Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2779

2024, 43(2) 20-33

Hybrid model for speech emotion recognition of normal and autistic children (SERNAC)

Maria Andleeb Siddiqui [a], Najmi Ghani Haider [b], Waseemullah [a, *], Syed Muhammad Nabeel Mustafa [a]

[a] *Department of Computer Science and Information Technology, NED University of Engineering and Technology, Karachi, Pakistan*

[b] *Department of Computer Science, Usman Institute of Technology University, Karachi, Pakistan*

[*] Corresponding author: Waseemullah, Email: waseemu@cloud.neduet.edu.pk

Received: 09 March 2023, Accepted: 28 March 2024, Published: 01 April 2024

K E Y W O R D S

Autistic

Angry

Emotions

Speech Recognition

Prosodic

A B S T R A C T

Since the last decade, autism spectrum disorder (ASD) has been used as a general term to describe a wide range of conditions, including autistic syndrome, Asperger's disorder, and pervasive developmental disability. This problem emerges as a decreased ability to share emotions and a greater difficulty understanding others' feelings, leading to increased social communication difficulties. To assist patients with ASD, we proposed a concept that incorporates speech emotion detection technologies, which are widely used in the field of human-computer interaction (particularly youngsters). An algorithm based on a novel method for classifying normal and autistic children's speech emotions is implemented in this article. The training data set is treated to a new approach after all features have been extracted. The technique discussed in this study is the creation of a hybrid algorithm that serves as a classifier for normal and autistic children's speech emotions. Voice emotion recognition can be identified accurately and with a lower error rate. The data collection includes speech samples from 200 normal and 250 autistic groups in four moods (Angry, Happy, Neutral and Sad). As per research findings, the implemented hybrid algorithm for Normal and Autistic Children Speech Emotions (SERNAC) outperformed the existing classifiers by increasing accuracy.

1. Introduction

Humans use a variety of sense processes to interact with their surroundings. Affective computing aims to improve human-computer interaction and tailor machine responses to human requirements. All of the information in the human body's environment is an input. As a result, the human body assesses this information and produces an emotional state [1].

Emotions have a critical role in all human experiences. Emotions with the advancement of technology require a more natural interface for human-computer interaction [2]. Affective computing aims to improve human-computer interaction and tailor machine responses to human requirements. Many researchers have been drawn to improving human-computer interaction over the last three decades. Research on vocal emotion analysis by C. Williams and K. Stevens in 1972 [3] was one of the first attempts. Speech emotion recognition systems are used for a variety of purposes, including 1) medical psychiatric diagnosis, 2) lie detection, 3) analysis of behavior study of call attendants where call center conversation is used to improve call attendant quality of service, 4) mental stress analysis during human conversation, 5) E-learning for

student emotional state, and 6) speech recognition system training to recognize stressed speech in aircraft cockpits to improve performance [4,5]. Every aspect of speech has emotions. The emotions expressed in speech are communicated between the sender and recipient throughout a conversation. The speaker's emotional state is easily triggered by the tone and manner of his or her speech. This is a process in which emotional states are transferred and shaped [6]. The acoustic element of speech carries vital emotion signs. The unique qualities of emotions are used to identify them. The Speech Emotions Recognition-SER is classified as a pattern recognition task with three main phases: feature extraction, feature selection, and pattern classification [7, 28]. The intensity, timing, pitch, articulation, and voice quality of a spoken signal are all strongly linked to the underlying emotion [8,29].

The majority of acoustic features employed in SER fall into two categories: prosodic and spectral features. The speaker's prosodic qualities are frequently used to offer key emotional clues. Prosodic features are produced from pitch and energy contour statistics [9], whereas spectral features are derived from the spectrum of speech. Spectral features express the frequency contents of the speech signal and offer matching information for prosodic features [10,11].

Finding the speaker's emotional states through speech is the goal of a speech emotion detection algorithm [12]. The most popular methods for classifying voice emotion include Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Hid-den Markov Model (HMM), and Neural Networks (NN). Different kinds of spectral and prosodic features on emotional speech classification have been researched in order to evaluate and compare the effects of GMM [13]. Investigations were conducted into the effects of the initial parameter settings (number of mixture components and number of used iterations) on the GMM training procedure. The accuracy of emotion categorization was then investigated in another study to see how the number and order of parameters in the input feature vector, as well as the computing complexity, affected the results [14]. German, English, Swedish, Mandarin, and Indian languages all have relevant databases. Three key components of the design of a voice emotion recognition system were highlighted in a research survey [15]. The first step was choosing the proper speech representation features. The creation of an appropriate classification scheme is the second problem, and creating an appropriate classification system correctly is the third problem. The third concern is how to properly prepare an emotional speech database for system performance evaluation [16]. Global and local prosodic qualities derived from sentences, words, and syllables have been proposed for speech emotion or affect recognition.

In this paper, a novel strategy known as the hybrid algorithm is applied to classify the speech emotions of typically developing and autistic children. On the Berlin Emotional Database (EMO DB), this algorithm is evaluated in comparison to the Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Neural Network (NN), and Dynamic Time Wrapping (DTW) classification methods. The data set includes speech samples from 200 normal and 250 autistic groups in all four emotional states—angry, happy, neutral, and sad—for the algorithm's implementation. This modelling approach was developed because it can be difficult to distinguish between speech samples using machine learning algorithms with small datasets, especially when trying to identify emotions that sound quite similar in features, such as speaking excitedly happy emotions, which can occasionally resemble angry emotions, etc. Therefore, modelling and analysis by hybrid algorithms will analyze the results of each individual speech sample.

2. Literature Review

In this study, which comprised over 40 depressed participants and 40 control participants (both male and female), the effects of various speech patterns on categorization outcomes were investigated. To quantitatively identify the characteristics of depressed speech, the analysis of variance (ANOVA) results is combined with the Gaussian Mixture Model (GMM) and the support vector machine. Similar to how the classifier configuration for speech emotion recognition was utilized to respond to the question, "How speech segments should be selected?" What characteristics allow for effective discrimination? What advantages might feature normalization offer, given the speaker-specific nature of mental illness? A database of 23 depressed and 24 healthy persons was built using audio and video data. The challenge of identifying depression from subject recordings utilizing speech processing and learning was examined [17]. This study discovered that when it comes to identifying depression from spoken words, the harmonic model's features work better than alternatives. There were 148 participants in total, including 50 men and 98 women, in this trial. Another study discovered that it was possible to reliably gather various

objective speech acoustic measurements over the phone from 35 depressed individuals who had been referred by their doctors. It also demonstrated that it is possible to get voice acoustic data that reflect the severity of depression [18]. G. Deshmukh looked into the two feature vector methodologies that were used as well as the effects of giving the classifier more feature vectors. It examines the precision of speech classification in Marathi, Hindi, and Indian English. 80% of Indians were found to speak English correctly [19]. K. Tarunika used k-nearest neighbour (k-NN) and a deep neural network (DNN) to identify emotions in speech, particularly when the subject was in a critical frame of mind. The system's primary application is in healthcare institutions [20]. Mel Frequency Cepstral coefficients (MFCC), speech signal energy, and the Berlin database of emotional speech are used as feature inputs in M. Ghai's technique. The Berlin Database of Emotional Speech was the database used in [21]. A speech emotion identification model was developed in this study to assist children with autism spectrum disorder (ASD) in identifying emotions in social interactions. The machine learning model, which is based on the Support Vector Machine (SVM), is created using the Python programming language. Speech processing inputs have been demonstrated to be relatively accurately classified by SVM. Individual audio data sets are being created with the intention of training emotion identification models. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is one such speech corpus that was utilized in this study to train the model. As part of the pre-processing procedure, acoustic feature extraction will be carried out using Python libraries. This method made use of the libROSA library. The machine learning model is trained using the zero-crossing rate (ZCR) and the first 26 Mel-frequency Cepstral Coefficients (MFCCs) as acoustic features. The final SVM model's test accuracy was 77 %. This model did well when there was a lot of background noise present in the RAVDESS audio recordings, producing a test accuracy of 64% [22]. One method was selected because it is more similar to how the human auditory system responds than other systems. The Support Vector Machine (SVM) was developed by Chervonenkis and Vapnik in the 1990s and is currently the most advanced method for classifying data. SVM is used to categorize human voice recognition in various research. The RBF kernel was the most widely used kernel from SVM Multi-Class in prior studies. This is so that SVM, which makes use of the Radial Basis Function (RBF) kernel, can increase accuracy. The study's maximum accuracy ratio was 72.5 % with a frame size of 0.001 seconds, 80 filter banks, [0.3-0.7] gamma, and 1.0 C values [23]. The article describes the audio processing of the samples, methods for including background noise, and feature extraction coefficients used in model creation and training. In one study, the utilization of all samples across all three datasets, the in-corporation of background sounds, and the performance of several modelling techniques are all examined. In order to evaluate the efficacy of the ensemble learning strategy, the optimum hyperparameters configuration for the models was chosen using the majority voting method. When it came to categorizing emotions, the ensemble learning model performed better overall than the MLP model, reaching a high accuracy of 66.5 % [24]. Mel Frequency Cepstral Coefficient and Discrete Wavelet Transform were used as the feature extraction method.

3. Methodology

The Hybrid algorithm is based on a novel way to categorize normal and autistic children's speech emotions is implemented. After extracting all features, the training data set is subjected to a new technique. Support Vector Machine (SVM), KNN (K-Nearest Neighbour), NN (Neural Network), and DTW were used to train and identify normal and autistic children's spoken emotions (Dynamic Time Wrapping speech samples from 200 normal and 250 autistic groups in all four moods, which will be used to develop the algorithm (Angry, Happy, Neutral and Sad). Support Vector Machine, K-Nearest Neighbour, Neural Network, and Dynamic Time Warping will be utilized to categories the emotions of anger, happiness, neutrality, and sadness for both normal and autistic children's verbal emotions. Fig. 1 depicts the flow diagram.

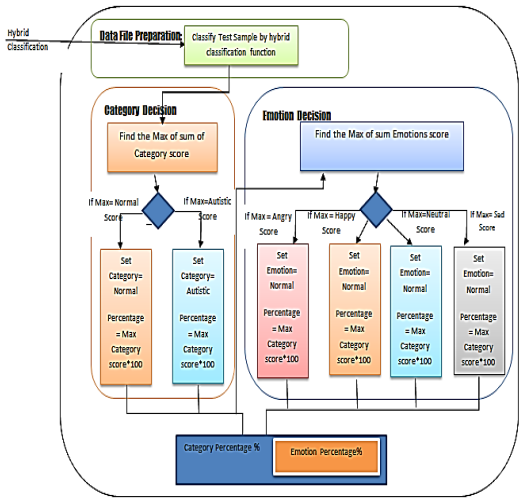


Fig. 1. Proposed Algorithm for Decision Making And Recognition

3.1 Flow of Algorithm

The algorithm is implemented in the following steps as follows:

Step.1 The extracted features discussed previously are used. They consist of seven prosodic features and three coding schemes. After the extraction of features, the threshold for each feature is set in order to carry out experiments for classification accuracy.

Step.2 Classification function is incorporated to classify the test samples. Maximum category score position is calculated as shown below:

if result of position rc=1

Set category to normal.

Write the percentage.

else

Set category to autistic.

Write the percentage.

[End of if structure]

Step.3 After category score percentage is calculated, the maximum.

emotion score is calculated as follows:

if result of position rn=1

Set emotion to Angry.

Write the percentage.

else if rn=2

Set emotion to Happy.

write the percentage.

else if rn=3

Set emotion to Neutral.

Write the percentage.

else Set emotion to Sad.

Write the percentage.

[end of if structure]

Step.4 The results of category and emotion label are saved, and percentage is extracted to display.

3.2 Speech Dataset

The data evaluated in this study were collected from 200 normal and 225 autistic children of age group 10 - 13 years of both genders in four emotions. The Urdu language sentences which consist of maximum suprasegmental Phonemes are used in this study with four different emotions (Angry, Happy, Neutral and Sad). The sentences consist of the following phonemes: /f/, /z/, /kh/,/gh/,/q/. 200 normal Children uttered the sentences in four emotions, so the total data set consists of 800 samples for normal children speech. Similarly, 225 autistic children uttered the sentences in four emotions, so the total data set consists of 900 samples for autistic children speech. To implement the speech emotion corpus, the sentences were chosen keeping in mind that they were suitable to utter by the subjects and contain maximum phonetic information. The description of training and validation samples is shown in Fig. 2.

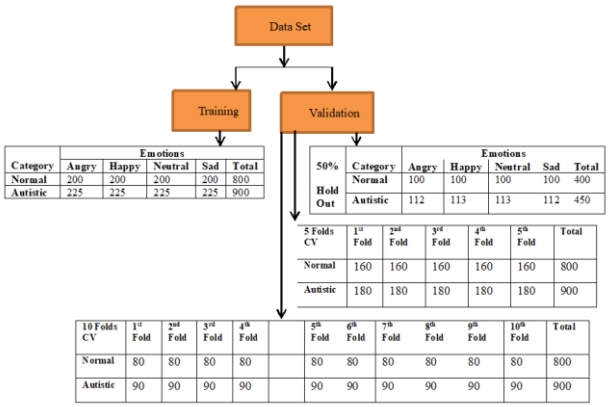


Fig. 2. Description Of Speech Samples for Training And Validation

For corpus recording with specifications, the following ITU recommendations have been applied: SNR> = 45 dB and 24 120 bps bit rate. The speaker's utterances were recorded using the microphone and sound recorder included into Windows 10 with a sensitivity of 56 dB 25 dB and a sampling rate of 48 KHz.

The total system flow diagram for the recognition and classification of speech emotion is shown in Fig. 1. Prior to creating the data file, the system flow performs the thresholding of the extracted features. The algorithm's decision-making process employs the technique for creating data files for each sample file. The first file to load is the speech.mat example file [25]. After the file has been loaded, the file's name, categories, and emotions are extracted by path. Then speech.mat is updated with the file name. Speech.mat now includes files with information on emotions and categories in addition to this data. The thresholding feature extraction is completed using the following function.

Features= allfeatures_extraction(wav_file);

Implementation of Decision-making Function for Category and Emotion percentage.

The function for decision making is shown below.

Function [Category, Category_Percentage, Emotion, Emotion_Percentage]
=Hybrid_Decision_Making(wav_file)

By employing a hybrid classification algorithm and locating the highest category score position, the test samples were categorized. The classification procedure is described as follows:

Step.1 If Result of position n = 1

Set category = Normal and Category_Percentage = Categories_output (1).

Else

category = Autistic and Category_Percentage = Categories_output (2).

Step.2 After the category is decided as normal or autistic, the emotion is labelled on the category according to the classification accuracy. The maximum score position is to be estimated using the following function.

[rn, cn] =find (strcmp (Emotion_outputs, max (Emotions_output)))

If Result of position n = 1

Set Emotion = Angry and Emotion _Percentage= Emotions_output (1).

Else If Result of position n = 2

Set Emotion = Happy and Emotion _Percentage= Emotions _output (2).

Else If Result of position n = 3

Set Emotion = Neutral and Emotion_Percentage= Emotions_output (3).

Else

Set Emotion = Sad and Emotion _Percentage= Emotions _output (4).

[End of If Structure]

1) Feature Extraction and selection

This step includes features extraction we extracted 14 features from categories mentioned below:

a) Prosodic features

The features which give information about emotional cues of speaker are the prosodic features.

i. Pitch contour
ii. Intensity
iii. Rate of acceleration
iv. Formant frequencies
v. Log energy
vi. Log power
vii. Zero crossing rate (ZCR)

b) Coding schemes used for extraction of spectral features.

i. Mel Frequency Cepstrum Coefficient (MFCC)
ii. Mel Frequency Cestrum Coefficient (MFCC)
iii. Relative Spectral Transform- Perceptual Linear Prediction (RASTA-PLP)

c) Correlation in data

The sampling technique used in this study is "Simple Random Sampling." There are four big institutions in Karachi that deal with autistic children. Normal children's speech data as control group is collected from different schools in Karachi. The speech samples of total 200 children of age group 10 - 13 years are involved in Normal speech corpus with four emotions (Angry, Happy, Neutral and Sad). This study focusses on Autistic children in comparison with the normal children's speech emotions to formulate an algorithm that can help the doctors dealing the children with autism spectrum disorder [26]. For the Autistic Children Speech Emotion Corpus, 225 children of the same age group are considered from five institutions in Karachi (45 samples from each institution). Group 1 data is collected from the Department of Psychology, Civil Hospital, Karachi; Group 2 data is collected from the Institute of Behavioral Psychology, Karachi; Group 3 data is collected from NUST School of Behavioral sciences, Karachi Group 4 data is collected from Special Children Education; and Group 5 data is collected from Institute of Clinical Psychology, Karachi. The Pearson's correlation coefficient of data samples is given in Table 1.

Table 1

Pearson's correlation coefficient among data samples of different groups

	Group 1	Group 2	Group 3	Group 4
Group 2	0.305			
Group 3	0.434	0.419		
Group 4	0.563	0.475	0.662	
Group 5	0.368	0.207	0.410	0.202

It is evident that the data obtained from different sources is correlated with each other. Group 1 has a

strong positive correlation between Group 3 and Group 4, whereas a medium positive correlation was found between Group 2 and Group 5. Group data is strongly correlated between Group 3 and Group 4, whereas it is weakly correlated with Group 5. Group 3 data is strongly and positively correlated between Group 4 and Group 5. Group 4 data is weakly and positively correlated with Group 5. All the Pearson's Correlation Coefficient values are positive for each group.

The algorithm is implemented in two ways.

a) Full multiclass classifier design

b) Divide and conquer classifier design

a) Full multiclass classifier design

The flow diagram of full multi class problem is shown in Fig. 3. The hybrid algorithm discussed in this section consists of classification modules that are; Artificial Neural Network, K-Nearest Neighbour and Support Vector Machine. These three classification modules are organized as nodes; it is a 3- layer network:

i. Input layer (Speech Features input)

ii. Hybrid layer (NN, KNN, SVM nodes)

iii. Output layer (probability function layer)

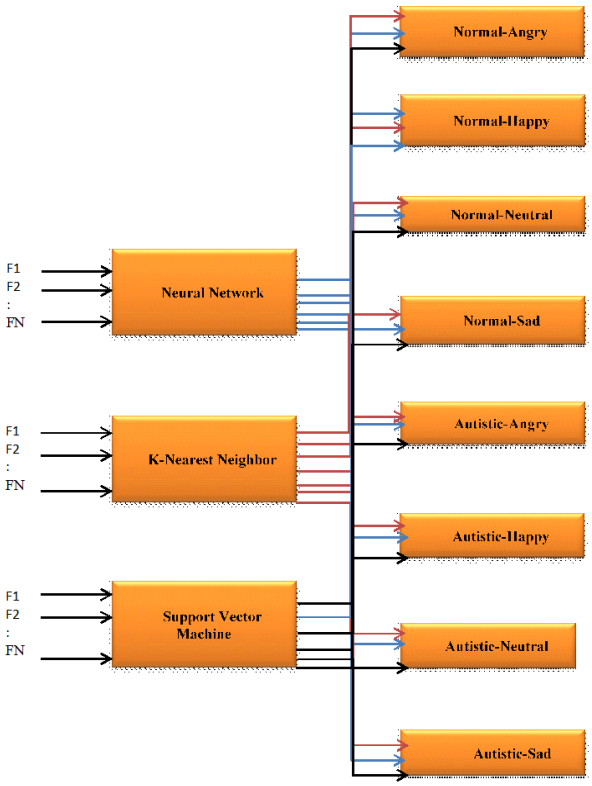


Fig. 3. Full Multiclass Problem for Speech Emotion Classification Of Normal And Special Children

b) Divide and conquer classifier design

The flow diagram of categorical analysis (normal and autistic), emotional analysis (angry, happy, neutral, and sad) and the combined strategy of the hybrid algorithm are shown in Fig. 4, Fig. 5 and Fig. 6, respectively.

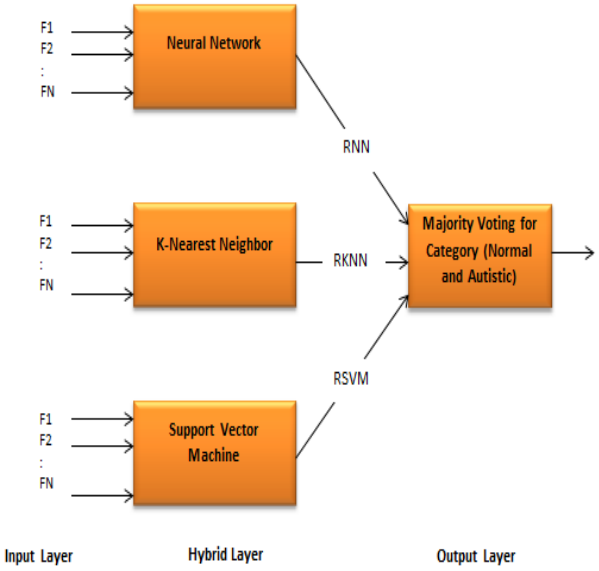


Fig. 4. Flow Diagram of Categorical Classification

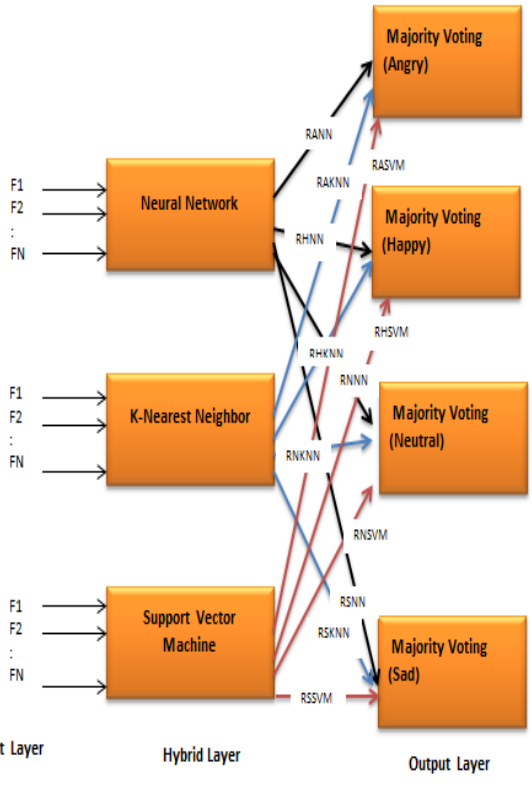


Fig. 5. Flow Diagram of Emotional Classification

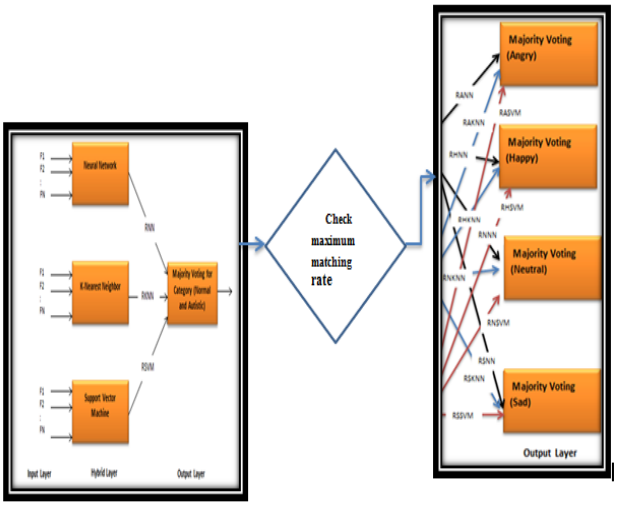


Fig. 6. Flow Diagram of Hybrid Algorithm Strategy

From the Literature Review discussed in Section 2, several machine learning algorithms have been found effective for speech emotion recognition using audio data set. Out of which the combination of following machine learning algorithms was studied and tested for the implementation of the hybrid model using ANOVA.

1) Support Vector Machine (SVM)

2) K-Nearest Neighbour (KNN)

3) Artificial Neural Network (ANN)

The proposed model will pass two outputs in which it provides a matching rate of speech sample with normal and autistic category. Input sample is going to three selected modules that is Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Neural Network (NN), the result of input layer is forward to hybrid layer of three modules and then the result of hybrid layer forward to output layer. Dynamic Time Warping was also selected due to its ANOVA statistics result, but it showed lower classification accuracy as shown in Table 2, so it is not included in the model.

Table 2

Classification accuracy of machine learning classifiers.

Classifiers	Normal Children				Autistic Children			
	Angry	Happy	Normal	Sad	Angry	Happy	Normal	Sad
SVM	100 %	83. %	83. %	66 %	66 %	83. %	83. %	100%
KNN	83. %	33 %	25 %	10 %	83. %	83. %	33. %	10 %
NN	83. %	66 %	83. %	66 %	50. %	66 %	83. %	0%

Here the combination of four supervised learning algorithms is used to deeply analyze the minor difference of data sample with different ways like to analyze features by SVM maximum geometrical margin method, KNN's maximum vote by neighbor's method with Euclidean distance calculation and Neural network Scaled conjugate gradient back propagation.

Here, the hybrid model consists of the following layers.

a) Input layer (Speech Features).

b) Hybrid layer (Feature extracted for NN, KNN, SVM)

c) Output layer (probability function layer).

a) Input layer (Feature Extraction)

First after basic steps or Signal acquisition, pre-processing, feature extraction the dataset organized for model training/testing will go in hybrid layer model to recognize the category of speech whether it is Normal or Autistic and matching rate by feature analysis. This feature part is completely discussed in speech processing and feature extraction in previous research [27].

b) Hybrid layer (Matching rate extraction by feature analysis)

Here R1 describes the matching rate according to specific model of hybrid layer and F is features, A is matrix in which we are storing our Weights according to input. Here W is Weight, F1, F2, F3……FN define the features, RSVM is weight from SVM per feature, RKNN is the weight from KNN per feature, RNN is the weight of NN per feature and a1, a2, a3 represent the nodes of hybrid layer. A is matrix in which we are storing our matching rates according to input as shown in Fig. 7.

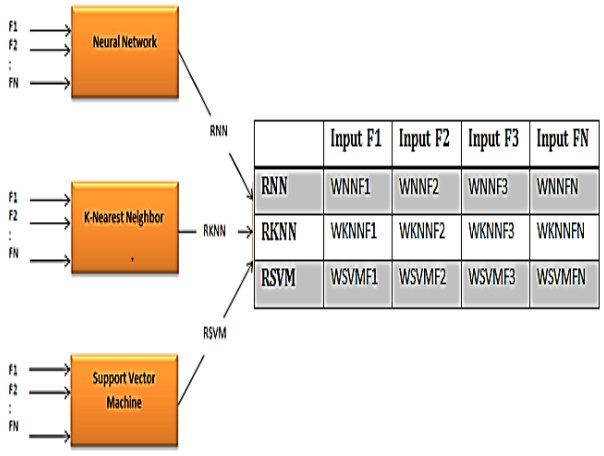


Fig. 7. Matching Rate Extraction in Hybrid Layer

After this the correlation method is applied to choose highly correlated features weights for training process. Now this matrix of highly correlated weights dataset will be input to the next

hidden layer for training. The mathematical representation of this layer has two output nodes and network for emotions recognition has four nodes. It can be seen in Fig. 8 that Matrix B with b1 to b4 is defining the node we are configuring according to the input 'a1-an' are the results which were extracted from the function that is forwarded to the next layer nodes. From the results shown in Table 3 (a), R1 describes the weights and result according specific of 1st hidden layer and F is features, is matrix in which we are storing our Weights according to input. Here W is Weight, R1, R2, R3……RN define the features, SVM, KNN, NN and DTW results and weights, RKNN is the weight from KNN per feature a1, a2, a3, a4 represent the per node of hybrid layer as shown in Tables 3(a) and 3(b).

a) Output layer

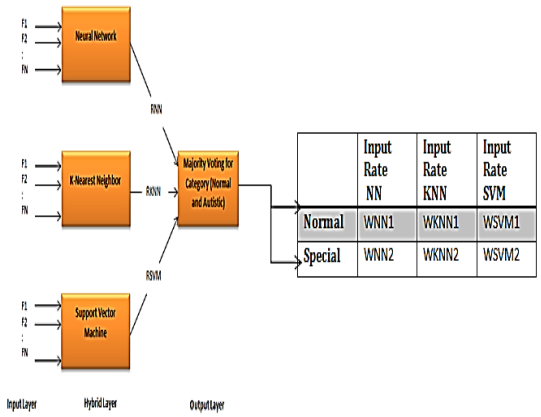
Then condition will check which category has the maximum matching rate then according to the matching rate data feature will be passed to emotions model.

Here in this layer the probability is calculated. As total Models in hybrid layer are 3 so the probability.

$$Probability = Number\ highest\ rate\ /\ Number\ of\ total\ Model$$

Here W is Weight, WSVM is weight from SVM per feature, WKNN is the weight from KNN per category, WNN is the weight of NN per feature and c1 and c2 represent the per category of hybrid layer. If the rate of any 2 or all 3 model of any category will be greater than 0.5 than that category will get highest rate. Here we are calculating probability of both categories and then for Emotions as shown in Table 4(a), 4(b) 4(c) and 4(d).

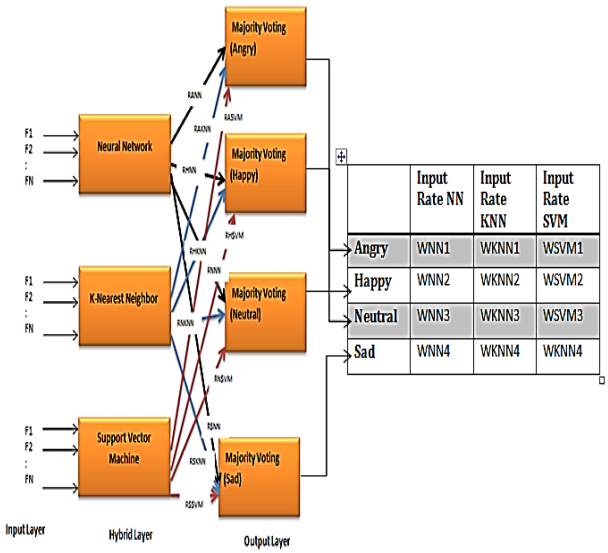
If the rate of any 2 or all 3 model of any category will be greater than 0.5 than that category will get highest rate. Here the probability of both categories is calculated as shown in Fig. 8.



	Input Rate NN	Input Rate KNN	Input Rate SVM
Normal	WNN1	WKNN1	WSVM1
Special	WNN2	WKNN2	WSVM2

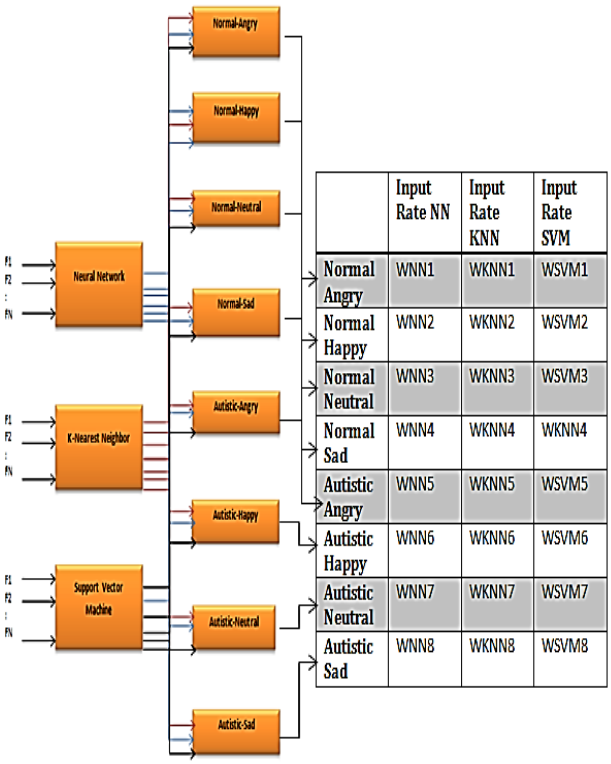
Fig. 8. Checking Of Maximum Matching Rate for Category

After recognition of category sample, it will be sent to the specific category emotion model where same process applied to emotion recognition process but with different features. By separating the emotion model of categories complex and minor differences will be detected properly. If the rate of any 2 or all 3 model of any category will be greater than 0.5 than that emotion will get highest rate. Here the probabilities of all emotions are calculated as shown in Fig. 9 and Fig. 10.



	Input Rate NN	Input Rate KNN	Input Rate SVM
Angry	WNN1	WKNN1	WSVM1
Happy	WNN2	WKNN2	WSVM2
Neutral	WNN3	WKNN3	WSVM3
Sad	WNN4	WKNN4	WKNN4

Fig. 9. Checking Of Maximum Matching Rate for Emotion Using Divide And Conquer



	Input Rate NN	Input Rate KNN	Input Rate SVM
Normal Angry	WNN1	WKNN1	WSVM1
Normal Happy	WNN2	WKNN2	WSVM2
Normal Neutral	WNN3	WKNN3	WSVM3
Normal Sad	WNN4	WKNN4	WKNN4
Autistic Angry	WNN5	WKNN5	WSVM5
Autistic Happy	WNN6	WKNN6	WSVM6
Autistic Neutral	WNN7	WKNN7	WSVM7
Autistic Sad	WNN8	WKNN8	WSVM8

Fig. 10. Checking Of Maximum Matching Rate for Emotion Using Multiclass Problem

Table 3(a)

Hybrid layer Matching Rate Matrix of Features

	1st Result Matching rate	2nd Result Matching rate
SVM result and Weight (a1)	$f(x1) = \sum \alpha_1 y_1 \; x_1^T x + b$	$f(x2) = \sum \alpha_2 y_2 \; x_2^T x + b$
KNN result and Weight (a2)	$D(x1,y1) = \sqrt{\sum_{i=1}^{k} (x_1 - y_1)^2}$	$D(x2,y2) = \sqrt{\sum_{i=2}^{k} (x_2 - y_2)^2}$
Neural Network result and Weight (a4)	$f(x1) = \sum_{i=1}^{n} w_1 x_1$	$f(x2) = \sum_{i=2}^{n} w_2 x_2$

Table 3(b)

Hybrid layer Matching Rate Matrix of Features

	3rd Result Matching rate	Nth Result Matching rate
SVM result and Weight (a1)	$f(x3) = \sum \alpha_3 y_3 \; x_3^T x + b$	$f(xN) = \sum \alpha_n y_n \; x_n^T x + b$
KNN result and Weight (a2)	$D(x3,y3) = \sqrt{\sum_{i=3}^{k} (x_3 - y_3)^2}$	$D(xN,yN) = \sqrt{\sum_{i=n}^{k} (x_n - y_n)^2}$
Neural Network result and Weight (a4)	$f(x3) = \sum_{i=3}^{n} w_3 x_3$	$f(xn) = \sum_{i=n}^{n} w_n x_n$

$$f(x) = \sum_{i=i}^{n} \left[\sum \alpha_i y_i \; x_i^T x + b \quad \sqrt{\sum_{i=i}^{k} (x_i - y_i)^2} \quad \sum_{i=1}^{n} w_i x_i \right] \tag{1}$$

$$wr = \sum_{i=i}^{n} \left[\sum \alpha i y i x i \quad \frac{1}{d^2} \quad \sum_{i=i}^{n} (w_i x_i) \right] \tag{2}$$

After this we applied correlation method to choose highly correlated features weights to input to the output layer.

Table 4(a)

Probabilities of Category (Normal and Autistic Speech)

	SVM W	KNN W
Normal (c1)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Special (c2)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$	$\frac{Obtained\ \ normal\ w}{Total\ weight}$

Table 4(b)

Probabilities of Category (Normal and Autistic Speech).

	NNW
Normal (c1)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Special (c2)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$

After recognition of category sample will be send to the specific category emotion model where same process applied to emotion recognition process but with different features. By separating the emotion model of categories complex and minor differences will be detected properly. Similarly, if weight of any 2 or all 3 model of any category will be greater than 0.5 then that emotion will get highest weights.

Table 4(c)

Probabilities of Emotions (Angry, Happy, Neutral, Sad)

	WSVM	WKNN
Angry (c1)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Happy (c2)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Neutral (c3)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Sad (c4)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$	$\frac{Obtained\ \ normal\ w}{Total\ weight}$

Table 4(d)

Probabilities of Emotions (Angry, Happy, Neutral, Sad)

	WNN
Angry (c1)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Happy (c2)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Neutral (c3)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$
Sad (c4)	$\frac{Obtained\ \ normal\ w}{Total\ weight}$

4. Results and Discussion

This section describes the classification accuracy obtained by the implemented algorithm and its performance comparison with other machine learning algorithms autistic children speech emotions recognition. The suggested technique was compared to established machine learning algorithms such as support vector machine (SVM), Neural Network (NN), K- Nearest Neighbour (KNN), and Dynamic Time Wrapping (DTW) classifiers in terms of classification accuracy. In this chapter, the implemented hybrid algorithm for Normal and Autistic Children Speech Emotions

(SERNAC) is compared to traditional classification algorithms in order to evaluate the implemented approach's performance. Support Vector Machine, K-Nearest Neighbour, Artificial Neural Network, and Dynamic Time Warping were among the classifiers utilized.

The classification performance is shown in Table 5.

Table 5

Comparison of classification accuracy (ca) results on different classifiers with SERNAC

Classifiers	Normal Children				Autistic Children			
	Angry	Happy	Normal	Sad	Angry	Happy	Normal	Sad
SVM	100%	83.%	83.%	66%	66%	83.%	83.%	100%
KNN	83.%	33%	25%	10%	83.%	83.%	33.%	10%
NN	83.%	66%	83.%	66%	50.%	66%	83.%	0%
DTW	83.%	53.%	10.%	0.%	10.%	50.%	10.%	0.%
SERNAC	80.%	66.%	100.%	75.%	66.%	100.%	83.%	100.%

It is shown that:

i. For normal speech samples SVM classifier shows 100% accuracy in angry emotion and for autistic children speech SVM classifier shows 100% in sad emotion.
ii. For normal speech samples KNN classifier shows the best performance of 83.3% in angry emotion while for neutral and sad emotion it shows little accuracy. For the autistic children speech, KNN shows best performance in angry and sad emotions.
iii. For normal Speech samples ANN classifier shows 83.3% accuracy in neutral and angry emotion of normal speech and for autistic speech, it shows best performance for neutral emotion.
iv. For normal speech samples DTW classifier shows highest accuracy for
v. angry emotion and shows no recognition of sad emotions.
vi. SERNAC performed better in Normal category on Angry with 80%, Happy with 66%, Neutral with 100% AND Sad with as well as Autistic category on Angry with 66.67%, Happy with 100%, Neutral with 83.33% and Sad with 100%.

Table 6

Error rate of SVM, DTW, KNN, ANN and SERNAC

Classifiers	Normal Children				Autistic Children			
	Angry	Happy	Normal	Sad	Angry	Happy	Normal	Sad
SVM	0%	16.%	16.%	33%	33%	16.%	16.%	0%
KNN	16.%	66%	75%	90%	16.%	16.%	66.%	90%
NN	16.%	33%	16.%	33%	50.%	33%	16.%	100%
DTW	16.%	50.%	90.%	100.%	90.%	50.%	90.%	0.%
SERNAC	0.%	33.%	0.%	0.%	33.%	0.%	16.%	0.%

It is shown in Table 6 that DTW classifier shows the most error rate in almost all emotions except angry emotion of normal speech. SERNAC algorithm produces the least error rate.

The significant impact of classifiers and emotions on classification is assessed using two-way Analysis of Variance (ANOVA) testing. In Table 7, a two-way analysis of variance (ANOVA) is used to investigate the impact of two independent factors on a single dependent variable. It looked at each independent variable's main effect as well as any possible interactions between them. The level of significance ‘α’ is taken as 0.05. The general model of equation of Two-way ANOVA is represented in Eq. (3)

$$y = \mu + \alpha i + \beta j + (\alpha\beta)ij \quad (3)$$

Where, i=1, 2….n, j=1, 2,…….m

y = Independent factor, consider as Classification Accuracy

μ = Constant Value

αi = Main effect of first factor, consider as (emotion_ angry, emotion happy, emotion neutral, emotion sad)

βj = Main effect of second factor, consider as (Classifiers, Classifier, Classifier ANN, Classified, Classified)

(αβ) ij = Two factor interaction effect.

The following preliminaries are defined in Two-way ANOVA testing.

- Degree of freedom (DF): The degree of freedom (or number of values) refers to how much of an ANOVA's final computation is subject to change. It is the maximum number that a dynamic system can move independently of any imposed constraints.

- Adjusted sum of squares (Adj SS): a measure of deviation from the mean is the. In an ANOVA, the sum of squares is used to express the overall variation that may be attributable to several factors.
- Adjusted Mean Squares (Adj MS): A mean squares value is obtained by multiplying the sum of squares value by the relevant degree of freedom. It is employed to ascertain whether or not certain variables have a significant impact.
- F-value: The F-value is the difference between the two mean squares. If the null hypothesis is correct, then this value is typically very near to 1. It establishes the significance of the specific factor.
- P-value: Each P-value has a corresponding F-value. The crucial value, which is 0.05, and is the p-value linked to specific F-statistics, serves as the standard for the alpha level of significance. P-value less than 0.05 indicates that a factor has a substantial impact. To evaluate the relevance of the model, two-way ANOVA statistics of Classification Accuracy versus classifiers and emotions are used.

Table 7

The two-way ANOVA statistics

Source	DF	SS	MS	F	P
Classifiers	4	2196.8	5490.45	12.05	0.0
Emotions	3	4367.1	1455.7	3.20	0.46
Interactions	12	8764.6	730.39	1.6	0.169
Error	20	9112	455.6		
Total	39	44205.5			

R-Sq = 79.39%

R-Sq (adj) = 59.81%

It shows that:

i. The effect of classifiers on Classification Accuracy is significant i.e., not all means are equal since $P<\alpha$.
ii. The effect of emotions on Classification Accuracy is significant i.e., not all means are equal since $P<\alpha$
iii. The interaction of both factors is not significant as $P>\alpha$.
iv. The R-Square value is an indication that the goodness of fit for the model is 79.39%.
v. The R-Square (adjusted) value is an indication of the proportion of variables in the dependent variables accounted for by the explanatory variables.

The residual plot of Classification Accuracy (CA) is shown in Fig. 11. It shows that.

i. The values of Classification Accuracy are normally distributed although there are some outliers.
ii. The residuals are divergent from the mean value; it means that the residuals have constant variance.
iii. The residuals are dependent on each other as they are showing a specific pattern.

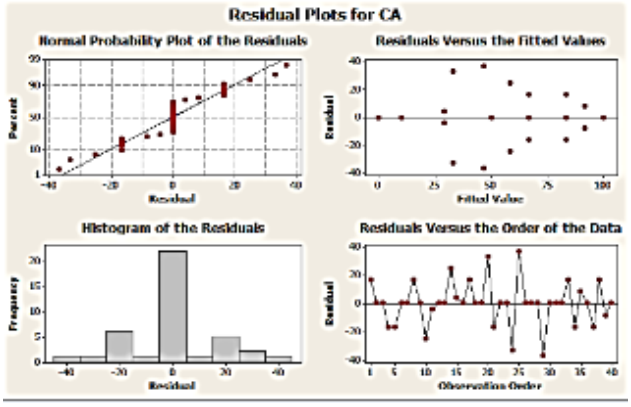


Fig. 11. Residual plot of classification accuracy (CA).

The main effect plot for Classification Accuracy (CA) with respect to the classifiers and emotions are shown in Fig. 12 (a) and Fig. 12 (b), respectively.

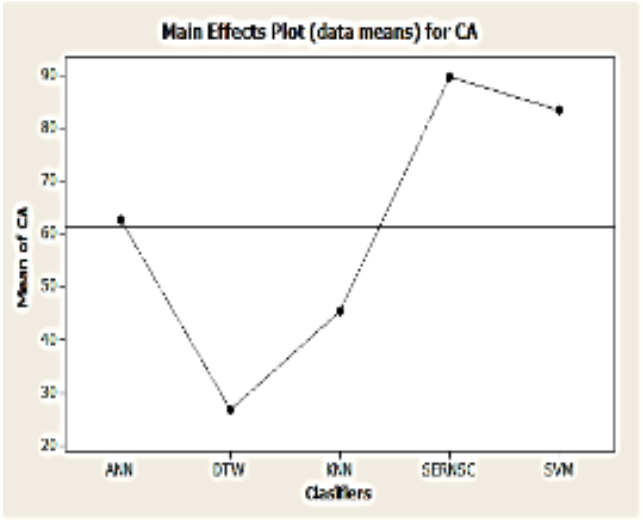


Fig. 12. (a): Main effect plot of classification accuracy with respect to classifiers

Fig. 12(a) shows that:

i. The best Classification Accuracy is shown by Speech emotion Recognition of Normal and Autistic children (SERNAC) algorithm.
ii. Support Vector Machine (SVM) classifier shows the Classification Accuracy around 85%.
iii. Artificial Neural Network (ANN) shows the Classification Accuracy of 62%.

iv. K-Nearest Neighbour (KNN) shows the Classification Accuracy of around 45%.
v. Dynamic Time Warping (DTW) shows the Classification Accuracy of around 25%.

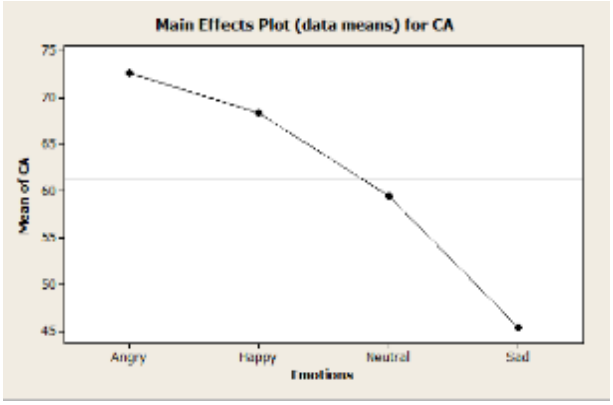


Fig. 12. (b): Main effect plot of classification accuracy (CA) with respect to emotions

Fig. 12(b) shows that.

i. Angry Emotion produces the best Classification Accuracy around 73%.
ii. The Happy emotion produces the Classification Accuracy of around 69%.
iii. The Neutral emotion produces the Classification Accuracy of around 60%.
iv. The Sad emotion produces the Classification Accuracy of around 45%.

The interaction plot of both the dependent factors i.e., Classifiers and Emotions on Classification Accuracy is shown in Fig. 13.

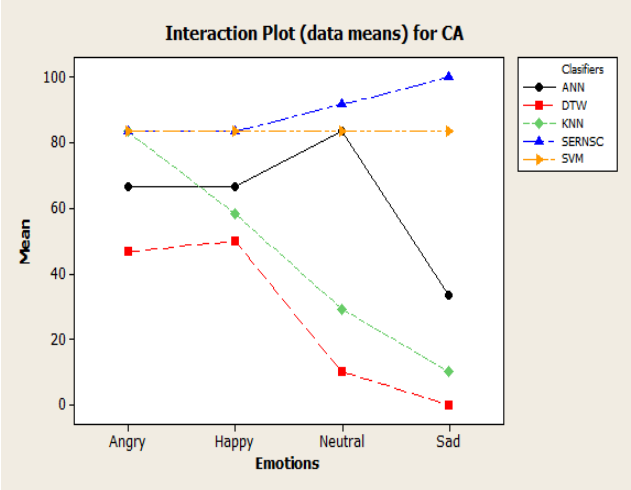


Fig. 13. Plot of emotions and classifiers on classification accuracy

Fig. 13 shows that

i. Dynamic Time Warping (DTW) classifier didn't interact with any other classifier on any emotion.
ii. Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifiers interact with other on Neutral emotion.
iii. K nearest neighbour (KNN), Support Vector Machine (SVM) and Speech emotion recognition for Normal and Autistic children (SERNAC) interact with each other on Angry emotion.
iv. Support Vector Machine (SVM) and Speech emotion recognition for Normal and Autistic children (SERNAC) interact with each other on Happy emotion.
v. Overall SERNAC curve is in the range in Normal category on Angry with 80%, Happy with 66%, Neutral with 100% AND Sad with as well as Autistic category on Angry with 66.67%, Happy with 100%, Neutral with 83.33% and Sad with100%.

5. Conclusion

The implemented hybrid algorithm for Normal and Autistic Children Speech Emotions (SERNAC) is explained in this article, and the results are compared to current classification algorithms to determine the efficiency of the technique. The classifiers used included Support Vector Machine, K-Nearest Neighbour, Artificial Neural Network, and Dynamic Time Warping. In comparison to typical machine learning classifiers, it displays greater classification accuracy in most of the emotions of both normal and autistic children. The data set was also tested on the Berlin Emotional Database (EMO-DB), with an overall Classification Accuracy of 87.5% on male children emotion and 83.2% on female children speech emotion, though the algorithm was not tested for Autistic children because there is no such data base available globally in this regard. According to the findings of this study, the Hybrid algorithm may greatly improve the accuracy of speech emotion recognition for ASD children's speech emotion datasets when the dataset is small as it performed better in Normal category on Angry with 80 %, Happy with 66%, Neutral with 100% and Sad with as well as Autistic category on Angry with 66.67%, Happy with 100 %, Neutral with 83.33% and Sad with 100 % which is over all.

6. References

[1] F. Y. Leung, V. Stojanovik, M. Micai, C. Jiang, and F. J. A. R. Liu, "Emotion recognition in autism spectrum disorder across age groups: A cross-sectional investigation of various visual and auditory communicative domains," 2023.

[2] T. C. Day, I. Malik, S. Boateng, K. M. Hauschild, M. D. J. J. o. A. Lerner, and D. Disorders, "Vocal emotion recognition in Autism: behavioral performance and event-related potential (ERP) response," pp. 1-14, 2023.

[3] C.E. Williams, K.N. Stevens," Emotions and speech: some acoustic correlates", The Journal of Acoustical Society of America, Vol.52, No.4, pp. 1238-1250, 1972.

[4] A. Landowska et al., "Automatic emotion recognition in children with autism: a systematic literature review," vol. 22, no. 4, p. 1649, 2022.

[5] E.Yuncu, "Speech emotion recognition using auditory models", MS dissertation, Dept. of Cognitive Science, The Graduate School of Informatics Institute of Middle East Technical University, September 2013.

[6] J. Tao, T.Tan, "Affective computing: A review", Proc. of International Conference on Affective Computing and Intelligent Interaction (ACII), Springer, pp. 981-995, Berlin, Heidelberg, 2005.

[7] M. E. Ayadi, M. S. Kamel and F. Karray, 'Survey on speech emotion recognition: features, classification schemes, and databases", Pattern Recognition, Vol. 44, No.16, pp. 572-587, 2011

[8] A.S Utane, S.L. Nalbalwar, "Emotion recognition through speech" International journal of applied information systems, Vol. 1, pp. 5-8, 2013.

[9] S. Farashi, S. Bashirian, E. Jenabi, and K. J. I. J. o. D. D. Razjouyan, "Effectiveness of virtual reality and computerized training programs for enhancing emotion recognition in people with autism spectrum disorder: a systematic review and meta-analysis," pp. 1-17, 2022.

[10] M. Milling et al., "Evaluating the impact of voice activity detection on speech emotion recognition for autistic children," vol. 4, p. 837269, 2022.

[11] T. Johnstone, K. Scherer, "The effects of emotions on voice quality", Proc. of the 14th international congress of phonetic sciences, pp. 2029-2032, San Francisco, 1999.

[12] S. Wu, T.H. Falk, W.Y. Chan, "Automatic speech emotion recognition using modulation spectral of features", Speech Communication, Vol.53, pp. 768-785, 2011.

[13] J. Pribil and A. Pribilova, "Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech", EURASIP Journal of Audio, Speech and Music Processing, Springer, Vol.1, No.1, pp. 1-22, 2013

[14] M. Milling et al., "Evaluating the impact of voice activity detection on speech emotion recognition for autistic children," vol. 4, p. 837269, 2022.

[15] X. Xiao, J. Li, E.S. Chng, H. Li, C.H Lee, "A study on the generalization capability of acoustic modes for robust speech recognition", IEEE Transaction on Audio, Speech and Language Processing, Vol. 18, No.6. August 2010.

[16] J. M. Garcia-Garcia, V. M. Penichet, M. D. Lozano, and A. J. U. A. i. t. I. S. Fernando, "Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions," vol. 21, no. 4, pp. 809-825, 2022.

[17] K. S. Rao, S.G. Koolagudi, R.R Vempada, "Emotion recognition from speech using global and local prosodic features", International Journal of Speech Technology, Vol.16, pp. 143–160, 2013.

[18] M. Asgari, I. Shafran, L. B. Sheeber, "Inferring clinical depression from speech and spoken utterances", Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1 – 5, Reims, France, 21-24 Sept. 2014.

[19] J.C. Mundt, P. J. Snyder, M.S. Cannizzaro, K. Chappie, D.S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology", Journal of neurolinguistic, Vol. 20, No. 1, pp.50-64, Jan 2007.

[20] G. Deshmukh, A. Gaonkar, G. Golwalkar and S. Kulkarni, "Speech based emotion recognition using machine learning," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 812-817, doi: 10.1109/ICCMC.2019.8819858.

[21] K. Tarunika, R. B. Pradeeba and P. Aruna, "Applying machine learning techniques for speech emotion recognition," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-5, doi: 10.1109/ICCCNT.2018.8494104.

[22] K. Soltani and R. N. Ainon, "Speech emotion detection based on neural networks," 2007 9th international symposium on signal processing and its applications, 2007, pp. 1-3: IEEE.

[23] R. Matin and D. Valles, "A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions," 2020 Intermountain Engineering, Technology and Computing (IETC), 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249147.

[24] R. A. A., M. Nasrun and C. Setianingsih, "Human emotion detection with speech recognition using mel frequency cepstral coefficient and support vector machine," 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), 2021, pp. 1-6, doi: 10.1109/AIMS52415.2021.9466077.

[25] D. Valles and R. Matin, "An audio processing approach using ensemble learning for speech-emotion recognition for children with ASD," 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0055-0061, doi: 10.1109/AIIoT52608.2021.9454174.

[26] Ali, S.A., Haider, N.G. and Andleeb, M., 2016. “Evaluating the performance of learning classifiers and effect of emotions and spectral features on speech utterances”. International Journal of Computer Science and Information Security, 14(10), p.406.

[27] J. M. Garcia-Garcia, V. M. Penichet, M. D. Lozano, and A. J. U. A. i. t. I. S. Fernando, "Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions," vol. 21, no. 4, pp. 809-825, 2022.

[28] M. S. Farooq, R. Tehseen, M. Sabir, and Z. J. S. R. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," vol. 13, no. 1, p. 9605, 2023.

[29] I. Voinsky, O. Y. Fridland, A. Aran, R. E. Frye, and D. J. I. J. o. M. S. Gurwitz, "Machine learning based blood RNA signature for diagnosis of autism spectrum disorder," vol. 24, no. 3, p. 2082, 2023.

