Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07261-x

ORIGINAL ARTICLE

GRaNN: feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals

Avishek Garain[1] · Biswarup Ray[1] · Fabio Giampaolo[2] · Juan D. Velasquez[3] · Pawan Kumar Singh[4] · Ram Sarkar[1]

Received: 22 July 2021 / Accepted: 29 March 2022

Abstract

Compared to other features of the human body, voice is quite complex and dynamic, in a sense that a speech can be spoken in various languages with different accents and in different emotional states. Recognizing the gender, i.e. male or female from the voice of an individual, is by all accounts a minor errand for human beings. Similar goes for speaker identification if we are well accustomed with the speaker for a long time. Our ears function as the front end, accepting the sound signs which our cerebrum processes and settles on our disposition. Although being trivial for us, it becomes a challenging task to mimic for any computing device. Automatic gender, emotion and speaker identification systems have many applications in surveillance, multimedia technology, robotics and social media. In this paper, we propose a Golden Ratio-aided Neural Network (GRaNN) architecture for the said purposes. As deciding the number of units for each layer in deep NN is a challenging issue, we have done this using the concept of Golden Ratio. Prior to that, an optimal subset of features are selected from the feature vector extracted, common for all three tasks, from spectral images obtained from the input voice signals. We have used a wrapper-filter framework where minimum redundancy maximum relevance selected features are fed to Mayfly algorithm combined with adaptive beta hill climbing (AβHC) algorithm. Our model achieves accuracies of 99.306% and 95.68% for gender identification in RAVDESS and Voice Gender datasets, 95.27% for emotion identification in RAVDESS dataset and 67.172% for speaker identification in RAVDESS dataset. Performance comparison of this model with existing models on the publicly available datasets confirms its superiority over those models. Results also ensure that we have chosen the common feature set meticulously, which works equally well on three different pattern classification tasks. The proposed wrapper-filter framework reduces the feature dimension significantly, thereby lessening the storage requirement and training time. Finally, strategically selecting the number units in each layer in NN help increases the overall performance of all three pattern classification tasks.

Keywords Multilayer perceptron · Golden ratio · Mayfly algorithm · RAVDESS · Gender classification · Emotion recognition · Speaker identification

1 Introduction

Identification of speech and voice in any form has various real-life applications. Also, gender identification from voice has applications like automatic gender identification in telephonic calls, video classification with labelling, multimedia indexing, gender-based advertisements, speech feeling acknowledgment, and programmed greetings and sound classification. Sometimes medicine suggestion and diagnosis using artificial intelligence (AI) bots like Alexa, Echo dot, etc., on the basis of the voice of the user can be more customized if the gender is recognized correctly. Similarly speaker and emotion identification via voice can be used in crime investigation purposes by pointing out the criminals from their telephonic conversations. Also, it can be used as part of security systems of the future generation as a voice-print biometric signature. Recognizing the emotion in the voice can help AI agents in setting up a more personalized experience for their users like playing motivational videos when detecting sadness in user's voice, playing music based on detecting the emotion in the user's voice, etc. Previously due to less availability of labelled data and computational constraints, semi-supervised learning algorithms came into existence as an alternative

Extended author information available on the last page of the article

Published online: 08 May 2022

approach to the commonly used supervised learning methods. With the increase in data availability, however, these constraints are no longer valid in many research fields, and such fields are reaching to a new high owing to successful use of the deep learning models.

The voice of an individual not only provides the meaning of words spoken, but also contains some characteristics of the speaker. For example, the gender, the identity, the age or the emotional state of the speaker can be understood from the voice of that person. A characteristic voice acknowledgment framework is the human ear, which is considered as a very efficient instrument. This can naturally recognize the sexual orientation of the speaker by voice and speech using the characteristics like recurrence and uproar. This very fact motivates the researchers to think whether a machine can be instructed to do likewise. Hence, methods are devised to build an automatic system with the information extracted from the voice data for the said purpose. These algorithms procure acoustic qualities, like length, force, recurrence and separation. In this paper, we have tried to develop a computerized gender, emotion and speaker identification system using voice information by transforming the human voice from the analog to the digital form in order to extract useful features and then constructed an identification model.

The rest of the paper has been organized as follows: Sect. 2 provides a literature survey about the works done on this topic. The methodology used is described in Sect. 3. This is followed by the results and concluding remarks in Sects. 4 and 5, respectively.

2 Literature survey

In this section, some recently published significant works on gender, emotion and speaker identification are discussed.

2.1 Gender and emotion identification

In recent years, several researchers have leveraged classic supervised machine learning classifier models to predict the gender and emotion from voice signals. In the work by Buyukyilmaz et al. [6], the results related to a frequency-based baseline model, logistic regression model, classification and regression tree (CART) model, random forest model, boosted tree model, support vector machine (SVM) model, XGBoost model, stacked model for Voice Gender dataset have been shown for comparison purpose with their architecture. According to the used models, the results are shown in Table 10.

The paper by Shafran et al. [39] has explored the problem of automated and accurate extraction of voice signatures from a speaker's voice. They basically have analysed dual approaches for extraction of speaker traits. The first approach focuses on general acoustic and prosodic features, while the second works on the choice of words used by the speaker. In the first approach, they have shown that standard speech or non-speech hidden Markov model (HMM), conditioned on speaker traits and evaluated on cepstral and pitch features, achieves accuracy well above chance for all examined traits. Another approach, using the SVM with rational kernels, has been applied to speech recognition lattices. This method has gained an accuracy of approximately 81% in binary classification of emotion. They have used speech data corpus collected from a deployed customer-care application (HMIHY 0300).

The authors Yacoub et al. [42] in their paper have reported the results of emotion recognition from speech signals. They have mainly focused on extracting emotion-based features from the short utterances that are important for interactive voice response (IVR) applications. It involves distinguishing neutral speech from anger, which finds its importance in call centre applications. They have also worked on classification of emotions comprising of boredom, happy, sadness and cold anger. Comparison of the results using decision trees, neural networks and SVM, K-nearest neighbours (KNN) has also been done. The database that they have used was created by the Linguistic Data Consortium at University of Pennsylvania, which consists of records by 8 actors expressing 15 emotions. Results conclude that hot anger can be distinguished from neutral utterances with above 90% accuracy.

Several researchers have also implemented various unsupervised and semi-supervised machine learning models for the task. Li Wei et al. [24] in their work have designed a feasible identification system for non-semantic voice information using language and gender. Their system is text and speaker independent and concatenated the language and gender identification models. It has utilized feasible acoustic features and an optimal model-training method. From the four different Gaussian mixture modelling (GMM) training approaches, they have evaluated the system performance in terms of recognition rate. Their analysis has shown that the method they have followed achieves accuracies of 85.25% and 93.2% for language and gender recognition tasks, respectively.

In the work by Livieris Ioannis et al. [27], the authors have presented a semi-supervised algorithm for the purpose of gender identification using voice information and named it iCST-Voting. Their algorithm has constituted an ensemble of the prominent self-labelled algorithms like Self, Tri and Co-training by using them as base learners. The contribution of their approach compared to the state-of-the-art approaches is that they have utilized an ensemble of classifiers as base learners in place of using a single learner as used in any other self-labelled algorithm resulting in better classification performance than classical supervised algorithms.

However, the techniques discussed above yielded unsatisfactory results in terms of accurately classifying the gender and emotion for the voice signals.

Hence, research models using a hybrid of both the neural network and machine learning models were also utilized for the task. Zvarevashe et al. [44] have worked on a gender recognition method from voice, which uses feature selection by using the random forest recursive feature elimination (RF-RFE) algorithm with gradient boosting machine (GBM) algorithm. The GBM algorithm has been evaluated against the feedforward neural network and extreme machine learning algorithm. The results obtained indicate that GBM outperforms all the algorithms compared against it in classification accuracy, and it is proved to be a suitable methodology for gender voice recognition.

Pahwa et al. [32] proposed a gender recognition system where they have used voice samples of 46 speakers using hybrid frameworks using both deep learning and machine learning models. Specifically, they made use of one of the most popular, dominating and most applied speech feature, Mel-frequency cepstral coefficients (MFCCs) and the derivatives of first and second orders. Their proposed model makes use of SVM and neural network classifiers using a stacking methodology.

Scherer et al. [36] have explained the reason behind the achievement of an accuracy level of studies investigating the ability of listeners to recognize different emotions from a wide variety of standardized vocal stimuli, which largely exceeds what can be called a chance expectation. The work summarizes the reporting of a series of studies related to recognition conducted by Scherer, Banse and Wallbott in nine different countries in the continents of the North America, Europe and Asia. They have made use of vocal emotion portrayals containing content-free sentences which were produced by professional German actors.

Nasef et al. [31] in the year 2021 presented two self-attention-based models to deliver an end-to-end voice gender recognition system under unconstrained environments. The first model consisted a stack of six self-attention layers and a dense layer. To the first model, a set of convolution layers and six inception-residual blocks were added before the self-attention layers in the second model. Mel-frequency cepstral coefficients (MFCCs) were used as a representation of the audio data. Also, logistic regression was used for classification. The experiments were performed under unconstrained environments such as background noise, different languages, accents and ages of the speakers.

2.2 Speaker recognition

In the past few years, researches have leveraged several neural network frameworks to classify the speaker from the voice signals.

Joon Son Chung et al. [9] in their paper have tried to solve the problem of speaker recognition under unconstrained noisy conditions. They have put forward a very large-scale audio–visual-based speaker recognition dataset which has been crawled from open-source media. Using a fully automated pipeline, they have curated the dataset VoxCeleb2, which consists of more than a million utterances by over 6000 speakers. They have also developed and made comparison of convolutional neural network (CNN) models and training pipelines that can effectively carry out the task of speaker identification from voice under different conditions. The models that have been trained on the VoxCeleb2 dataset have superseded the performance of previous works on benchmark datasets by a considerable amount.

Zhenhao Ge et al. [15] in their work have built the speaker recognition system based on neural networks with the network parameters being optimized with grid search. They have made use of regularization parameters, which are dynamically reduced in order to avoid termination of training in any local minima. In speaker verification, performance has been improved by making use of normalizing of prediction score, which works by rewarding the speaker-identity indices with distinct peaks and penalizing the weak ones with high scores. The corpus used here was the TIMIT corpus with a sampling rate of 8K. From the corpus, 200 male speakers have been used in order to train and test the classification performance. For validation purposes, the testing data has been used as registered speakers from within domain, while data from the remaining 126 male speakers have been used as test data representing speakers from out of domain.

In 2021, Biswas et al. [4] made use of an enhanced approach to identify singer voice using neural network. The neural network model takes into consideration some songs to create the training data. The efficiency was observed for the detection of new and unknown signer to be detected. The neural network models used for classification was ANN and MLP. Both of these models were tested, and the better performing model was used to predict the speakers/singers for the song test data.

Though the neural network-based models provide state-of-the-art results, they are generally very computationally expensive, which sometimes create a major problem in a resource-constraint environment.

Frameworks were also implemented by using machine learning models. Bisio et al. [3] have developed an android

SPEech proCessing plaTform as smaRtphone Application (SPECTRA) for gender, speaker and language recognition by making use of multiple unsupervised SVM classifiers. Generally, every classifier is trained by making use of a fixed training set. This makes it difficult for making decisions based on new data. The authors in their work have overcome this problem. Every time an user acquires a new audio, all the feature vectors are extracted and sent to the web server for retraining purposes and passed through the SVMs for classification purposes. But the framework failed to provide accurate results for the classification purposes.

Several review papers by comparing various methods for the speaker recognition task have also been done by researchers. In his paper, Douglas A. Reynolds [34] provided a brief overview of the area of speaker recognition. He has described the various applications, underlying techniques used and some indications of performance. Following this overview, he has discussed some of the advantages and loopholes of current speaker recognition technologies. He has also outlined some potential future trends in research, development and applications.

In the article by Hansen et al. [19], they have reviewed various works based on speaker recognition by machines imitating humans. They have emphasised on state-of-the-art speaker-modelling techniques that have emerged in the last decade. Discussion has been made based on different aspects of automated architectures. This includes voice-activity detection (VAD), speaker models, features, datasets for standard evaluation, and performance metrics. Speaker recognition has been discussed involving forensic speaker-recognition methods and illustrating how a naive listener performs this task from a neuroscience perspective.

The above discussions reveal that over the past few years, the traditional machine learning approaches have been explored by the researchers for classification of emotion, gender and speaker from voice signals. However, the major downfalls of such models are that only a few number of models performed all the three tasks single-handedly. This brings in the need for developing an efficient model, which can serve all the said classification tasks and combine them into a single and effective system. There is also a major disarray regarding the features to be used from the voice signals in order to feed the models. Hence, it brings a need for an efficient and accurate feature selection method on the common set of features extracted from the voice clips.

2.3 Motivation and contributions

The effectiveness of the classifiers is greatly determined by the quality of the features available in training data. Therefore, evoking useful voice features plays an important role in developing an efficient model since the human voice is liable for many non-useful and extraneous information. Works on improvement of the efficiency of voice classification systems are abundant, particularly on analysing the process of efficient feature extraction from the voice data, which include identifying the language content of voice signal components and removing non-useful contents such as background noise. Over the years, several methods have been developed on gender and emotion identification as well as speaker identification, and many of those methods have managed to gain very precise results. However, to the best of our knowledge, there are only a few number of works where the researchers performed all the three tasks single-handedly. However, it is quite challenging to design a single model for gender, emotion and speaker identification, which can work on the same set of features, and yields precise output. This inspires us to design a model, which can serve all these said purposes by a single system. Here, we have extracted features from the voice clips, applied feature selection on the feature vector and fed them to the customized neural network model. In a nutshell, we have contributed in the following way:

- We have converted the input voice signals into spectral images and extracted a common set of features for emotion, gender and speaker identification tasks.
- We have used a wrapper-filter-based feature selection framework, where minimum redundancy maximum relevance (mRMR) selected features are fed to the meta-heuristic Mayfly algorithm combined with adaptive beta hill climbing (AβHC) algorithm for selecting the optimal feature subset.
- The final classification model, called GRaNN (Golden Ratio-aided Neural Network), has a backbone of a deep multilayer perceptron (DMLP). As deciding the number of units for each layer in DMLP is a challenging issue, we have done this using the concept of Golden ratio.
- We have evaluated our model on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [28] dataset and Voice Gender dataset for emotion and gender identification purposes. We have also performed speaker identification on the RAVDESS dataset, which has not done in literature till date.
- Our experimental results reveal that the proposed model performs better than many state-of-the art methods.

3 Proposed methodology

The flowchart of the whole pipeline is shown in Fig. 1. Our proposed methodology has four main modules—(a) Feature engineering, (b) Scaling, (c) Feature selection and (d) Classification. These modules are explained in the following subsections in detail.

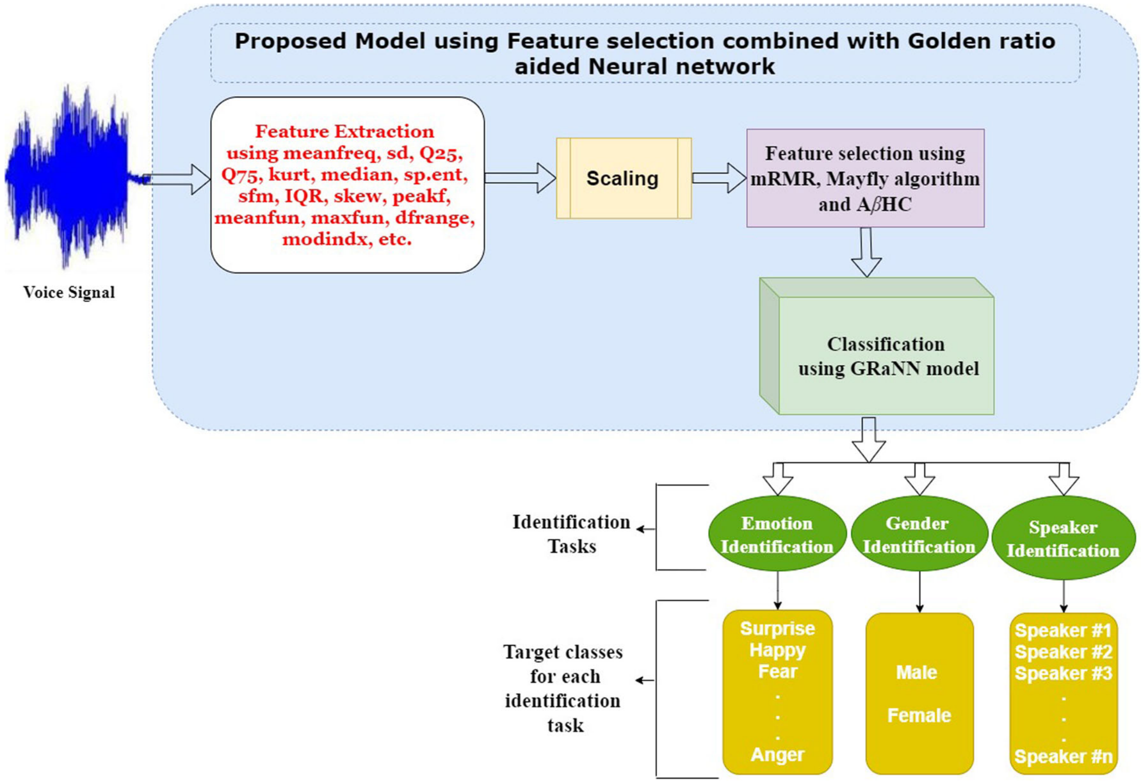

Proposed Model using Feature selection combined with Golden ratio aided Neural network
Voice Signal
Feature Extraction using meanfreq, sd, Q25, Q75, kurt, median, sp.ent, sfm, IQR, skew, peakf, meanfun, maxfun, dfrange, modindx, etc.
Scaling
Feature selection using mRMR, Mayfly algorithm and AβHC
Classification using GRaNN model
Identification Tasks
Emotion Identification
Gender Identification
Speaker Identification
Target classes for each identification task
Surprise
Happy
Fear
Anger
Male
Female
Speaker #1
Speaker #2
Speaker #3
Speaker #n

Fig. 1 Working pipeline of our proposed feature selection using GRaNN model

3.1 Feature engineering

It is worth mentioning that there are a large number of features [14] that can be extracted from a voice signal, so we need to meticulously extract the features and process them according to our requirements. For loading the audio files and preparing it for feature extraction, we have made use of the library Librosa [30]. The following acoustic properties of each voice in both the datasets are measured and applied as features for classification purpose of all kinds:

- meanfreq: Mean of frequencies (in kHz)
- sd: Standard deviation of frequency
- Q25: First quartile of frequency(in kHz)
- Q75: Third quartile of frequency(in kHz)
- kurt: Kurtosis
- median: Median of frequencies (in kHz)
- sp.ent: Spectral entropy
- sfm: Spectral flatness (calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum)
- mode: Mode of frequencies
- centroid: Frequency centroid
- IQR: Interquartile range (in kHz)
- skew: Skewness
- peakf: Peak frequency (frequency having highest energy)
- meanfun: Mean of all fundamental frequencies measured across acoustic signal
- dfrange: Range of all dominant frequencies measured across acoustic signal
- minfun: Minimum fundamental frequencies measured across acoustic signal
- meandom: Mean of all dominant frequencies measured across acoustic signal
- mindom: Minimum of all dominant frequencies measured across acoustic signal
- maxdom: Maximum of all dominant frequencies measured across acoustic signal
- maxfun: Maximum fundamental frequencies measured across acoustic signal
- modindx: Modulation index (calculated as the per frequency range, accumulated absolute difference between adjacent measurements of fundamental frequencies)

For the RAVDESS dataset, some additional features incorporated are listed below:

- MFCCs (Mel Frequency Cepstral Coefficients)
- Spectral bandwidth
- Spectral contrast
- Spectral rolloff

- Tonnetz

Total number of features: For Voice Gender dataset, we have used 21 features in total, and for RAVDESS [28] dataset, we have extracted 1000 features which are then passed on through feature selection algorithms to get the optimal feature subset.

3.2 Scaling

All the features are scaled and normalized before feeding to the classification model. For this purpose, we have used Scikit learn's Min-Max-Scaler function. All the features are fed to this transformer column-wise and output are feature values belonging to the range [0,1]. These feature values are then fed to the classification model, which is discussed in the next section.

3.3 Feature selection

The feature engineering procedure may create some irrelevant and redundant features in the feature space. It is essential to remove such irrelevant feature attributes, which helps not only to enhance the overall classification accuracy but also to minimize the computational overheads. In the proposed work, we have used a two-stage feature selection approach, where in the first stage a filter called, mRMR, is used to remove the irrelevant and redundant features from the feature space, and in the second stage, a modified wrapper method, called Mayfly algorithm with AβHC algorithm, is used. This approach selects most relevant features from the original feature set so that the classification accuracy can be improved. The adopted approach has been mentioned in detail in the following subsections.

3.3.1 Minimum redundancy maximum relevance

The mRMR is a filter method, which uses a statistical or probabilistic approach to rank the feature attributes on the basis of the assigned score. The feature attributes then could be either selected or removed based on their scores/ranks. It tends to select the features with a high correlation with the class (output) and a low correlation between themselves. The initial idea behind mRMR is as follows. A feature space S with m features, Xi highly correlated features with the output class C, should be considered. Implementation of the maximum relevance is done using mean value of mutual information of all feature Xi with class C. The mutual information between two random variables X and Y whose joint probability is $p(x, y)$ and $p(x)$, $p(y)$ as their individual probabilities can be stated formally as:

$$\text{MutualInfo} = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

The measure of global relevance of the variables in S with respect to C is:

$$\text{GR}_I(S) = \frac{1}{|S|}\sum_{X_i\in S}\text{MutualInfo}(C,X_i) \tag{2}$$

Here, $\text{MutualInfo}(C,X_i)$ is the mutual information of feature X_i with class C. Another idea behind mRMR is that for the class variable, the maximum relevance criterion of the features should be supplemented by the use of a minimum redundancy among features. Minimum redundancy should be implemented without disturbing its relevance since if only relevance is implemented, there is a high chance that the dependency between features could be increased. The formula to get minimum redundancy between features is as follows:

$$W_I(S) = \frac{1}{|S|^2}\sum_{X_i\in S}\sum_{X_j\in S}\text{MutualInfo}(X_i,X_j) \tag{3}$$

Here, $\text{MutualInfo}(X_i,X_j)$ is the mutual information of feature X_i with X_j. The mRMR criterion combines the above two constraints to optimize relevance and redundancy in order to obtain a good subset of features:

$$\max\phi(\text{GR}_I(S), W_I(S)) \tag{4}$$

where, $\phi = (GR_I(S) - W_I(S))$

3.3.2 Mayfly algorithm

Mayfly algorithm has been developed by Zervoudakis and Tsafarakis [43]. Many researchers have used it for various research purposes. Liu [26] proposed a novel multiobjective version of the Mayfly algorithm to estimate the optimal weight coefficients for integrating the forecasting values of the sub-series in order to design a Forecasting System for Short-Term Wind Speed. Liu and Chai [25] also proposed a novel resonance demodulation frequency band selection method to diagnose the bearing fault with the help of a modified Mayfly algorithm. Guo [17] used a modified Mayfly algorithm for optimizing the component size and operation strategy of a high-temperature proton exchange membrane fuel cell (PEMFC)-powered combined cooling, heat and power (CCHP). Mayfly algorithm has been recently used successfully in the domain of feature selection. Bhattacharyya [2] has proposed a hybrid of Mayfly algorithm and harmony search and used the combined algorithm to perform feature selection task on various datasets. The Mayfly algorithm performs the necessary modifications to the existing algorithms such as PSO, which is likely to get stuck in a local optimum, especially for feature space having a higher dimension, hence enabling the algorithm to have a better performance across both small- and large-scale feature sets. The Mayfly algorithm is based on the mating process of male and female mayfly insects. Most male adults assemble in a swarm a few metres above the water to attract the females. They perform a nuptial dance, which involves characteristic up and down movements, thereby generating a pattern. Female mayflies go to the swarm for mating. The mating process lasts only for a few seconds after which they drop the eggs in the water and the cycle continues. The Mayfly algorithm have three components given as follows:

- **Movement of male mayflies:** The new position acquired by a male mayfly from a present position of x_i^t is updated by:

  $$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{5}$$

  where, v_i^{t+1} is the velocity of the male mayfly. The new velocity of male mayfly m in dimension j at time t having a previous velocity of v_{mj}^t is updated by:

  $$\begin{aligned} v_{mj}^{t+1} = g \times v_{mj}^t + a_1 \times e^{-\beta r_p^2} \times (\text{opbest}_{mj} - x_{mj}^t) \\ + a_2 \times e^{-\beta r_g^2} \times (\text{cpbest}_j - x_{mj}^t) \end{aligned} \tag{6}$$

  where, x_{mj}^t is the position of the mayfly at time t, positive attraction constants $a1$ and $a2$ are used for measuring the contribution of the cognitive and social components, respectively. β is a fixed visibility coefficient used to limit a mayfly's visibility to others and g is a gravitational coefficient. $opbest_m$ is the most optimal position that the particular mayfly m has ever visited and $cpbest_j$ is the jth component of the position of the best male mayfly. And r_p is the Cartesian distance between x_m and $opbest_m$, while r_g is the Cartesian distance between x_m and $cpbest_j$. Due to being a minimization task, the $opbest_m$ is updated as:

  $$\text{opbest}_m = \begin{cases} x_m^{t+1} \\ \text{if} \quad \text{fit}(x_m^{t+1}) \quad < \text{fit}(\text{opbest}_m) \end{cases} \tag{7}$$

  where, $\text{fit}(x_m^t)$ provides the fitness value of a position or the quality of a solution. A stochastic element is present in the algorithm as it is necessary that at a particular time the best mayflies keep performing the nuptial dance. Mathematical representation of this dance is given by:

  $$v_{mj}^{t+1} = g \times v_{mj}^t + d \times rval \tag{8}$$

  where, d is the coefficient of nuptial dance and r is a random value such that $rval \in [-1, 1]$. Also, d reduces from an initial value d_0 by a random value $\delta \in [0, 1]$ as $d_{\text{itr}} = d_0 \times \delta^{itr}$, where, itr is the current number of iteration.

- **Movement of female mayflies:** The new position acquired by a female mayfly, which moves towards a male mayfly from a present position of x_i^t, is updated by:

  $$y_i^{t+1} = y_i^t + v_i^{t+1} \tag{9}$$

  where, v_i^{t+1} is the velocity of the female mayfly. The velocity of a female mayfly is updated based on conditions since the attraction process between mayflies depends on the quality of the current solution. Therefore, the new velocity of mth female mayfly in jth component at time t having a previous velocity of v_{mj}^t is updated by:

  $$v_{mj}^{t+1} = \begin{cases} \text{if } \text{fit}(y_m) > \text{fit}(x_k) \\ g \times v_{mj}^t + a_2 \times e^{-\beta r_g^2} \times (x_{mj}^t - y_{mj}^t) \\ \text{else if } \text{fit}(y_m) \leq \text{fit}(x_k) \\ g \times v_{mj}^t + wc \times \text{rval} \end{cases} \tag{10}$$

  where, x_{mj}^t is the jth component position of the mayfly at time t, positive attraction constants $a1$, β and g are previously defined in Eq. 6. r is a random value and $r \in [-1, 1]$ and wc is a random walk coefficient in the case when a female is not being attracted by a male.

- **Crossover between mayflies:** In order to perform the crossover operation, initially, a male mayfly is selected and then a female mayfly. This selection is done on the basis of their fitness values, i.e. the best male breeds with the best female. Following equation shows two offspring produced after applying the crossover operation:

  $$\begin{aligned} \text{offspring}_1 &= \text{rval}_{\text{of}} \times \text{pmale} + (1 - \text{rval}_{\text{of}}) \times \text{pfemale} \\ \text{offspring}_2 &= \text{rval}_{\text{of}} \times \text{pfemale} + (1 - \text{rval}_{\text{of}}) \times \text{pmale} \end{aligned} \tag{11}$$

  Here, pmale and pfemale are the parent male and female mayflies, respectively, and rval_{of} is a stipulated value between 0 and 1. Initial velocities of 0 are set for the offspring.

- **Mutation of mayflies:** To enhance the exploration ability of the algorithm, the newly generated offspring are mutated by simply adding a normally distributed random number to the offspring's variable.

3.3.3 Adaptive β-hill climbing

AβHC is a meta-heuristic algorithm, which is an adaptive version of the βHC, which itself is a modified version of the Hill climbing (HC) algorithm. HC algorithm is a comprehensible form of local search method. But it often gets stuck in local optima. To overcome this limitation,

βHC was proposed. Given a solution $S == (s_1, s_2, \ldots, s_D)$ The βHC iteratively generates $S'' == (s''_1, s''_2, \ldots, s''_D)$ an improved solution based on β operator and N (neighbourhood operator), which randomly chooses a neighbour $S' = = (s'_1, s'_2, \ldots, s'_D)$ of the solution S defined by the equation:

$$s'_i = s_i \pm U(0,1) \times N \quad \text{such that } i \in [1, D] \tag{12}$$

where, a random value of i is selected. D is the dimension of the problem under consideration. The neighbourhood operator N denotes the maximum distance possible between current solution and its neighbour. Values are assigned either from the current solution or randomly from the corresponding range with probability value $\beta \in [0,1]$ for the new solution given by:

$$s_{i''} = \begin{cases} s_r & \text{if rnd} \leq \beta \\ s_i & \text{else} \end{cases} \tag{13}$$

where, rnd and s_r are random values and $rnd \in [0,1]$ and s_r depends on the dimension of the problem. Hence, it can be visualized that the outcome of this βHC largely depends on the values chosen for N and β, which requires extensive experiments to be conducted. Hence, to avoid this issue AβHC was proposed. The values for N and β are expressed as a function of iteration number in AβHC. $N(k)$, i.e. N for k^{th} iteration, is defined by:

$$N(k) = 1 - \frac{k^{\frac{1}{T}}}{\text{IterMax}^{\frac{1}{T}}} \tag{14}$$

where, T is a constant and IterMax is the maximum number of iterations. Also, $\beta(k)$, i.e. β for kth iteration, is defined by:

$$\beta(k) = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \times \frac{t}{\text{IterMax}} \tag{15}$$

Now, S is replaced with S'' if the generated neighbour S'' is better than the current solution S.

3.3.4 The proposed Mayfly algorithm combined with AβHC

In the first stage, the filter method, called mRMR, is used. k number of features is filtered out whose values have been taken optimally from a graphical analysis shown in Fig. 2. The figure illustrates a variation of the k value with the classification accuracy for the RAVDESS dataset on which the analysis for mRMR has been done. Hence, according to the graphs presented in Fig. 2, 500 insignificant features for both the gender and emotion classification tasks, whereas 600 insignificant features for the speaker classification task are filtered out of overall previously extracted features, which are then used as input to the combination of Mayfly algorithm and AβHC algorithm for each of the classification tasks.

The proposed wrapper method is present in Algorithm 1. Conversion of continuous space to binary space is required as feature selection uses a binary feature space. The S-shaped transfer function is used in order to achieve such conversion. The probability of whether to choose a particular feature in a solution vector is given by the function. It is a general and reliable function, which has been previously used by many researchers. The equation for the S-shaped transfer function used in this algorithm is:

$$Sf(x) = \frac{1}{1 + e^{-}x} \tag{16}$$

The updated agent (or feature subset) during the binary conversion is given by :

$$Fs_d^{t+1} = \begin{cases} 1 & \text{if } Sf(Fs_d^{t+1}) > \text{randval} \\ 0 & \text{if } Sf(Fs_d^{t+1}) \leq \text{randval} \end{cases} \tag{17}$$

The fitness function used for the learning algorithm consists of the classification error and the number of features selected. This has been done in the intent of simultaneously increasing the accuracy and reducing the number of features. Thus for this purpose, classification error is used instead of accuracy. Combining these two reduces the fitness function to a single objective function. The equation for the fitness function is:

$$\downarrow \text{fit} = \gamma \times E + (1 - \gamma) \times \frac{|\text{Ns}|}{|\text{Nf}|} \tag{18}$$

where, $|\text{Nf}|$ is the number of features in the given dataset, $|\text{Ns}|$ is the number of features in the feature subset, E is the classification error and $\gamma \in [0,1]$ is a parameter that gives a relative contribution between the classification error and the number of features. In order to find the optimal feature subset, the method needs to find the global optima which requires proper exploration and exploitation of the search space. Hence, AβHC has been used to enhance the exploitation ability of the Mayfly algorithm.

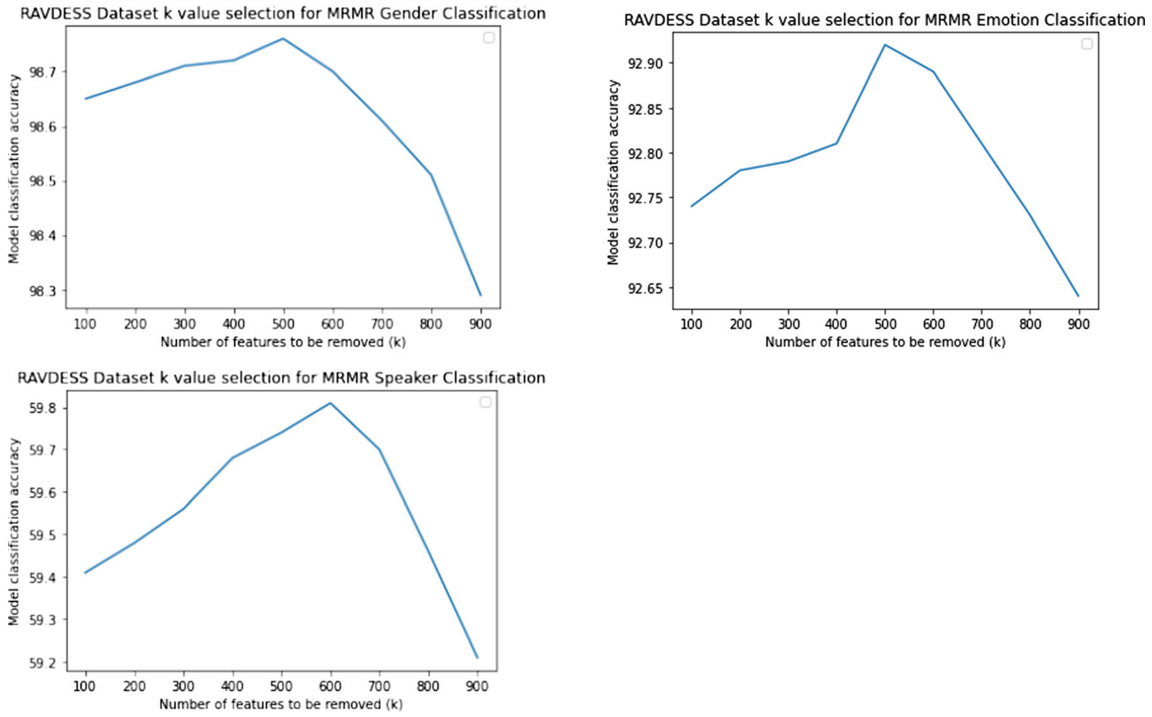


Fig. 2 Graph used to show the selection of k value for mRMR for gender, emotion and speaker recognition tasks on RAVDESS dataset

Algorithm 1: Proposed Mayfly algorithm combined with AβHC method

```
Input : PopSize, IterMax
Result: Best agent $A = (a_1, a_2, ..., a_d)$
Randomly initiate population and velocity of male and female mayflies;
Evaluate population and then find opbest ;
for $itr \leftarrow 1$ to $IterMax$ do
    for $i \leftarrow 1$ to $PopSize$ do
        Update opbest;
        Evaluate and update the velocities of male and female mayflies;
    end
    Sort and rank the mayflies;
    Generate male and female offspring by crossover;
    Mutate the offspring;
    Replace worst mayflies with new best offspring generated;
    Perform A$\beta$HC on male mayflies;
    Update opbest;
end
```

3.4 GRaNN: golden ratio-aided neural network

In the present work, a variant of MLP is used which is deep as it consists of 15 hidden layers. The next subsection describes it in detail. It is to be noted that our main aim in this section revolves around the significance of Golden ratio. We have not made any changes to the architecture of MLP. Here, a framework is required where the units could be varied without hindering the results much. We have selected this architecture because of its lightweight characteristic and lesser computational complexities, thereby proving the effectiveness of the Golden ratio.

3.4.1 Definition

MLP is a special class of network belonging to the class of feed-forward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to any feed-forward ANN, sometimes strictly referring to networks that are composed of multiple layers of perceptrons with threshold activation.

A basic MLP unit (hypothetical representation shown in Fig. 3) consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses an activation function which is nonlinear. Generally, it applies a supervised learning technique called back propagation for training. Its multiple layers and nonlinear activation help it to distinguish data that are not separable by using linear techniques.

3.5 Optimization of MLP layers

It has been already mentioned that deciding the number of nodes in the hidden layers of DMLP is always challenging and generally it requires a huge amount of trial and error. To overcome this, in this work, we have used the concept of Golden ratio which helps us to set the number of nodes in the hidden layers of DMLP used here. In this section, we have first defined the idea of Golden ratio followed by its use in various applications specifically in an optimization technique called the Golden Section search method. Finally, we have described how it is incorporated in our architecture for optimization purposes.

3.5.1 Golden ratio

In mathematics, two quantities are said to be in the Golden ratio if their ratio is the same as the ratio of their sum to the larger of the two quantities. Expressed algebraically, for quantities a and b with a > b > 0,

$$\frac{a+b}{a} = \frac{a}{b} \overset{\text{def}}{=} \varphi, \tag{19}$$

where the letter phi (φ or ϕ) depicts the Golden ratio. It is an irrational number that is a solution to the quadratic equation $x^2 - x - 1 = 0$, with a value of:

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.6180339887.\ldots \tag{20}$$

3.5.2 Use of golden ratio

The characteristic of the Golden ratio in any number systems that include any positive integral radix (base),

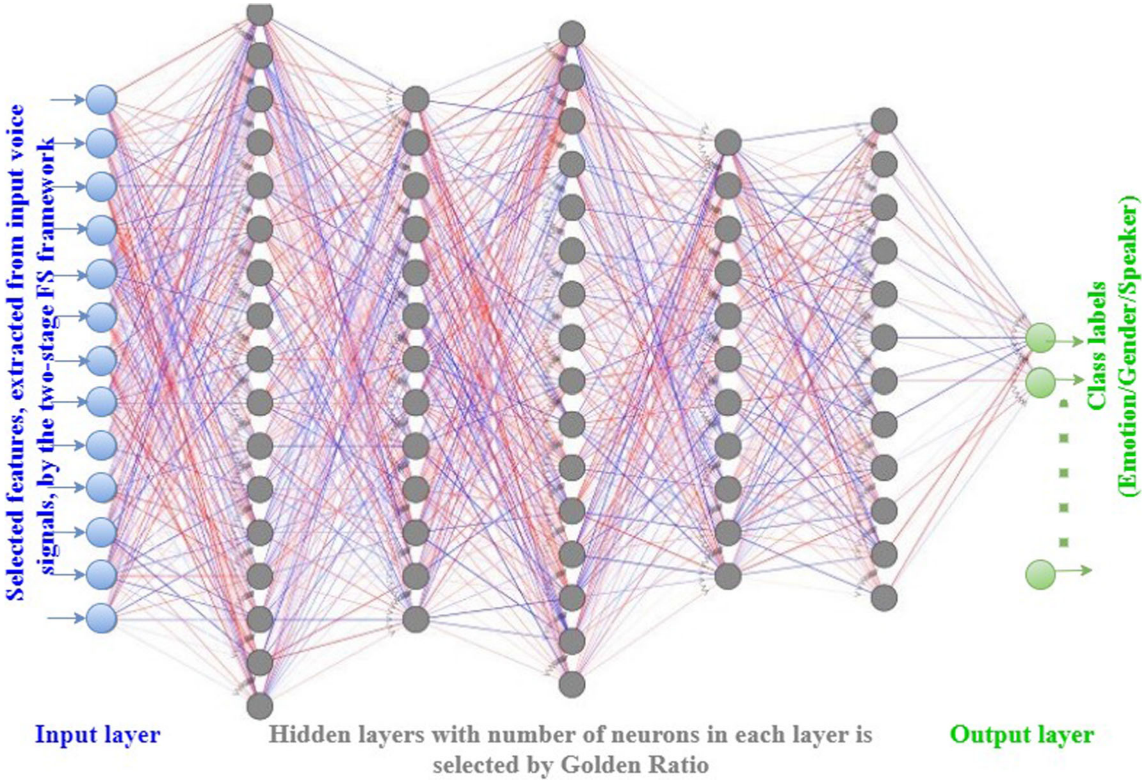

Selected features, extracted from input voice signals, by the two-stage FS framework
Class labels (Emotion/Gender/Speaker)
Input layer
Hidden layers with number of neurons in each layer is selected by Golden Ratio
Output layer

Fig. 3 Hypothetical representation of deep MLP

negative radix, and variable radix possibly remains invariant [38]. Specifically, digits that have been chosen even randomly or systematically from consecutive digits or consecutive blocks of digits of the Golden ratio can be used as a source of uniformly distributed random numbers. Unlike any of the several quasi- and pseudo-random number generators using various methods, we do not need to follow any stringent method here. We only have to select the consecutive or non-consecutive blocks of digits from the stored Golden ratio, thereby making it a fastest means of obtaining random numbers. This idea of obtaining random sequences can be considered as quite an efficient way of solving numerous optimization problems including the NP-hard problems by polynomial time heuristics and randomized algorithms. Also, whether these random numbers that are sieved out of the Golden ratio are more uniformly distributed (quasi) or pseudo-random numbers may be studied including its scope among other random number generators.

3.5.3 Golden section search method

The Golden Section Search method [18] is used to find out the maximum or minimum of any unimodal function which is a function containing only one maximum or minimum value in any interval [a,b].

There are many methods for finding the local maximum or local minimum. The equal-interval search method is one of such simplest methods. Referring to Fig. 4, an interval of ϵ is chosen over which assumption of occurrence of the maximum is done. Then, $f\left(\frac{a+b}{2}+\frac{\epsilon}{2}\right)$ and $f\left(\frac{a+b}{2}-\frac{\epsilon}{2}\right)$ are computed.

If $f\left(\frac{a+b}{2}+\frac{\epsilon}{2}\right) \geq f\left(\frac{a+b}{2}-\frac{\epsilon}{2}\right)$, then the maximum occurs in the interval, $\left[\frac{a+b}{2}-\frac{\epsilon}{2}, b\right]$, else it occurs in $\left[a, \frac{a+b}{2}+\frac{\epsilon}{2}\right]$. This helps in reducing the interval of occurrence of the local maximum. These steps are repeated until the interval is reduced to the precision of desired choice.

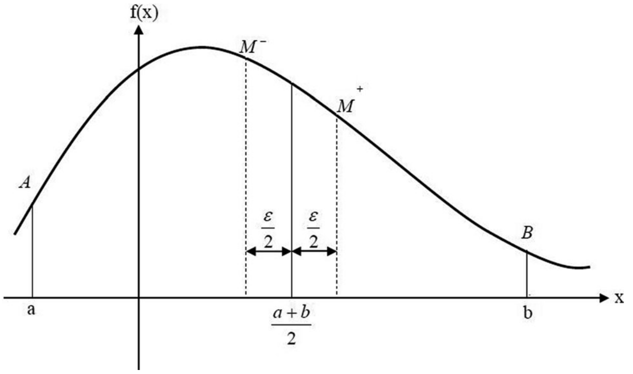


Fig. 4 Equal-interval search method

The equal-interval search method has its drawbacks, being the inefficiency caused when the interval is small, resulting in an undesirably long time to find the maximum of a function. To improve this inefficiency, the Golden Section Search method was introduced.

As Fig. 5 shows, three points x_l, x_1 and x_u $(x_l < x_1 < x_u)$ are chosen along the X-axis with corresponding values of the function $f(x_l)$, $f(x_1)$ and $f(x_u)$, respectively. Since $f(x_1) > f(x_l)$ and $f(x_u) > f(x_1)$, the maximum ought to lie between x_l and x_u. Now a fourth point denoted by x_2 is chosen such that it lies between the larger of the two intervals of $[x_l, x_1]$ and $[x_1, x_u]$. Making assumption that the interval $[x_l, x_1]$ is larger than $[x_1, x_u]$, $[x_l, x_1]$ is chosen as the interval in which x_2 is chosen. If $f(x_2) > f(x_1)$, then the new three points would be $(x_l < x_2 < x_1)$; else if $f(x_2) > f(x_1)$, then the points are $(x_2 < x_1 < x_u)$. This procedure is repeated until the distance between the outer points is sufficiently small and matches the desired precision. Now question arises that how the intermediate points in the Golden Section Search are determined. Here, comes the role of the Golden ratio [1].

First the intermediate point x_l is chosen in order to equalize the ratio of the lengths as shown in Eq. (19) where a and b are distances as present in Fig. 5. It should be noted that a+b is equal to the distance between the upper and lower boundary points x_u and x_l.

The second intermediate point x_2 is chosen similarly in the interval a to satisfy the following ratio in Eq. (21) where the distances of a and b are shown in Fig. 6.

$$\frac{b}{a} = \frac{a-b}{b} \tag{21}$$

The ratios in Eqs. (19) and (21) are equal and have a value equal to the Golden ratio.

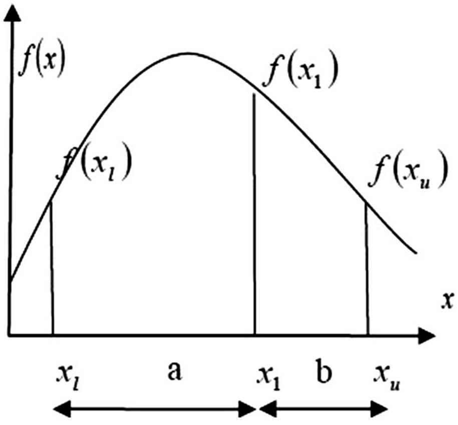


Fig. 5 Determination of the first intermediate point

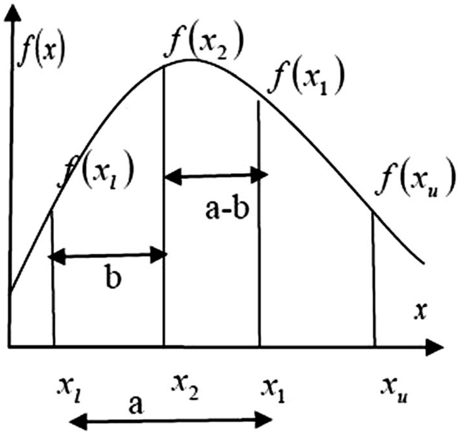


Fig. 6 Determining the second intermediate point

This method deals with finding minimum and maximum of functions, and the nonlinear functions involved in the bases of the MLP architecture have a requirement of finding the same at various steps of learning. Hence, we have been motivated to experiment and incorporate the Golden ratio in our model, and the results are quite promising.

3.5.4 Final architecture

The number of units in each layer of DMLP has been set keeping the Golden ratio in mind. The model consists of 1 input layer, 1 output layer and 15 hidden layers. The usage of 15 hidden layers is purely experimental. Our main aim is to preserve the Golden ratio in deciding the number of hidden units. In the process, we ensure that our model does not overfit pertaining to a large number of hidden layers or underfit pertaining to a lesser number of hidden layers. The magic numbers used here are 110, 68, 42 corresponding to number of units in the hidden layers. As we can see,

$$110/68 = 1.61764 \approx \phi$$
$$68/42 = 1.61904 \approx \phi$$

They are basically multiples of numbers, which are part of the famous Fibonacci sequence, and it is a result of the Golden ratio itself. The pattern in the hidden neurons has been incorporated to bring uniformity throughout the network while preserving the Golden ratio. Either linearly increasing or decreasing values of the number of units in a hidden layer are avoided in the present work.

3.6 Output

The output layer has a Softmax activation. The class with max argument value is predicted as the output class. For the purpose of gender classification, the shape of output layer is set to fit 2 label classes. For the purpose of emotion identification, the shape of output layer is changed to fit 8 classes of emotion. Lastly, for speaker identification purpose, the shape has been modified to fit 24 classes of speakers.

The schematic diagram of our optimized GRaNN architecture used in the present work is shown in Fig. 7.

4 Experimental results

The organization of the result section is performed as follows: Sect. 4.1 presents the evaluation metrics used to evaluate our proposed GRaNN model, whereas Sects. 4.3, 4.4 and 4.5 report the detailed results related to gender, emotion and speaker identification tasks, respectively. The overall accuracies (rounded off to 2 decimal places) of the GRaNN model on both RAVDESS and Voice Gender datasets are depicted in Table 4. However, the results obtained using feature selection framework for both RAVDESS and Voice Gender datasets are given in Tables 5 and 6, respectively. It is to be noted that we have applied our model for gender classification problem on both the RAVDESS and Voice Gender datasets, whereas for emotion identification problem, we have considered only RAVDESS dataset. To test the effectiveness of using both feature selection process and Golden ratio in our proposed model, different classification accuracies attained before and after applying feature selection and Golden ratio for gender classification task are shown in Table 7. The gender identification results obtained by our GRaNN model are illustrated in Table 8. We have also compared our performance of gender classification with a semi-supervised algorithm K-RMS[13] in Table 9. Performance comparison of our proposed model with some existing methods for Voice Gender dataset is shown in Table 10.

The performance (evaluated in terms of accuracy and number of features used) for emotion identification problem using our feature selection framework is illustrated in Table 11. The different classification accuracies achieved before and after applying feature selection and Golden ratio for emotion identification task are shown in Table 12. The performance of the emotion identification produced by our feature selection-based GRaNN model is shown in Table 13, whereas comparison with some existing methods for RAVDESS dataset is illustrated in Table 14.

However, to check the versatility of our model, we have made use of the RAVDESS dataset for speaker identification purpose by changing the shape of the output layer while keeping the rest of the model intact. The performance (measured in terms of accuracy and number of features used) for speaker identification problem using our feature selection framework is illustrated in Table 16. The different classification accuracies produced before and

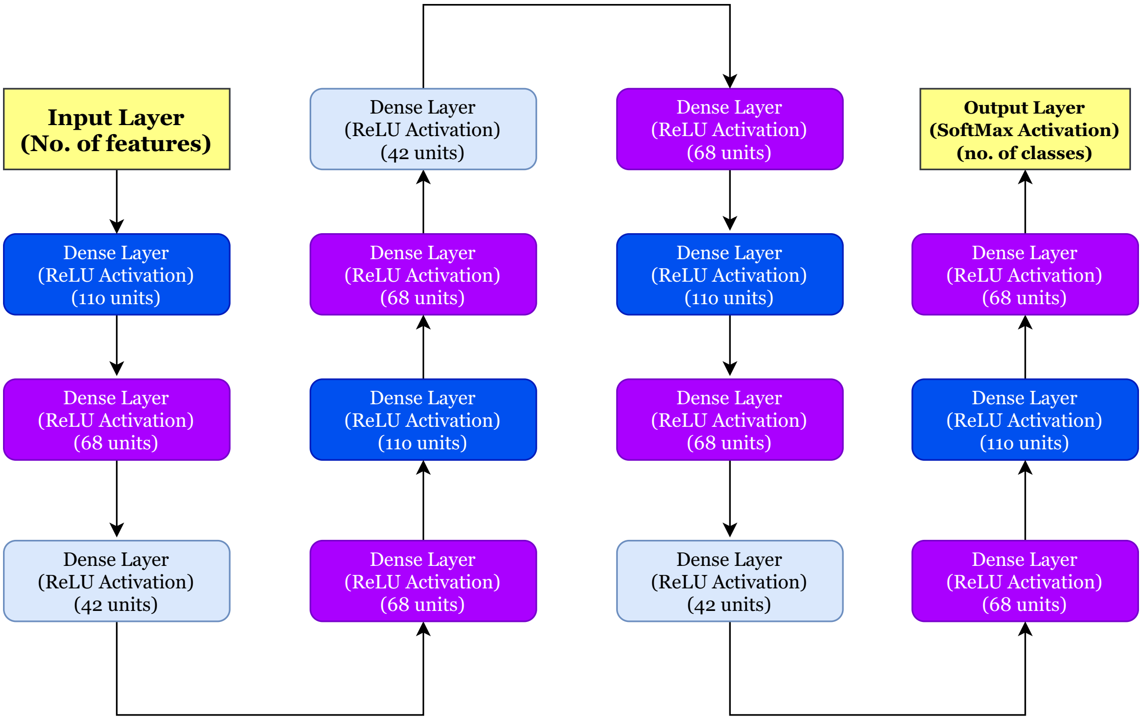


Fig. 7 Optimized GRaNN architecture used in the present work

after applying feature selection and Golden ratio for speaker identification task are shown in Table 17. The detailed individual speaker-wise performance is shown in Table 15, whereas the confusion matrix generated by the proposed feature selection-based GraNN model is illustrated in Table 19. We have also compared the performance of the speaker identification task with logistic regression [41] model in Table 20 owing to the fact that it gives quite impressive results for multiclass classification tasks.

To convey the basis of the selection of the proposed configuration for the model, the experiments have been conducted by varying network configurations of hidden units on both the datasets for all the three classification tasks. The overall results are shown in Table 18.

Table 1 Usage stats of datasets for different purposes

Dataset	Problem
RAVDESS	Gender classification
	Emotion identification
	Speaker identification
Voice gender	Gender classification

4.1 Datasets used

The different tasks performed by our model for the datasets are given in Table 1.

4.1.1 RAVDESS dataset

The RAVDESS dataset [28] consists of 7356 data files. The database contains voice of 24 professional actors. Out of these actors, 12 are female and 12 are male. They have vocalized two lexically matched statements in a neutral North American accent. The Song files of the dataset contain emotions, namely ”Happy”, ”Sad”, ”Angry”, ”Calm” and ”Fearful”. The speech data include expressions of ”Calm”, ”Happy”, ”Sad”, ”Angry”, ”Fearful”, ”Surprise”, and ”Disgust”. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The emotion dataset contains 4948 instances. The work of gender and speaker identification is based on the speech files of the dataset, which contain 1440 files made out of 60 trials per actor with 24 actors. The detailed division is shown in Table 2.

4.1.2 Voice Gender dataset

This dataset [12] is created to serve the purpose of identifying a human speech signal as male or female, based upon certain acoustic properties of the voice and speech.

Table 2 Training, validation and test data division of RAVDESS dataset for emotion, gender and speaker identification

Dataset	Label	Train samples	Validation samples	Test samples
RAVDESS	Emotion	3463	990	495
	Speaker	972	324	144
	Gender	972	324	144

Table 3 Training, validation and test data division of Voice Gender dataset for gender identification

Dataset	Label	Train samples	Validation samples	Test samples
Voice gender	Male	1063	354	167
	Female	1076	358	150

The dataset consists of 3168 voice samples, which were recorded and collected from male and female speakers. The voice samples are pre-processed using speech signal analysis with the help of several voice processing packages. The analysed frequency range is that of 0Hz–280Hz, which is the human vocal range. Detailed division is shown in Table 3

4.2 Evaluation metrics used

For analysing the performance of our model on the datasets, we have calculated some well-known metrics, namely Accuracy, Precision, Recall [7] and F1-score values with corresponding class support division.

4.3 Gender identification results

A comparison of the accuracy and number of features has been made between the model without any feature selection method, with only Mayfly and AβHC and with mRMR, Mayfly and AβHC feature selection method for the gender classification task for the RAVDESS dataset in Table 5. From the table, it can be clearly seen that after applying the mRMR, Mayfly and AβHC feature selection method, the accuracy has increased along with a decrease in the number of features, which is the favourable outcome of any feature selection method. It is also to be noted that with the use of the mRMR method, the number of features has a huge reduction along with a fair increase in the accuracy. This indicates that reducing the dimension of the feature space using a filter method has a major positive impact on the achieved results.

Similarly, Table 6 presents a comparison between the accuracy and number of features for the Voice Gender dataset between the model without using any feature selection method and with only Mayfly and AβHC. Being a small dataset of just 21 features, the mRMR filter method is not needed to reduce the dimensionality of the feature space. From Table 6, it can be noticed that there has been a

Table 6 Results achieved before and after applying feature selection for gender classification task for Voice Gender dataset

Method used	Accuracy (%)	Number of features
Without any feature selection	95.32	21
With Mayfly AβHC	95.68	6

Table 4 Comparison of accuracies for gender, emotion and speaker identification using our proposed GRaNN model after feature selection

Entity	Train accuracy (%)	Validation accuracy (%)	Test accuracy (%)
Gender (RAVDESS)	100	98.5	99.3
Gender (VoiceGender)	95.34	94.32	95.68
Emotion	100	94.71	95.27
Speaker	100	67.01	67.17

Table 5 Results achieved before and after applying feature selection for gender classification task for RAVDESS dataset

Method used	Accuracy (%)	Number of features
Without any feature selection	98.611	1000
With Mayfly AβHC	99.000	254
With mRMR and Mayfly AβHC	99.306	101

Table 7 Results achieved before and after applying feature selection and Golden ratio for gender classification task on both RAVDESS and Voice Gender datasets

Dataset	Combination of methods used	Accuracy (%)
RAVDESS	Without any feature selection or golden ratio	91.37
	With feature selection and without golden ratio	95.53
	Without feature selection and with golden ratio	98.611
	With feature selection and golden ratio	99.306
Voice gender	Without any feature selection or golden ratio	83
	With feature selection and without golden ratio	91.44
	Without feature selection and with golden ratio	95.32
	With feature selection and golden ratio	95.68

Table 8 Gender identification results obtained by GRaNN model on RAVDESS and Voice Gender datasets

Dataset	Class	Precision	Recall	F1-score	Support
RAVDESS	Male	0.97	1.00	0.99	69
	Female	1.00	0.97	0.99	75
	Weighted avg	0.99	0.99	0.99	144
Voice gender	Male	0.97	0.93	0.95	167
	Female	0.93	0.97	0.95	150
	Weighted avg	0.95	0.95	0.95	317

Table 9 Gender identification results obtained by K-RMS algorithm on RAVDESS and Voice Gender datasets

Dataset	Class	Precision	Recall	F1-score	Support
RAVDESS	Male	0.54	0.49	0.51	720
	Female	0.53	0.58	0.56	720
	Weighted avg	0.54	0.54	0.53	1440
Voice gender	Male	0.73	1.0	0.84	1584
	Female	1.0	0.63	0.77	1584
	Weighted avg	0.87	0.82	0.81	3168

Table 10 Comparison of accuracies on Voice Gender dataset using our GRaNN model with state-of-the art methods

	Accuracy (%)	
Model	Train	Test
Frequency-based baseline [23]	61	59
Logistic regression [33]	72	71
CART [10]	81	78
Random forest [37]	100	87
Boosted tree [23]	91	84
SVM [5]	96	85
XGBoost [8]	100	87
Stacked [22]	100	89
Our model	95.34	95.68

huge reduction in the number of features after applying the Mayfly and the AβHC feature selection method and a slight increase in the overall accuracy. This further validates the efficacy of our proposed model with feature selection method.

Table 7 shows the classification accuracies attained before and after applying feature selection and Golden ratio for gender identification task. It can be seen from Table 7 that the combination of feature selection with Golden ratio-based optimization comes out to be the winner among all configurations producing an accuracy of 99.306% and 95.68% over RAVDESS and Voice Gender datasets, respectively. Furthermore, the Golden ratio plays an important role than feature selection as it shows more increment in accuracy compared to using feature selection only. It is to be noted from Table 7 that using Golden ratio, the model attains accuracies of 98.611%, 95.32%, whereas using feature selection, the model produces accuracies of 95.53%, 91.44% over RAVDESS and Voice Gender datasets, respectively.

Our feature selection-based GRaNN model performs quite well for gender identification purposes on both the datasets as evident from Tables 8, 9 and 10. It is evident from Table 8 that the overall gender identification accuracy on the RAVDESS dataset is found to be almost 99%, which is quite good pertaining to the fact that the speech data have absence of any kind of noise in it. In case of Voice

Gender dataset, the F1-measure is 95% which is relatively low. This may be due to the presence of zero values in some instances of features used for training as compared to the RAVDESS dataset. It is to be noted that we have not applied any feature selection procedure on this dataset because the creators of the dataset have already performed it before finalizing the dataset. Hence, redundant features are found to be almost absent for this dataset, thus making it suitable for passing it directly through the classifier for training purposes. Moreover, size constraints of dataset are also a limitation, and the presence of noise also hinders the overall performance too. Another form of hindrance affecting the performance is the random presence of silent zone in audio segments.

In the present work, the performance of gender classification task is also measured using a semi-supervised K-RMS[13] algorithm, and the overall results are tabulated in Table 9. It is observed from Table 9 that the unsupervised K-RMS algorithm performs quite well, considering it works on unlabelled data but our model, being a supervised algorithm outperforms it. As seen in Table 10, most of the existing gender classification models have over-fitted to the data, resulting in poor test accuracies. However, both the train and test accuracies of the proposed model are approximately the same and quite high, depicting proper training and absence of any kind of either over-fitting or under-fitting.

4.4 Emotion identification results

For the RAVDESS dataset, a comparison of the accuracy and number of features has been done considering three cases: (a) between the method without any feature selection method, (b) with only Mayfly and AβHC and with mRMR, and (c) Mayfly and AβHC feature selection method for the emotion classification task, which are depicted in Table 11. It can be visualized from Table 11 that after applying the mRMR, Mayfly and AβHC feature selection method, the accuracy has increased along with a decrease in the number of features, which is the favourable outcome of any feature selection method. It is also to be noted that with the use of the mRMR method, the number of features has a huge reduction along with a fair increase in the accuracy. Indicating the major positive impact on the achieved results, a filter method has by reducing the dimension of the feature space even for a multiclass problem.

Table 12 Results achieved before and after applying feature selection and Golden ratio for emotion classification task on RAVDESS dataset

Combination of method used	Accuracy (%)
Without any feature selection or golden ratio	84.85
With feature selection and without golden ratio	88.56
Without feature selection and with golden ratio	92.727
With feature selection and golden ratio	95.27

Table 13 Emotion identification results obtained by GRaNN model on RAVDESS dataset

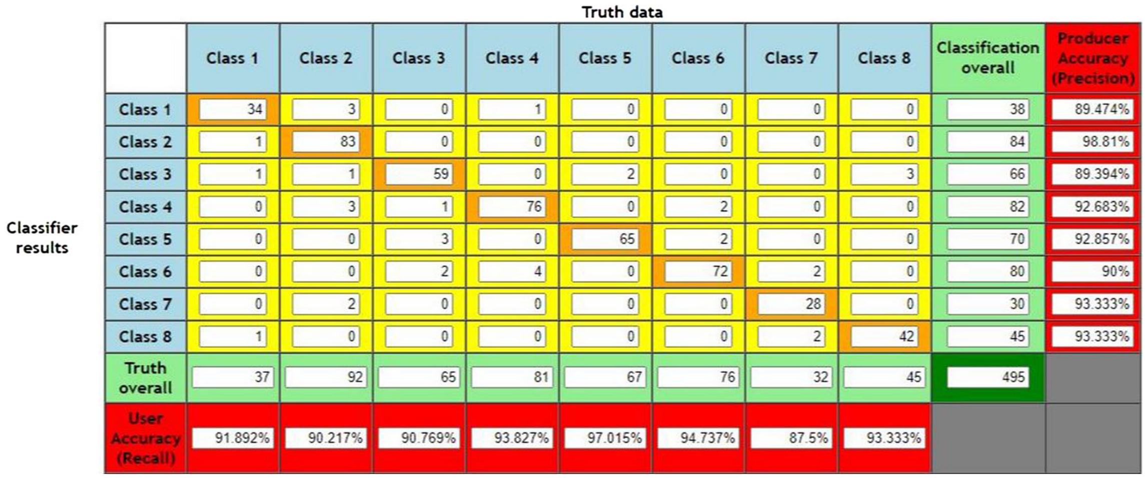
Class	Precision	Recall	F1-score	Support
Neutral	0.92	0.89	0.91	38
Calm	0.90	0.99	0.94	84
Happy	0.91	0.89	0.90	66
Sad	0.94	0.93	0.93	82
Angry	0.97	0.93	0.95	70
Fearful	0.95	0.90	0.92	80
Disgust	0.88	0.93	0.90	30
Surprised	0.93	0.93	0.90	45
Weighted avg	0.93	0.93	0.93	495

From Table 12, it can be said that the combination of feature selection with Golden ratio-based optimization comes out to be the winner among all configurations with an accuracy of 95.27% for RAVDESS dataset. However, the Golden ratio concept plays an important role than feature selection as it yields better results compared to using feature selection only. Moreover, an accuracy of 92.727% is achieved using Golden ratio, whereas an accuracy of 88.56% is only achieved using feature selection.

Our feature selection-based GRaNN model also performs almost equally well for emotion identification purpose on RAVDESS dataset. All the emotion classes have been classified with approximately same precision, which is quite high as seen in Table 13.

Table 11 Results achieved before and after applying feature selection for emotion classification for RAVDESS dataset

Method used	Accuracy (%)	Number of features
Without any feature selection	92.727	1000
With Mayfly AβHC	93.121	267
With mRMR and Mayfly AβHC	95.27	140



Truth data

Classifier results

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Classification overall	Producer Accuracy (Precision)
Class 1	34	3	0	1	0	0	0	0	38	89.474%
Class 2	1	83	0	0	0	0	0	0	84	98.81%
Class 3	1	1	59	0	2	0	0	3	66	89.394%
Class 4	0	3	1	76	0	2	0	0	82	92.683%
Class 5	0	0	3	0	65	2	0	0	70	92.857%
Class 6	0	0	2	4	0	72	2	0	80	90%
Class 7	0	2	0	0	0	0	28	0	30	93.333%
Class 8	1	0	0	0	0	0	2	42	45	93.333%
Truth overall	37	92	65	81	67	76	32	45	495	
User Accuracy (Recall)	91.892%	90.217%	90.769%	93.827%	97.015%	94.737%	87.5%	93.333%		

Fig. 8 Confusion matrix for emotion identification on RAVDESS dataset using GRaNN model. (Class 1 = ”Neutral”, Class 2 = ”Calm”, Class 3 = ”Happy”, Class 4 = ”Sad”, Class 5 = ”Angry”, Class 6 = ”Fearful”, Class 7 = ”Disgust” and Class 8 = ”Surprised”)

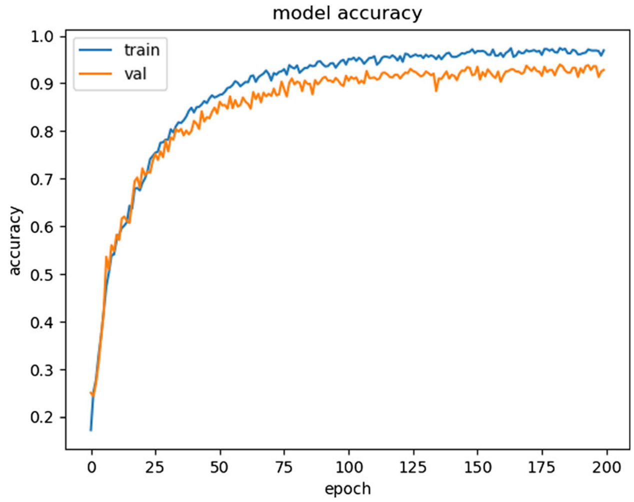


Fig. 9 Graph showing the variation of training and validation epochs with accuracies on RAVDESS dataset for emotion classification problem

Figure 8 shows the confusion matrix of classification of emotions on RAVDESS dataset using our GRaNN model. It can be observed from Fig. 8 that the emotions like ”Surprise”, ”Happy” and ”Disgust” have lower F1-score values compared to others. This is due to the fact that the emotion class ”Calm” coincides with ”Happy” class in many cases. Similarly, the emotion class ”Disgust” also coincides with both the emotions classes ”Angry” and ”Sad”. Again, the emotion class ”Surprised” coincides with ”Fearful” in terms of voice notes. This leads to lower precision and recall values, hence leading to comparatively poorer results. From the identification results for each emotion classes presented in Table 13, it can be seen that the feature selection-based GRaNN model performs reasonably well for all of the emotion classes.

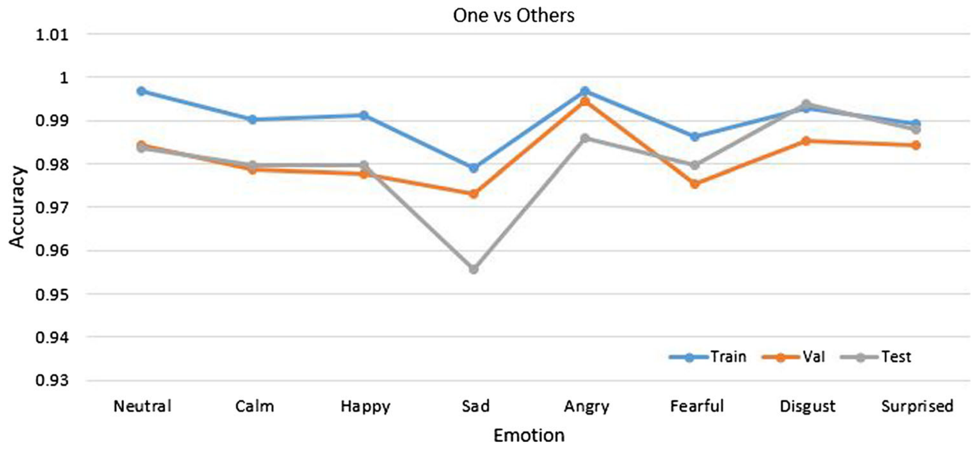


Fig. 10 Illustration of emotion identification accuracies decomposing multiclass classification into one emotion versus others binary classification

Table 14 Comparison of our proposed GRaNN model with other previously proposed models on RAVDESS dataset for emotion identification problem

Method	Accuracy (%)
Supervector + SVM [20]	36.3
Supervector + CNN [40]	34.6
eGeMAPS+CNN [11]	33.0
(Feature*) + capsule networks [21]	25.5
(Feature*) + BiLSTM [16]	63.9
(Feature*) + capsule routing [35]	69.4
(FB128)+ capsule routing [29]	68.1
(Feature*)+ capsule networks+BiLSTM [21]	35.4
(F0+MFCCs) + BiLSTM+CNN [21]	51.3
Our method	95.27

Table 15 Speaker identification results obtained by GRaNN model on RAVDESS dataset

Speaker	Precision	Recall	F1-score	Support
#1	0.80	1.00	0.89	8
#2	0.40	0.50	0.44	4
#3	0.80	0.63	0.70	7
#4	0.43	0.43	0.43	7
#5	1.00	0.78	0.88	9
#6	0.60	0.60	0.60	5
#7	0.75	1.00	0.86	3
#8	0.87	0.73	0.79	6
#9	1.00	0.83	0.91	6
#10	0.57	0.80	0.67	5
#11	0.33	1.00	0.50	3
#12	0.60	0.50	0.55	6
#13	0.25	0.50	0.33	4
#14	0.17	0.17	0.17	6
#15	1.00	0.50	0.67	4
#16	0.80	0.67	0.73	6
#17	0.83	0.56	0.67	9
#18	0.38	0.30	0.33	10
#19	0.43	0.60	0.50	5
#20	1.00	0.60	0.75	5
#21	0.75	0.60	0.67	5
#22	0.70	0.70	0.70	10
#23	0.80	0.67	0.73	6
#24	0.60	0.60	0.60	5
Weighted avg	0.7	0.67	0.68	144

Figure 9 shows the curve depicting the variation of the model's training and validation accuracy over the number of epochs trained. It can be inferred that the learning is quite gradual and involves very less spikes justifying the robustness of our model from any noisy data present in the dataset.

Figure 10 shows the training, validation and test accuracies by considering the multiclass classification problem as an amalgamation of binary classification keeping one emotion static and considering every other emotion as another class. The emotion ”Sad” provides comparatively worse results, the reason being lower amplitude of voice in the corresponding speech samples leading to loss of features.

Table 14 reports a comparison (in terms of accuracy) between previously proposed emotion classification models and our proposed feature selection-based GRaNN model. It is almost clear from Table 14 that our model outperforms all the other models for the emotion recognition task.

4.5 Speaker identification results

Our model also performs satisfactorily in classifying the speakers. However, a huge drop in accuracy is noticed when it comes to the purpose of speaker identification as evident from Table 15. It is noticed from Table 15 that in case of speaker IDs 2, 4, 11, 13, 14, 18 and 19, the values of both precision and F1-score are pretty low. This may be due to lower amplitude, absence of maintenance of constant pitch, volume, and inflection, and continuous fluctuation in the vocal tract musculature. As same set of features is being used for all the purposes so, accuracy in case of speaker identification task is not found to be as high as expected. However, this has been done in order to check the flexibility of our proposed GRaNN model for handling other speech classification problems.

A comparison of the accuracy and number of features for the RAVDESS dataset has been performed among three methods: (a) method without any feature selection method, (b) with only Mayfly and AβHC and (c) with mRMR, Mayfly and AβHC feature selection method for the speaker classification task. The results are shown in Table 16. Even though the accuracy are lower due to the fact of such high number of target classes present in the dataset still it can be visualized that after applying the mRMR, Mayfly and AβHC feature selection method, there is a comparative increase in accuracy along with a decrease in the number of features, which is the favourable outcome of any feature selection method. It is also to be noted from Table 16 that with the use of the mRMR method, the number of features has a huge reduction along with a fair increase in the accuracy. Indicating the major positive impact on the achieved results, a filter method has been implemented by reducing the dimension of the feature space even for a problem having such a high number of speaker classes.

From Table 17, it is to be noted that the combination of feature selection with Golden ratio-based optimization

Table 16 Results achieved before and after applying feature selection for speaker classification for RAVDESS dataset

Method used	Accuracy (%)	Number of features
Without any feature selection	59.370	1000
With Mayfly AβHC	63.136	316
With mRMR and Mayfly AβHC	67.172	170

Table 17 Results achieved before and after applying feature selection and Golden ratio for speaker identification task on RAVDESS dataset

Combination of method used	Accuracy (%)
Without any feature selection or golden ratio	51.29
With feature selection and without golden ratio	56.8
Without feature selection and with golden ratio	59.370
With feature selection and golden ratio	67.172

comes out to be the winner among all configurations with an accuracy of 67.172% for speaker classification on RAVDESS dataset. Similar to gender and emotion classification tasks, the use of Golden ratio proves to be more significant than using the concept of feature selection as it shows better accuracy as compared to using feature selection only. Here, the model with Golden ratio produces an accuracy of 59.37%, whereas with feature selection, an accuracy of 56.8% is achieved.

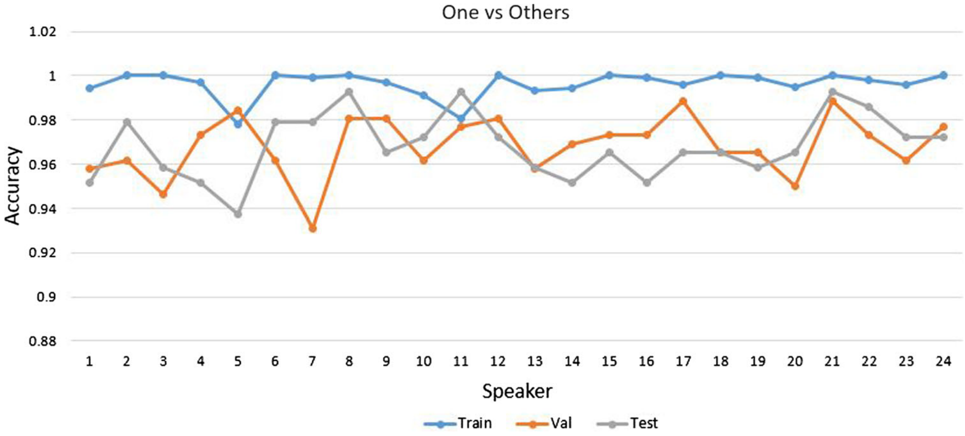


Fig. 11 Speaker identification accuracies decomposing multiclass classification into one speaker versus others binary classification

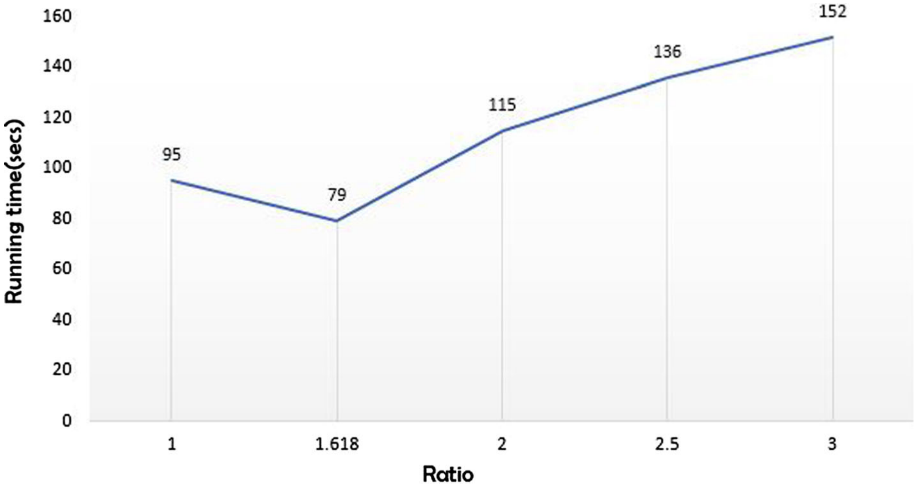


Fig. 12 Comparison of running times for various network configuration ratios on RAVDESS dataset

Table 18 Results for various network configurations of hidden units on both datasets

Dataset			Voice gender	RAVDESS
Task	Configuration	No. of hidden layers	Accuracy	Accuracy
Emotion	233, 144, 89, 55, 34, 21, 13	7	N.A.	0.84
Emotion	13, 21, 34, 55, 89, 144, 233	7	N.A.	0.82
Emotion	55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55	15	N.A.	0.85
Speaker	233, 144, 89, 55, 34, 21, 13	7	N.A.	0.53
Speaker	13, 21, 34, 55, 89, 144, 233	7	N.A.	0.54
Speaker	55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55	15	N.A.	0.61
Gender	233, 144, 89, 55, 34, 21, 13	7	0.83	0.88
Gender	13, 21, 34, 55, 89, 144, 233	7	0.85	0.88
Gender	55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55	15	0.89	0.93

The accuracies obtained while training, validating and testing the identification of speakers by decomposing the multiclass problem into binary classification of one speaker versus others is shown in Fig. 11.

It can be seen from Fig. 11 that the accuracies are quite exceptionally high compared to multiclass classification accuracies, showing the existence of feature overlaps between voice of various speakers in terms of tonal features like pitch, entropy, etc. Our model performs relatively worse for speaker ids 3, 5, 7 and 20 compared to others. The reasons are low amplitude and frequencies of the voice in the speech samples, which were checked and verified manually.

The running times of the model consisting of training, validation and testing for 200 epochs of training for various ratio configuration of the model's units on RAVDESS Dataset are shown in Fig. 12. The accuracies for all the configurations have a standard deviation of 1-2%. The lowest running time for Golden ratio proves that it helps in more efficient and faster training and testing of our model.

As it can be observed from Table 18 that the Golden ratio has a significant effect on the accuracy of the overall model. For strictly increasing or strictly decreasing hidden unit configurations, the maximum number of units in a hidden layer is kept lesser than the number of features, and its minimum number is greater than the number of classes. Maximum number of hidden layers supporting these constraints in the given range of numbers is 7. For the configuration with uniform number of units in the hidden layers throughout, more number of layers lead to better training. It has been set to 55, keeping in mind, the average number of units for our experimentation.

The confusion matrix generated using our GRaNN model for solving speaker identification problem on RAVDESS dataset is shown in Table 19.

Moreover, the size constraints of the data contribute to this problem. Since this type of experiment has been done for the first time on RAVDESS dataset, it is not possible to compare the results with other state-of-the art methods. Since no other models are available for comparison purposes for speaker identification on the RAVDESS dataset hence, we have compared our performance with a logistic regression model (described in [41]) only. For this purpose, we have used the features before feature selection and not included any kind of optimization. This is pertaining to the fact that logistic regression model performs quite well in classification tasks. The results obtained using the logistic regression model (where number of iterations taken as 500) are shown in Table 20.

5 Conclusion

In this paper, we have proposed a deep neural network model, called GRaNN, which is optimized using the concept of Golden ratio for the purpose of emotion, gender and speaker recognition from human speeches. We have elaborated the relevance and usage of Golden ratio in optimization by discussing its use in other domains and the role it plays in our architecture. We have used feature engineering and converted the voice clips of the RAVDESS dataset into a set of features and used features already present in the Voice Gender dataset, which are passed through feature selection algorithms and then fed to our model for classification purposes. We have applied a wrapper-filter-based feature selection technique, for the said purpose aiming for improved learning by the model. Though we have aimed for gender and emotion identification tasks primarily, so in order to check the versatility of our model, we have exploited the RAVDESS dataset (comprising voices of 24 different actors) for speaker

Table 19 Confusion matrix for speaker identification problem using our GRaNN model

Truth values														
Speaker	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14
Predicted values														
#1	5	0	0	0	1	0	0	0	0	0	0	0	1	0
#2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
#3	1	0	3	0	0	0	1	0	0	0	0	0	0	0
#4	0	0	0	2	0	1	0	0	0	0	0	0	0	1
#5	0	0	1	0	3	0	0	0	2	0	0	0	2	0
#6	1	0	0	0	0	3	0	0	0	0	0	0	0	0
#7	0	0	0	0	0	0	4	0	0	0	0	0	0	0
#8	0	0	0	0	0	0	1	5	0	0	0	0	0	0
#9	0	0	0	0	0	0	1	0	2	0	0	0	1	0
#10	0	0	0	1	0	0	0	1	0	3	0	0	0	0
#11	0	0	0	0	0	0	0	0	0	0	2	0	0	0
#12	0	1	0	0	0	1	0	1	0	0	0	3	0	0
#13	0	0	0	0	0	0	1	0	2	0	1	0	0	0
#14	0	0	0	1	0	0	0	0	0	0	0	0	0	2
#15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#16	1	0	0	1	0	0	0	0	0	0	0	0	0	0
#17	0	0	0	0	0	0	0	0	0	0	1	0	0	0
#18	0	0	0	3	0	0	0	0	0	0	0	0	0	0
#19	0	0	2	0	0	0	0	0	0	1	0	1	1	0
#20	0	0	0	0	0	0	0	0	0	0	0	0	0	1
#21	0	0	0	0	1	0	0	0	0	0	0	0	0	0
#22	0	0	0	0	0	0	0	0	0	0	0	0	0	0
#23	1	0	0	0	0	0	0	0	0	0	0	0	0	0
#24	0	0	0	2	0	0	0	0	0	0	0	0	0	1

Truth values										
Speaker	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24
Predicted values										
#1	0	0	0	0	0	0	0	0	1	0
#2	0	1	0	0	0	0	0	0	0	1
#3	0	0	1	0	1	0	0	0	0	0
#4	0	1	0	0	0	1	0	0	0	1
#5	0	0	0	0	1	0	0	0	0	0
#6	0	0	0	0	0	1	0	0	0	0
#7	0	0	0	0	0	0	0	0	0	0
#8	0	0	0	0	0	0	0	0	0	0
#9	0	0	0	0	0	0	1	0	0	0
#10	0	0	0	0	0	0	0	0	0	0
#11	1	0	0	0	0	0	0	0	0	0
#12	0	0	0	0	0	0	0	0	0	0
#13	0	0	0	0	0	0	0	0	0	0
#14	0	0	0	0	0	1	0	0	0	2
#15	3	0	1	0	0	0	0	0	0	0
#16	0	3	0	0	1	0	0	0	0	0
#17	1	0	7	0	0	0	0	0	0	0
#18	0	0	0	4	0	0	0	1	0	2

Table 19 (continued)

Truth values										
Speaker	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24
#19	0	0	0	0	0	0	0	0	0	0
#20	0	0	0	0	0	3	0	0	0	1
#21	0	0	1	0	0	0	3	0	0	0
#22	0	0	0	0	0	0	0	7	1	1
#23	0	0	0	0	1	0	0	0	4	0
#24	0	0	0	0	0	0	0	1	0	2

Table 20 Comparison of speaker identification results with logistic regression on RAVDESS dataset

Speaker	Precision	Recall	F1-score	Support
#1	0.67	0.75	0.71	8
#2	0.40	0.50	0.44	4
#3	0.50	0.29	0.36	7
#4	0.40	0.29	0.33	7
#5	0.33	0.22	0.27	9
#6	0.60	0.60	0.60	5
#7	0.40	0.67	0.50	3
#8	0.83	0.83	0.83	6
#9	0.20	0.17	0.18	6
#10	0.33	0.40	0.36	5
#11	0.60	1.00	0.75	3
#12	0.25	0.17	0.20	6
#13	0.11	0.25	0.15	4
#14	0.57	0.67	0.62	6
#15	0.11	0.25	0.15	4
#16	0.50	0.33	0.40	6
#17	0.88	0.78	0.82	9
#18	1.00	0.30	0.46	10
#19	0.29	0.40	0.33	5
#20	0.33	0.20	0.25	5
#21	0.50	0.60	0.55	5
#22	0.67	0.80	0.73	10
#23	0.67	0.67	0.67	6
#24	0.60	0.60	0.60	5
Weighted avg	0.53	0.49	0.48	144

identification purpose. However, we achieved impressive results for speaker identification task as shown in Sect. 4. We have presented results related to one versus rest binary classification accuracies for the tasks of speaker and emotion recognition tasks. We have compared our model with a semi-supervised model and other state-of-the-art models, and our model outperforms those in terms of classification accuracy.

Availability of large-sized datasets and investigation of better methods of feature extraction can lead to more accurate results in future. Due to the absence of work on speech identification on the RAVDESS dataset, we are unable to show much comparison, but we have included the results for logistic regression. So, it can be considered as the part of our future work to apply other approaches for speaker identification on this dataset and check their performances. We can consider the use of various other nature-inspired feature selection techniques available as part of our future works and check the effectiveness of the same. Since our experiments provide quite promising results, we intend to focus on extending our experiments by application of the proposed algorithm to other voice signal-based datasets for speaker identification.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-022-07261-x.

Acknowledgements The authors would like to thank the Centre for Microprocessor Applications for Training, Education and Research (*CMATER*) research laboratory of the Computer Science and Engineering Department, Jadavpur University, Kolkata, India, for providing the infrastructural support.

Author Contributions AG and PKS conceived and designed the experiments; BR performed the experiments; PKS and AG analysed the data; RS contributed reagents/materials/analysis tools; AG, BR, FG, JDV, PKS, and RS wrote the paper; writing-review and editing, supervision, PKS and RS; data acquisition, FG and JDV All authors read and approved the final manuscript.

Funding The authors gratefully acknowledge financial support from ANID PIA/APOYO AFB180003. Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Availability of data and materials Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

1. Avriel M, Wilde DJ (1966) Optimally proof for the symmetric fibonacci search technique. Fibonacci Q J 265−269
2. Bhattacharyya T, Chatterjee B, Singh PK, Yoon JH, Geem ZW, Sarkar R (2020) Mayfly in harmony: a new hybrid meta-heuristic feature selection algorithm. IEEE Access 8:195929–195945. https://doi.org/10.1109/ACCESS.2020.3031718
3. Bisio I, Lavagetto F, Marchese M, Sciarrone A, Frà C, Valla M (2015) Spectra: a speech processing platform as smartphone application. In: 2015 IEEE international conference on communications (ICC), pp 7030–7035
4. Biswas S, Solanki S (2021) Speaker recognition: an enhanced approach to identify singer voice using neural network. Int J Speech Technol. https://doi.org/10.1007/s10772-020-09698-8
5. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167
6. Buyukyilmaz M, Cibikdiken AO (2016) Voice gender recognition using deep learning. In: 2016 International conference on modeling, simulation and optimization technologies and applications (MSOTA2016). Atlantis Press. https://doi.org/10.2991/msota-16.2016.90
7. Carterette B (2009) Precision and recall. Springer, Boston, pp 2126–2127. https://doi.org/10.1007/978-0-387-39940-9_5050
8. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794
9. Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: deep speaker recognition. CoRR abs/1806.05622
10. De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11):3178–3192
11. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS et al (2015) The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202
12. Garain A (2020) Gender recognition from voice. https://doi.org/10.21227/v62v-g267
13. Garain A, Das D (2020) K-rms algorithm. Procedia Comput Sci **167**, 113 – 120. ; International conference on computational intelligence and data science. https://doi.org/10.1016/j.procs.2020.03.188
14. Garain A, Singh PK, Sarkar R (2021) Fuzzygcp: a deep learning architecture for automatic spoken language identification from speech signals. Expert Syst Appl 168:114416. https://doi.org/10.1016/j.eswa.2020.114416
15. Ge Z, Iyer AN, Cheluvaraja S, Sundaram R, Ganapathiraju A (2017) Neural network based speaker classification and verification systems with enhanced features. In: 2017 intelligent systems conference (IntelliSys), pp 1089–1094
16. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw 18(5–6):602–610
17. Guo X, Yan X, Jermsittiparsert K (2021) Using the modified mayfly algorithm for optimizing the component size and operation strategy of a high temperature pemfc-powered cchp. Energy Rep 7:1234–1245. https://doi.org/10.1016/j.egyr.2021.02.042
18. Golden search selection method. http://mathforcollege.com/nm/mws/gen/09opt/mws_gen_opt_txt_goldensearch.pdf
19. Hansen JHL, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag 32(6):74–99
20. Hu H, Xu MX, Wu W (2007) Gmm supervector based svm with spectral features for speech emotion recognition. In: 2007 IEEE international conference on acoustics, speech and signal processing-ICASSP'07, vol 4. IEEE, pp IV–413
21. Jalal MA, Loweimi E, Moore RK, Hain T (2019) Learning temporal clusters using capsule routing for speech emotion recognition. In: Proceedings of the Interspeech, vol 2019, pp 1701–1705
22. Kazienko P, Lughofer E, Trawiński B (2013) Hybrid and ensemble methods in machine learning j. ucs special issue. J Univ Comput Sci 19(4):457–461
23. Kushwah S, Singh SK, Vats K, Nemade V (2019) Gender identification via voice analysis
24. Li W, Kim D, Kim C, Hong K (2010) Voice-based recognition system for non-semantics information by language and gender. In: 2010 third international symposium on electronic commerce and security, pp 84–88
25. Liu Y, Chai Y, Liu B, Wang Y (2021) Bearing fault diagnosis based on energy spectrum statistics and modified mayfly optimization algorithm. Sensors 21:2245. https://doi.org/10.3390/s21062245
26. Liu Z, Jiang P, Wang J, Zhang L (2021) Ensemble forecasting system for short-term wind speed forecasting based on optimal sub-model selection and multi-objective version of mayfly optimization algorithm. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.114974
27. Livieris IE, Pintelas E, Pintelas P (2019) Gender recognition by voice using an improved self-labeled algorithm. Mach Learn Knowl Extr 1(1):492–503. https://doi.org/10.3390/make1010030
28. Livingstone SR, Russo FA (2018) The Ryerson audio-visual database of emotional speech and song (RAVDESS). Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583. https://doi.org/10.5281/zenodo.1188976
29. Löllmann HW, Vary,P (2008) Low delay filter-banks for speech and audio processing. In: Speech and audio processing in adverse environments. Springer, pp 13–61
30. McFee B, McVicar M, Raffel C, Liang D, Nieto O, Moore J, Ellis D, Repetto D, Viktorin P, Santos JF, Holovaty A (2015) librosa: v0.4.0. https://doi.org/10.5281/zenodo.18369
31. Nasef M, Mausad A, Nabil M (2021) Voice gender recognition under unconstrained environments using self-attention. Appl Acoust 175:107823. https://doi.org/10.1016/j.apacoust.2020.107823
32. Pahwa, A., Aggarwal, G.: Speech feature extraction for gender recognition (2016)
33. Peng CYJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96(1):3–14
34. Reynolds DA (2002) An overview of automatic speaker recognition technology. In: 2002 IEEE international conference on acoustics, speech, and signal processing, vol 4, pp IV–4072–IV–4075

35. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, pp 3856–3866
36. Scherer KR (2000) A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In: Sixth international conference on spoken language processing
37. Segal MR (2004) Machine learning benchmarks and random forest regression
38. Sen S, Agarwal RP (2008) Golden ratio in science, as random sequence source, its computation and beyond. Comput Math Appl 56(2):469–498. https://doi.org/10.1016/j.camwa.2007.06.030
39. Shafran I, Riley M, Mohri M (2003) Voice signatures. In: 2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No.03EX721), pp 31–36
40. Tripathi S, Ramesh A, Kumar A, Singh C, Yenigalla P (2019) Learning discriminative features using center loss and reconstruction as regularizer for speech emotion recognition. arXiv:1906.08873
41. Wright RE (1995) Logistic regression
42. Yacoub S, Simske S, Lin X, Burns J (2003) Recognition of emotions in interactive voice response systems. In: Eighth European conference on speech communication and technology
43. Zervoudakis K, Tsafarakis S (2020) A mayfly optimization algorithm. Comput Ind Eng 145:106559. https://doi.org/10.1016/j.cie.2020.106559
44. Zvarevashe K, Olugbara OO (2018) Gender voice recognition using random forest recursive feature elimination with gradient boosting machines. In: 2018 International conference on advances in big data, computing and data communication systems (icABCD), pp 1–6 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Avishek Garain[1] · Biswarup Ray[1] · Fabio Giampaolo[2] · Juan D. Velasquez[3] · Pawan Kumar Singh[4] · Ram Sarkar[1]

✉ Fabio Giampaolo
fabio.giampaolo@unina.it

Avishek Garain
avishekgarain@gmail.com

Biswarup Ray
raybiswarup9@gmail.com

Juan D. Velasquez
jvelasqu@dii.uchile.cl

Pawan Kumar Singh
pksingh.it@jadavpuruniversity.in

Ram Sarkar
ram.sarkar@jadavpuruniversity.in

[1] Department of Computer Science and Engineering, Jadavpur University, 188, Raja S.C. Mallick Road, Kolkata, West Bengal 700032, India

[2] Department of Mathematics and Applications "Renato Caccioppoli", University of Naples Federico II, Via Cinthia, 80126 Naples, NA, Italy

[3] Department of Industrial Engineering, Faculty of Physical and Mathematical Sciences, Instituto Sistemas Complejos de Ingeniería (ISCI), University of Chile, Santiago, Chile

[4] Department of Information Technology, Jadavpur University, Jadavpur University Second Campus, Plot No. 8, Salt Lake Bypass, LB Block, Sector III, Salt Lake City, Kolkata, West Bengal 700106, India