Deep Neural Networks based Music System using Facial Emotion

Bharani D [1], LakshmiPriya V [2], Santhi P[3]

[1,2,3]*Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India*

Abstract. In recent days, Facial emotion Recognition has become an interesting sector and it has fascinated many developers and other common people for its own prominent features. Even though it has many difficulties in recognizing the emotions of human individuals due to the diversity of emotions and ambiguity. Our paper discusses a prominent model developed to overcome the drawbacks of conventional Facial Emotion recognition using Deep Learning. The model has been employed with Convolution Neural Networks to recognize spatial hierarchies and patterns which leads to extracting the features of the face from the image being taken in real-time and subsequently, it categorizes the emotion. Pre-trained CNN models such as VGG, ResNet, and MobileNet are utilized to enhance the performance of the model. The dataset used in the development of the model contains 35,000+ images for training and testing purposes which categorize human emotion into seven types such as anger, disgust, fear, happy, neutral, surprise, and sad. Here, the color image is converted into a grayscale image followed by data augmentation is used to improve the dataset for a better accuracy rate. Overall, the model distinguishes the human emotion detected and emphasizes multiple applications. This paper not only discusses Facial Emotion Recognition but also incorporates with Multimedia system for a real-time application. It established the Automatic music player handled by Facial Emotion recognition using deep learning. The music player has a list of music assigned for each emotion and it dynamically plays the music based on the user's emotional state. Collectively, the fusion of Facial Emotion Recognition with a music player offers a prominent solution for future breakthroughs in technology.

Keywords: *Facial Emotion Recognition, Convolution Neural Network, Pre-trained model, Music Player.*

1 Introduction

Facial Emotion Recognition (FER) is becoming an important user interaction system involving Artificial Intelligence which has foremost benefits in computer Vision, Emotional-aware Technology, and psychological research. It overcomes the conventional difficulties such as diversity in emotion and ambiguity faced during the emotion recognition time. Deep learning Techniques are utilized to transform the Facial Emotion Recognition model to figure out patterns and detect the face from the image. This method outperforms existing handcrafted feature extraction approaches, enabling automated emotion identification across varied populations, cultures, and circumstances. The introduction of convolutional neural networks (CNNs) into FER has significantly advanced the discipline, allowing for the collection of spatial hierarchies and context-sensitive data from face pictures. Various CNN designs, including VGG and customized networks, are being investigated for their ability to recognize a variety of facial expressions, from pleasure and sadness to rage and surprise. Furthermore, this study clarifies the obstacles and new research directions in FER, such as dealing with occlusions, addressing cross-cultural variances, and adapting to real-world, dynamic situations. Transfer learning and fine-tuning strategies are presented to improve the ability to generalize and real-time performance of FER systems.

The use of Artificial Intelligence in multimedia applications has changed user experience in recent years. This paper describes a revolutionary Music Player that operates by Facial Emotion Recognition (FER) and is powered by Deep learning algorithms. The system leverages a sophisticated Convolutional Neural Network (CNN) for real-time emotion analysis. Emotions such as anger, disgust, fear, happy, neutral, surprise, and sad can be detected by the model by extracting and evaluating the feature from the image that is being captured in real-time. These emotions evaluated by the model are assigned to the playlist of music where each emotion is assigned a set of songs suitable for that particular emotion. It automatically selects a song from the playlist based on the emotion that is being detected by the model which is the present state of the user. Additionally, the model has the capacity to adjust the playlist based on the individual's suggestion over time. Collectively, the combination of Facial Emotion Recognition and neural networks offers a feasible solution for future development in case psychological research as well as human and computer interaction.

2 Literature Overview

Human emotions may be correctly categorized into seven emotions. (Happiness, anger, sadness, fear, surprise, disgust, and neutral). The manipulation of complex face muscles conveys human facial emotions. This frequently subtle and complex indication of speech also provides a wealth of information about the condition of our thinking. We can evaluate the impact of the resources and services made available to consumers using a straightforward and affordable technique based on facial expressions. This study uses a model based on the Eigenface technique, which computes the Euclidean and eigenspace separation between pictures, to convert faces into grayscale images in order to determine the emotion and categorize it into five categories. utilized the HAAR Cascade classifier in order to identify a face in an image. Moreover, MobileNet, a pre-trained model, was utilized to increase the model's accuracy[1]. The paper employed statistical data to assess facial changes over time, record those characteristics, and categorize them into six emotional states. To train the model, a two-layered Feedforward neural network was used. The Scaled Conjugate Gradient back propagation technique is employed to test the model, and it achieves a success rate of 92.2%[2]. This paper proposes a two-level strategy in which the primary level merely recognizes the face in the picture and the secondary level determines the facial characteristics in the image. It categorizes the face into five emotions based on the retrieved characteristics. It used FERC to increase accuracy [3].

This paper offers a Deep CNN model utilizing a transfer learning approach in which the model is replaced with dense layers of data and a Pipeline strategy is employed in which each layer modifies the pre-trained DCNN model, resulting in increased accuracy. It also used VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, and DenseNet-161 pre-trained models. Working with several photos posed challenges for the model. DenseNet-161 outperformed KDEF and JAFFE datasets in terms of accuracy, with 96.51% and 99.52%, respectively[4]. This paper discusses some of the challenges associated with traditional FER, such as mutual optimization of feature extraction and classification. To improve the model's accuracy, it presented a strategy using Transfer learning and other pre-trained models such as Inception V3, MobileNet, ResNet50, and VGG90. In addition, they used their own fully connected layers instead of ConvNet completely connected layers for each instruction, resulting in an accuracy of 96% for the research utilizing the CK+ database[5]. The paper created a hybrid model combining Deep CNN and HAAR Cascade Architecture to identify emotion in real-time and static images. It includes a ReLU activation function as well as several kernels to improve feature extraction. HAAR Cascade, on the other hand, is used to extract features from real-time pictures. Also includes a comparison of other standard model performances, with the conclusion that the suggested technique improves performance while taking less time to execute (2098.8s)[6]. This paper develops an Automatic Emotion Recognition system that may be used in a variety of situations using Deep Learning. It featured several architectures for improved performance, as well as the database utilized, and offered a thorough progress via comparative analysis of the suggested technique and the results acquired[7].

This paper created a FER App to collect real-time photos that can be accessed from a desktop, mobile, IOS, or Android device. It used CNN to train the dataset to recognize the emotion and choose an appropriate music that is suited for the user's emotion with a greater accuracy rate. It built an interface that was linked to the backend and trained with 28000 photos to categorize emotions with an accuracy of 85% in training and 83% in testing[8].This paper developed a CNN model to recognize aspects of the face, which aggregates the identified features to determine the emotion and automatically adjusts the songs, resulting in less time spent selecting a song based on the user's emotion. It used CNN techniques to analyze the FER 2013 dataset[9]. This paper created a CNN model using a five-layer model with global average pooling. It used CNN with a Transfer learning model and certain pre-trained models to boost accuracy, such as ResNet50, SeNet50, and VGG16. It merged the FER model with a music player, which dynamically plays music when emotion is identified in real-time [10]. Noises in the images are removed using filtering-based approach [11]. Correlation and the classification plays an vital role in the machine learning approach [12]. Bi-LSTM is combined with CNN for giving the better accuracy in heart disease prediction system [13]. Deep belief network is for the purpose of disease prediction. This method produced the better accuracy [14].

3 Problem Statement

Facial Emotion Recognition (FER) is a crucial component of affective computing that attempts to offer robots the ability to recognize and respond to human emotions via facial expression analysis. In recent years, the interaction of emotion and music has become a significant component of human experience. Music has the ability to elicit, enhance, and reflect a wide range of emotions. The main objective of the project is to develop a Facial Emotion Recognition based music player where it can be used as a real-time solution in many cases. The model will automatically select a song from the playlist based on the emotion that is being detected. It dynamically adjusts the playlist based on the user's suggestion over time. it was possible by combining the Convolution Neural network model. Pre-trained models were utilized to increase the performance as well as to decrease the execution time.

4 Research Motivation

The motivation to do research and develop a Facial Emotion Recognition-based music player started from a thought "Why can't music change automatically based on the changes in the mood of a person". The main goal of this project was to access the music in an effortless manner and also it should suit the user's mood. As the mood of the person changes, the model has been designed to change the music immediately. The model will update itself based on the user's suggestion over the time provided by the user. Primarily it was designed to take less time to play a piece of music than the manual procedure of selecting a music based on the user's choice. The practical aspect is investigating cross-modal integration, which involves mixing expressions on the face with other signals from the surroundings. Aside from the study, the system's economic feasibility and possible influence on the music broadcast business have a practical reason. The system's capacity to attract and keep customers via its unique and customized features has the potential to position it as a game changer in the industry.

5 Proposed Method

For our project, we used a dataset of static images to train the model to detect the face and analyze the emotion in real time and it plays music according to the emotion it detects. The Proposed method of the project is mentioned below :

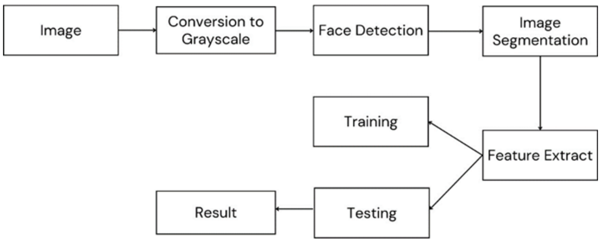


Fig.1.Proposed method

The flowchart effectively describes the whole procedure. The image is taken from the dataset first, followed by data preprocessing, which includes grayscale conversion from RGB, identification of faces, and visual scaling, all of which are key aspects of data preparation. After that method is accomplished, the flow proceeds with segmentation of the image and feature extraction. Some data is used for training, while others are used for testing. The classifier is used to train and test, and the results are shown. the ultimate outcomes of the recognition. After receiving the emotion detection result from the real-time video, it analyses the emotion and plays music appropriate / suited to the emotion identified.

6 Methodology

The accuracy was obtained by using the described dataset with the goal of moving the approach for deciding on the data, evaluating the findings, training the data to choose the appropriate method that would deliver high precision and remain relevant regarding the issue, and extracting the characteristics from the image that was provided with the goal of creating a list of characteristics that could assist the model, using stimulation and extension for the model to perform well.

6.1 Data Preprocessing

The first process that is to be undertaken is the acquisition of the dataset that supports the model, we are supposed to select an appropriate dataset that will be helpful during the simulation to give preciseness and accuracy which can also apply in other places. We start by importing some predefined Python libraries to do the data preparation. The following is libraries used for particular tasks. Now we need to import the dataset that's been specified in the model. The face detection algorithms, that we employ, are very powerful in locating and extracting the facial regions from photos or frames. This information is used to accurately align and normalize the face images so as to maintain a consistent kind of input while training our neural network.

6.2 Segmentation

This technique scans for certain areas of interest within facial images hence producing higher precision and accuracy in recognizing emotions. Instead of the whole face region as in other techniques, the attention focuses on other specific features, that is, the eyes, brows, nose, and lips plus others which are imperative in emotion identification. This segmentation technique helps us to extract fine granularity characteristics, and it is useful in detecting small differences in expressions. The segmented components form the very basis of our neural network model. Each of the separated and standardized components now harbors abundant depth of information as regards to the presented complex expressions evident in the face area. This degree of granularity, no doubt, will highly enhance detection capabilities of such subtle emotional cues.

6.3 Model development

This effort starts with a model design being critically selected, and naturally in this case especially, Convolutional Neural Networks (CNNs) would easily be the natural choice since they perform much better than many other kinds of models on any kind of image-related tasks. Specify the number and type of layers between input and output, including feature extraction by means of convolutional layers, spatial down sampling by means of pooling layers, and high-level feature integration through fully connected layers. Represent in the output layer the number of emotional classes which we would like to recognize. Categorical cross-entropy serves as a guide to the training process since it ranks in one of the most suitable loss functions in multi-class classification problems. It ensures that model optimization is carried out at acceptable rates. After compilation, The model has been prepared for being trained. We monitor its progress using the pre-processed training data keeping metrics on the training and validation sets. This allows us to assess the model's performance and find areas for improvement.

.6.4 Features Extraction

In this step we extract information, from processed facial photos. This helps our model recognize patterns and variations in emotions. It plays a role in reducing data complexity while retaining information necessary for identifying emotions. By organizing layers, within our network design we enable the model to distinguish subtle textures, shapes and spatial relationships among facial images.

6.5 Training and Testing the Model

These processes are meticulously planned to ensure that our neural network not has the ability to identify emotions but also provides swift results instantaneously.

Throughout the training phase the neural network undergoes a process of understanding patterns and features in pre-processed facial photos. We provide the model with a set of training examples that encompass emotional expressions. By going through backward passes the network adjusts its internal parameters gradually enhancing its capacity to predict emotions. As the training progresses, we consistently assess the model’s performance, on both the training and validation sets. The evaluation conducted simultaneously ensures that our neural network not learns from the training data but also expands its expertise to unfamiliar samples.

6.6 Face & Emotion Detection

Our project relies heavily on technology, for face and emotion detection. The robust system, for identifying faces meticulously analyses data to identify and isolate facial components. This initial phase can be likened to shining a flashlight on a painting of emotions uncovering the canvas where our neural network works its magic. Once our system detects faces it utilizes a method, for detecting emotions. This component has been carefully calibrated to recognize patterns and characteristics in expressions allowing it to discern a wide range of emotions including pleasure, sadness, surprise and more.

6.7 Music Connection

The integration of music serves as a connection, between technology and human emotions. This dynamic collaboration transforms the Music Player into a personalized experience that goes beyond music listening. As our neural network analyzes time expressions to determine emotions it triggers a series of musical selections tailored to the user’s current emotional state The dataset we utilized for this project functions as a guiding compass, for our Music Player allowing it to resonate in harmony with the user's emotions. This creates an experience that goes beyond music playback. The dataset itself is a collection of musical choices each labeled with corresponding emotions. What makes this approach truly remarkable is its ability to adapt and evolve. The seamless interaction between music and emotions surpasses the limitations of an interface forging a deep connection, between technology and the human spirit. This learning process ensures that the Music Player not only selects music that aligns with the expressed emotion but also does so in a way that feels truly personalized to the user.

7 Results and Discussion

We have developed the model to detect the face from an image that is being captured and identify the emotion automatically. It has been trained and tested over 100 epochs using the dataset of 37000+ images for analyzing the emotion. We developed this project using Visual Studio 2022 IDLE, and since it uses real-time snapshots to identify the emotion, a camera must be present in the system to capture the emotion. Additionally, music is played while the emotion is being detected using a playlist dataset, where each emotion has a specific number of songs assigned to it. These are the seven emotions that our algorithm can recognize in real-time and identify overall.

- Angry
- Disgust
- Fear
- Happy
- Neutral
- Sad
- Surprise

Also, we have employed certain pre-trained models in our project. To train our model in a more effective way

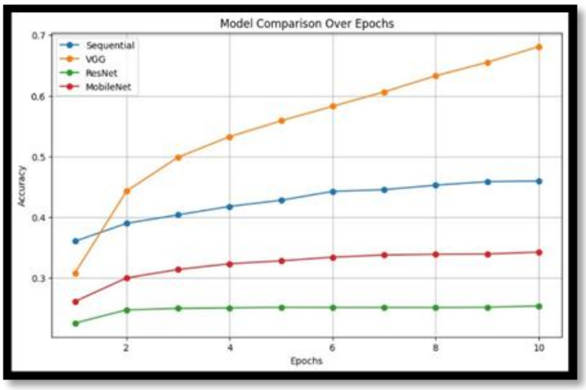


Fig.2. Accuracy of Training

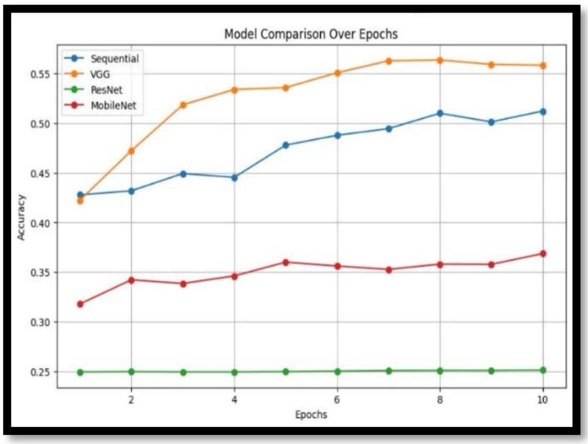


Fig.3. Accuracy of Validation

Finally, we performed a comparison between Sequential and several pre-trained models. The accuracy comparison done between the models with variable epochs is indicated here as the key requirement to be examined.

Table 1. Comparison of accuracy obtained from various CNN models - Training

Epoch	Sequential	VGG	ResNet	MobileNet
1	0.3604	0.3082	0.2256	0.2612
2	0.3899	0.4430	0.2474	0.3000
3	0.4039	0.4988	0.2498	0.3139
4	0.4178	0.5327	0.2507	0.3234
5	0.4280	0.5589	0.2516	0.3282
6	0.4426	0.5828	0.2514	0.3342
7	0.4454	0.6065	0.2514	0.3378
8	0.4528	0.6330	0.2512	0.3390
9	0.4586	0.6552	0.2517	0.3394
10	0.4597	0.6811	0.2538	0.3426

Table 2. Comparison of accuracy obtained from different CNN models - Validation

Epoch	Sequential	VGG	ResNet	MobileNet
1	0.4278	0.4224	0.2494	0.3179
2	0.4319	0.4723	0.2497	0.3422
3	0.4493	0.5185	0.2494	0.3385
4	0.4454	0.5339	0.2494	0.3461
5	0.4779	0.5358	0.2497	0.3600
6	0.4878	0.5508	0.2502	0.3561
7	0.4945	0.5628	0.2508	0.3527
8	0.5100	0.5637	0.2510	0.3580
9	0.5013	0.5592	0.2510	0.3578
10	0.5122	0.5584	0.2513	0.3686

From the above tables, we would see a detailed accuracy comparison of different CNN models such as Sequential, VGG, ResNet, MobileNet at different epoch in their Training and Validation. So From above comparison, we can conclude that VGG has a better accuracy rate than other model used for comparison. Here the mathematical formula used to calculate the accuracy of these model are mentioned below.

1. Calculate the accuracy of the Sequential model

$$A(\%) = \frac{\sum_{i=1}^{N} \delta(y_i\, \hat{y}_i)}{N} * 100 \qquad (1)$$

2.Calculate the accuracy of the VGG model

$$A_{VGG}(\%) = \frac{\sum_{i=1}^{N} \delta(y_i\, \hat{y}_{VGG,i})}{N} * 100 \qquad (2)$$

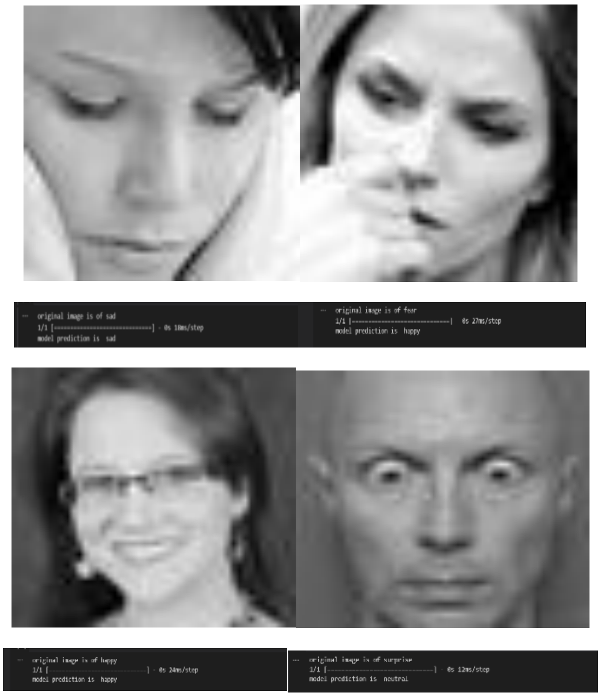
3.Calculate the accuracy of the ResNet model

$$A_{ResNet}(\%) = \frac{\sum_{i=1}^{N} \delta(y_i\, \hat{y}_{Resnet,i})}{N} * 100 \qquad (3)$$

4.Calculate the accuracy of the MobileNet model

$$A_{MobileNet}(\%) = \frac{\sum_{i=1}^{N} \delta(y_i\, \hat{y}_{MobileNet,i})}{N} * 100 \qquad (4)$$

Based on the accuracy, we got from the CNN models it produces an output of emotion detected from dataset and those results are mentioned below:



```
...  original image is of sad
     1/1 [==============================] - 0s 18ms/step
     model prediction is  sad
```

```
...  original image is of fear
     1/1 [==============================]   0s 27ms/step
     model prediction is  happy
```

```
...  original image is of happy
     1/1 [==============================] - 0s 24ms/step
     model prediction is  happy
```

```
...  original image is of surprise
     1/1 [==============================] - 0s 12ms/step
     model prediction is  neutral
```

Fig.4. Sample of Outputs

8 Conclusion

In the realm of interaction between humans and computers, our Facial Emotion Recognition (FER) Music Player project has passed a significant milestone. Through a comprehensive and iterative development process, we successfully built a system that not only recognizes emotions from expressions on the face in real-time but also tailors the audio experience to the user's emotional state. Our adventure began with the collecting of a broad dataset that had been properly labelled with emotional categories. We guaranteed that our model was accurate and resilient by deliberately constructing the architecture, employing rigorous data preparation approaches, and implementing regularization procedures.

The incorporation of real-time facial emotion detection into the Music Player interface is a quantum leap forward in user experience. The Music Player is able to dynamically curate playlists that connect with the user's current emotional state because to this seamless marriage of cutting-edge technology and emotional intelligence. It connects with the listener on an emotional level and provides a highly individualized aural experience. It not only showcases the power of AI in enhancing user experiences but also highlights the profound impact of technology on our emotional well-being. This project serves as a testament to the endless possibilities that emerge at the intersection of artificial intelligence and human.

9 References

1. Tatikonda Lakshman, Sai Sathvik Yadlapalli, Ramineni Vikramaditya, Bhimineni Rakesh Chowdary, Akshay Reddy Katakam, “FACIAL EMOTION DETECTION AND RECOGNITION” .2022

2. R.R. Londhe, ‘‘Analysis of facial expression and recognition based on statistical approach”, Int. J. Soft Comput. Eng. IJSCE 2012.

3. Bhoomika, Pushpalatha, “Facial Emotion Recognition using CNN”.2021

4. M.A.H.Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, Tetsuya Shimamura ,”Facial Emotion Recognition Using Transfer Learning in the Deep CNN”.2021

5. M. K. Chowdary, Tu N. Nguyen, D. Hemanth,”Deep learning-based facial emotion recognition for human–computer interaction applications”.2021

6. Ozioma Collins Oguine, Kaleab Alamayehu Kinfu, anyifeechukwu Jane Oguine, Hashim Bisallah,” Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction”.2022

7. Wafa Mellouk, Wahida Handouzi,” Facial emotion recognition using deep learning: review and insights”.2020

8. Ahmed hamdy Ahmed Ibrahim, ‘‘Emotion-based music player,” researchgate, 2019

9. Sekaran, R., Munnangi, A. K., Ramachandran, M., & Gandomi, A. H. (2022). 3D brain slice classification and feature extraction using Deformable Hierarchical Heuristic Model. Computers in Biology and Medicine, 149, 105990-105990.

10. Ramesh, S. (2017). An efficient secure routing for intermittently connected mobile networks. Wireless Personal Communications, 94, 2705-2718.

11. Sekaran, R., Al-Turjman, F., Patan, R., & Ramasamy, V. (2023). Tripartite transmitting methodology for intermittently connected mobile network (ICMN). ACM Transactions on Internet Technology, 22(4), 1-18.

12. S.k. Sana, G. Sruthi , D. Suresh , G. Rajesh , G.V. Subba Reddy ,”Facial emotion recognition based music system using convolutional neural networks”.2022

13. Sulaiman Muhammad; Safeer Ahmed; Dinesh Naik,”Real Time Emotion Based Music Player Using CNN Architectures.2021.

14. T.M. Nithya , P. Rajesh Kanna , S. Vanithamani , P. Santhi, “ An Efficient PM - Multisampling Image Filtering with Enhanced CNN Architecture for Pneumonia Classfication” , Biomedical Signal Processing and Control, Volume 86, Part C, 2023.

15. Palanisamy, Santhi, K. Deepa, and M. Sathya Sundaram. "Implementation of Machine Learning Models for Analyzing the Correlation and Classification of Complications in Pregnancy Using Amniotic Fluid." In Technological Tools for Predicting Pregnancy Complications, edited by D. Satishkumar and P. Maniiarasan, 289-302. Hershey, PA: IGI Global, 2023. https://doi.org/10.4018/979-8-3693-1718-1.ch017.

16. B D, M L, R A, Kallimani JS, Walia R, Belete B. A Novel Feature Selection with Hybrid Deep Learning Based Heart Disease Detection and Classification in the e-Healthcare Environment. Comput Intell Neurosci. 2022 Sep 28;2022:1167494. doi: 10.1155/2022/1167494. PMID: 36210997; PMCID: PMC9534609.

17. Madhavan, M., Gopakumar, G. DBNLDA: Deep Belief Network based representation learning for lncRNA-disease association prediction. Appl Intell 52, 5342–5352 (2022). https://doi.org/10.1007/s10489-021-02675-x.

18. Saravanan, T., & Saravanakumar, S. (2022, February). Modeling an Energy Efficient Clustering Protocol with Spider Cat Swarm Optimization for WSN. In 2022 IEEE VLSI Device Circuit and System (VLSI DCS) (pp. 188-193). IEEE.

19. A. S. Kumar, S. Annamalai, M. Kumaresan, P. Manikandan, R. Sekaran and H. A. Pai, "CNN-Based Analysis of Ultrasound Images for PCOS Diagnosis," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 347-350, doi: 10.1109/ICTACS59847.2023.10390451.