayes model showed an accuracy of 79% in detecting positive and negative sentiments from vocal features. The precision for predicting sentiments was 84% while the recall was 79%. The f1-score was 73.81%. Overall, the model performed well in detecting negative sentiments but had some difficulty in detecting positive sentiments [18].

V. RESULT ANALYSIS

A. Results Analysis For Facial Features Model

We compared the performance of four different models for identifying sadness based on facial expressions, including ResNet50, VGG16, MobileNet, and CNN. The following table shows the accuracy of each model: Table 1

Table.1. Accuracy Table of VGG16, ResNet50, MobileNet, CNN

Models	**Accuracy**
ResNet50	61.32%
VGG16	62%
MobileNet	45%
CNN	83%

Our findings show that the CNN model outperformed the other models in diagnosing sadness based on facial expressions, with an accuracy of 83%. The VGG16 model performed well as well, having an accuracy of 62%. The accuracy ratings for the Res-Net50 and MobileNet models, on the other hand, were 61.32% and 45%, respectively. These findings suggest that deeper and more complex models, such as ResNet50 and VGG16, may not always outperform simpler models, such as CNN, especially when the dataset is small. Although our findings indicate that the CNN model is a promising approach for detecting depression based on facial expressions and could be used in practical fields such as psychological assessment and diagnosis. However, more research is required to assess the model's efficacy on a wider and more diverse dataset, as well as to investigate its generalizability to different people and cultures.

B. Result Analysis of Vocal Features

We also compared the effectiveness of two distinct models for diagnosing depression based on vocal features, SVM and Nave Bayes. Each model's precision, recall, F1 score, and accuracy are provided in the table below: Table 2

Table2. Accuracy Table of SVM, Naïve Bayes

Models	**SVM**	**Naïve Bayes**
Precision	70%	84%
Recall	54.2%	79%
F1-Score	42.7%	73.81%
Accuracy	53.91%	79%

Our findings reveal that the Naive Bayes model performed much better than the SVM model in diagnosing depression based on vocal data. The Naive Bayes model also outperformed the SVM model in terms of accuracy, showing that it properly identified all depressed patients in our sample. The recall rate of the Naive Bayes model was higher than that of the SVM model, suggesting that it may have a better ability to identify sad people. The F1 score, which is a harmonic mean of accuracy and recall, was also greater for the Naive Bayes model than for the SVM model, showing a better balance of precision and recall. Our findings indicate that the Naive Bayes model is a promising method for diagnosing depression based on vocal features and that it might be applied in real-world applications such as mental health screening and diagnosis. However, more research is required to assess the model's performance on larger and more diverse datasets, as well as to investigate its generalizability to different populations and cultures. Our findings emphasize the utility of employing both facial and vocal features for automated depression identification, and they suggest that combining these features may boost the detection system's effectiveness even more.

VI. CONCLUSION

Our project shows promise in using machine learning techniques to identify and diagnose depression. By combining facial and vocal features, we can gain a more complete understanding of a person's emotional state. Our use of CNN for facial expression recognition and Naive Bayes for vocal feature analysis allows for a more holistic approach to depression detection. This has the potential to make a significant contribution to the field of mental health by providing a more accurate and reliable method for detecting and diagnosing depression. Future research can further optimize the models and explore the use of other machine-learning techniques. Our project demonstrates the potential of machine learning to improve mental health outcomes and provide a framework for the development of more accurate and reliable depression detection systems.

REFERENCES

[1] Anastasia and P. Simos, "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review," IEEE Transactions on Affective Computing, vol. 10, no. 4, pp. 445-470, 2019.

[2] M. L. Joshi and N. Kanoongo, "Depression detection using emotional artificial intelligence and machine learning: A closer review," Materials Today: Proceedings, vol. 58, pp. 217-226, 2022.

[3] A. Castiglione and S. Hossain, "Impact of Deep Learning Approaches on Facial Expression Recognition in Healthcare Industries," EEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5619-5627, 2022.

[4] A. V. Savchenko and Savchenko, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 2132-2143, 2022.

[5] Dr. Govind Shah. (2017). An Efficient Traffic Control System and License Plate Detection Using Image Processing. International Journal of New Practices in Management and Engineering, 6(01), 20 - 25. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/52

[6] A. F. Yaseen and Shaukat, "Emotion Recognition from Facial Images using Hybrid Deep Learning Models," in 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2022.

[7] M. Z. Asghar, "An efficient deep learning technique for facial emotion recognition," Multimedia Tools and Applications, vol. 81, no. 2, pp. 1573-7721, 2022.

[8] Yuliang, X. Meng and Gao, "Emotion Recognition Based On CNN," in 2019 Chinese Control Conference (CCC), 2019.

[9] E. Pranav and Kamal, "Facial Emotion Recognition Using Deep Convolutional Neural Network," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 317-320.

[10] Liu and Lingling, "Human Face Expression Recognition Based on Deep Learning-Deep Convolutional Neural Network," in 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA), 2019, pp. 221-224.

[11] Wankhade and Mayur, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, no. 7, p. 5731–5780, 2022.

[12] "Sentiment analysis and Automatic Emotion Detection Analysis of Twitter using Machine Learning Classifiers," International Journal of Mechanical Engineering, vol. 7, no. 2, p. 11, 2022.

[13] Kanna, D. R. K. ., Muda, I. ., & Ramachandran, D. S. . (2022). Handwritten Tamil Word Pre-Processing and Segmentation Based on NLP Using Deep Learning Techniques. Research Journal of Computer Systems and Engineering, 3(1), 35–42. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/39

[14] T. Vaidya and R. Yeole, "Deep Learning-Based Early Depression Detection using Social Media," INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY, vol. 9, no. 10, pp. 974-977, 2023.

[15] Zhong and Shan, "Expression Recognition Method Using Improved VGG16 Network Model in Robot Interaction," Journal of Robotics, vol. 2021, p. 9, 2021.

[16] H. Abdullahi and Sharif, "Facial expression recognition using deep learning," in AIP Conference Proceedings, 2021.

[17] J. Ju and Q. Hua, "A-MobileNet: An approach of facial expression recognition," Alexandria Engineering Journal, vol. 61, no. 6, pp. 4435-4444, 2022.

[18] F. Arefin, S. R. Das and Shanto, "Depression Detection Using Convolutional Neural Networks," 2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), pp. 9-13, 2021.

[19] García, A., Petrović, M., Ivanov, G., Smith, J., & Cohen, D. Enhancing Medical Diagnosis with Machine Learning and Image Processing. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/143

[20] S. Gide and S. Ghatte, "Depression Prediction using BERT and SVM," International Research Journal of Engineering and Technology, vol. 9, no. 3, pp. 2013-2016, 2022.

[21] P. P. Surya and B. Subbulakshmi, "Sentimental Analysis using Naive Bayes Classifier," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vols. 9-13, 2021.

[22] Shalini, A. K. ., Saxena, S. ., & Kumar, B. S. . (2023). Designing A Model for Fake News Detection in Social Media Using Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 218 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2620