ML Approach for Musical Therapy Using Facial Expression and Voice Emotion Recognition

Mrs. S. Merena[1], Assistant professor,
Harini[2], R.Madhubala[2], G.Mounika[2]
[1]Assistant Professor, [2]Student, Department of Information Technology
Vivekanandha College of Engineering for Women, Elayampalayam, Tiruchengode, India.

ABSTRACT - In order to personalize musical therapy sessions, this research investigates a novel strategy that combines machine learning approaches with face expressions and speech emotion identification. The work makes use of audio analysis with librosa for voice emotion recognition and computer vision with OpenCV for facial feature extraction. In order to extract relevant features, the methodology preprocesses a heterogeneous datasets made up of facial and voice samples. After that, a machine learning model is trained with these attributes to enable automatic identification of emotional cues from both vocal intonations and facial expressions. By using parameters like accuracy, precision, and recall, rigorous testing is used to assess the efficiency of the suggested techniques. Findings point to a potentially useful ability to identify emotional states, providing the groundwork for a framework for individualized musical treatment. The implications of these findings for tailored therapeutic approaches are explored in the research. The use of voice and facial expression emotion detection in musical therapy creates new opportunities for therapies to be customized to the requirements of individuals, adding to the dynamic field of technology-assisted emotional well-being.

KEYWORDS
Emotional states, Musical treatment, OpenCV, Librosa, Speech emotion recognition, Facial Expressions.

1. INTRODUCTION

It has long been known that musical therapy is an effective means of promoting emotional expression, mental stimulation, and general well-being. A growing number of people are interested in utilizing technological developments to improve the therapeutic effectiveness of musical treatments due to the complex relationship that exists between music and human emotions. Within this framework, the combination of machine learning methods with facial expression and speech emotion detection shows promise for a more customized and adaptive method of musical therapy. The integration of machine learning with traditional musical therapy offers a data-driven dimension, whereas traditional musical therapy depends on the competence of therapists to interpret and modify therapies based on observed emotional cues.

The goal of the research is to close the gap that exists between the objective insights provided by computational analysis and the subjective subtitles of emotional states. With the integration of spoken emotion detection through audio analysis and facial expression recognition through computer vision, create a comprehensive system that can automatically detect and recommend a suitable song to the users to relax them. The method is explained, along with how to extract facial features and auditory characteristics. The combination of machine learning model training with these modalities. [1] We hope to clarify the efficacy of our method in identifying emotional states through empirical evaluation and results discussion, ultimately advancing musical therapy's development into a more flexible and responsive type of intervention. We hope that this research will contribute to our understanding of how machine learning could enhance the emotional impact of musical therapy and create a deeper relationship between technology and the nexus between technology and therapeutic practice.

Artificial intelligence and CNN use improve recognition accuracy; however, using large networks also increases computing costs. The introduction may draw attention to the rising body of research on the therapeutic uses of technology and the promise of machine learning (ML) for comprehending and addressing human emotions. In the context of musical therapy, talk about the importance of facial expressions and vocal emotion identification, highlighting their function in assessing emotional states. [1] Mention how ML integration aims to improve musical therapy's efficacy by offering individualized, real-time emotional support based on these inputs. Furthermore, recognize the current gaps in conventional therapeutic techniques that machine learning (ML) can fill in order to provide people seeking musical therapy with a more personalized and dynamic experience.

2. LITERATURE REVIEW

In order to better understand emotional states during musical therapy sessions, this paper aims to develop reliable emotion recognition models, personalized therapy recommendations, multimodal fusion, and explore possible relationships between modalities. Additionally, it aims to create a simple interface for therapists, enabling them to interpret and apply machine- generated insights in the context of therapeutic practices, as well as long-term emotional tracking and collaboration with experts.

2.1 KMT dynamics and review of emotion recognition

It aims to assess the state-of-the-art research on emotion identification from KMT dynamics and identify key research possibilities, challenges, and a roadmap for future research that may be used as a reference. Furthermore, this work addresses the six research questions listed below: Which emotion recognition databases and emotion elicitation techniques are most frequently used? Based on KMT dynamics, which emotions might be identified? Which distinguishing characteristics work best for identifying several distinct emotions? Do certain classification schemes work well for certain emotions? How is emotion recognition using KMT dynamics being applied? (6) Which application scenarios should worry us the most? [2].

2.2 Musical therapy for disabled persons

Disabled people have physical, sensory, or communication limitations; sometimes they have greater difficulty expressing themselves and interacting with others nonverbally. To give the music therapist unbiased information in real-time regarding the user's adaptability to the techniques and interfaces used in their sessions, the ECG and EDA signals have been used to evaluate the person's emotional state. [3]

2.3 Clustering-based speech-emotion recognition

Choose one key segment from the entire cluster that is close to the cluster centroid and represents the remaining segments. SER is based on sequence choices and extraction using a non-linear RBFN- based approach to determine the similarity level in clustering. [4]

2.4 FCN approach for fixing size problems

In order to manage fixed inputs of variable size, several researchers have created fully conventional networks (FCNs) with the use of CNNs. When it came to time series classification tasks based on fixed input variable size, the FCNs performed well. [12]

2.5 Facial expression detection

The residual masking network proposed by Pham et al. on FEB2013 datasets, a new segmentation block was devised and used to extract more relevant characteristics maps, yielding a classification precision of 75.97%. [13]

2.6 Voice recognition for customer satisfaction

The physiological makeup of each person's vocal tract varies, and the frequency spectrum of speech signals may be utilized for a variety of speech applications, such as speaker identification and speech. Spectral feature extraction is used to convert raw speech into compressed signals for efficient speech recognition. [9]

2.7 Multi-modal emotion recognition using hybrid fusion

In [14], it was suggested to use written, graphical, and sound characteristics in a multidisciplinary sentiment analysis. The authors extracted features using a Convolutional Neural Network and audio and visual features using a three-dimensional CNN model. After concatenating the features from each modality, the support vector machine is used to classify the data in the end.

2.8 Accuracy of emotion recognition

When dealing with naturally misleading facial expressions, the model that is being provided is especially helpful in identifying the appropriate emotional state. This study outperforms or compares favorably to the references subject-independent multi-modal emotion recognition studies published in the literature in terms of the accuracy of emotion recognition.

3. PROBLEM DEFINITION

The field of musical therapy has demonstrated its efficacy in enhancing mental well-being, but there remains a significant gap in tailoring interventions to individual emotional states in real-time. This project aims to address this gap by leveraging machine learning techniques for facial expression and voice emotion recognition in the context of musical therapy. The primary problem is to develop a robust and adaptive system that can analyze an individual's facial expressions and voice emotions to recommend or generate personalized music that aligns with their emotional state during a therapy session

.

Design a facial expression recognition model capable of accurately identifying a range of emotions, considering variations in facial expressions across individuals and cultural differences. Implement a voice emotion recognition system that effectively captures and classifies emotional cues in speech, accounting for nuances in tone, pitch, and intensity. Develop a seamless integration mechanism for combining facial expression and voice emotion data to create a comprehensive emotional profile for the user. Enable the system to perform real-time analysis of facial expressions and voice emotions during a therapy session, providing immediate feedback and interventions. Design algorithms that recommend or generate music based on the individual's emotional state, ensuring a personalized and therapeutic musical experience.

4. PROPOSED SYSTEM

Through the use of voice and facial expressions for emotion identification, musical therapy seeks to improve therapeutic outcomes by leveraging the deep relationship between music and human emotion. Through the use of cutting-edge technologies that can recognize subtle facial expressions and voice tones, this method aims to provide a more profound comprehension and interpretation of the emotional states of therapy participants. Therapists can better meet the unique emotional requirements of their clients by customizing musical interventions based on real-time analysis of facial and voice intonations. The goals include encouraging emotional expression and awareness, facilitating communication, and developing relationships based on empathy and connection between the therapist and the patient.

4.1 LOGIN PAGE

A login page is the entrance point to a digital platform or service in order to access it, users must first authenticate themselves. Usually, it contains fields asking for the user's password and username or email address; occasionally, other security features like two-factor authentication are added. The system checks the submitted credentials against its database upon submission, allowing access if they match. On the other hand, inaccurate input result in error message that advise users to try again or, if required, reset their password. The user experience is a top priority.

In the context of musical therapy, a login page might be the point of entry for patients or therapists to utilize apps or online platforms made to make musical therapy sessions or associated activities easier. Users would normally be required to authenticate themselves on the login page in order to guarantee that only those who are permitted can access the resources and functions offered by the platform.

4.2 DATA COLLECTION

The COHN-KANADE-AU (Action units) dataset is a facial expression database containing images of posed facial expressions. Data collection involved capturing video recordings of posed facial expression from 210 adult subjects. Subjects were instructed to perform specific facial action units as defined by facial action coding system. The dataset contains both neutral and expressive facial images, capture under controlled lighting conditions. The COHN-KANADE AU dataset is widely used in research related to facial expression recognition, affective computing, and computer vision.

We used the IEMOCAP and EMODB datasets for spontaneous emotional data, and we also assessed the model's effectiveness on the RAVDESS datasets, compared to ten in the IEMOP and EMODB corpus. We split the data into an 80.20 percent ratio depending on the number of speakers using the fifth-fold cross-validation technique; the remaining data are used for model testing, and the remaining 80% of the data are utilized for model training.

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is multi-modal database containing audio and video recordings of actors performing various emotional expressions. Data collection involved twenty-four professional actors (twelve males and twelve females) who were instructed to produce vocalizations and facial expressions representing different emotions. Actors perform scripted vocalizations with emotional cues, ensuring consistency across recordings.

Both audio and video recordings were captured in a controlled studio environment, ensuring high quality and consistency. The RAVDESS dataset is widely used in research related to emotion recognition, speech processing, and affective computing, providing valuable resources for developing and evaluating algorithms and models.

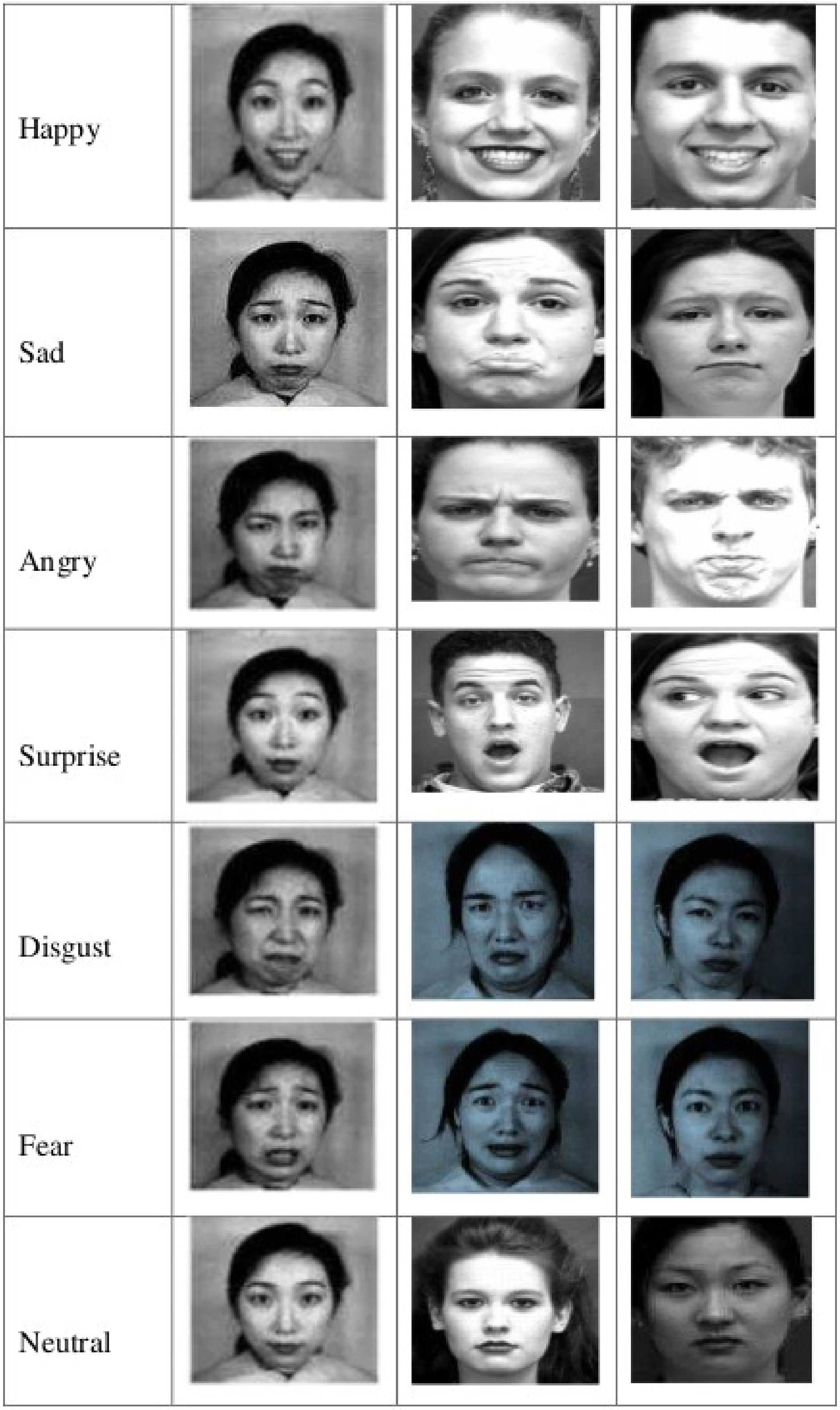


Fig 1: shows 8 expressions with each from the different subjects [8]

4.3 FACIAL EXPRESSION

In computer vision, the use of the Convolutional Neural Networks (CNNs) for facial expression identification has proven to be a potent and successful method. CNNs are a great option for face expression analysis since they excel at image-based tasks. The CNN architecture performs exceptionally well in this situation, automatically deriving hierarchical features from face photos and identifying complex pattern and spatial correlations that are essential for precise emotion categorization. Usually, the network is composed of several convolutional layers for methodically extracting pertinent characteristics and then fully linked layers for making decisions. A varied collection of face photos annotated with appropriate emotions is fed into CNN during its training process for facial expression recognition. [15] This enables the network to pick up on and adjust its settings so that it can identify minute differences in emotions like joy, sorrow, surprise, and sad. Then, unseen face photos may be processed by the trained CNN.

4.4 VOICE EMOTION RECOGNITION

In field of machine learning, voice emotion recognition entails modals and algorithms to examine audio information specifically human speech, and identify the underlying emotional states conveyed by the speaker. It involves training models to analyze audio signals and identify the emotional content conveyed through speech. Techniques include extracting features like pitch, intensity, and duration to classify emotions such as happiness, sadness, or anger. Common approaches include using deep learning models, like neural networks, to achieve accurate emotion recognition from voice data.

4.5 MUSIC RECOMMENDATION SYSTEM

A music recommendation system is an automated computational framework that provides tailored song recommendations by examining user preferences, actions, and past musical interactions. This system processes large amounts of data, such as listening history, user ratings, and genre preferences, using machine learning techniques. The algorithm can forecast the user's taste and suggest songs that fit their musical tastes by finding patterns and correlations in this data. In these systems, methodologies like content-based filtering, hybrid models, and collaborative filtering are frequently employed. A hybrid model improves accuracy by combining aspects of both.

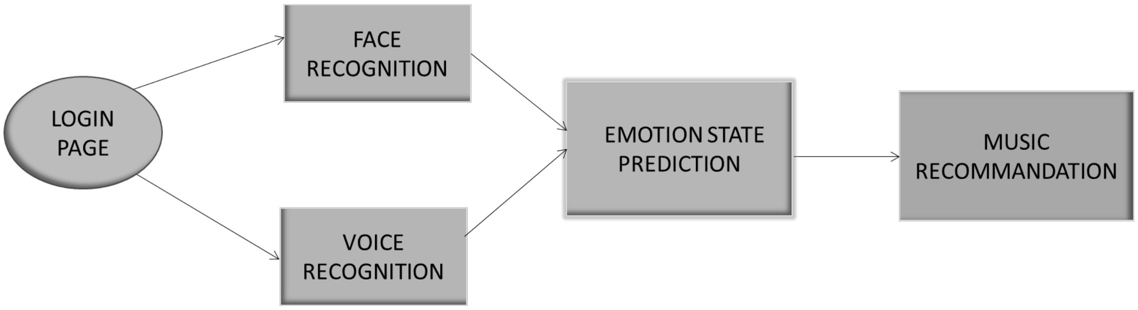


Fig 2: Flow chart for provided system

4.6 Local Binary Pattern

LBP is a method for extracting features. The original LBP operator uses decimal numbers, often known as LBPs or LBP codes, to indicate the location of each pixel in an image. These numbers represent the local structure surrounding each pixel. In a 3X3 neighborhood, each pixel is compared with its eight neighbors by deducting the value of the central pixel. As a result, numbers that are negative are encoded with 0 and all other values with 1. By adding together all of these binary values in a clockwise manner, beginning from one of a pixel's top-left neighbors, a binary number for each pixel is produced. The provided pixel is then labeled using the binary number's corresponding decimal value. The terms "LBPs" or "LBP codes" refer to the resulting binary numbers.

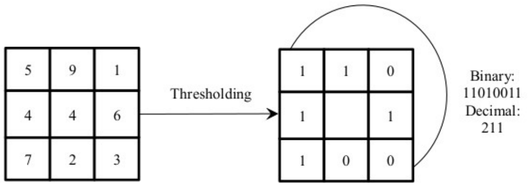


Fig 3: The Simple LBP operator

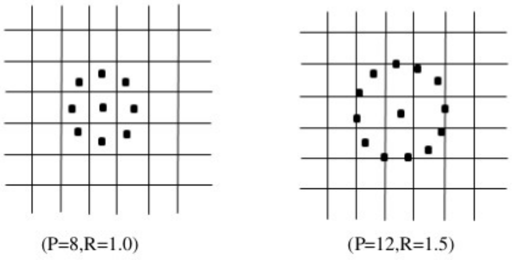


Fig 4: Two extended LBP instances

4.7 Support vector machine

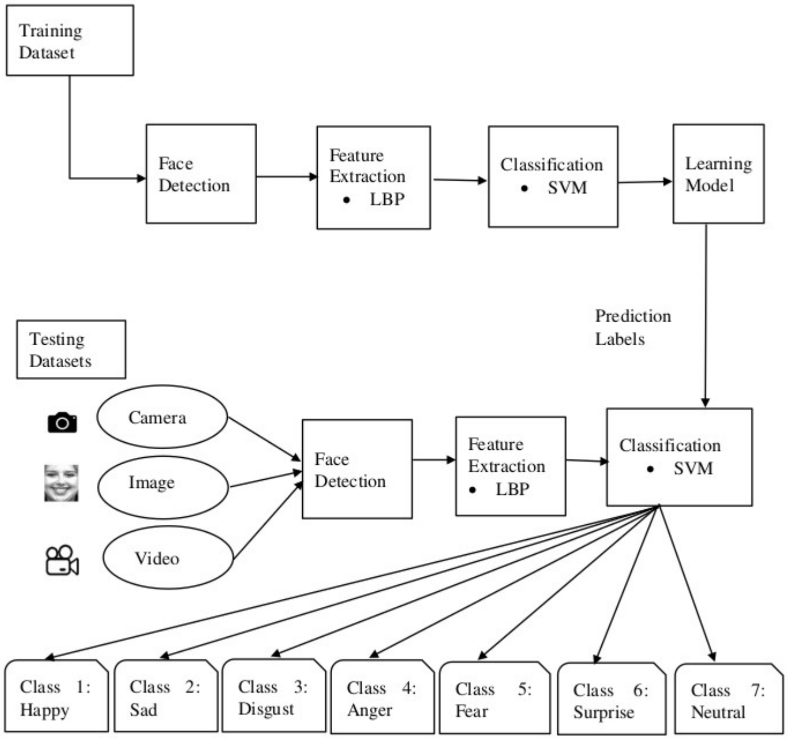
SVM is widely used for various pattern recognition tasks. SVM is a state-of-the-art machine learning approach based on modern statistical learning theory. SVM can achieve near-optimum separation among classes. SVMs are trained to perform facial expression and voice emotion using the features proposed. In general, SVM is the maximal hyperplane classification method.

5. SYSTEM DESIGN

Technology must be integrated into the design of a musical therapy system that detects emotions in speech and facial expressions in order to improve the therapeutic outcome. First, the user's emotional cues will be captured and evaluated through their facial movements using facial expression recognition technologies. This could entail identifying emotions on the face, including stress, and happiness, using computer vision algorithms.

The user's voice's emotional tonality, pitch, and intensity will each be evaluated concurrently by speech emotion detection algorithms. These two data streams will be integrated by the system to provide a comprehensive picture of the user's emotional conditions.

For example, the system can reply with calming or upbeat music to improve the user's spirits if the facial expression recognition recognizes signals of stress or unhappiness and the vocal emotion detection suggests similar emotional states.



Music

Fig 5: Block drawing of Provided System

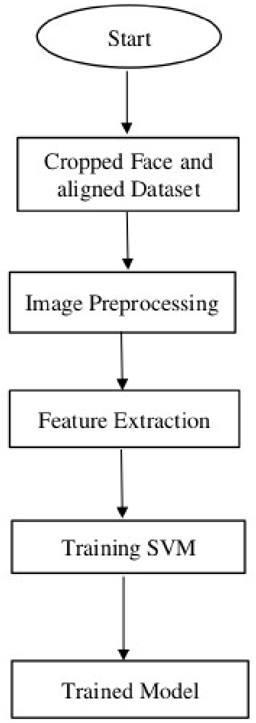


Fig 6: Flowdiagram of Training

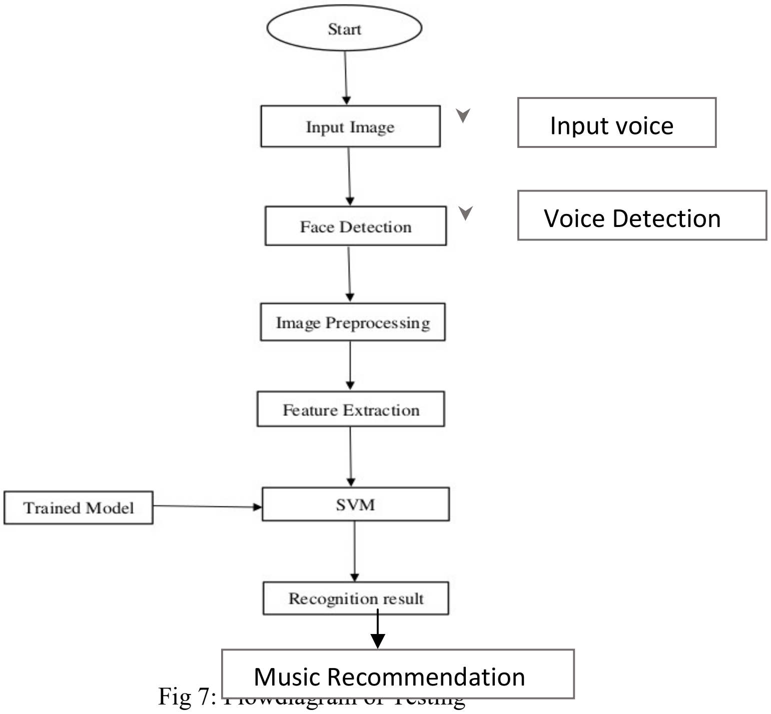


Fig 7: Flowdiagram of Testing

6. CONCLUSION

By offering a more sophisticated knowledge of clients' emotional states, the integration of facial expression and voice emotion recognition into musical therapy improves its effectiveness. This novel strategy promotes a closer relationship between the therapeutic process and people looking to use music for emotional well-being by enabling tailored and responsive interventions.

This project analyzes seven distinct facial expressions from photos of various people taken from various datasets. In this research, facial expressions are preprocessed, then features are extracted using Local Binary Patterns, and finally, speech emotions and face expressions are classified using training datasets of support vector machine-based facial images. The datasets were split into training samples and testing samples in an 8:2 ratios for both training and testing purposes. The dataset's corresponding precision, recall, and F-score were 91.8986%, 83.6142% and 88.9955% respectively.

REFERENCES

[1] alecologists. BioScience. 2002; 52: 19-30. 2. Yang S, Lho H-S and Song B. Sensor fusion for obstacle detection and its application to an unmanned ground vehicle. ICCAS-SICE, 2009. IEEE, 2009, p. 1365-9.

[2] YOUNG J, ELBANHAWI, E., and SIMIC, M. Developing a Navigation System for Mobile Robots. Intelligent Interactive Multimedia. Springer, 2015.

[3] Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision. 2004; 60: 91-110.

[4] Ke Y and Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors. Computer Vision and Pattern Recognition, 2004 CVPR 2004 Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, p. II-506-II-13 Vol. 2.

[5] Al-Smadi, M., Abdulrahim, K., Salam, R.A. (2016). Traffic surveillance: A review of vision-based vehicle detection, recognition and tracking. International Journal of Applied Engineering Research, 11(1), 713–726

[6] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.

[7] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.

[8] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor &amp; Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.

[9] C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.

[10] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. "Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter" Journal of VLSI Design Tools & Technology. 2022; 12(2): 34–41p.

[11] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya "Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay" Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.

[12] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756

[13] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, "Perfromance Investigation of T-Source Inverter fed with Solar Cell" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749

[14] C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007

[15] M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022

[16] M Suganthi, N Ramesh, CT Sivakumar, K Vidhya, "Physiochemical Analysis of Ground Water used for Domestic needs in the Area of Perundurai in Erode District", International Research Journal of Multidisciplinary Technovation, pp: 630-635, 2019

[17] Radhakrishnan, M. (2013). Video object extraction by using background subtraction techniques for sports applications. Digital Image Processing, 5(9), 91–97.

[18] Qiu-Lin, L.I., & Jia-Feng, H.E. (2011). Vehicles detection based on three-frame-difference method and cross-entropy threshold method. Computer Engineering, 37(4), 172–174.

[19] Liu, Y., Yao, L., Shi, Q., Ding, J. (2014). Optical flow based urban road vehicle tracking, In 2013 Ninth International Conference on Computational Intelligence and Security. https://doi.org/10.1109/cis.2013.89: IEEE

[20] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation, In 2014 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10. 1109/cvpr.2014.81: IEEE.

[21] Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M. (2013). Selective search for object recognition. International Journal of Computer Vision, 104(2), 154–171.

[22] Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(9), 1904–16

[23] Zhe, Z., Liang, D., Zhang, S., Huang, X., Hu, S. (2016). Traffic-sign detection and classification in the wild, In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) https://doi.org/10.1109/cvpr.2016.232: IEEE.

[24] Krause, J., Stark, M., Deng, J., Li, F.F. (2014). 3d object representations for fine-grained categorization, In 2013 IEEE International Conference on Computer Vision Workshops. https://doi.org/10.1109/iccvw.2013.77: IEEE.

[25] Yang, L., Ping, L., Chen, C.L., Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification, In 2015 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr. 2015.7299023 (pp. 3973–3981): IEEE.

[26] Zhen, D., Wu, Y., Pei, M., Jia, Y. (2015). Vehicle type classification using a semi supervised convolutional neural network. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2247–2256.