MULTIDISCIPLINARY SCIENCE JOURNAL

CONFERENCE PAPER
Published Online: August 10, 2023
https://doi.org/10.31893/multiscience.2023ss0108

Advancing Engineering, Science, and Technology: Innovations and Developments in Research and Development

Spoken emotion recognition through human-computer interaction using a novel deep learning technology

Manju Bargavi S. K.[a] ✉ | **Pawan Bhambu**[b] | **Mohan Vishal Gupta**[c]

[a]Jain (deemed to be) University, Bangalore, India, Professor, Department of Computer Science and Information Technology.
[b]Vivekananda Global University, Jaipur, India, Associate Professor, Department of Computer Science and Engineering.
[c]Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Assistant Professor, College of Computing Science and Information Technology.

Abstract The paradigm of textual or display-based control in human-computer interaction (HCI) has changed in favor of more understandable control methods, such as gesture, voice, and imitation. Speech in particular contains a large quantity of information, revealing the speaker's inner state as well as his or her goal and intention. The speaker's request can be understood through language analysis, but additional speech features show the speaker's mood, purpose, and intention. As a consequence, in modern HCI systems, emotion identification from speech has become crucial. Additionally, it is challenging to aggregate the results of the many professionals engaged in emotion identification. There have been several methods for analyzing sound in the past. However, it was impossible to analyses people's emotions during a live speech. Studies on real-time data are now more prominent than ever because of the advancement of artificial intelligence and the great performance of deep learning techniques. This research uses a cutting-edge deep-learning technique to identify emotions in human speech. The research made use of the open-source Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. More than 2000 fragments of data were captured by 24 performers as speeches and songs for the RAVDESS dataset. The actors' responses to eight distinct moods were recorded. It was designed to find various emotion classifications. In this study, a novel neuro-fuzzy swallow swarm-optimized deep convolutional neural networks (NFSO-DCNN) approach for classification was suggested. The performance of the suggested model was compared to that of similar research, and the outcomes were assessed. Employing the suggested example on the RAVDESS dataset, an overall accuracy of 98.5% was attained for categorizing emotions.

Keywords: HCI, RAVDESS, SER, NFSO-DCNN

1. Introduction

The technique or method of recognizing and comprehending emotions expressed through speech or spoken language is known as SER. To determine a speaker's emotional state, many auditory variables are analyzed, including pitch, quantity, language rate, and spectrum characteristics. SER systems propose to identify and describe different feelings people experience at the moment or using previously captured speech. Call centers, voice-controlled gadgets, mental health monitoring, virtual assistants, and market research are objective some of the many places these technologies may be placed into action (Nayak et al 2021). The term HCI is used to describe the research and development of systems. It includes all aspects of a person's interaction with an electronic device, program, website, or other digital medium. HCI is the study of how to design technologies to ensure they can be easy to use by the average person. HCI is a multi-disciplinary field that studies how people communicate with and make sense of technological systems. Its goal is to enhance the usability of a product or service by tailoring the user experience to the specific requirements of individual consumers. HCI encompasses a broad variety of digital devices, including tablets, virtual reality systems, wearables, smartphones, and IoT gadgets, in addition to conventional PCs. To improve the design and usability of digital interfaces, it aims to comprehend how consumers engage with these gadgets (Alnuaim et al 2022). HCI-based SER is a fast-developing topic that focuses on the creation of technology capable of comprehending and interpreting spoken expressions of human emotions. With the use of artificial intelligence and natural language processing, this ground-breaking field of study enables computers to recognize and react to the emotional information included in spoken words. Sophisticated algorithms may determine useful psychological information from a speech by analyzing its numerous acoustic and linguistic characteristics, including tone, intensity, pitch, and word selections. To find patterns and connections between audio variables and certain emotional states, these systems

use machine learning methods and train on vast datasets of labeled emotional speech (Rapp et al 2021). HCI, medical care, and other fields all benefit greatly from facial expression recognition. There are six primary feelings happiness, sadness, surprise, fear, and rage. They demonstrated that people from different backgrounds have the same emotional experience. Both valence and arousal are independent variables along which emotions may be described. Both the valence and the arousal levels are complex ranging from calm to excitement. They are sensitive enough to capture minute shifts in facial expression, allowing them to recognize broad categories of emotion but also differentiate between degrees of feeling within each. HCI, social robotics, market research, affective computing, psychological studies, healthcare, and the diagnosis and monitoring of mental health issues constitute only a number of the many fields that may benefit from facial expression recognition (Chen et al 2021). Speech emotion analysis also has important applications in the fields of customer service and market research. By gaining insight into consumers' feelings during interactions, organizations can adapt to their wants and requirements. Companies may learn a lot about their customers' contentment, preferences, and purchasing habits through the study of emotional patterns. With this information, businesses can create more satisfying products, implement more effective marketing methods, and better manage their relationships with customers (Karanchery and Palaniswamy 2021). LinkedIn, and YouTube. Integration of SMD with a healthcare.

2. Related Work

The study (Santhoshkumar and Geetha 2019) employed a feedforward deep convolution neural network architecture with varying parameters to determine an emotional state based on patterns of whole-body movements. The benefit of emotion detection based on body movements is that it may be used to determine a person's emotional state regardless of the angle from which the camera observes individuals. The research (Ren and Bao 2020)provided HCI and intelligent robots will face several significant obstacles, as this study also predicts. The study emphasizes the variety of tools that are presently available for reading, speaking, writing, and using additional senses in human connection. The research (Abbaschian et al 2021) conducted a comprehensive literature review on the topic of emotion recognition in discrete speech. Comparison of existing SER methods and records is essential for finding feasible options and getting a better grasp on that open-ended challenge, especially in light of recent developments in artificial neural networks and the constant need for correct and close to real-time SER in HCI. The article (Tsiourti et al 2019) examined the way people recognize and react to emotions expressed by the physical appearance and speech of humanoid robots, with a focus on the impacts of incongruence. They do this by integrating current discoveries from psychology, neurology, HCI, and HRI. The role of humanoid social robots in modern society is growing. The research (Pustejovsky and Krishnaswamy 2020) provided a simulation environment for the study and development of Embodied HCI (EHCI). VoxWorld is a multilingual conversation platform that facilitates focused task conversations through the use of voice, movement, gesture, expressions on the face, and gaze detection. The study (Yun et al 2021) presented a novel graphical method for making business-critical decisions. HCI is increasingly important in high-tech systems including brain-machine interfaces, human action recognition, telemedicine, and somatosensory games. Extensive testing demonstrates the superiority of the suggested strategy over competing approaches. The research (Wu et al 2020) enhanced manufacturing effectiveness and flexibility while also realizing the benefits of the multi-variety and small-batch assembly via direct interaction among technology and humans. The machine vision technique explored in this article can successfully filter out background noise and produce the desired picture. The study (Al Mahdi et al 2019) investigated how HCI factors into the development of e-learning platforms. Graphics has several important uses in the field of education technology. The field of higher education greatly benefits from the use of interactive multimedia, often known as Hypermedia. The research (Tsai et al 2020) offered a cheap HCI system that can recognize hand gestures. Multiple forms of vision are included in this system. To separate the region of interest from the backdrop, skin and motion detection are employed. To find the object's midpoint, an approach called linked component labeling is presented. The study (Qi et al 2019) focused on optimizing the time differences in surface electromyogram (sEMG) pattern recognition. HCI shows an essential role in bridging the gap in the implementation of information technology in contemporary cities, serving as the interface between people and smart cities. The research (Xu et al 2021) provided a novel method for improving SER accuracy has been identified, and it's been named head fusion. This method takes advantage of the multiple attention heads mechanism. They introduce sounds of differing intensities, alter the noises over time, and combine different types of noise to see how well our model holds up under stress. The study (Chattopadhyay et al 2023) approached assistance in decreasing the feature dimension but also boosts the learning model's classification precision. In addition to its central position in human communication, speech is also the primary information exchange channel in HCI. The study (Atmaja and Akagi 2021) suggested a two-stage late-fusion technique for fusing acoustic and textual information. It's important to note that deep learning algorithms first train audio and text characteristics independently. The second step involves a support vector machine (SVM) that predicts the final regression score based on the outcomes of the deep learning systems' predictions. The paper (Li et al 2021) developed a compact network model for instantaneous emotion classification, to identify emotions in students' faces in real-time. The HCI has benefited greatly from the development of the detection of facial expressions, which has been boosted by the use of deep learning technology. The research (Heracleous et al 2020) identified multilingual speech emotions using authentic emotional speech that has been taken from English, Italian, and Spanish movies. The information

fusion-based approach resulted in a 73.3% unweighted average recall (UAR). This outcome is encouraging and outperforms the UAR determined by evaluation by humans.

3. Proposed Methodology

In this paper, we propose an innovative NFSO-DCNN technique for categorization. The suggested block diagram is shown in Figure 1.

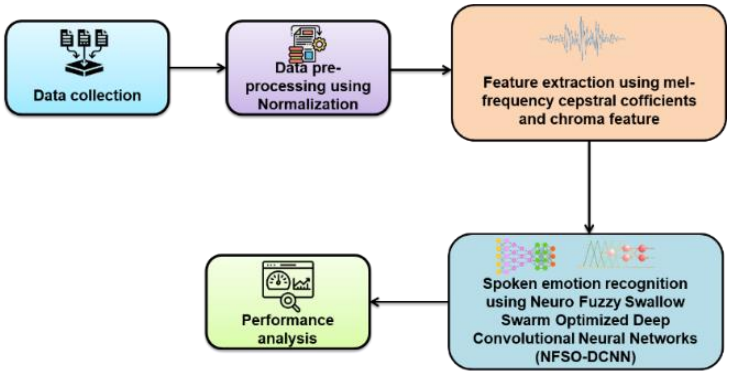

Figure 1 Block diagram of proposed.

3.1. Data collection

The initial dataset for the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consists of 7356 files, including video and audio recordings of speeches and songs. It was chosen to utilize the Kaggle reduced version of the dataset since our study relies on voice and the initial data set size is rather enormous at 24.8 GB. All of the audio files in this edition of the dataset are 16-bit, 48-kilohertz.wav speech files. There are a total of 1440 voice samples in this dataset, with preparations made by 24 distinct actors and 12 male and 12 female changers. There are 60 iterations for each actor, resulting in 1440 files rather than 7356. Due to the absence of video content, the information amount is reduced to a more manageable 590 MB. To enhance efficiency, song files are additionally incorporated into our application. The song has 1012 files and encompasses feelings of fear, sadness, happiness, calmness, and anger. Actor 18's song file is missing. Therefore, there are 2452 audio recordings of emotions in the collection. The dataset was built using two neutral North American-accented speeches to reduce the impact of words and increase the focus on emotions. The name of files also includes identifying codes for things like modality, voice channel, emotion, emotional intensity, statement, repetition, and actor.

3.2. Pre-processing of Normalization

The goal of the field of study known as HCI is to create machines that can identify and interpret human emotions from speech. Emotion identification relies heavily on the pre-processing of speech signals for accuracy and reliability. Collecting high-quality speaking data is the starting point of the pre-processing phase. To accomplish this, samples of spoken language must be recorded using suitable equipment in sterile settings free of distractions. To guarantee the system's efficacy in a wide variety of settings, the data-gathering procedure must contain a spectrum of feelings shown by individuals from many different groups, cultures, and language backgrounds. To minimize differences across speakers and recordings without compromising the feature's capacity to discriminate, feature normalization is a critical process to do. To improve the features' capacity to generalize, feature normalization is used. Both the function and the corpus levels are available for normalization. Among normalization procedures, z-normalization is the most common. The formula for z-normalization, given the means μ and standard deviations of the data, is:

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

3.3. Feature Extraction by using Mel-Frequency Cepstral Coefficients (MFCC) and Chroma Feature

Feature extraction makes extensive use of the NumPy libraries, Libros, and Panda. Mel spectrum pictures were also acquired for use with a convolutional neural network; however, the outcomes were insufficient for processing. Images of spectra are not processed, but rather one-dimensional characteristics are collected.

3.3.1. Mel-Frequency Cepstral Coefficients (MFCC)

A speech recognition system's initial stage is feature extraction, which reveals the parts of an audio signal most suited for recognizing language while filtering out irrelevant parts (such as emotion, and noise). MFCC is among the popular methods for extracting features. The audible path falls within the scope of the short-term power spectrum. MFCCs are used

to represent this limit. MFCCs were first created by Davis and Murmelstein in the 1980s for use in the automatic identification of speakers and recognition. The Mel-frequency Cepstrum is a linear cosine transform of a log power spectrum on a nonlinear Mel-frequency scale that represents the sound's short-term power spectrum. The cepstral representation of the audio sample is included in the Mel-frequency cepstral coefficients. The MFCC differs from a standard Cepstrum in that its frequency bands are equally separated on the Mel scale, simulating the regularly separated bands of frequencies observed in a conventional spectrum, more closely resembling how the human auditory system responds.

3.3.2. Chroma Feature

Chroma features are an interesting and powerful visualization of musical sound, in which the spectrum's characteristics are displayed in 12 separate boxes that correspond to the 12 halftones that make up a musical octave. It is possible to successfully classify musical instruments and sounds according to pitches using chroma-based features, commonly known as pitch class profiles. The harmonic content of the limited audio frame is intended to be reflected in particular through chromatic qualities. Chroma characteristics are connected to the concept of harmony in music and could be very resilient despite changes in volume and pitch. The chroma feature vector, constant-Q transforms (CQT), chroma energy normalized statistics (CENS), and short-time Fourier transform (STFT) are some of the methods used to analyze music.

3.4. Neuro-Fuzzy Swallow Swarm Optimized Deep Convolutional Neural Networks (NFSO-DCNN)

The term neural-fuzzy describes a hybrid computing framework that combines the benefits of neural networks with fuzzy logic systems. While neural networks are great machine learning models for recognizing patterns and making approximations, fuzzy logic systems are great at handling uncertainty and imprecision in data. In a neuro-fuzzy system, the language variables and fuzzy sets used to describe and modify uncertain or subjective information are managed using a rule-based method provided by the fuzzy logic system. Fuzzy logic is a method of thinking and making decisions based on incomplete or ambiguous data that allows for the modeling of fuzzy concepts.

The swallow swarm drives the main concept of our innovative optimization approach. This method uses three different kinds of particles.

- Explorer particle (ei)
- Aimless particle (oi)
- Leader particle (li)

These particles are constantly interacting with one another as they travel in parallel directions:

3.4.1 Explorer particle

The exploration particles include the majority of the colony's people. Their primary role is to investigate the issue domain. If this location is the most important one in the issue space, this particle acts as a Head Leader (HLi) by simply directing the group there with the use of a different sound. However, when a particle is beneficially situated in comparison to its surrounding particles, this is known as a local leader (LLi); else ei valuing VHLi, VLLi, and experience of the opposition of both of these carrying creates an arbitrary move.

3.4.2 Aimless particle

The particulates do not initially have a favourable location for different particles, and the magnitude they have f (o_i) is poor. Their objective is to conduct a haphazard and enquiring search. They proceed to move at random and were unrelated to HL_i and LL_i's positions.

3.4.3 Leader particle

In the SSO algorithm, there are a few particles with the designation Leader. In a colony, the most powerful leader is called the Leader Head, and there are numerous smaller leaders known as Local Leaders.

$$V_{HLi+1} = V_{HLi} + \alpha_{HL} rand\ ()(e_{best} - e_i) + B_{HL} rand\ ()(HL_i - e_i) \quad (2)$$

$$V_{LLi+1} = V_{LLi} + \alpha_{LL} rand\ ()(e_{best} - e_i) + B_{LL} rand\ ()(LL_i - e_i) \quad (3)$$

Where, V_{LL} = velocity of a local leader, V_{HL} =Velocity of head leader, e_{best}= best position of the explorer particle, e_i= current location of the explorer particle

Recalculate the speed using the current formula,

$$V_{i+1} = V_{HLi+1} + V_{LLi+1} \quad (4)$$

The particle's worth is evaluated as:

$$e_{i+1} = e_i + V_{i+1} \quad (5)$$

In the field of artificial intelligence, DCNNs are a special kind of neural network developed for handling and analyzing visual input like photographs and movies. Convolutional Neural Networks (CNNs) are a kind of artificial neural network that performs complicated tasks such as pooling and convolution over several layers of data. DCNNs are an example of multilayer neural networks that typically have an input layer, a convolutional layer, a pooling layer, and an output layer. Two such layers that are concealed are the layer of convolution and the pooled layer. The relationship between layers in a DCNN is an explanation of the propagation that occurs between its input layers to the output layer. To avoid the drawbacks of a model based on linearity, it is necessary to enhance the activity of neurons in every group using a measure of irregular activity in the forward process. There are no activation functions in the first layer since it merely gets pronounced dimensions from the data. Starting with the second layer's use of nonlinear functions for activation, they may write down the expression for the lth layer's output as follows.

$$\left.\begin{array}{c} z^l = W^l * x^{l-1} + b^l \\ a^l = \sigma(z^l), \end{array}\right\} \quad (6)$$

Where l is the layer number and the process of convolution * is symbolized by a symbol. For l=2, the visual vector is $x^{2-1} = x^1$, and for l > 2, the generated characteristic mapping vector for the (l2)th layer is $x^{l-1} = a^{l-1} = \sigma\ (z^{l-1})$. Load vector W^l, biases vector b^l, and balanced input z^l all pertain to the l_{th} layer, whereas is the σ neural stimulation value. If Layer L is the ultimate exit layer, then a^L is the true production vector.

Parameters of W^l and b^l are often repeatedly updated using the backpropagation (BP) algorithm, an automated learning approach. It constructs a cost estimator from the observed and expected amounts and subsequently employs gradient descent (GD) to adjust the parameters in the path of a negative gradient of the expense equation. Here's how it goes down step-by-step in detail.

3.4.4 Cost Function Selection:

Error cost functions are often chosen as quadratic functions. It is possible that correcting a mistake made by the neurons during DCNN training would take some time. As a result, instead of using a quadratic function as the error cost function, we use cross entropy ($E_0{}^L$). Where N is the number of neurons in the result layers and n is the total number of training sets. As a result, DCNN is eventually separated into N classes. The kth output layer neuron's actual output value is represented bya_K^L, whereas its desired value is represented byt_K^L.

$$E_0{}^L = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{N}[t_k^L\ In\ a_k^L + (1 - t_k^L) In\ (1 - a_k^L)] \quad (7)$$

3.4.5 Calculation of Error Vectors:

The error vector for the kth neuron in the product stage has been established for all layers and is represented as follows:

$$\delta^L = \frac{\partial E_0^L}{\partial z^l} \quad (8)$$

The back propagation (BP) Procedures, δ^L used to make reversible deductions $\delta^L - 1$. Correspondingly, assume δ^L and δ (l + 1) is the error vectors of the lth and (l+ 1) th layers respectively. Then, according to Equations 1 and 3 the chain rule, δ^l is written as follows:

$$\delta^1 = W^{1+1} + \delta^{1+1} \odot \sigma'(z^l), \quad (9)$$

The gradients of W^l and b^l are $\frac{\partial E_0^L}{\partial W^l}$ and $\frac{\partial E_0^L}{\partial b^l}$ correspondingly. ∂ (·) stands for the action of partial derivatives. Reduced Equivalent of ($E_0{}^L$) to $W^l$and b^l can be proportional to Equations 6 and 8

$$\left.\begin{array}{c} \frac{\partial E_0^L}{\partial W^l} = \frac{\partial E_0^L}{\partial a^l} \odot \frac{\partial a^l}{\partial W^l} = \delta^1 \odot x^{l-1} \\ \frac{\partial E_0^L}{\partial b^l} = \frac{\partial E_0^L}{\partial a^l} \odot \frac{\partial a^l}{\partial W^l} = \delta^1 \end{array}\right\} \quad (10)$$

The revised estimates of $W^l$and b^l are characterized by ΔW^l and Δb^l, and the optimal solution is obtained by reducing the sum of all possible jobs in each scenario.

$$\left.\begin{array}{c} \Delta W^l = -\eta \frac{\partial E_0^L}{\partial W^l} \\ \Delta b^l = -\eta \frac{\partial E_0^L}{\partial b^l} \end{array}\right\} \quad (11)$$

Where η denotes the learning rate

4. Performance Analysis

4.1. Results

In this part, the suggested system's effectiveness is evaluated. The performance indicators used for assessment are accuracy, precision, recall, f1-measure, and efficiency. Classical Support Vector Machine (C-SVM), Fully Connected (FC), and Decision Tree (DT) are the existing methods used for comparison.

4.1.1. Accuracy

A difference between the result and the true number is caused by inadequate precision. The percentage of actual outcomes reveals how balanced the data is overall. Accuracy is assessed using an equation (12).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (12)$$

Figure 2 shows the comparable values for the accuracy measures. When compared to existing methods like C-SVM, which has an accuracy rate of 94.31%, DT, which has an accuracy rate of 95.09%, and FC, which has an accuracy rate of 97.58%, the recommended method's NFSO-DCNN value is 99.3%. The suggested NFSO-DCNN performs with higher accuracy than other methods. Table 1 displays the proposed method's accuracy.

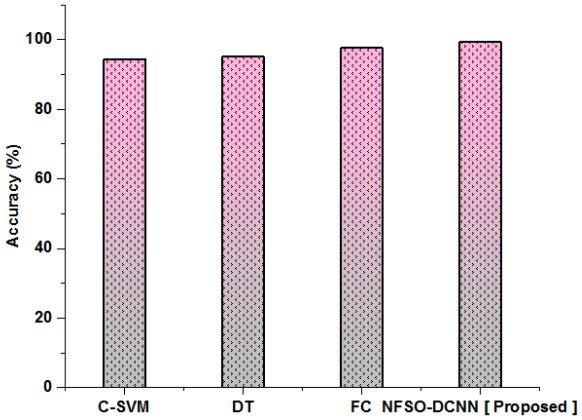

Figure 2 Accuracy comparisons between the suggested and current approaches.

Table 1 Comparison of Accuracy.

Methods	Accuracy (%)
C-SVM	94.31
DT	95.09
FC	97.58
NFSO-DCNN [Proposed]	99.3

4.1.2. Precision

The most crucial standard for accuracy is precision, it is clearly defined as the percentage of properly categorized cases to all instances of predictively positive data. Equation (13) is used to compute the precision.

Comparable values for the precision measures are shown in Figure 3. This proves the suggested strategy may provide performance results that are superior to those obtained by the current study methods. The precision of the proposed approach is 99.42%, which performs better than existing outcomes. Include C-SVM, DT, and FC precision rates are 95.2%, 95.94%, and 98.05%. Table 2 shows the precision of the suggested method is contrasted with the existing methods.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (13)$$

Table 2 Comparison of Precision.

Methods	Precision (%)
C-SVM	95.2
DT	95.94
FC	98.05
NFSO-DCNN [Proposed]	99.42

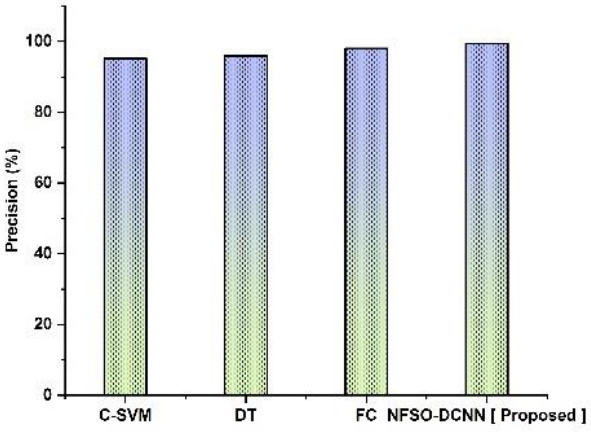

Figure 3 Precision comparisons between the suggested and current approaches.

4.1.3. Recall

The potential of a model to identify each important sample within a data collection is known as recall. The percentage of TPs divided by the sum of True Positive and False Negative is how it is statistically defined. The recall is calculated using equation (14).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (14)$$

Figure 4 shows the comparative data for the recall metrics. Recall rates for C-SVM are 94.42%, DT 95.93%, FC 97.99%, and NFSO-DCNN 99.5%. The proposed method performed better than the current results with a recall of 99.5%. Table 3, the recall of the suggested method is contrasted with the existing methods.

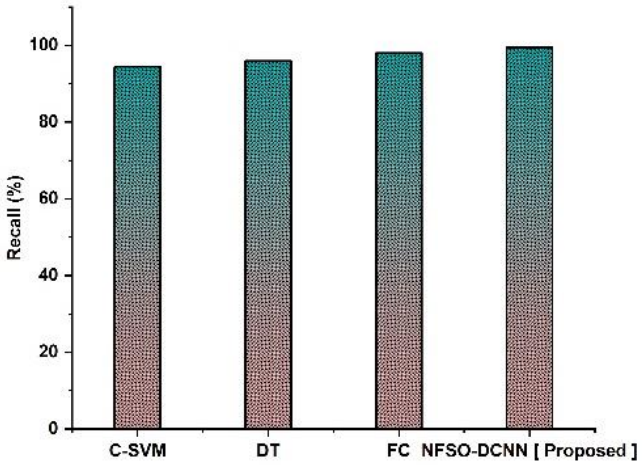

Figure 4 Recall comparisons between the suggested and current approaches.

Table 3 Comparison of Recall.

Methods	Recall (%)
C-SVM	94.42
DT	95.93
FC	97.99
NFSO-DCNN [Proposed]	99.5

4.1.4. F1-Score

The F1-measure is often used while assessing information. It is possible to alter the F1-measure so that accuracy is prioritized above recall, or vice versa. The recommended technique has a higher level of F1-measure when measured against the currently used methods. In Figure 5 the F1-measure of the suggested method is contrasted with the traditional methods.

$$F1 = \frac{2 * (precision * recall)}{(precision+recall)} \qquad (15)$$

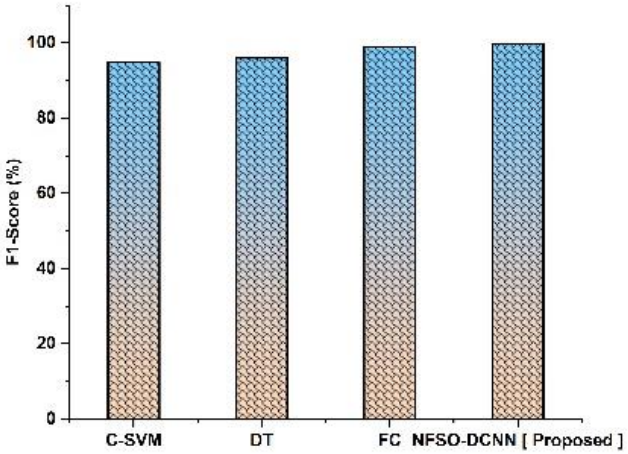

Figure 5 F1-Score comparisons between the suggested and current approaches.

When compared to existing methods like C-SVM, which has an F1-measure of 94.81%, DT, which has an F1-measure of 95.96%, and FC, which has an F1-measure of 98.82%, the recommended method's NFSO-DCNN value is 99.67%. The suggested (NFSO-DCNN) performs with higher accuracy than other methods. Table 4 displays the proposed method's F1-Score.

Table 4 Comparison of F1-Score.

Methods	F1-Score (%)
C-SVM	94.81
DT	95.96
FC	98.82
NFSO-DCNN [Proposed]	99.67

5. Conclusion

In conclusion, while great progress in SER using HCI; additional research is required to raise the reliability and sturdiness of emotion recognition models. To effectively achieve the promise of this technology and to ensure its appropriate and ethical usage, it will be essential to address issues relating to subjectivity, cultural diversity, delicate emotions, and privacy. This study aims to enhance the effectiveness of a machine learning model using the Ryerson Audio-Visual Database of Emotional Speech and Song, a speech dataset. There are a total of 2452 audio recordings in the collection, covering a range of emotions from happy to sad. In this approach, feature extraction takes up the majority of the time, taking longer than model training. The MFCC features' mean is determined. As a result, we introduced the NFSO-DCNN for the recognition of spoken emotion. Performance metrics like accuracy, precision, Recall, efficiency, and F1-measure, are evaluated and compared with existing technologies like FC, C-SVM, and DT. SER in the context of HCI may be improved by using NFSO-DCNN. The accuracy, robustness, and flexibility of emotion detection systems may be improved by using the optimization skills of FSSO and the characteristic learning capabilities of DCNNs, allowing more efficient and customized interactions between people and computers. These new research avenues will help the field advance and realize its full potential in a variety of applications.

Ethical considerations

Not applicable.

Declaration of interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References

Abbaschian BJ, Sierra-Sosa D, Elmaghraby A (2021) Deep learning techniques for speech emotion recognition, from databases to models. Sensors 21:1249.

Al Mahdi Z, Rao Naidu V, Kurian P (2019) Analysing the Role of Human Computer Interaction Principles for E-Learning Solution Design. In Smart Technologies and Innovation for a Sustainable Future: Proceedings of the 1st American University in the Emirates International Research Conference - Dubai, UAE 2017, pp. 41-44. Springer International Publishing.

Alnuaim AA, Zakariah M, Shukla PK, Alhadlaq A, Hatamleh WA, Tarazi H, Sureshbabu R, Ratna R (2022) Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. Journal of Healthcare Engineering 2022.

Atmaja BT, Akagi M (2021) Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM. Speech Communication 126:9-21.

Chattopadhyay S, Dey A, Singh PK, Ahmadian A, Sarkar R (2023) A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm. Multimedia Tools and Applications 82:9693-9726.

Chen X, Cao M, Wei H, Shang Z, Zhang L (2021) Patient emotion recognition in human-computer interaction system based on machine learning method and interactive design theory. Journal of Medical Imaging and Health Informatics 11:307-312.

Heracleous P, Mohammad Y, Yoneyama A (2020) Integrating language and emotion features for multilingual speech emotion recognition. In Human-Computer Interaction. Multimodal and Natural Interaction: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22, pp. 187-196. Springer International Publishing.

Karanchery S, Palaniswamy S (2021) Emotion recognition using one-shot learning for human-computer interactions. In 2021 International Conference on communication, control and information sciences (ICCISc), Vol. 1, pp. 1-8. IEEE.

Li Q, Liu YQ, Peng YQ, Liu C, Shi J, Yan F, Zhang Q (2021) Real-time facial emotion recognition using lightweight convolution neural network. In Journal of Physics: Conference Series 182:012130. IOP Publishing.

Nayak S, Nagesh B, Routray A, Sarma M (2021) A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences. Computers & Electrical Engineering 93:107280.

Pustejovsky J, Krishnaswamy N (2020) October. Embodied human-computer interactions through situated grounding. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents 1-3.

Qi J, Jiang G, Li G, Sun Y, Tao B (2019) Intelligent human-computer interaction based on surface EMG gesture recognition. Ieee Access, 7, pp.61378-61387.

Rapp A, Curti L, Boldi A (2021) The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. International Journal of Human-Computer Studies 151:102630.

Ren F, Bao Y (2020) A review on human-computer interaction and intelligent robots. International Journal of Information Technology & Decision Making 19:5-47.

Santhoshkumar R, Geetha MK (2019) Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks. Procedia Computer Science 152:158-165.

Tsai TH, Huang CC, Zhang KL, (2020) Design of hand gesture recognition system for human-computer interaction. Multimedia tools and applications 79:5989-6007.

Tsiourti C, Weiss A, Wac K, Vincze M (2019) Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots. International Journal of Social Robotics 11:555-573.

Wu S, Wang Z, Shen B, Wang J H, Dongdong L (2020) Human-computer interaction based on machine vision of a smart assembly workbench. Assembly Automation 40:475-482.

Xu M, Zhang F, Zhang W (2021) Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. IEEE Access 9:74539-74549.

Yun Y, Ma D, Yang M (2021) Human–computer interaction-based decision support system with applications in data mining. Future Generation Computer Systems 114:285-289.