Synthesis: A Multidisciplinary Research Journal
Volume 3, Issue 4, October – December 2025, Pp: 9-20

Research Paper

Sparse Coding–Driven Deep Learning for Robust Emotional State Recognition from Multichannel Speech Signals

[1*] Ch.Suneetha, [2] Vijay Keerthika, [3] M. Harshini

[1] Assistant Professor, Department of IT, Anil Neerukonda Institute of Technology & Science (A), Visakhapatnam, Andhra Pradesh, India*
Email: maanash11@gmail.com
[2] Assistant Professor, Department of CSE (AI & ML), MLR Institute Of Technology, Dundigal, Hyderabad, India
Email: vijaykeerthika@gmail.com
[3] Assistant Professor, Department of IT, MLR Institute of Technology, Dundigal, Hyderabad, India
Email: harshinimacherla90@gmail.com

**Corresponding Author(s): maanash11@gmail.com*

Article Info

Received: 12/10/2025
Revised: 19/11/2025
Accepted:28/12/2025
Published:31/12/2025

Abstract

Human emotions can be read from a person's face, words, actions (gesture/posture), or even their heart rate. Due to recent advancements in Machine Learning and data fusion, we can now equip computers with the ability to comprehend, identify, and evaluate human sentiment. Emotional state recognition and Stress disorder diagnosis from speech signals have both been concerns for the recent decade. An increasingly useful computer-aided method for identifying emotional disorders is emotion recognition based on multichannel neurophysiologic inputs, a difficult pattern recognition challenge. Correlation information between channels and frequency components is underutilized by conventional fusion techniques. This paper reveals that deep neural networks trained on emotion data can align with prior domain knowledge and acquire representations that are more accurate than those obtained using hand-crafted techniques. Emotional state identification was the focus of this dissertation, which develops the proposed model named Sparse Coding Technique-Deep Learning (SCT-DL) network models. This is done through two methods named the Convolutional-Recurrent Neural Network (CR-NN) which is a deep learning model that can extract task-related characteristics, extract correlated data between channels and incorporate the contextual information gained from this analysis. Due to the complexity of deep belief networks, limited data sets such as the voice database are incompatible with this type of model. Hence the second method named Knowledge Transmission (KT) which is implemented to deal with the issue of limited data. The purpose is to enhance learning by drawing information from multiple source tasks and applying it to a single target activity. The proposed models have statistically and experimentally been proven to be more effective than most state-of-the-art techniques currently available for recognizing emotional states.

Keywords: Sparse coding, Deep Learning, Neural network, CNN, KT Emotion recognition

1. Introduction

Emotions play a crucial part in human interaction and decision-making. The internal experience (such as happiness, sadness, fear, anger, sympathy, disappointment) prompted by different environmental factors is called an emotion, often known as an affect or a mood [1]. It's crucial to human life in many ways, including conversation,

organization, imagination, logic, behavior, and more. The goal of analyzing a speaker's speech for emotional cues is to determine the speaker's emotional state. When human-machine or human-computer interaction becomes increasingly common, the ability to infer emotional state from spoken language will become increasingly useful [2]. Several outward bodily manifestations, including as a person's pulse, blood pressure, sweat level, shaking, skin tone, facial expression, and voice intonation, can be used to deduce an individual's emotional state [3].Anger, grief, and terror are only few of the human emotions that can be detected by vocal signals, as has been shown [4].

Most modern human-computer interaction (HCI) systems, however, are not up to snuff when it comes to processing and understanding feelings and emotions [5]. They cannot recognize human emotions and apply that knowledge to make decisions or take action [6].Emotion recognition and response from the computer system is a straightforward method of incorporating the human [7]. Advertising, human-robot interaction (HRI), and medical diagnostics are among existing applications that employ or would benefit from an emotion recognition system [8]. Emotion recognition that is trustworthy, accurate, flexible, and robust is a crucial component of intelligent HCI [9]. Many researchers in the field of AI are focusing on affective computing and emotion identification in an effort to give computers emotions [10].

Judgment, thinking, and decision making are all influenced by one's emotions, hence they are increasingly being considered an integral aspect of intelligence [11]. Many experts in the field of AI think it's impossible for a computer to become truly intelligent unless it can understand and respond to human emotions [12]. This demand has led to the development of Affective Computing (AC), a new area of AI research. Our experience with many forms of Human-Computer Interaction (HCI) like video games, search engines, driver-assistance systems, e-learning, and more can be enhanced with the help of Affective Computing (AC), which aims to give computers the ability to recognize, understand, express, and respond appropriately to human emotions [13].Being a difficult pattern identification challenge, automatic emotion recognition has attracted growing attention from a wide variety of academic disciplines [14].The medical professions of psychiatry and neurology have recently adopted the observed emotional states of patients as a sign of various organic or functional emotional diseases, such as post-traumatic stress disorder and major depression [15].

Emotion recognition, in its broadest sense, is a pattern recognition task that involves observing and analyzing a person's outward displays of emotion [16]. In most cases, people's overt physical actions (such as their facial expressions, body language, and voice) are acted out sub consciously [17]. These are the primary channels via which we share intimate details about our lives and our feelings with others [18]. Vocal tract characteristics, prosodic features, and excitation features are the three main types of speech features found in the literature [19]. Vocal tract features are significantly associated with articulator movement and vocal tract morphology, and are best defined in the frequency domain [20]. Mel-frequency cepstral coefficients (MFCC), formants, and other similar metrics are all examples of tract characteristics. Glottal characteristics and linear prediction coefficients (LPC) are two types of excitation features [21]. Prosodic aspects are the patterns of duration, intonation, and intensity that arise during the creation of human speech. Maximum, minimum, average, variation, and standard deviation in signal energy and pitch are all examples of prosodic characteristics [22].

The major contributions of the paper are as follows:

1. Acquire and recognize emotional states with high accuracy by training a model with the neural network architecture known as Sparse Coding Technique-Deep Learning (SCT-DL).
2. Convolutional-Recurrent Neural Network (CR-NN) is a deep learning model that may be used to recognize task-related features, mine inter-channel correlation, and include contextual information from those frames.
3. Knowledge Transmission (KT), is used to address the problem of insufficient information for the speech database and the small data size challenge.

In the following part, will examine this work in relation to similar efforts. Section 2 covers the literature view regarding the research. Sparse coding technique deep learning (SCT-DL) with a convolutional neural network is the topic of Section 3. The results are provided in Section 4, and Section 5 concludes the research.

2. Related Work

Humans and computers work together in many settings, thus experts in ergonomics and intelligent systems are constantly attempting to find ways to make this collaboration more effective and adaptable (HCI). For this type of HCI system to work, the computer needs to be flexible, and it's essential that it be able to decipher human communication styles in order to respond appropriately. Humans have the ability to communicate their intentions through a wide range of vocal and nonverbal actions and expressions.

A study [23] developed hybrid deep learning model [HDL], extracting task-related features and mining inter-channel and inter-frequency correlation is done by the CNN, while the RNN is layered on top to incorporate context from the frame cube sequence. Emotion detection performance is evaluated using the DEAP benchmarking dataset in a trial-and-error setting. In terms of both Valence and Arousal, the experimental results show that the suggested framework outperforms the conventional methods.

The authors from the study [24] introduced models of emotional state recognition [ESR] and PTSD diagnosis based on deep belief networks were introduced. There is a supplemental transferring-knowledge approach for identifying PTSD. The complexity of models like deep belief networks means that they struggle with limited data sets like the PTSD speech database. As a result, we relied on transfer learning to deal with the issue of limited data. Learning is intended to be enhanced by the transfer of knowledge from one work to another. It has proven effective when there is a dearth of high-quality training data for the desired task.

The research [25] proposed Facial Action Units (FAU) using the gradient as a saliency map, we additionally display the facial features that have the most influence on the final forecast. Finally, we investigate the potential benefits of multimodal emotion recognition by fusing our model with those trained on acoustic and physiological data. The user will accept a system if the action taken in the action stage is appropriate for the emotion recognized in the recognition stage.

Previous study [26] elaborated both EEG (both single- and multi-channel) and other physiological signals in terms of their utility in emotion identification. Physiological cues allow for more objectivity and reliability in emotion identification. EEG signals can provide important aspects of emotional states because they respond more sensitively to fluctuations in affective states and in real time than peripheral neurophysiological data.

The study [27] incorporated improved objectivity and trustworthiness in emotion recognition [ER] can be achieved through the use of physiological signals. Signals from the brain's electroencephalogram (EEG) respond more sensitively to fluctuations in affective states and in real time than those from the periphery, making them a valuable source of information about emotional states. Conventional approaches rely on in-depth domain expertise to create and extract a wide array of characteristics from single or multiple channel signals. Furthermore, the conventional feature fusion strategy makes insufficient use of channel-correlation information.

All of the currently available methods, such as HDL, ESR, FAU, SM-EEG, and ER, face the same problems when used to CNN, including the precision rate, prediction rate, accuracy ratio, and performance analysis. It was discovered that SCT-DL is far more effective than previous methods. Research has looked into many non-traditional ways and strategies for dealing with these problems.

3. Emotion Recognition and Stress Disorder Detection

Emotion is being recognized as a key component of intelligence due to the fact that it permeates all stages of the cognitive process. As a result, many experts in the field of AI contend that a machine cannot become truly intelligent until it learns to understand and respond to human emotions. In addition to its applications in HCI, emotion recognition and monitoring have promise in the field of auxiliary diagnosis of a variety of mood disorders. Mental health professionals, such as psychiatrists, are in dire need of improved tools to aid in the early detection and treatment of mental health issues like depression, PTSD, anxiety, and etc.

Emotion recognition is a pattern recognition activity that generally involves seeing and analyzing a person's outward behavior. Physical actions (such as smiles, gestures, and tone of voice) are often subconsciously expressed. They are the primary channels via which we share our innermost thoughts and feelings with other people. There are times when we may actively hide, fake, or even amplify such expressions of feeling. This suggests that methods of recognition based on observable manifestations alone may not be as reliable or stable as others. Specifically, the emergence of an emotion is the result of cooperation between the CNS and the ANS. It implies that we can investigate a computational approach to emotion based on neurophysiological measurements associated with the central nervous system (EEG, FMRI, etc.) or the autonomic nervous system (metrics including blood pressure, heart rate, skin conductance, body temperature, and a host of others.).

Feature extraction is typically performed on an input voice signal in an emotion identification system. As an added bonus, it occasionally performs feature selection to zero in on the best features. At long last, it assigns feelings to the voice signal. Several speech corpora, characteristics, and classifiers have been used in the past for emotion recognition in the literature.It's possible to use this to create a speech-based detection system. The acquisition of speech does not involve any risk to the subject and can be accomplished in a non-personal manner using means such as the telephone or recording media. Furthermore, it can track how well a patient is responding to their treatment. The restricted size of Speech Corporation is a key disadvantage of speech-based diagnosis, which has a number of disadvantages.

This dissertation has attempted to address these limitations by creating a system for emotion identification and diagnosis that draws on sparse coding, deep belief network models, and a transfer learning strategy. To combat the issue of limited data, the concept of transfer learning has been presented. Recently, sparse coding methods have demonstrated state-of-the-art performance in a wide variety of settings. Sparse coding first involves a dictionary of basic functions that are learned from the data. All the original data samples were then reconstructed using the dictionary as a building component. The reconstruction weights, which were then used for classification, were essentially new representations of the original data. Since sparse coding's basis functions are data-driven, the feature extraction procedure is dynamic.

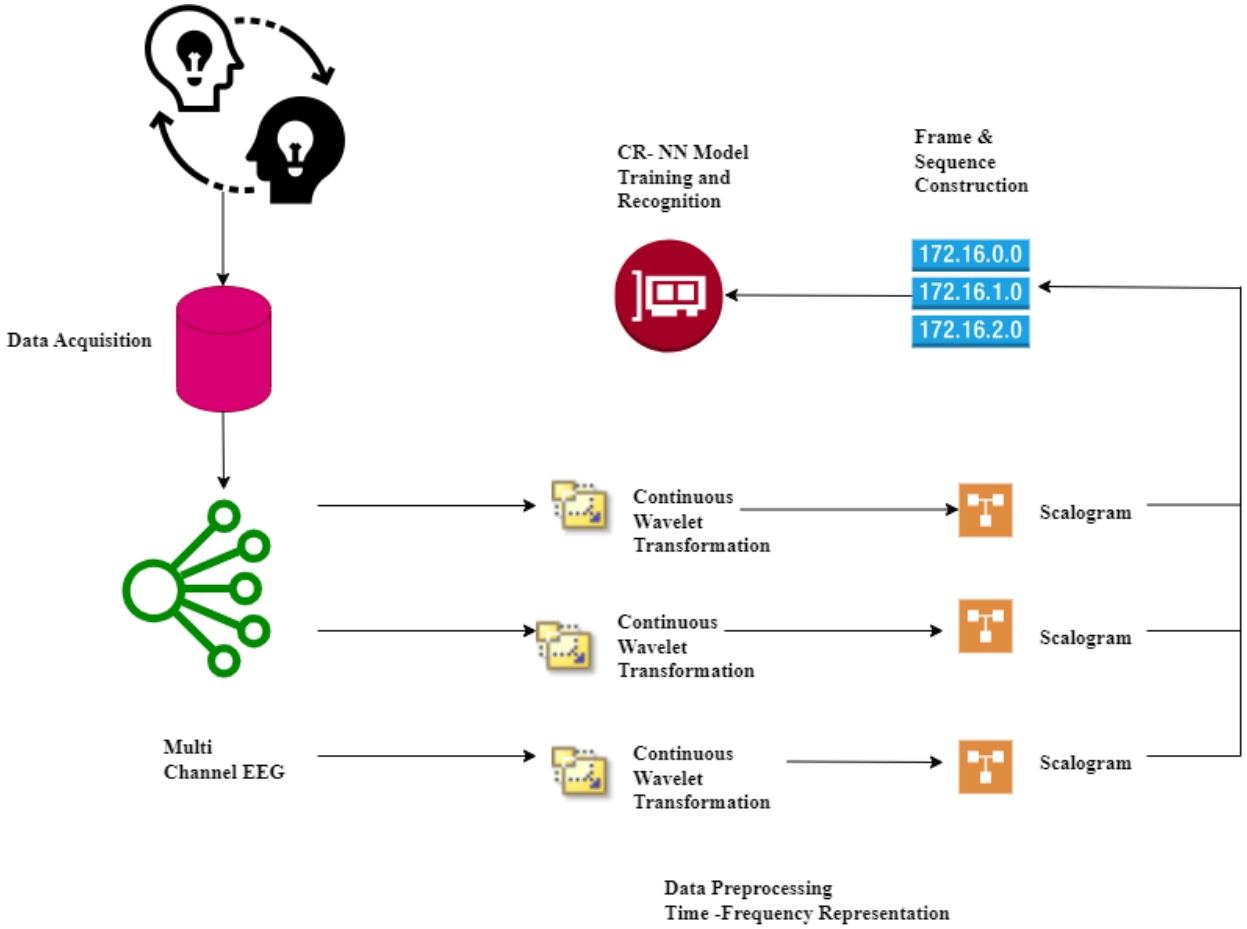

Fig.1. Built-In Framework for Multi-Channel Emotion Detection

Our architecture for emotion identification using various channels of neurophysiological information is depicted in a simplified flowchart in Figure 1.Two research issues are addressed by our proposed approach. Problem one is figuring out how to model something before it has been preprocessed and represented in a suitable way using multiple channels. The second is utilizing this prepared data to model and identify emotions. This study propose a 2D frame-based representation method and a hybrid deep learning approach.

Creating an accurate representation or approximation of a signal is of utmost importance in signal processing and associated pattern recognition activities. Signal processing, machine learning, computer vision, etc. all benefit greatly from the "Sparse Representation" (SR). It helps learner systems acquire a compact structure and learn higher-level implicit semantic information from raw data. Several types of spectral representation, such the Wavelet transform (WT), are usually viewed as subsets of the broader spectral representation (SR). In contrast to conventional SR approaches like Sparse Coding. Wavelet analysis relies on a predetermined mother wavelet rather than a learnable dictionary ∂. After scaling s and translation u a group of wavelet basis functions denoted by $\partial_{s,u}$

$$\partial_{s,u}(t) = \sqrt{s}\,\partial\left(\frac{t-u}{s}\right), u \in R, s > 0 \qquad (1)$$

Equation (1) represents that for each time intervalt , we build a two-dimensional data structure that we refer to as a "frame," denoted by M_t, It is a representation of the wavelet energy at different scales and frequencies (a term from signal processing, where each scale represents a different coarseness of the signal).

$$CWT\, f(s,u) = \int_{-\infty}^{+\infty} f(t) * \partial_{s,u}(t)\, dt \qquad (2)$$

The above equation (2) shows the CWT which is the continuous wavelet transform where $f(s,u)$ is the original EEG signal.After the CWT, the signals from each one-dimensional channel are represented in terms of a time scale based on the wavelet coefficients.

The design's main strength is in its ability to encapsulate and centralize data pertaining to signals across all channels in a coherent and understandable framework. So, these multi-channel signals are ideal for the multi-channel EEG signal based data mining tasks since they can be further processed as a whole, and the inner relationship of the different channels may be mined, given that the signal on each channel is a composite of electrical activity originating from a number of different brain regions. Emotional experiences are accompanied by a flurry of dynamic activity that changes across different regions of the cortex, and this is represented in the progression of images.

$$\mathrm{SG}(s) = \int_{-\infty}^{+\infty} \sqrt{|\, CWT\, f(s,u)|^2\, du} \qquad (3)$$

The above equation (3) represents the scalogram denoted by SG (s) following CWT, the signals in each one-dimensional channel are represented in terms of a time scale based on the wavelet coefficients, a representation known as a scalogram as shown in the above equation (3) and the sensitivity is gained.

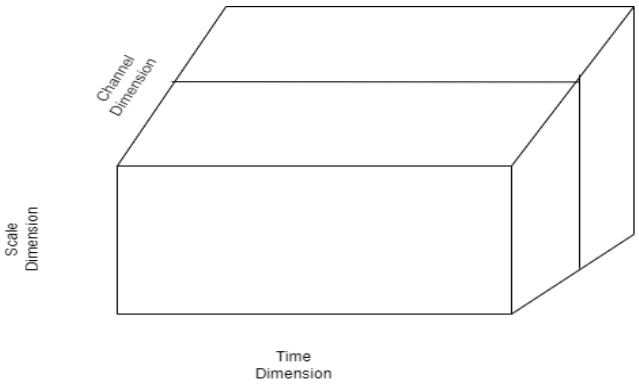

Fig.2. Information Modeling Using a Cube-Shaped Frame

3.1 Building of a Cube-Shaped Frame

Once the scalogram of the signal for each channel has been collected, the frame cubes can be built. It can be seen in Figure 2 that the spectral energy distribution in the C channels and the S selected scales across an L-length time window is represented by a cube-shaped C S L structure in each frame cube. There are three primary steps involved in building a frame cube:

1. The time period represented by a frame cube, denoted by the letter L, needs to be measured. With a time window of 1 second in duration with no overlap between succeeding windows, we can produce 60 frame cubes for a 60-second test, for example.
2. By simultaneously dragging a temporal window from the starting point across all of these scalograms, each of which is S, L in size which are able to extract frames.
3. A frame cube is built by stacking the frames retrieved from each channel signal at the current time step t. In this case, the energy distribution at the current time step t can be represented by a cube-shaped frame with dimensions C, S, and L.

Steps 1-3 are repeated until all the frame cubes for one trial have been created, at which point we progress to the next time step t + 1 by sliding right with a length L.

3.2 Emotional Voice Reconstruction

This section presents a synopsis of the work done to date on emotion recognition, outlining the various speech parts and classification strategies that have been applied. Data and affective context are both carried by speech signals. Since the human voice is dynamic, the loudness of the signal changes over time. Frames are the discrete time intervals used in speech processing, where the signal is assumed to be stationary for the purposes of analysis. Local features are characteristics of a speaker's voice that are retrieved from individual frames of speech. In contrast, an utterance's global features are a statistical tally of all of its speech features. Although opinions vary as to which type of feature is better for emotion recognition, most studies have found that global features perform better when using cross validation and feature selection algorithms.

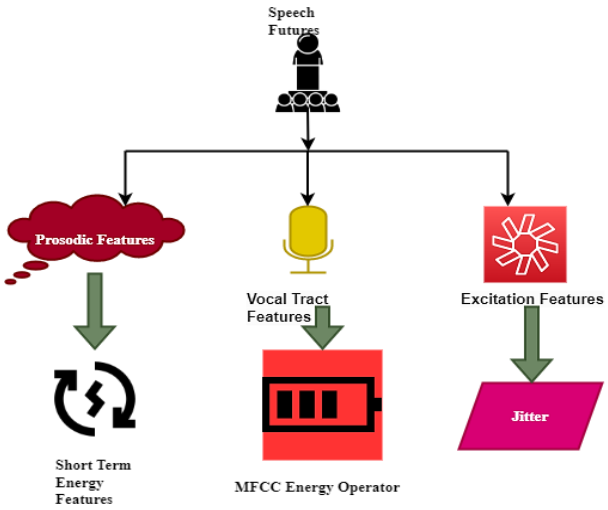

Fig.3. Classification of Speech features

They can be implemented more quickly than native capabilities. Studies have shown that global features are superior, they are only able to tell the difference between high-arousal feelings like anger and terror and low-arousal feelings like sadness. Segmenting speech signals into their constituent parts is another method for feature extraction, initial phonemes as a starting point, and then create a feature vector for each phoneme that has been isolated. This strategy is based on research that compares the spectral forms of calls made from the same phone when the user is experiencing a range of emotions. Prosodic, vocal-tract, and excitation aspects constitute the three broad groups into which speech characteristics fall. The process of extracting characteristics from animated conversational agents' word-level utterances. There were a total of twenty-two acoustic and prosodic features. Statistics at the utterance level were calculated, and they were found to be strongly connected to the fundamental frequency.

$$\mathrm{E}\,(s) = \frac{Energy\,(s)}{Entrophy(s)} \qquad (4)$$

Energy to Shannon Entropy Ratio (EER) is a useful metric for selecting representative time scales. The above equation describes the optimal scales to use when the spectral energy is high and the Shannon entropy is low (4). Efficiency is achieved in the above equation.

$$\mathrm{Energy\ (s)} = \sum_{j=1}^{n} |\mathrm{WC(s)}|^2 \qquad (5)$$

Calculating the's' scale's energy is as simple as adding up the 'n' wavelet coefficient's energy, as shown in equation (5) above.

$$\mathrm{Entrophy(s)} = -\sum \mathrm{R_i} \log \mathrm{R_i} \qquad (6)$$

The Shannon entropy quantifies the degree of dispersion in the energy spectrum on the s-scale. More information is included in a given scale with a lower entropy, as shown in equation (6) where R_i is the probability of energy distribution of coefficient WC in scale s.

$$\mathrm{R_i} = \frac{|\mathrm{WC(s)}|^2}{\mathrm{Energy\ (s)}} \qquad (7)$$

All channel signals' EERs can be determined using the aforementioned formula in equation (7).

3.3 Convolutional Recurrent Neural Network (CR-NN)

The accuracy of automatic emotion recognition systems has increased thanks to the incorporation of feature selection techniques that have allowed for the creation of features that yield reasonably good accuracy. Nonetheless, these systems can typically only capture linear relationships between the features. Neural networks and deep learning techniques can model complex non-linear feature interactions in the data. As a result, compared to baselines that do not use these models, deep belief network models perform better in terms of classification accuracy.

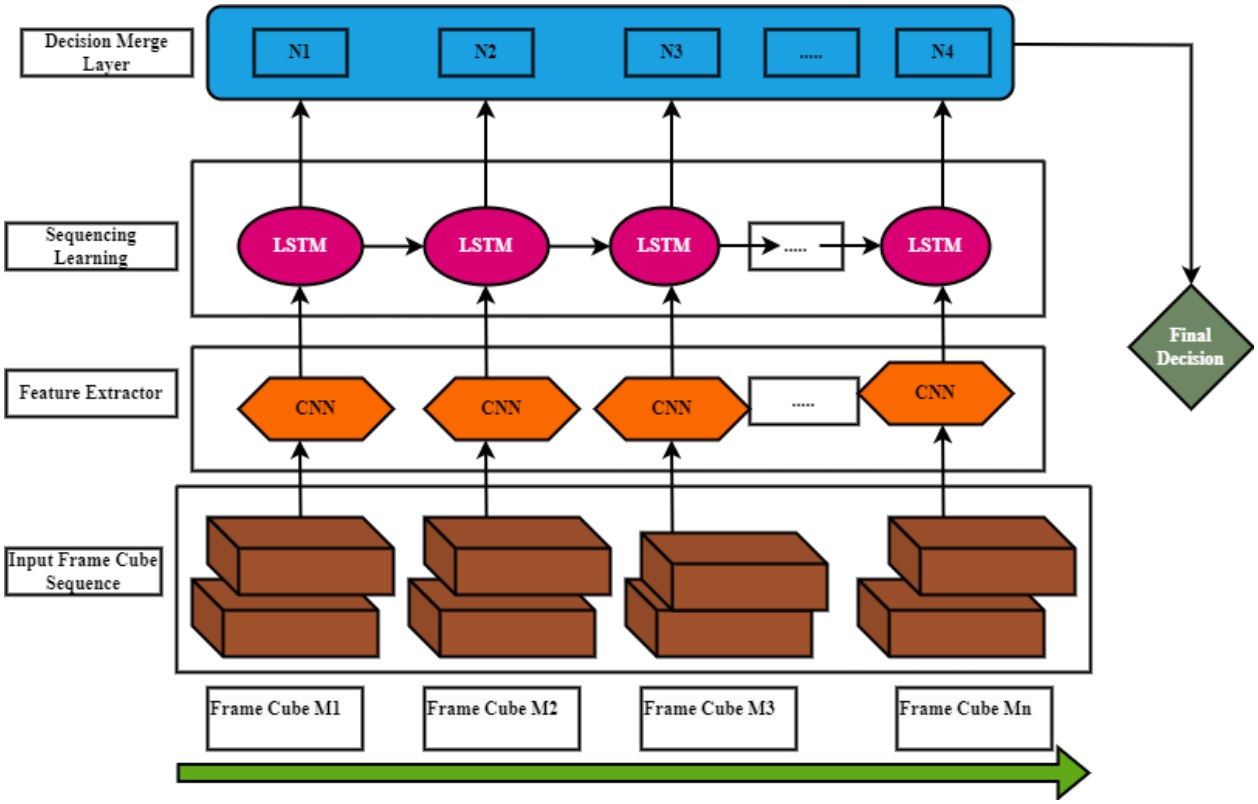

Fig.4. CR-NN Model for Emotion Recognition

In addition to the aforementioned pre-processing technique, Convolutional recurrent neural networks (C-RNNs) are a new type of deep learning model we introduce here for emotion recognition. Figure 4 shows that the model is a combination of two different types of deep learning algorithms. It combines the RNN's proficiency with sequential data processing with the robust capability of the CNN's grid-like architecture. The Convolutional Neural Network (CNN) module is responsible for feature extraction from frame cubes as well as channel and frequency correlation mining. Modifications to the RNN design known as Long Short-Term Memory (LSTM) units are used to model the context of sequences of arbitrary length. This mixed approach works very well for handling sequential information in either two or three dimensions.

Finally, the convolutional filter's size was set that correlations between channels and between granularities could be mined (frequencies).The frontal convolutional neural network (CNN) recovers a feature sequence, and this model uses the LSTM-based RNN's output at each time step to determine if an emotion has been recognized. As a result of the construction and operation of CNNs, neural networks may now deal with data in either a two- or three-dimensional structure. Many features can be automatically extracted using the established convolutional filters. It is common for a CNN to have multiple convolutional layers that are stacked on top of one another. The three phases that make up a convolutional layer are the convolution stage, the detector stage, and the pooling stage. CNN's underlying mechanism and design make it possible for neural networks to analyze input organized in a grid format. Convolutional filters were created to aid in the automatic extraction of many types of features. It is common for a CNN to have multiple convolutional layers that are stacked on top of one another. Convolution, detection, and pooling are the three processes that make up a standard convolutional layer.

Because of their modest size, bodily measurements typically lag behind the development of new emotions. The RNN is well-suited to resolving the delayed effect because of its capacity to accumulate the properties of the weak signal in each time step. The hybrid model gains time series learning capabilities after incorporating the RNN component. When compared to common deep neural networks (DNNs), RNNs excel at sequential modeling. Weights parameters in an RNN are reused at each time step, unlike in a DNN, thus, the size of the input sequence does not proportionally increase the number of parameters. The RNN's recurring structure is what makes it effective at modeling context from sequences of varied or constant length.

3.4 Sparse Coding For Emotion Recognition

Sparse coding has been shown to be superior to the state-of-the-art in many areas, including computer vision.

Input vectors are roughly approximations in sparse coding, which uses a small set of "basic" functions to be linearly combined with weights. Since the number of basic functions in this set is typically more than the dimension, it is able to capture a wide variety of recurring structures in the input data.

Dictionary learning is crucial to the sparse coding scheme because it is what will ultimately identify the data building components. The process of education, however, is laborious and time-consuming. The randomized dictionary learning strategy will be used in this paper. The input data samples were encoded using the learned dictionary to produce the desired representations.

For this reason, many corrected recurrent units have been utilized to replace the typical units of the traditional RNN in an effort to decrease the difficulties in acquiring a long-term dependency. The gate's self-looping design permits the gradient to flow for extremely extended periods of time without the gate recollecting that the information has already been used. Involvements in activities such as recognizing handwriting, voice, machine translation, captioning images, parsing, etc. have all seen considerable improvements to the use of these gated RNNs.

$$O_t = \alpha(W_f * [h_{t-1}, y_t] + b_t) \quad (8)$$

In the above equation (8) h_{t-1}, represents hidden state from LSTM cell and input is represented by y_t , and weights are denoted by W_f of the gate , output vector is noted as O_t prior cell state is denoted by h_{t-1} .Accuracy is achieved in the above equation.

The tangent layer provides candidate data, denoted by S_t, for the sigmoidal layer to evaluate. The sigmoidal layer then makes the final decision on which pieces of data to use, denoted by the decision vector d_t

$$d_t = \alpha(W_f * [h_{t-1}, y_t] + b_t) \quad (9)$$

$$S_t = \tanh(W_f * [h_{t-1}, y_t] + b_t) \quad (10)$$

As a result, the current chain's cell state, denoted by the symbol S_t, is a fusion of the preserved past data represented by S_{t-1} and the most up-to-date data drawn from S_t.

$$S_t = S_{t-1} * d_t + S_t * d_t \quad (11)$$

Updated information is selected from S_t as shown in equation (11).

$$O_t = \alpha(W_o * [h_{t-1}, y_t] + b_o) \quad (12)$$

$$h_t = \tanh(S_t) * O_t \quad (13)$$

By multiplying vector O_t denotes the information selected from the pool of possibilities. S_t, as illustrated in Equations (12) and (13), the algorithm determines which of the hidden states h_t in the current chain will be output. Performance is achieved greatly in the above equation.

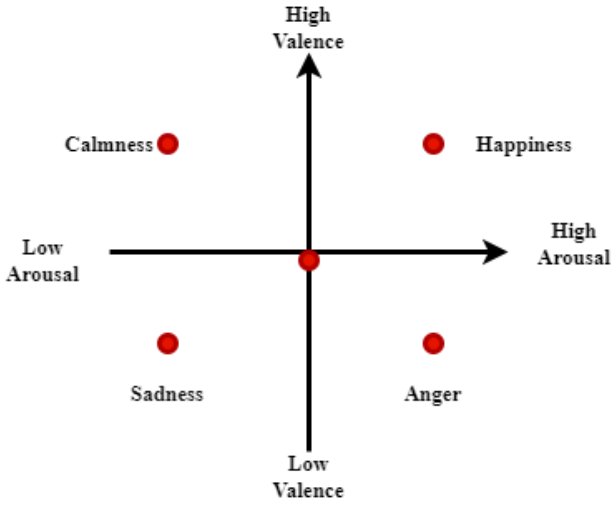

Fig.5. Emotion in a Two-Dimensional Model

The number of distinct emotional states has long been a contentious topic in the field of psychology. Traditionally, psychologists have used two distinct approaches to modeling emotions: Figure 5 depicts the differences between basic emotion theory, which classifies feelings into discrete buckets, and multi-dimensional theory, which classifies feelings along a variety of axes. According to the theory of basic emotions, humans experience a range of feelings including joy, sorrow, fear, anger, contempt, and surprise. These fundamental emotions are the building blocks of more complex feelings like exhaustion, anxiety, satisfaction, confusion, and irritation. There are distinct psychological, behavioral, and physiological signatures associated with each emotion type.

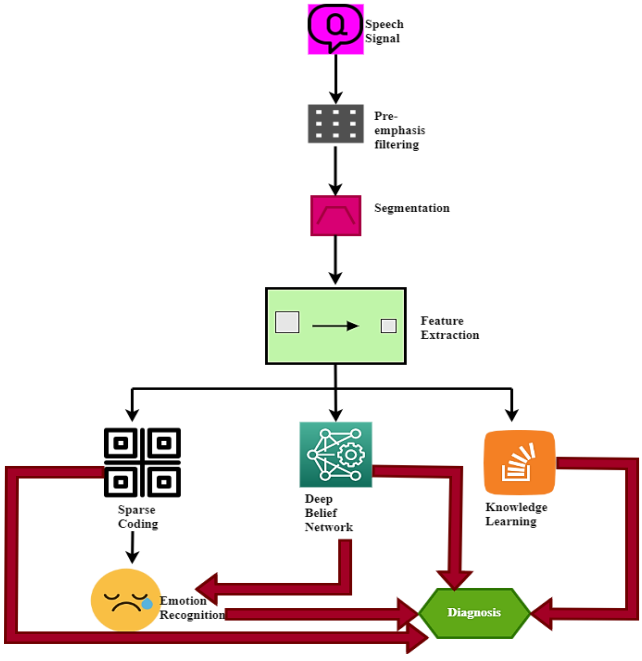

Fig.6. Proposed Model

The schematic for the suggested setup is depicted in Figure 6. Prior to further processing, the input voice signal undergoes pre-emphasis filtering. Next, it takes the speech segments that have already been processed and separates out the spoken frames using feature extraction. These audio samples are then used in conjunction with sparse coding, deep belief network, or transfer learning models for emotion

detection and diagnosis.

A feature pooling approach was then applied to the new feature space to minimize its high dimensionality. The feature pooling approach divided the feature vector into four parts, and the average of those parts was then computed as new features, thereby halving the dimensionality.

Emotional recognition and post-traumatic stress disorder are two of the most significant issues that can be pinpointed with the aforementioned methods. Using Convolution to Improve Precision, Sensitivity, Efficiency, Performance, Computational Complexity, and Determination The suggested model uses a deep learning network to produce a recurrent neural network.

4. Results and Discussion

This research describes a convolutional neural network (CNN) for recognizing emotions and diagnosing stress disorders, and it uses simulations to define and run the network for maximum accuracy, performance, and efficiency. Here, compare this model to others in terms of prediction, technical Performance, sensitivity, and accuracy (such as HDL, ESR, FAU, SM-EEG, and ER).

4.1 Data collection and analysis:

Dataset is taken from the site given as a reference purpose. https://medium.com/analytics-vidhya/emotion-recognition-datasets-8a397590c7d1

4.1.1 Accuracy Analysis

No.of Samples	HDL	ESR	FAU	SM-EEG	ER	SCT-DL
10	25.3	48.33	35.14	37.56	32.76	50.22
20	25.21	49.45	41.24	30.6	43.65	52.01
30	36.36	50.15	34.19	50.99	54.11	59.1
40	40.21	52.47	35.76	52.62	71.65	61.43
50	42.56	56.33	59.33	59.34	72.98	64.34
60	43.98	47.14	26.54	60.98	47.26	78.44
70	51.54	21.89	67.39	66.41	63.61	84.65
80	67.22	65.25	75.25	69.89	57.98	88.98
90	72.36	84.66	76.15	84.57	79.1	97.2

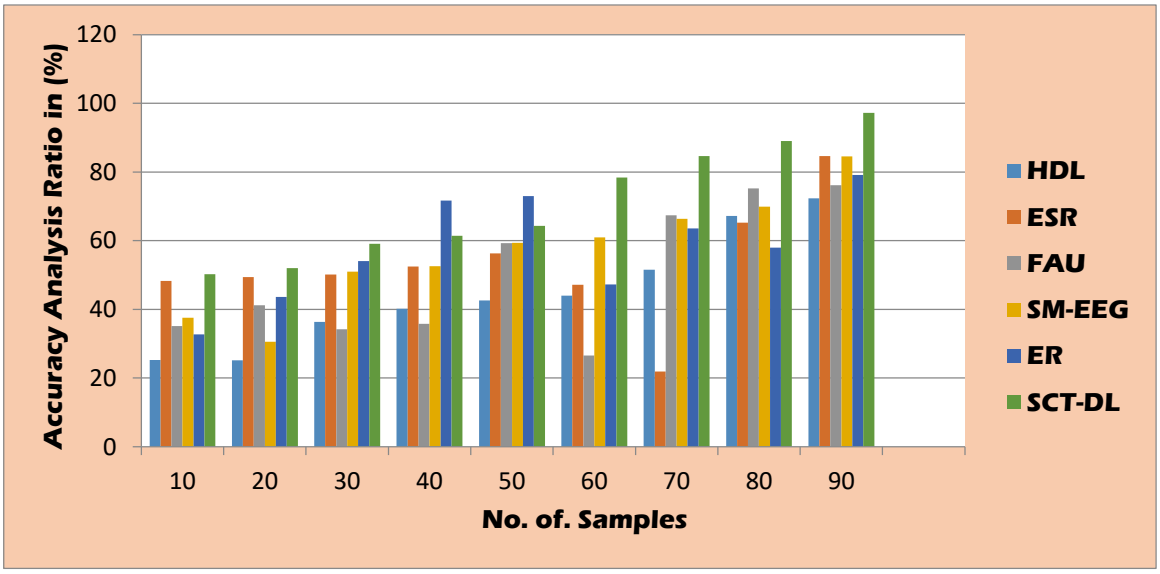

Fig.7. Accuracy Analysis

Method for verifying the accuracy analysis ratio is depicted in Figure 7. Number of datasets is on the X-axis, while accuracy is measured along the Y-axis. When evaluating and forecasting incoming data signals, the Accuracy of SCT-DL shows accuracy performance vs the temporal variation factor. Accuracy factor, which aids in meeting the aforementioned conditions, is denoted by equation (8).

4.1.2 Sensitivity Analysis

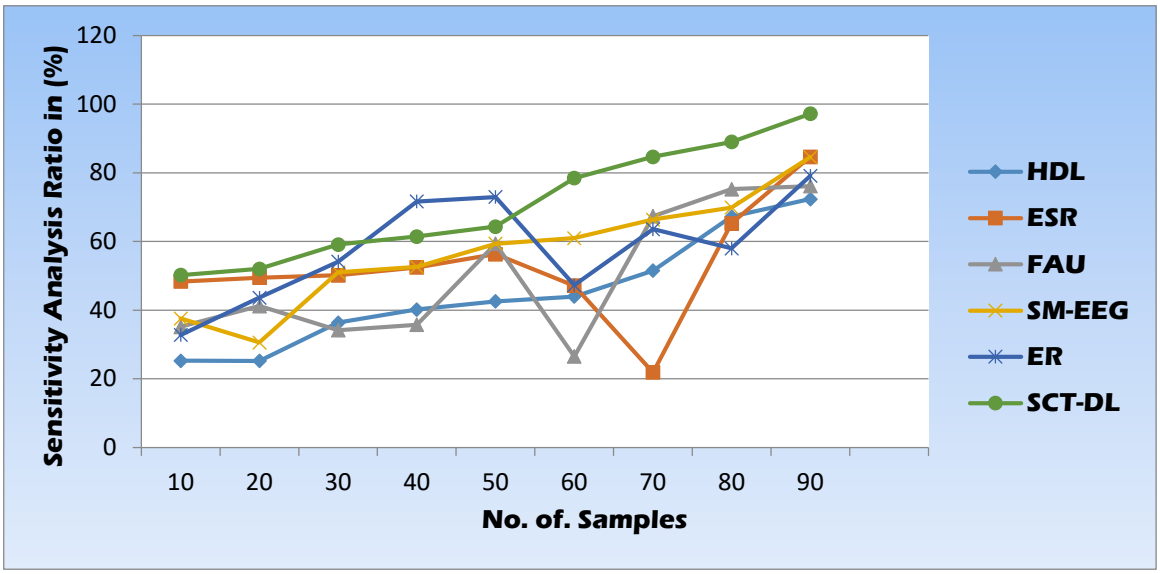

Fig.8. Sensitivity Analysis

Calculating the Sensitivity analysis ratio is shown in Figure 8. Sensitivity analysis is displayed along the y-axis, and sample size is displayed along the x-axis. SCT-DL When evaluating and anticipating input data signals, the performance versus the temporal variation element is critical. As seen in Equation (3), the Sensitivity indicator helps with the aforementioned goals**.**

4.1.3 Efficiency Analysis

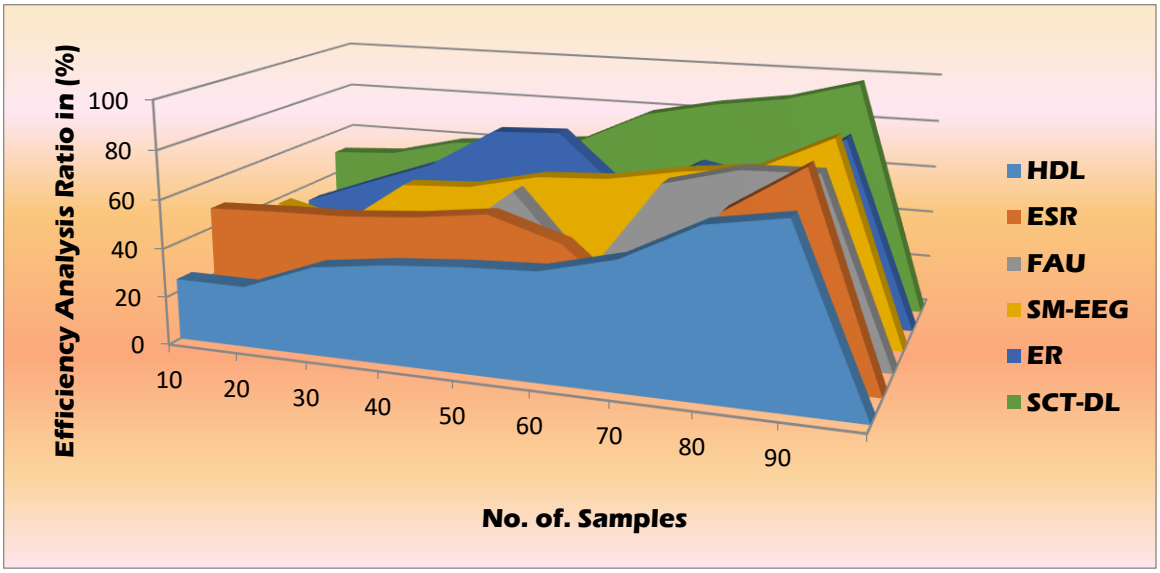

Fig.9. Efficiency Analysis

Figure 9 depicts one way to calculate the Efficiency analysis ratio. The graph's X-axis represents the total number of samples taken, while the Y-axis displays the Efficiency analysis's findings. SCT-DL When assessing and forecasting incoming data signals, the performance vs the temporal variation element is crucial. To make sure these requirements are met, we can use the Efficiency measure, which can be written as Equation (4).

4.1.4 Performance Analysis

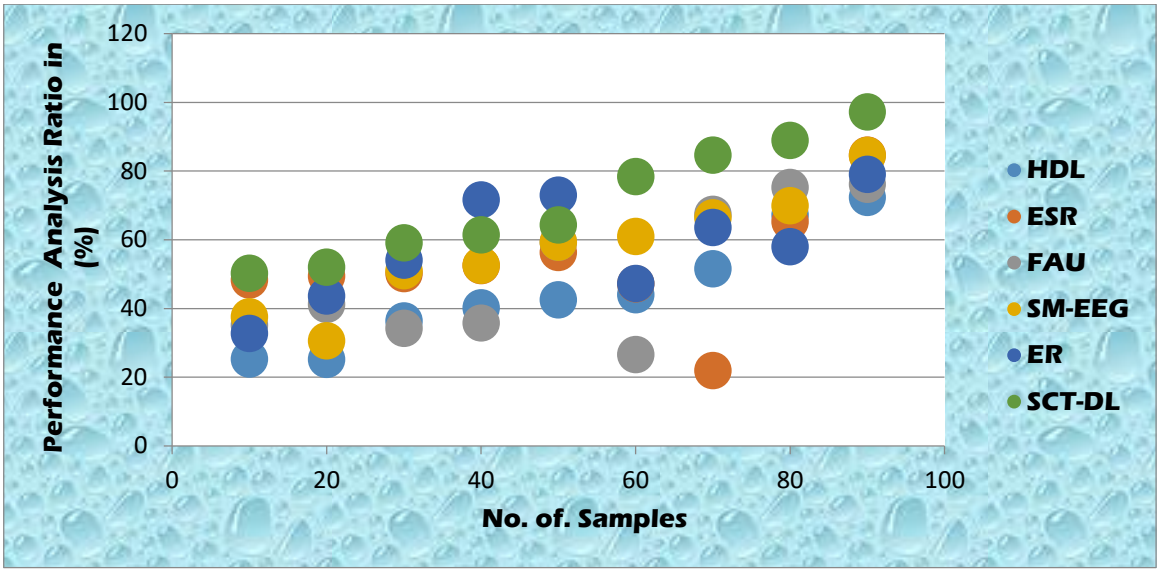

Fig.10. Performance Analysis

Several measures of the Performance analysis ratio are displayed in Figure 10 for your perusal. One measure of effectiveness is shown against the number of samples along an X-Y axis. In order to accurately assess and anticipate incoming input signals, CNN performance vs the temporal variation component, the SCT-DL Efficacy in Performance is crucial. Performance, which may be expressed as Equation. (13), helps meet these standards.

4.1.5 Energy Analysis

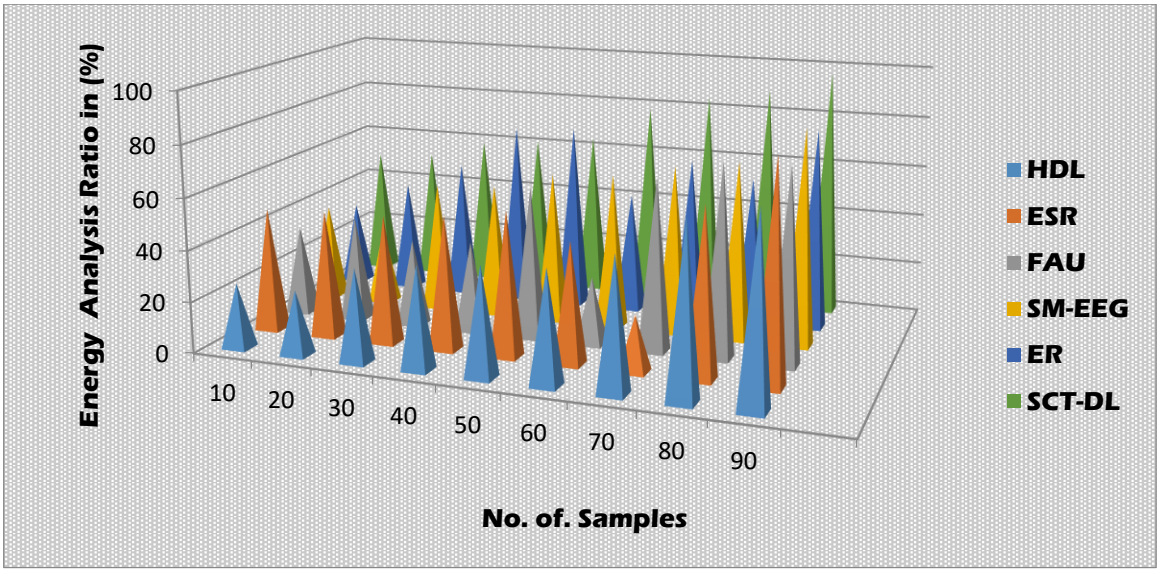

Fig.11. Energy Analysis

Figure 11 depicts one method for assessing the Energy analysis ratio. The Y axis shows how well the system is performing relative to the X axis, which indicates the number of samples. To accurately assess and predict incoming input signals, CNN performance vs the temporal variation component, the SCT-DL Efficacy in Energy is crucial. Performance, which may be expressed as Equation. (5), helps meet these standards.

Results from the simulations in the research lend credence to the claim that combining CR-NN with KT can boost prediction precision, speed, sensitivity, and user friendliness. Many points of comparison are made between the models, such as the factors taken into account by each (such as as HDL, ESR, FAU, SM-EEG, and ER).

5. Conclusion

Here, we provide a sparse coding approach to emotion recognition that is driven by speech and has proven to be highly effective. The SCT-DL data set was used to compare state-of-the-art algorithms, the suggested system performed exceptionally well.The first framework for stress detection using sparse coding and deep belief networks is presented. It overcame a shortcoming of conventional diagnostic procedures used in clinics, which rely mainly on structured

interviews for patient assessment. Detection of stress was improved by cutting-edge feature extraction methods. The excitation feature category was shown to be useless, while the vocal-tract feature category and the prosodic feature category were found to be superior at identifying tension and emotions. The issue of limited data size was overcome by a transfer learning approach. Statistically significant gains were shown when using transfer learning in place of the other models in the suggested technique, demonstrating the method's overall efficacy in detection. Current clinical diagnosis accuracy was surpassed by the proposed detection technology. As a whole, the presented models performed well and are suggested for diagnosis.

Based on our findings, we introduce a C-RNN model that combines CNN and RNN for emotion recognition and monitoring with multi-channel EEG inputs. Particularly, the CNN part may fuse, mine, and choose data based on correlations across channels and frequencies. On the other hand, the RNN (i.e. LSTM) based model structure can make use of the constructed frame cube sequences to learn long-term dependencies and contextual information. In the future, deep convolutional networks could be used for emotion recognition and diagnosis, expanding the scope of this work. The deep convolutional networks could benefit from Transfer Learning as well. Emotion recognition and diagnosis are two further areas where recurrent neural networks could be investigated.

Author Contributions: Ch. Suneetha conceptualized the study, designed the overall methodology, and supervised the research work. Vijay Keerthika was responsible for model development, implementation of the Sparse Coding Technique–Deep Learning framework, experimentation, and performance evaluation. M. Harshini contributed to data preprocessing, analysis of results, literature review, and manuscript drafting. All authors jointly discussed the results, contributed to refining the methodology, reviewed and edited the manuscript, and approved the final version of the paper.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes

References

[1] Saganowski, S. (2022). Bringing emotion recognition out of the lab into real life: Recent advances in sensors and machine learning. *Electronics*, *11*(3), 496.

[2] Fahad, M. S., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. *Digital signal processing*, *110*, 102951.

[3] Bhangale, K. B., &Mohanaprasad, K. (2021). A review on speech processing using machine learning paradigm. *International Journal of Speech Technology*, *24*, 367-388.

[4] Houssein, E. H., Hammad, A., & Ali, A. A. (2022). Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. *Neural Computing and Applications*, *34*(15), 12527-12557.

[5] Kwon, S. (2021). Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Applied Soft Computing*, *102*, 107101.

[6] Islam, M. R., Moni, M. A., Islam, M. M., Rashed-Al-Mahfuz, M., Islam, M. S., Hasan, M. K., ...&Lió, P. (2021). Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques. *IEEE Access*, *9*, 94601-94624.

[7] Banerjee, D., Islam, K., Xue, K., Mei, G., Xiao, L., Zhang, G., ...& Li, J. (2019). A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. *Knowledge and Information Systems*, *60*, 1693-1724.

[8] Arora, T. K., Chaubey, P. K., Raman, M. S., Kumar, B., Nagesh, Y., Anjani, P. K., ...&Debtera, B. (2022). Optimal facial feature based emotional recognition using deep learning algorithm. *Computational Intelligence and Neuroscience: CIN*, *2022*.

[9] Onyema, E. M., Shukla, P. K., Dalal, S., Mathur, M. N., Zakariah, M., & Tiwari, B. (2021). Enhancement of patient facial recognition through deep learning algorithm: ConvNet. *Journal of Healthcare Engineering*, *2021*.

[10] Dey, A., Chattopadhyay, S., Singh, P. K., Ahmadian, A., Ferrara, M., & Sarkar, R. (2020). A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition. *IEEE Access*, *8*, 200953-200970.

[11] Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, *53*, 209-221.

[12] Al Machot, F., Elmachot, A., Ali, M., Al Machot, E., &Kyamakya, K. (2019). A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors. *Sensors*, *19*(7), 1659.

[13] Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P. D., Murugappan, M., ...& Acharya, U. R. (2022). Automated emotion recognition: Current trends and future perspectives. *Computer methods and programs in biomedicine*, 106646.

[14] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., &Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, *7*, 19143-19165.

[15] Tiwari, P., &Darji, A. D. (2022). A novel S-LDA features for automatic emotion recognition from speech using 1-D CNN. *International Journal of Mathematical, Engineering and Management Sciences*, *7*(1), 49.

[16] Zhang, S., Liu, R., Tao, X., & Zhao, X. (2021). Deep cross-corpus speech emotion recognition: Recent advances and perspectives. *Frontiers in neurorobotics*, 162.

[17] Halim, Z., Waqar, M., & Tahir, M. (2020). A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowledge-based systems*, *208*, 106443.

[18] Das, S., Lønfeldt, N. N., Pagsberg, A. K., &Clemmensen, L. H. (2021). Towards Interpretable and Transferable Speech Emotion Recognition: Latent

Representation Based Analysis of Features, Methods and Corpora. *arXiv preprint arXiv:2105.02055*.

[19] Alex, S. B., Mary, L., &Babu, B. P. (2020). Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features. *Circuits, Systems, and Signal Processing*, *39*(11), 5681-5709.

[20] Joshi, M. L., &Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings*, *58*, 217-226.

[21] Kshirsagar, S. R., & Falk, T. H. (2022). Quality-aware bag of modulation spectrum features for robust speech emotion recognition. *IEEE Transactions on Affective Computing*, *13*(4), 1892-1905.

[22] Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-Garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, *46*, 147-170.

[23] Li, X., Song, D., Zhang, P., Hou, Y., & Hu, B. (2017). Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring. *International Journal of Data Mining and Bioinformatics*, *18*(1), 1-27.

[24] Banerjee, D. (2017). *Speech based machine learning models for emotional state recognition and ptsd detection* (Doctoral dissertation, Old Dominion University).

[25] Khorrami, P. R. (2017). How deep learning can help emotion recognition.

[26] Zhang, J., Yin, Z., Chen, P., &Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103-126.

[27] Li, X., Song, D., Zhang, P., Yu, G., Hou, Y., & Hu, B. (2016, December). Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 352-359). IEEE