A rough set theory and deep learning based predictive system for gender recognition using audio speech

Ghazaala Yasmin, Asit Kumar Das*, Janmenjoy Nayak, S Vimal, Soumi Dutta

St. Thomas' College of Engineering and Technology, Kolkata-700023, India
Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711103, India
Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh-532201
Ramco Institute of Technology, Rajapalayam, India
Institute of Engineering and Management, Saltlake, Kolkata-700091, India

Abstract

Speech is one of the most delicate medium through which gender of the speakers can easily be identified. Though the related research has shown very good progress in machine learning but recently, deep learning has imparted a very good research area to explore the deficiency of gender discrimination using traditional machine learning techniques. In deep learning techniques, the speech features are automatically generated by the reinforcement learning from the raw data which have more discriminating power than the human generated features. But in some practical situations like gender recognition, it is observed that combination of both types of features sometimes provides comparatively better performance. In the proposed work, we have initially extracted and selected some informative and precise acoustic features relevant to gender recognition using entropy based information theory and Rough Set Theory (RST). Next, the audio speech signals are directly fed into the deep neural network model consists of Convolution Neural Network (CNN) and Gated Recurrent Unit network (GRUN) for extracting features useful for gender recognition. The RST selects precise and informative features, CNN extracts the locally encoded important features, and GRUN reduces the vanishing gradient and exploding gradient problems. Finally, a hybrid gender recognition system is developed combining both generated feature vectors. The developed model has been tested with five bench mark and a simulated dataset to evaluate its performance and it is observed that combined feature vector provides more effective gender recognition system specially when transgender is considered as a gender type together with male and female.

Keywords: Acoustic features, Feature selection, Rough set theory, Information theory, Machine learning, Deep neural network

1. Introduction

Gender recognition from speech and images has always remained a challenging task. It is a very common and needful requisite in all areas including health care section, forensic lab, and any industrial area. Speech and image both are important data to identify the gender. Speech is the medium through which gender can be easily identified. It is known as a physiological signal which represents information at multiple levels such as linguistic content (like language, word, accent etc.) and paralinguistic content (like gender, age, emotion etc.). Beside it, speech also carries important information of acoustic nature of sound. As the technology is enhancing and the use of electronic devices (such as mobile phone Google assistant, alexa) has been already introduced and in the peak demand as per the market value is concerned, the trade for

*Corresponding author, Tel: +91-9830342574
Email addresses: ghazaala.yasmin@gmail.com (Ghazaala Yasmin), akdas@cs.iiests.ac.in (Asit Kumar Das), mailforjnayak@gmail.com (Janmenjoy Nayak), svimalphd@gmail.com (S Vimal), soumi.it@gmail.com (Soumi Dutta)

the development of speech and audio analytic tools is kept on increasing. The research is going on not only in the linguistic areas [1] such as extracting message and working on words but also in paralinguistic areas such as Automatic identification of speaker [2], emotion analysis [3], [4] from speech. This area has a wide range of applications including telecom industry. Gender has yet been devoted only with two types, male and female. There was no stand given to transgender. In this current era, the transgender has also getting right and now itself getting privileged with a unique gender. Plenty of work has been done for male and female gender identification from speech using machine learning [5], [6], [7]. The techniques are still enhancing the system for better performance. The traditional supervised and unsupervised machine learning techniques generally used higher label features extracted by the human from the speech for categorization of speakers. These higher label features may not be sufficient for optimal categorization. The lower label features extracted using various deep neural networks [8], [9] are more effective for this purpose. In this paper, we have explored two different deep neural network models, namely, Convolution Neural Network (CNN) [10] and Gated Recurrent Unit Network (GRUN) [11] and have demonstrated that they perform better than the traditional machine learning approaches.

1.1. Motivation

Deep learning is the subset of machine learning where various layers of networks provide different interpretation on the feeding. The main benefit of deep network is that it does not require higher label structured data for classification, rather it uses the raw input data all the way through different layers. Each hierarchy of layers of network defines specific set of features just as similar to human brain solve any problem hierarchically passing queries to the concept of related queries. After processing the data within different layers of deep neural network, the system computes the appropriate identifiers for classification of data. Gender recognition has many applications such as, improving the intelligence of a surveillance system, analyzing the customer's demands for store management, allowing the robots to perceive gender, and so on. Though many works have been introduced for gender identification using deep learning but transgender has not been considered as a gender in most of them. As transgender is very difficult to distinguish from male and female based on the speech, so the concept of rough set theory and information theory is very helpful for distinguish them from other class of genders. This motivate us to propose a hybrid model integrating Rough Set Theory (RST) and deep neural networks, namely CNN and GRUN to select the minute features from speech which can differentiate transgender, male and female speakers more effectively. Thus we extract features in different forms which are complementary to each other. The classification model is learned using this multi-view dataset to make full use of the hidden information. The RST selects precise and informative human extracted features which are very important to distinguish transgender from others; on the other hand, deep learning model captures the advantage of extracting the locally encoded important features with the help of CNN and long-term dependencies with the help of GRU.

1.2. Literature Survey

Gender recognition from speech is a very well known topic among researchers from past decades. Plenty of research has been undergone for this problem. As time passed the technique to develop the gender recognition system get improved to enhance the performance. Earlier lot of research had done for gender recognition [12], [13], [14] using machine learning, data mining and pattern recognition. Bisio et al. [15] represented gender driven emotion recognition system from audio signal allowing effective human-computer intelligent interaction. Their system consists of two subsystems, one is gender recognition and other is emotion recognition. Gender recognition is done based on pitch extraction from the audio signal and a Support Vector Machine (SVM) based emotion classification model is developed. Zeng et al. [16] proposed Gaussian Mixture Model (GMM) based approach for gender classification by applying the combined parameters of speech and relative spectral perceptual linear predictive coefficients to model the characteristics of male and female speech. The model provides very good accuracy even if sufficiently noisy speech is considered. The paper [17], [18] proposed speaker, age and gender recognition system using acoustic and prosodic level features. Their work also adopted the approach of GMM to train the classification model. Yasmin et al. [19] had proposed a new system of gender recognition using acoustic features from speech signal. The work

has adopted perceptual features like pitch, MFCC, tempo and other low level acoustic features. Ahmad et al. [20] proposed a technique for gender recognition using MFCC for telephonic voice application and the performance has been compared with different well known existing classifiers, where SVM has been found to come up with better result for classifying male and female. Harb et al. [21] also introduced gender recognition system using audio signal with the help of first order spectrum statistics of 1 second windows. They have used neural networks as classifiers. To improve the performance, the research is now going on in the area of deep learning. Levi et al. [22] proposed a simple convolution net architecture to estimate age and gender of the persons based on images. They claimed that the model is very effective even if the amount of learning data is limited. Alkhawaldeh et al. [23] described a gender classification model with the help of one dimension convolution neural network. They used the features, such as Mel Spectrogram, Mel Frequency Cepstral Coefficients, as the single dimension input sequence to CNN for training of the model. Kabil et al. [24] proposed a work for gender recognition from raw speech signal using convolution neural network. In their work, the audio data itself has been supplied to the convolution model to train the model for gender identification. Mansanet et al. [25] also classified male and female from image using local deep neural netwrok. The local deep neural network used local features of image to train the model for gender classification. Dehghan et al. [26] described the details of Sighthound's fully automated age, gender and emotion recognition system. The backbone of the system consists of several deep convolution neural networks that are not only computationally inexpensive, but also provide state-of-the-art results on several competitive benchmarks. Rajeev et al. [27] presented an algorithm for simultaneous face detection, landmarks localization, pose estimation and gender recognition using deep convolution neural networks. The proposed method fuses the intermediate layers of a deep CNN using a separate CNN followed by a multi-task learning algorithm that operates on the fused features. It exploits the synergy among the tasks which boosts up their individual performances. Wolfshaar et al. [28] applied deep convolution neural networks on gender classification by fine-tuning a pretrained neural network. They explored the performance of dropout support vector machines by training them on the deep features of the pretrained network as well as on the deep features of the fine-tuned network. Wang et al. [29] proposed a speech emotion and age/gender recognition system using deep neural networks. They have used deep neural networks to encode each utterance into a fixed-length vector by pooling the activation of the last hidden layer over time. The feature encoding process is designed to train the utterance-level classifier for better classification and a kernel extreme learning machine is further trained on the encoded vectors for better utterance-level classification. Markitantov et al. [30] presented a novel approach in the paralinguistic field of age and gender recognition by speaker's voice based on deep neural networks. The training and testing of proposed models were implemented on the German speech corpus aGender. They have conducted experiments using different network topologies, including neural networks with fully-connected and convolution layers. Their method provides better result of speaker age recognition than speaker gender recognition in comparison to existing traditional classification methods. Sánchez-Hevia et al. [31] dealt with joint gender recognition and age group classification from speech for improving the functionalities of interactive voice response systems. Due to the discriminative and representation capabilities of deep neural networks, they have used it in speech processing problems for features extraction and selection. They have presented various neural network architectures and compared themselves using Mozilla's 'Common Voice' dataset, an open source speech corpus. Gupta et al. [32] proposed a stacked machine learning technique for gender recognition through voice using the acoustic parameters of voice sample. The performance of their work is compared with some traditional and useful existing classifiers to demonstrate the effectiveness of their models. Ertam et al. [33] proposed an effective deeper LSTM networks based gender recognition system using audio data set. Initially, they have selected 10 most effective features and subsequently applied a double-layer LSTM architecture based deep learning networks. Based on the performance, authors claims that their model is an effective and fast approach for gender recognition. In the best of the author knowledge, most of the gender recognition systems developed by the researchers are capable of classifying male and female voice using audio signals. As discussed, although much effort has been dedicated to improve the performance of gender recognition, the noted algorithms suffer from the following limitations and challenges.

- The earlier works handles only two different gender types, male and female. But in presence of

transgender, it is very difficult to recognize all genders separately. This limitation is tried to overcome by devising a novel feature selection algorithm using information theory and rough set theory, which helps to select only informative, and precise features from the audio speech.

- Either CNN or RNN are used by the researchers for gender recognition. The combination of these two models is sometimes beneficiary. The CNN is responsible for capturing the locally encoded important features and GRU is used to consider long-term dependencies among the features. Thus it is one of the challenging tasks to construct a hybrid deep model for gender recognition.

- The previous works either use human extracted features or machine extracted features but not the both for gender recognition, which may not properly learn the model, specially when the transgender comes into the picture. This challenge and limitation is handled by developing a hybrid gender recognition model integrating information theory, rough set theory, and deep neural networks.

The proposed work explores about how the transgender can be distinguished from male and female. A hybrid deep neural network model together with the concepts of RST and information theory is framed for this purpose in the paper.

1.3. Contribution

Speech is produced by humans using a natural biological mechanism in which lungs discharge the air and convert it to speech passing through the vocal cords and organs including the tongue, teeth, and lips. In general, a speech and voice recognition system can be used for gender identification. Gender recognition is a technique to identify the gender category of a speaker by processing speech signals based on the extracting acoustic features such as duration, intensity, frequency and filtering. Recently, many machine learning techniques are available for gender recognition. But, transgender is a different gender of human being which is very difficult to identify using face recognition and speech recognition system. Traditional machine learning techniques are not so capable to accurately classify the audio data that contain all three genders, i.e., male, female, and transgender. Deep learning based gender recognition system provides comparatively better performance than traditional machine learning techniques by giving training to the model using huge volume of audio data. It has been observed that different deep neural networks provide different performance, no single model always provide the best result for all audio data. At the same time, machine extracted features are not self sufficient for gender recognition, as many imprecise and ambiguous features exist in the dataset due to the mixture of transgender voice with male and female vices. In the paper, we have proposed a novel hybridized deep neural network model combining CNN and GRU together with RST to develop a gender recognition system. We have considered human extracted acoustic features and applied them in a proposed RST based feature selection algorithm to filter out the redundant and irrelevant features and select only informative and precise features of audio speech. The CNN is used to capture the locally encoded important features and GRU is used to consider long-term dependencies among the features. Thus we extract features in different forms which are complementary to each other. The classification model is learned using this multi-view dataset to make full use of the hidden information. It is observed that, the method not only recognize the male and female accurately, but also recognize equally the transgender based on audio speeches. The main contributions of this paper are concluded by the following steps and depicted in Figure 1.

1. A possible set of acoustic features are extracted from audio speech and a novel feature selection algorithm is devised using the concepts of information theory and rough set theory to select only the informative, precise, and unambiguous features relevant for gender recognition.
2. The deep neural network architectures of CNN and GRU are explored for developing gender recognition system. A hybrid deep neural model combining CNN and GRU together with RST is framed considering the selected acoustic features. The selected acoustic features are combine with deep neural model extracted features and the resultant feature vector is fed into the model for gender recognition.

3. All the developed models (i.e., CNN, CNN and GRUN, CNN and GRUN and RST) are experimented using both sample and benchmark audio datasets to evaluate them. It is observed that RST based hybrid deep neural network model has the higher capability to recognize the genders specially when transgender is present in the dataset. The method is also compared with some popular gender recognition algorithms to demonstrate the effectiveness of the proposed model.

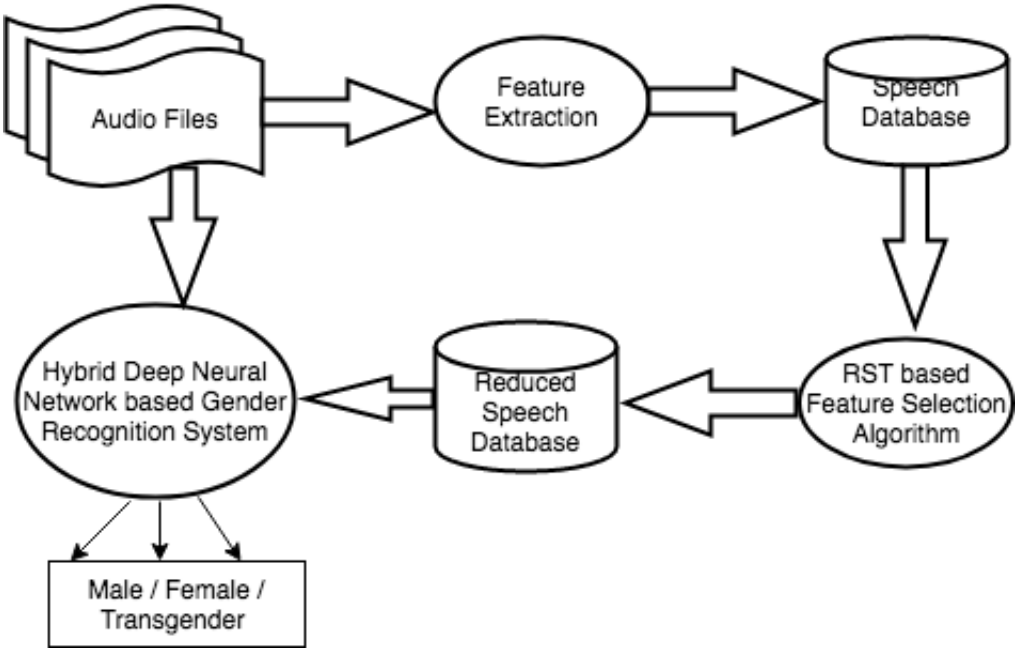


Figure 1: The work flow diagram of the proposed gender recognition system

1.4. Summary

The remaining part of the paper is structured in the following ways. Section 2 describes details about various human extracted features of audio signal and Section 3 describes informative optimal feature subset selection technique based on information theory and RST. Section 4 describes the proposed hybrid model combining CNN and GRUN architecture together with RST developed for recognizing genders based on audio speeches. Section 5 provides the experimental setup and empirical result analysis and finally, section 6 gives the brief conclusions and the future scopes of the paper.

2. Description of human extracted features

Gender information is a distinctive and the most important property in a speech. Determination of this information from a speech signal is a substantial subject. Gender information used for various purposes in many applications including speech emotion recognition, human to machine interaction, sorting of telephone calls by gender categorization, automatic salutations, muting sounds for a gender and so on. Gender identification can improve the prediction of other speaker traits such as age and emotion, either by jointly modeling gender with age or in a pipelined manner. Speaker verification systems also implicitly or explicitly use gender information. In general, identification of a speaker gender is important for increasingly natural and personalized dialogue systems. The acoustic features of the speech signal are very much helpful for the gender recognition. There are a set of features used for recognizing the voice gender. The most common features utilized for voice gender recognition are Frequency, Pitch, mel-frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma), and tempo features.

1. **Frequency:** The resonance structure of the vocal tract can be easily examined by drawing a smooth line above the spectrum, as shown in Figure 2. It gives the macro-shape of the spectrum of a speech signal, which is often used to model speech signals.

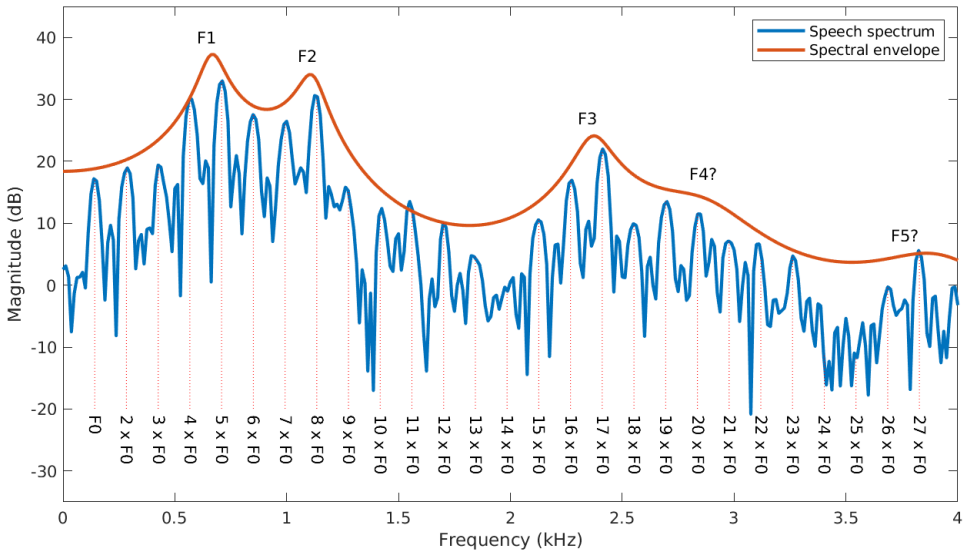


Figure 2: The resonance structure of the spectrum of a speech signal

From Figure 2, it is observed that the speech signals have many frequency features. We have considered only the frequencies, $F0$, $F1$, and $F2$, where $F0$ is the fundamental frequency, and $F1$, $F2$ are the first two formants, providing the two lowest resonances of the vocal tract. The frequency features include the statistics of fundamental frequency $F0$ and the first 2 formants, $F1$ and $F2$. The $F0$, $F1$, and $F2$ are computed over windows of 20 ms with overlaps of 10ms. The overlapping time is considered as the speech signal generally remains stationary in this time scale. The statistical properties, such as Mean, maximum, minimum, median and the standard deviation of all three frequencies are used as extracted frequency based features of the speech signals. Thus, as a whole 15 frequency features are used. The $F0$ is computed by auto-correlation method, and $F1$ and $F2$ are computed by solving the roots of the Linear Predict Coding (LPC) polynomial using *PRAAT* [34], an open-source toolkit for speech analysis. The frequencies are only computed through the vowels periods, and for the consonants, they are assumed as 0, and are not considered in the statistics.

2. **Pitch:** Pitch is termed as the degree of shrillness and harshness of a voice. Pitch is described as the fundamental frequency of glottal pulse. Precisely, the quality of any tone can be dictated by the rate of vibrations through which it is generated. The main motive of using the pitch feature for gender recognition is that the average fundamental frequency (i.e., reciprocal of pitch period) for men is typically in the range of 100 Hz to 146 Hz, and that for women is 188 Hz to 221 Hz [35]. But, an overlap of the pitch values between male and female voices naturally exists as shown in Figure 3.
We have estimated the pitch period of a speech sample as sum of amplitude modulation - frequency modulation (AM-FM) formant models [36]. AM components represent the envelope of the short-time speech signals which only contains information within a certain bandwidth, which reduces the noise effect and the distortion effect of the speech signals. In the proposed work, 88 pitch based features have been detected. The speech signal has been distributed into 88 frequency bands. The short-time mean-square power (STMSP) has been calculated for each band. Therefore, the average of STMSP of each band, computed using equation (1), has been considered as a single pitch feature obtained from

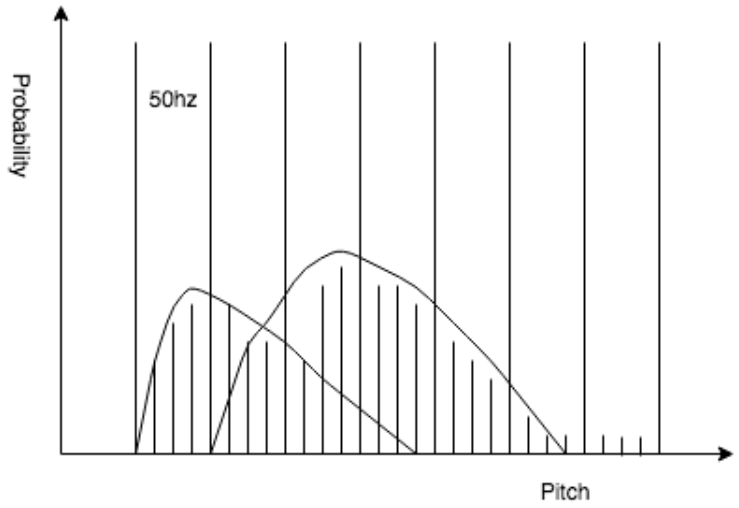


Figure 3: The overlapping of pitch values between male and female

that band.

$$P_s = \frac{1}{q}\sum_{f=1}^{q}(x_f)^2 \tag{1}$$

As a result, total 88 pitch features are extracted from each speech signal. In equation (1), P_s represents the STMSP of s-th band, q is the total number of samples of each band in frequency domain and x_f is the sampled value of each band. Here, the value of s is 88.

3. **Cepstral Coefficients:** In addition to Frequency statistics and Pitch features, we have explored the use of Mel-frequency cepstral coefficients (MFCCs) as features for gender detection. MFCCs are coefficients that collectively make up an MFCC feature of a signal. In MFCC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, which takes important role for gender recognition. For calculating MFCCs, speech signal is divided into collection of frames of 20 ms duration. We have generated 26 MFCCs for each frame and computed the mean and standard deviation of each coefficient. Thus 52 features are computed for each speech signal.
4. **power spectrogram chroma:** Chroma based feature is an interesting and powerful representation of audio in which the entire spectrum is projected into 12 bins representing 12 distinct semitones or Chroma. Chroma depended trait is so powerful acoustic feature that it can be used to study the characteristics of tones of a speech. Chroma features can reflect melodic and harmonic nature of a speech signal. Chroma features indicate the intensity of different frames of a signal. Chroma traits reflect perceptual dissimilarities among different speech because of which this feature is considered for gender recognition.
5. **Tempo:** Each speech acquires its own speed which can be measured by its tempo feature. It is measured in terms of beats per minute. Since the nature of speech signal is clearly reflected by tempo based feature, this feature has been extracted as a suitable one for gender recognition. Tempo feature is being calculated by the help of the novelty curve from the given input speech signal. Novelty curve is described as type of detection function whose peaks convey the note onsets. In order to extract these features, the novelty curve has broken into non-overlapping tempo windows, each having 20 ms duration. Then, from every window, first 30 Fourier coefficients have been calculated. Finally, the

mean and standard deviation of each Fourier coefficient of all windows is computed, which provides a 60-dimensional tempo based feature vector of the audio signal.

Thus, 227 high level human extracted audio features (frequency (15) + Pitch (88) + MFCC (52) + Chroma (12) + Tempo (60)) are considered from the speech signal using the feature extractor [37].

3. Feature Selection using Rough Set Theory

After feature extraction, a dataset $DS = \{S, F, C\}$ is obtained based on set of speech signals, where $S = \{S_1, S_2, \cdots, S_n\}$ is the set of speeches, $F = \{F_1, F_2, \cdots, F_m\}$ is the set of extracted features, and $C = \{C_1, C_2, \cdots, C_k\}$ is the set of classes representing gender types (here, it is of 3 classes, male, female and transgender). All these higher level human extracted features may not be important and some may be redundant during gender recognition. So feature selection algorithm provides us a minimum set of informative features. To find the most informative subset of features, we have used the information theory [38] and rough set theory [39] together. Information theory [38] is applied to rank all the features based on the ascending order of their entropy, and rough set theory [39] is used to apply the quick reduct [40] generation algorithm for feature selection. The traditional quick reduct algorithm is modified by incorporating the concept of information theory and the step-wise floating forward selection and backward removal concepts [41]. So before describing the proposed feature selection algorithm, we discuss the relevant concepts used for feature selection.

3.1. Information Theory

Information theory [38], discovered by Claude Shannon, has quantified entropy. This is key measure of information which is usually expressed by the average number of bits needed to store or communicate one symbol in a message. Information gain calculates the reduction in entropy from transforming a dataset in some way. Entropy measures the level of impurity in a group of samples. The higher the entropy the more the information content. If a feature say, F_i in F contains v discrete values $\{x_1, x_2, \cdots, x_v\}$, then the entropy of the feature F_i is given by equation (2), where p_j is the probability of occurrence of discrete value x_j of F_i in dataset.

$$H(F_i) = -\sum_{j=1}^{v} p_j log_2(p_j) \tag{2}$$

From equation (2), it is observed that, if feature F_i assumes only one value, then the entropy of this feature becomes zero, which implies that it is not a good feature for learning a classifier. Similarly, if all discrete values are of equal number in the dataset, then the entropy becomes maximum. Thus higher the entropy value implies more important the corresponding feature is and vice versa. In this section, we want to determine which feature in a given training feature set F is most useful for discriminating between the classes (i.e., gender type) to be learned. Here, we have used entropy of the features with respect to the class attribute (i.e., decision feature). In this slightly different usage, the calculation is referred to as mutual information between each condition feature and the decision feature. Mutual information calculates the statistical dependence between a condition feature and a decision feature and is the name given to information gain when applied to feature selection. Information gain tells us how important a given feature is for classifying the samples of different classes. Let, there are k classes, $C_1, C_2, \cdots, C_k$ (in our application of gender recognition, we have considered three different classes, male, female and transgender) in a dataset. Let $S = \{S_1, S_2, \cdots, S_n\}$ be the set of speech signals in the dataset. Let, s_i be the number of speeches of class C_i, $\forall\ i = 1, 2, \cdots, k$. Therefore, $\sum_{i=1}^{k} s_i = n$. Let p_i be the probability that an arbitrary speech, say S_i in S belongs to class C_i. So, p_i is estimated by $p_i = \frac{s_i}{n}$. The amount of information, needed to decide if an arbitrary speech in S belongs to any of the class C_j is defined by equation (3).

$$I(s_1, s_2, \cdots, s_k) = -\sum_{i=1}^{k} \frac{s_i}{\sum_{j=1}^{k} s_j} log_2 \frac{s_i}{\sum_{j=1}^{k} s_j} \tag{3}$$

Assume that using feature F_x of F, the speech set S is partitioned into sets $\{P_1, P_2, \cdots, P_v\}$, where P_i contains s_{ij} number of speeches of class C_j, for $j = 1, 2, \cdots, k$ and $i = 1, 2, \cdots, v$. Then the entropy or the expected information needed to classify the speeches in all subsets P_i is computed using equation (4), where $x = 1, 2, \cdots, m$.

$$E(F_x) = \sum_{i=1}^{v} \frac{\sum_{j=1}^{k} s_{ij}}{\sum_{j=1}^{k} s_j} I(s_{i1}, s_{i2}, \cdots, s_{ik}) \tag{4}$$

By the entropy theory, the encoding information gained by classifying the speeches using the feature F_x is given by equation (5), for $x = 1, 2, \cdots, m$.

$$g(F_x) = I(s_1, s_2, \cdots, s_k) - E(F_x) \tag{5}$$

So, for any two features, F_x and F_y, $g(F_x) > g(F_y)$ implies that feature F_x is more informative than F_y for classification of speeches.

3.2. Rough Set Theory

Rough Set Theory (RST) [39] is a very important concept purely based on mathematics which is frequently used in data mining and knowledge discovery. The dependency of a feature on another feature is easily determined using the indiscernibility relation, a preliminary but very powerful concept of RST. In the work, we are interested to find the dependency of each feature F_x in F on the decision feature C using the indiscernibility relation. Indiscernibility relation is an equivalence relation defined over a subset of features which gives the equivalence classes of speeches such that the speeches in an equivalence class are indiscernible from each other based on the considered feature subset. All the extracted features in feature set F of decision system $DS = \{S, F, C\}$ are real valued which are not suitable for discriminating the speeches using RST. So, the features are discretized using a popular modified *chi*2 algorithm [42]. Thus F becomes the condition feature set of discrete values. The indiscernibility relation *IND(P)* is defined in (6), where $P \subseteq F$.

$$IND(P) = \{(S_i, S_j) \in S^2 : S_i(x) = S_j(x) \forall x \in P\} \tag{6}$$

From the definition of indiscernibility relation, it can easily be proved that it is an equivalence relation, which induces a partition of equivalent classes. Each equivalence class contains a subset of speeches of S which are indiscernible from each other. Each speech S_i in S provides an equivalence class using the equivalence relation $IND(P)$, computed using equation (7).

$$[S_i]_P = \{S_j \in S : (S_i, S_j) \in IND(P)\} \tag{7}$$

The equivalence class obtained using equation (7) is a set of speeches indistinguishable from each other with respect to the feature subset $P \subseteq F$. Similarly, any speech from remaining speech set $S - [S_i]_P$ is selected arbitrarily to compute its corresponding equivalence class. This process is repeated until each speech is placed in any one equivalence class. In our work, we have partitioned the speech set S based on the single feature F_i in F, $\forall i = 1, 2, \cdots, m$. So, using $P = \{F_i\}$ in equation (7), we get a set of equivalence classes, $\forall i = 1, 2, \cdots, m$. One of the most important aspects of feature selection is the discovery of feature (or attribute) dependencies, that describe which features are strongly related to which other features. As the given system has the decision feature (i.e., class variable), so we measure the dependency of each of the condition feature in F on the decision feature C. Let us consider one condition feature F_i of F and the decision feature C to measure the degree of dependency between them, for which following steps are applied:

- The indiscernibility relation, defined in equation (6), induces a partition $\{[x]_{\{F_i\}}\}$ of equivalence classes for $P = \{F_i\}$ and partition $\{[x]_C\}$ for $P = C$.

- Let Q_j is an equivalence class in partition $\{[x]_C\}$. The lower approximation of the target set Q_j is a set $S\underline{\{F_i\}}(Q_j)$ of all speeches positively belong to the target set Q_j and is defined by equation (8), where obviously $[S_k]_{\{F_i\}} \in \{[x]_{\{F_i\}}\}$.

$$S\underline{\{F_i\}}(Q_j) = \{S_k \in S : [S_k]_{\{F_i\}} \subseteq Q_j\} \tag{8}$$

- The positive region $POS_C(F_i)$ is the region which contains all the speeches definitely belong to the equivalence classes of partition $\{[x]_{\{F_i\}}\}$ and is defined by equation (9).

$$POS_C(F_i) = \cup_{Q_j \in \{[x]_C\}} S\underline{\{F_i\}}(Q_j) \tag{9}$$

  It basically contains all the speeches obtained by taking the union of lower approximations, $S\underline{\{F_i\}}$ with respect to feature F_i for all target set Q_j in partition $\{[x]_C\}$.

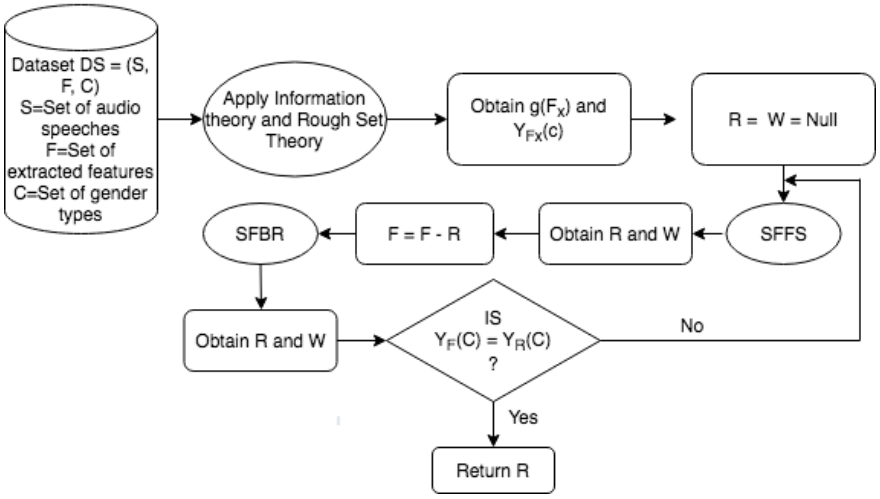
- The dependency of target feature C on feature F_i is the ratio of number of speeches in the positive region to the total number of speeches in the dataset. This dependency is denoted by $\gamma_{\{F_i\}}(C)$ and is defined by equation (10), where S is the set of all speeches in the dataset.

$$\gamma_{\{F_i\}}(C) = \frac{|\cup_{Q_j \in \{[x]_C\}} S\underline{\{F_i\}}(Q_j)|}{|S|} \tag{10}$$

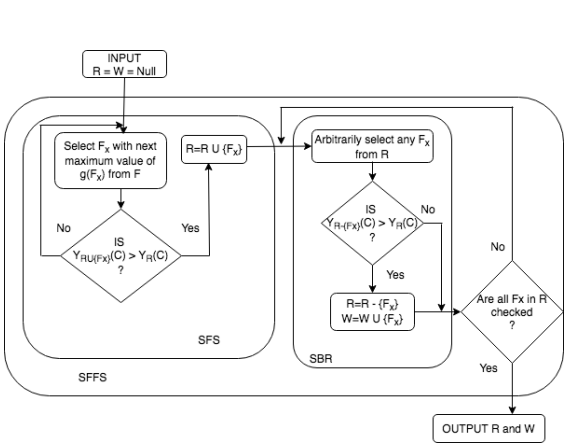
  Obviously, this dependency value ranges from 0 to 1. The larger the value means more the feature C is dependent on feature F_i. The feature on which C is more dependent is considered as more important feature for classification.

3.3. Step-wise floating forward selection and backward removal

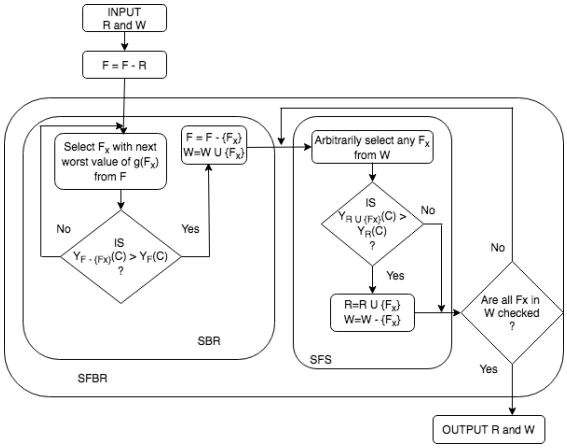
Feature selection [43] is a process of selecting a subset of the original features, yet produce similar or almost similar analytical results. It tries to select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features. But, selection of minimum set of features is an NP-hard [44] problem, and so different heuristics are applied by different researchers to select it. The objective of all the heuristics is to search the space of possible feature subsets that is optimal or near optimal with respect to some objective function(s). In the proposed work, we have used two objective functions, namely information gain based on information theory, and feature dependency based on rough set theory. The most common generic heuristic search algorithms are based on step-wise forward selection (SFS) and step-wise backward removal (SBR) of features based on the objective functions. Both are the iterative process either select or remove one feature in each iteration based on the objective functions. The process terminates in the proposed algorithm, when the value of the feature dependency based objective function fulfil a certain condition. Both the processes are very time consuming as only one feature is examined to be selected or removed in an iteration. To make it more efficient, bidirectional search (BDS) is used where both the SFS and SBR are applied simultaneously. SFS is performed from the empty feature set and SBR is performed from the full feature set. The main limitation of these three algorithms is that once a feature either selected or removed cannot be respectively removed or selected further. It is required because, in case of SFS, some features selected previously may become non useful after the addition of other features, and similarly, in case of SBR, some features previously removed cannot be allowed to reevaluate its usefulness. To overcome such limitation, we have used the concept of Step-wise floating forward selection (SFFS) and Step-wise floating backward removal (SFBR). SFFS starts from the empty feature set and after performing each SFS step, it performs multiple SBR steps as long as the value of the objective function improves. Similarly, SFBR starts from the full feature set, and after each SBR, it performs multiple SFS steps as long as the value of the objective function improves. The proposed feature selection algorithm performs bidirectional search repeatedly applying SFFS followed by SFBR together with the concepts of information theory and RST by defining objective functions.



(a) The higher level workflow



(b) The workflow of SFFS technique



(c) The workflow of SFBR technique

Figure 4: The flowchart of the proposed feature selection algorithm

3.4. Proposed Feature selection Algorithm

As discussed in the subsections of this section, the heuristics applied for feature selection are based on information theory, RST, forward selection and backward removal techniques. The proposed feature selection algorithm is described by Figure 4, where Figure 10a gives the overall higher level details, and Figure 10b, and Figure 10e gives the details work flow of SFFS and SFBR techniques. From the figures, it is noticed that, in SFFS, SFS performs once and SBR performs many times, similarly, in SFBR, SBR performs once and SFS performs many times. Initially, R and W are empty sets, and take as input to SFFS algorithm, which provide new values of R and W. These new values of R and W are input of SFBR algorithm and modified values are returned by this algorithm, which are again inputed to SFFS, the same process is repeated. Finally, the process terminates while attribute dependency of C with respect to F (i.e., $\gamma_F(C)$) is equal to attribute dependency of C with respect to R (i.e., $\gamma_R(C)$) and the algorithm returns the informative and precise set R of features.

As our dataset contains the target variable (i.e., gender type), so we have proposed a supervised feature selection algorithm to select the features which are mostly dependent on the class or target variable. As the

selected features have dependency with the class variable, it is expected that the classifiers generated by the selected features would be more effective for predicting the class labels of the objects. In the paper, we have computed information gain of each feature over the class variable to measure its dependency with the class variable. The feature with maximum information gain is considered as the most informative feature, and vice versa. Next, we follow the bidirectional search repeatedly using SFFS followed by SFBR. For selection of feature we consider the maximum information gain and for removal we select the feature with worst information gain.

To perform SFFS, initially, we start with the empty feature set R and select the feature with maximum information gain. If the feature dependency increases compare to the previous value then we insert it into the set R, otherwise we discard it and select a feature with the next highest information gain. After adding one feature, we repeatedly select an arbitrary feature from R and if its removal from R increases the feature dependency then only it is removed from R and stored in set W, otherwise, the same process is done for next feature in R and continued it until all features in R are exhausted. This process gives the feature subset R selected finally by one execution of SFFS.

Next, we perform SFBR, where we initially consider whole feature set F to select the worst feature with respect to the information gain. But as SFFS has already selected feature subset R they will not be further removed, so the initial feature set considers for SFBR is $F = F - R$, from which the worst feature is selected. If the removal of this feature from F increases the feature dependency compare to that of F then it is removed and stored into W, otherwise next worst feature is selected from F and the same process is repeated. Finally, when one feature is removed from F and stored into W, we repeatedly select an arbitrary feature from W and if its addition into R increases the feature dependency then only it is inserted into R and removed from W, otherwise, the same process is done for next feature in W and continued it until all features in W are exhausted. This process gives the modified feature subset R.

This process of performing SFFS, followed by SFBR is repeated until the feature dependency of R on C is similar to that of F on C. Thus, following this process, once a feature is either selected or removed gets the opportunity to remove or select into the final feature subset R. The pseudo code of the proposed bidirectional floating forward selection and backward removal (BFFSBR) algorithm is described in Algorithm 1.

4. Multi-layer Deep Neural architecture for Gender Recognition

An audio is of two different types, (i) audio signal which is Amplitude v/s Time, and (ii) Spectrogram which is Frequency Content v/s Time. The amplitudes are not very informative, as they give only the loudness of the audio recording. The frequency domain provides the better understanding of the audio signal, which gives different frequencies present in the signal. In the work, we have used spectrogram form of audio signal to extract the useful features. We have fed the spectrogram image form of audio signal into our proposed multi-layer deep neural architecture based model of gender recognition. Recently, among various machine learning techniques, deep learning models have gained popularity for classification of objects. We have employed deep learning model by combining Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). The CNN is used for extracting the locally encoded important features to capture the non-linearity of the data and the RNN is used to provide a memory for capturing the long term dependencies. The simple RNN model mainly suffers from the vanishing gradient problem where the gradient becomes extremely low and the exploding gradient problem where the gradient becomes extremely high. To overcome these problems, we have used a variant of RNN, known as Gated Recurrent Unit network (GRUN).

4.1. Spectrogram image of audio signal

Before generation of the multi-layer deep neural architecture based gender recognition system from audio signal, preprocessing of audio signal is very important, which analyse the signal and convert it into spectrogram of $2 - D$ image. We have used Fourier Transform to convert a continuous audio signal from time-domain to frequency-domain. Fourier transform not only gives the frequencies but also magnitude of each frequency present in the signal. In a spectrogram representation of audio signal, one axis represents

Algorithm 1: BFFSBR(Bidirectional Floating Forward Selection and Backward Removal)

Input : DS = (S, F, C), where S is the set of speeches, F is the set of extracted features, and C is the set of gender types
Output: Feature subset R
begin
  $R = W = \phi$;
  $F' = F$;
  $n = |S|$ /* No. of speeches in S */ ;
  for *each features* $F_x \in F$ **do**
    Compute information gain $g(F_x)$ using eq.(3) to (5) ;
  end
  repeat
    /* Perform SFFS algorithm */
    Select F_x from F with maximum $g(F_x)$;
    while $(\gamma_{R \cup F_x}(C) \leq \gamma_R(C))$ **do**
      Select F_x from F with next maximum $g(F_x)$;
    end
    $F = F - \{F_x\}$;
    $R = R \cup \{F_x\}$ /*forward selection*/ ;
    /* Repeated backward removal*/
    for *each* F_x *in* R **do**
      if $(\gamma_{R - F_x}(C) > \gamma_R(C))$ **then**
        $R = R - \{F_x\}$;
        $F = F \cup \{F_x\}$
      end
    end
    /* Perform SFBR algorithm */
    $F = F - R$;
    Select F_x from F with minimum $g(F_x)$;
    while $(\gamma_{F - F_x}(C) \leq \gamma_F(C))$ **do**
      Select F_x from F with next minimum $g(F_x)$;
    end
    $W = W \cup \{F_x\}$;
    $F = F - \{F_x\}$ /*backward removal*/ ;
    /* Repeated forward selection */
    for *each* F_x *in* W **do**
      if $(\gamma_{F - F_x}(C) > \gamma_F(C))$ **then**
        $F = F \cup \{F_x\}$;
        $W = W - \{F_x\}$;
      end
    end
  until $(\gamma_R(C) == \gamma_{F'}(C))$;
  Return(R) ;
end

the time and the other axis represents frequencies, where the colors represent amplitude of the observed frequency at a particular time. To create the spectrogram, we break the audio signal into different frames of smaller sizes and perform Discrete Fourier Transform (DFT) on each frame to get its frequency. We consider all the frames of the signal in order, i.e., frame-1 first, then frame-2, and so on. So, frame number represents the time. We have considered the frames in overlapping way so that all the frequencies are captured. In spectrogram calculation, We have considered frame duration of 25 ms long as human can't generally speak more than one phoneme in this time frame and allowed an overlapping of 40% among two consecutive frames. As our signal is sampled at 16 KHz (average number of samples in one second), each frame is of amplitude $(16000 \times 25 \times 0.001) = 400$. As there is an overlapping of 40%, so a particular frame contains $(400 \times \frac{40}{100}) = 160$ amplitude from the next frame, i.e., there is an overlapping of 160 amplitude between every two consecutive pair of frames of the signal. As the frames are overlapping on each other, spectral resolution is important. To achieve the better resolution, we have used Hanning window, which is, in general, effective in 95% of cases. It has good frequency resolution and reduced spectral leakage. Next, Hanning window is multiplied with amplitudes and passes to the Fourier Transform function. The output of the Fourier Transform algorithm is a list of complex numbers of $size = \frac{400}{2} = 200$, which represents amplitudes of different frequencies within the frame. Thus, we get a list of 200 amplitudes of frequency bins which represent frequencies from 0 Hz — 8 KHz, as our sampling rate is 16K. The absolute values of those complex-valued amplitudes are calculated and normalized, for each frame. For each frame of the signal, we perform the same algorithm and finally, we get the spectrogram of the audio signal in the form of a $2-D$ matrix, where rows and columns represent frame number and frequency bin, respectively and the values in the cells of the matrix represent the strength of the frequencies. So, we can consider the signal, which is transformed to a spectrogram, as an image and easily apply it in our Deep Neural architecture for gender recognition.

4.2. Convolution Neural Network

The main objective for designing CNN is to use the concept of convolution, which generates filtered feature maps. The CNN models train on the basis of many layers where each audio signal input passes through a series of convolution layers along with filters (i.e., Kernals), and Pooling to extract the informative features of the audio signal. Deep neural network performs two different functions, namely feature engineering and classification [45], [46], [47], [48]. The feature engineering process [45], [46] automatically extracts useful and nonlinear features from the raw data using convolution and pooling layers by optimizing the weights between the layers. In classification [47], [48], the useful features are transformed into a vector and fed into a fully connected layer, followed by an activation function such as softmax function to classify the speech signals into different groups based on the gender type. The function of softmax is to transform the output with probabilistic values between 0 and 1. The CNN sequence for gender recognition is shown in Fig. 5 and the functionalities of the model are discussed below.

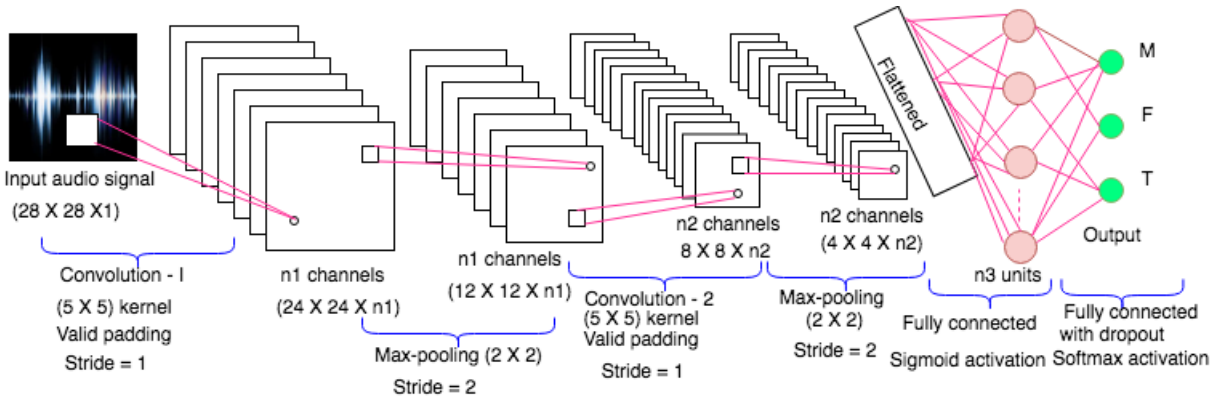


Figure 5: The CNN sequence for gender recognition (M=Male, F=Female, T=Transgender)

- **Convolution Layer:** This is the first layer where raw waveform of audio speech is fed. Convolution maintains the relationship among the extracted features of the signal with the help of small square windows of the input data. Thus, we feed spectrogram (i.e., the image form of audio signal) as input in convolution layer-I and a filter or kernel of size 5×5 undergoes throughout the input image to get the convolved feature matrix. The convolved feature is reduced in dimension as compared to that of the spectrogram by applying valid padding. The activation function used in this layer is the Rectified Linear Unit (ReLU), which invokes non-linearity of the data. This activation function is mostly used in CNN as it is known that the real world data needs the network to learn non-negative linear values.

- **Pooling Layer:** Similar to the Convolution Layer, the Pooling layer is responsible for reducing the dimension of the convolved features, which reduces the time complexity of the model to process the data. The main advantage of using this layer is that it extracts dominant features which are rotational and positional invariant, which helps to train the model effectively. The most frequently used pooling operations are max pooling and average pooling. Max pooling selects the maximum pixel value from the region of the image covered by the Kernel and the average pooling returns the average value of all the pixels in the region of the image covered by the kernel. There is no hard and fast rule about which pooling operation performs better, as it is both data and application dependent. The average pooling flattens the input image so the sharp and dark features cannot be identified properly. It simply performs dimension reduction as a noise suppressing mechanism. On the other hand, max pooling selects the bright pixels from the image, and thus suppress the noise. It ignores the noisy activation and performs de-noising along with the dimension reduction of the dataset. Hence, we can say that max pooling performs better than average pooling. The energy of the audio signal changes frequently with respect to pixel value, so sharp feature need to be identified, and that is why, max pooling mechanism is chosen for the proposed method.

- **CNN Layer:** The convolution layer and the pooling layer, together form one CNN layer. Depending on the complexities in the images, the number of CNN layers may be increased to capture low-level details of the image. But increase of number of CNN layers increases the computational complexity of the model, and so there must be a trade off between them. In the proposed method, we have considered two CNN layers, where each convolution layer uses a kernel of size 5×5 with valid padding, and each max pooling layer uses a kernel of size 2×2 with valid padding. The valid padding means that no padding is required and all the dimensions are valid so that the input image gets fully covered by the filter.
  After passing the audio signal through the two CNN layers, the model is successfully enabled to learn about the features.

4.3. Recurrent Neural Network

A recurrent neural network (RNN) is a kind of artificial neural network where connections between nodes form a directed graph along a temporal sequence. RNN can use their internal state to process variable length sequences of inputs, that is why we generally applying it in gender recognition from speech. RNNs have additional stored states, and the storage can be under direct control by the neural network. The storage can also be replaced by another network to incorporate time delays or feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of gated recurrent units (GRUs), which has the capability of reducing the vanishing gradient and exploding gradient problems, by which the Vanilla RNN [49] suffers. In our proposed work, the output of the CNN is flattened and subsequently fed into the Gated Recurrent Unit Network (GRUN) as input.
Let, in any current instant of time t, x_t be the input sequence and y_t be the output of GRU and s_{t-1} is the internal state of GRU at previous time instant. The equation (11) computes update gate output z_t at time instant t using the sigmoid function, $\sigma()$, where W_z and U_z are the weights. The update gate provides information from the previous time instant $(t-1)$ required for further processing.

$$z_t = \sigma(W_z x_t + U_z s_{t-1}) \tag{11}$$

Similarly, the reset gate output r_t at time instant t is computed using Equation (12), where W_r and U_r are the weights. Reset gate decides how much of the previous information need to be forgotten.

$$r_t = \sigma(W_r x_t + U_r s_{t-1}) \tag{12}$$

Finally, the output of the GRU is computed using Equation (13), (14), and (15), where h_t is the memory content and s_t is the internal state at time instant t. The operation $\odot$ performs the dot product of two vectors, and $\sigma()$ and tanh() are the sigmoid and tan-hyperbolic activation functions, respectively. The functionality of the GRU network is shown in Fig. 6.

$$h_t = \tanh(W_h x_t + U_h(s_{t-1} \odot r_t)) \tag{13}$$

$$s_t = (1 - z_t) \odot h_t + z_t \odot s_{t-1} \tag{14}$$

$$y_t = V s_t \tag{15}$$

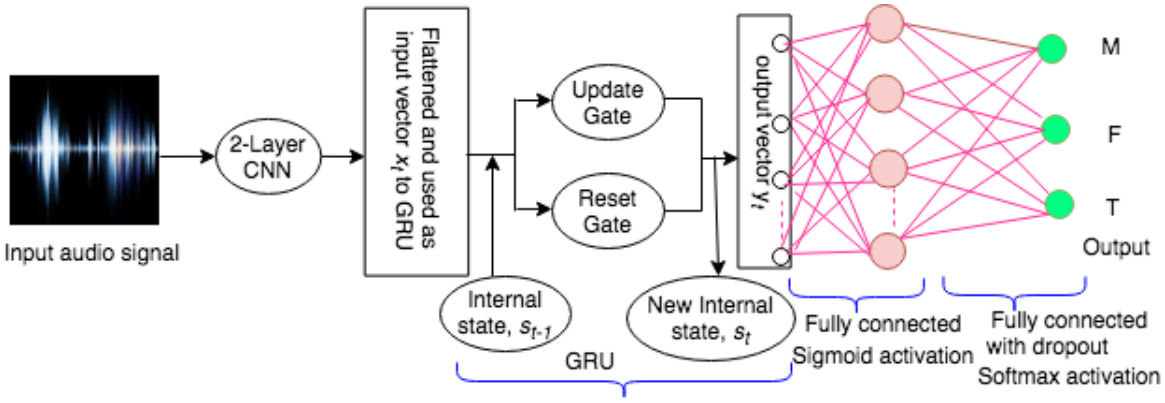


Figure 6: Gender Prediction using CNN and GRU Network

4.4. Gender Recognition

The proposed gender recognition system is developed in three different ways: (i) using CNN, (ii) using combination of CNN and GRUN (i.e., CNN + GRUN), and (iii) using combined feature selection algorithm, CNN and GRUN (i.e., BFFSBR + CNN + GRUN). After passing the audio signal through the two CNN layers, the CNN model is successfully enabled to learn about the features. The outputs of second CNN layer are flattened into the output vector y_t of CNN model. In case of $CNN + GRUN$ model, the output of CNN once flattened is fed into GRU network and it provides the output vector y_t, as shown in Fig. 6. In $BFFSBR + CNN + GRUN$ model, the output vector obtained from GRU is merged with the feature vector obtained using $BFFSBR$ feature selection algorithm, and the combined feature vector is used as output vector y_t of the model.

In all the three models, the output y_t is fed into a fully connected feed forward neural network that is having an intermediate layer consisting of 64 neurons with sigmoid activation function and a final dense layer consisting of 3 neurons (each one for a gender type) with softmax activation for classification purpose. After a sufficient number of epochs (we have used 500 epochs), the model is able to distinguish between dominating and certain low-level features and finally, pass them through a fully connected layer, where dropout probability of 0.2 and Adam optimizer are used, to classify them using the Softmax Classification technique. The dropout or regularization is used to remove the over fitting problem of the model. The output O_j^c of $j-$th neuron of a current dense layer is computed using Equation (16), where O_i^p is the output

of the i^{th} neuron of the previous layer, W_{ij} is the weight between the i^{th} neuron of the previous layer and j^{th} neuron of the current layer, and b_j^c is the bias attached with the j^{th} neuron of the current layer.

$$O_j^c = \sigma(\sum_i O_i^p W_{ij} + b_j^c) \tag{16}$$

Finally, the Softmax operation on the output layer is defined using Equation (17), where O_j^f is the output of the $j-$th neuron of the final layer.

$$softmax(O_j^f) = \frac{e^{O_j^f}}{\sum_{k=1}^{3} e^{O_k^f}} \tag{17}$$

Thus, the proposed gender recognition model is described in terms of a block diagram, as described in Fig. 7. The input audio signal is used to select some higher level human extracted features using rough set theory and information theory based feature selection algorithm (BFFSBR), which yields output feature vector F_3 for each audio file. The input audio signal is directly applied on CNN to extract the output feature vector F_1. In CNN based gender recognition system, F_1 is considered as output vector y_t and used for gender recognition. But in case of $CNN + GRUN$ based gender recognition system, output of CNN is applied on GRUN which provides output feature vector F_2. In this recognition system F_2 is used as output vector y_t. In $BFFSBR + CNN + GRUN$ based gender recognition system, output feature vector F_2 of GRUN and F_3 of BFFSBR are merged and considered as output vector y_t. Thus, we have developed three different gender recognition models, as shown in Fig. 7.

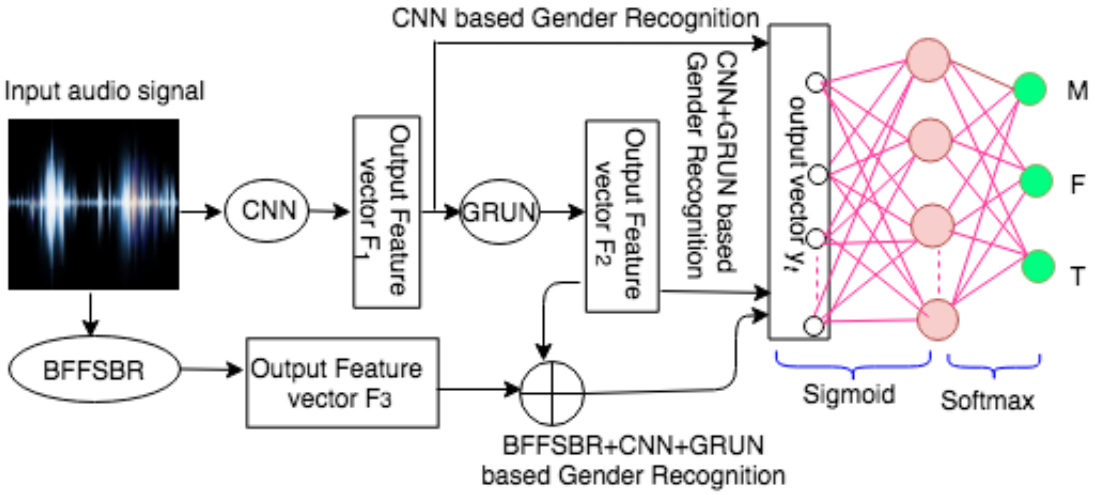


Figure 7: The proposed Gender Recognition System

5. Experimental Results and Discussions

The experiments are carried out on Google Colab virtual platform consists of Nvidia $K80/T4$ GPU,12 GB RAM & 358 GB of hard disk. The python language is used for implementation of the proposed methodologies. The Keras with Tensorflow in backend is used for Deep Learning implementation. We train the neural network for 500 epochs by backpropagation algorithm. During training, we employ dropout rate equals to 0.2. The extensive experiments are done for the evaluation of proposed gender recognition systems using six different datasets. The details of the datasets and performance comparisons are given in following subsections. Some experiments are done using WEKA tool [50] and other experiments including the proposed BFFSBR feature selection method are carried using python programming language.

5.1. DataSet Collection

In the proposed work, following six different speech datasets have been taken into account.

1. The simulated speech dataset, DS_1, is collected from different website from the internet and recorded with some of the speakers. The dataset contains 1500 audio files having 600 female, 500 males, and 400 transgender speeches. The duration of each file is 1 minute 30 seconds. We have collected the speeches of two different languages, Hindi and English. Out of 600 male speech files, 400 are of English language and remaining 200 are of Hindi language. Out of 500 female speech files, 350 are of English language and remaining 150 are of Hindi language, and for 400 transgender speech files, 250 are of English language and remaining 150 are of Hindi language. We have found 26 transgenders out of which 16 transgenders are chosen for providing English speeches and 10 transgenders provide Hindi speeches. Similarly, 50 male speakers and 50 female speakers are selected to record both the English and Hindi speeches.
2. Dataset, DS_2, is of Multi-lingual Indian Language dataset (generated by Indian Institute of Technology, Kharagpur), used in Reddy et al. [51]. The dataset had been collected by recording the speech through broadcast television channels using DISH-TV. It contains total 28 well known Indian languages [52]. For few languages, availability of TV broadcast channels were not present, where broadcast radio channels are utilized for grabbing the speech. This speech corpus contains news, interviews and live shows. There are 10 speakers taken into account for each language and the duration of each audio speech is about 5–10 minutes. We have labeled each language with a gender type based on the independent opinion of five different experts.
3. The dataset, DS_3, a bench mark dataset, is the collection of Ryerson Audio-Visual Dataset for Emotional Speech and Song [53]. The dataset contains 1440 audio files of eight type of emotion. The speech files of the database are taken by the voice of 24 professional actors where 12 are male and 12 are female with 60 trial per actor. North American accent has been used in the speech files.
4. The dataset, DS_4 consists of 90,000 raw audio waveforms collected from the audio domain of $VoxForge$[1] database. VoxForge is an open-source speech recognition corpus which consists of recorded samples, submitted by users using their own microphone. The dataset consists of six languages [52], each of 1500 samples.
5. The dataset, DS_5 is a benchmark Ryerson Audio-Visual Dataset for Emotional Speech and Song (RAVDESS) [53]. The dataset contains 8 different types of emotion. This Database of Speech acquires 1440 audio files of eight type of emotion. The speech files of the database are taken by the voice of 24 professional actors where 12 are male and 12 are female with 60 trial per actor. North American accent has been used in the speech files. The speech has been spoken with different emotions include calm, happy, sad, angry, fearful, surprise, disgust and neutral. All these expressions have been performed with two levels of energy, normal and strong. We have used this dataset to verify that our gender recognition model performs well for such versatile data too.
6. The dataset, DS_6 is the VocalSet [54] dataset which contains recordings from 20 different singers (11 are male and 9 are female) performing a variety of vocal techniques. The dataset has recordings of 10.1 hours of professional singers performing different types vocal techniques. It consists of diverse set of voices using different vocal techniques sung on the basis of different scales. The dataset diversify our range and variety of songs. Thus the proposed gender recognition model is also applied on singing speech data.

5.2. Evaluation of proposed BFFSBR feature selection method

The proposed RST and Information theory based feature selection algorithm (BFFSBR) has been compared with some recently published feature selection algorithms using all six dtasets. The comparison is done based on number of features selected and the accuracy of the classifiers used. The feature selection algorithms used are (i) Rough-spanning tree based feature selection algorithm (RMST) [43], (ii) Classification of vocal

[1]Free speech recognition (linux, windows and mac), `http://www.voxforge.org/`, accessed on 16 Jul 2019

and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS) [55] (iii) Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm (RFSA) [56], (iv) Acoustic feature selection for automatic emotion recognition from speech (AFSS) [57], (v) Exploring boundary region of rough set theory for feature selection (RSFS) [52], and (vi) Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm (SFGA) [58]. To measure the accuracy of the classifiers based on reduced feature set, we have considered eight different classifiers, namely Support vector machine (SVM), K-nearest neighbors (KNN), Decision tree (DT), Neural network (NN), Random forest (RF), Naïve Bayes (NB), Adaboost (BST), and Sequential minimal optimization (SMO). SVM is used with RBF kernal, K value of KNN is set to the square root of sample size of data and 10-fold cross validation is used to measure the performance of the classifiers. The classification accuracy is measured using WEKA tool [50] and the results are listed in Table 1 for all the datasets. The best result in each dataset obtained are marked by bold face font in Table. From the table, it is observed that, in all cases, the proposed BFFSBR feature selection method selects the minimum number of features, where $RMST$ and $RFSA$ also selects the minimum number of features in case of DS_2 and DS_3, respectively. But only minimum number of feature selection is not the criteria of a good feature selection algorithm. Thus we have measured the accuracy of eight different classifiers based on the reduced datasets obtained by the feature selection algorithms. In case of DS_1, $RSFS$ algorithm performs better than the proposed method with respect to the accuracy of decision tree (DT) classifier only. Similarly, for DS_2, $RSFS$ algorithm performs better than the proposed method with respect to the accuracy of SVM and NB classifiers. For DS_5, the proposed method performs better than all other feature selection algorithms with respect to the accuracy of all eight classifiers, and for all other datasets, the proposed method is dominated by one or two feature selection algorithms with respect to one or two classification accuracies. Thus the proposed method is superior than the other methods in terms of both number of features selected and accuracy of the classifiers, which shows the efficacy of the method. Wlicoxon's rank sum test [59], a non-parametric statistical test, is performed for independent samples with p-value of 0.05 (or a significance label of 5%) to evaluate whether the results obtained by the proposed BFFSBR algorithm differs from the other feature selection algorithms in a statistically significant manner. The test confirms that the probability that the proposed algorithm is statistically and significantly different from other algorithms is at least 0.95, as the performance(i.e., accuracy) of the classifiers for the reduced datasets obtained by the proposed algorithm is differing from the best result with a p value equal or less than 0.05. If the p value is greater than 0.05 between the best algorithm and the other algorithm then a $'\ddagger'$ symbol is used for the second one to indicate that the difference is statistically significant, otherwise the two performances are considered equivalent, i.e., the difference between the corresponding algorithms is not statistically significant and we use a $'\approx'$ symbol, as shown in Table 1. Thus, the experimental results show that the proposed BFFSBR algorithm is comparatively better than other feature selection algorithms.

5.3. Evaluation of proposed Gender recognition system

The proposed gender recognition system is evaluated using 10-fold cross validation method by computing some performance validation indices, such as Accuracy (A), Precision (P), Recall (R), and F-measure (F), which are defined in Equation (18) to (21), respectively, where TP, TN, FP and FN are known as true positive, true negative, false positive and false negative, respectively.

$$A = \frac{|TP| + |TN|}{|TP| + |FP| + |FP| + |FN|} \tag{18}$$

$$P = \frac{|TP|}{|TP| + |FP|} \tag{19}$$

$$R = \frac{|TP|}{|TP| + |FN|} \tag{20}$$

$$F = \frac{2.P.R}{P + R} \tag{21}$$

True positive (TP) is the set of objects of a dataset which are actually of positive class and the classifier also predicts them as positive class, True negative (TN) is the set of objects of the dataset which are actually of negative class and the classifier also predicts them as negative class. False positive (FP) is the set of objects of the dataset which are actually of negative class but the classifier predicts them wrongly as positive class and False negative (FN) is the set of objects of the dataset which are actually of positive class but the classifier predicts them wrongly as negative class. First, we have computed these performance metrics of our three proposed gender recognition models, namely CNN, $CNN+GRUN$, and $BFFSBR+CNN+GRUN$ using all six datasets. Out of all the datasets, only DS_1 contains three different classes, Male (M), Female (F), and Transgender (T), where all other datasets contain binary class, M and F. For these three models the performance metrics are computed as follows:

- All the datasets, except DS_1, have two class labels, M and F. To compute the values of TP, TN, FP, and FN, first M is considered as positive class and F is considered as negative class and computed all these four values. Based on these values, using Equation (18) to (21) four performance metrics are calculated. Next, F is considered as positive class and M is considered as negative class and computed all these four values, and similarly, the metrics are computed. Finally, the average values are considered as the accuracy (A), Precision (P), Recall (R), and F-Measure (F) of the model for the dataset.

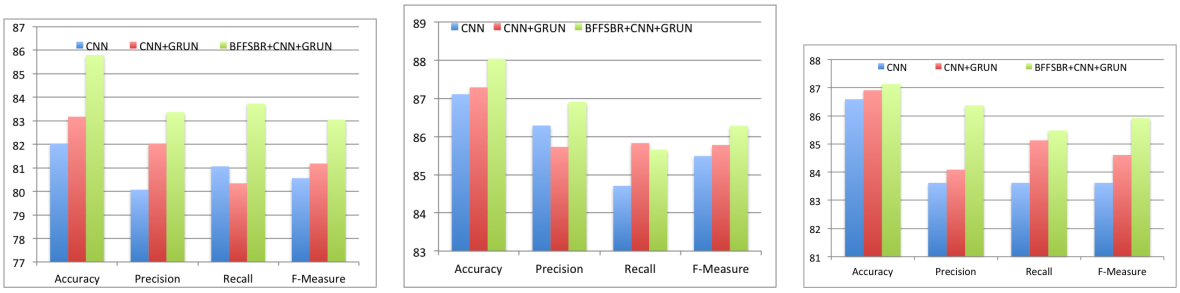
- Dataset DS_1 contains three class labels, namely M, F, and T. In this case, once M is considered as positive class and rest two as negative class and the values of TP, TN, FP, and FN are computed and accordingly, performance metrics, A, P, R, and F are calculated. Similarly, considering F as positive class and rest as negative class and T as positive class and rest as negative class, the metrics values are computed. Finally, the average of all three values provides the resultant accuracy (A), Precision (P), Recall (R), and F-Measure (F) of the model for the dataset DS_1.

The computed performance metrics of our proposed three gender recognition models for all six datasets are listed in Table 2. From this table, it is observed that, both the accuracy and F-Measure of the proposed $BFFSBR+CNN+GRUN$ model are the highest for all six datasets. In some cases, like for the datasets, DS_2 and DS_4, the $CNN+GRUN$ model also provides good performance. For DS_2, it gives the best Recall value and for DS_4, it provides the best Accuracy and Recall values. Though the proposed $BFFSBR+CNN+GRUN$ performs better than the other two proposed models, but the values of the performance metrics are quiet closed to each other, so statistical test is done using Wlicoxon's rank sum test [59] by considering both accuracy (A) and F-Measure (F) separately. Similar to Table 1, two symbols $'\ddagger'$ and $'\approx'$ are used to imply that the two models (i.e., CNN, and $CNN+GRUN$) are different or equivalent to $BFFSBR+CNN+GRUN$ model, respectively. From the observed symbols, we can say that the model $BFFSBR+CNN+GRUN$ is superior and statistically and significantly different from the other two proposed models. This result demonstrates that combining some human extracted features of audio signals with deep neural networks, it is possible to develop more accurate and effective gender recognition system.

The performance metrics of three different proposed gender recognition systems are also visualized using Figure 9 for all six datasets. It is observed that, accuracy, precision, and $F-$Measure of the proposed hybrid model (i.e., $BFFSBR+CNN+GRUN$) is the highest for all datasets, only recall of $CNN+GRUN$ is slightly higher than the model $BFFSBR+CNN+GRUN$ for DS_2. Thus, we consider the deep hybrid model with rough set theory and information theory (i.e., $BFFSBR+CNN+GRUN$) as the best proposed model for gender recognition.

5.4. Comparison of proposed model with other related models

Based on the results given in Table 1 and Table 2, it is observed that the proposed three deep neural network based models work better than the traditional classifiers used in Table 1, simply because these classifiers are applied on the datasets described by the human extracted higher label features. The proposed models used the concepts of deep neural networks where machine itself extracts the features from the raw speech signals. Among the three proposed models, the model $BFFSBR+CNN+GRUN$ which combines both human extracted and machine extracted features performs better than the other two models,



(a) Performance metrics for DS_1 (b) Performance metrics for DS_2 (c) Performance metrics for DS_3

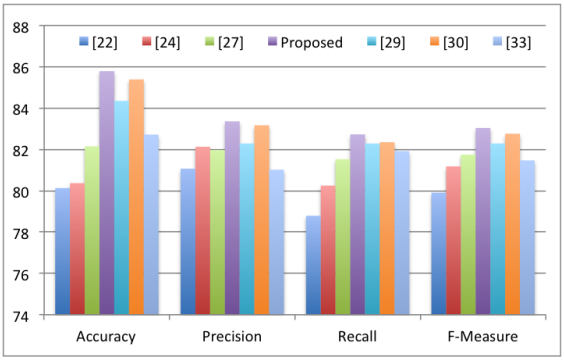
(d) Performance metrics for DS_4 (e) Performance metrics for DS_5 (f) Performance metrics for DS_6

Figure 8: Performance comparison of proposed gender recognition systems

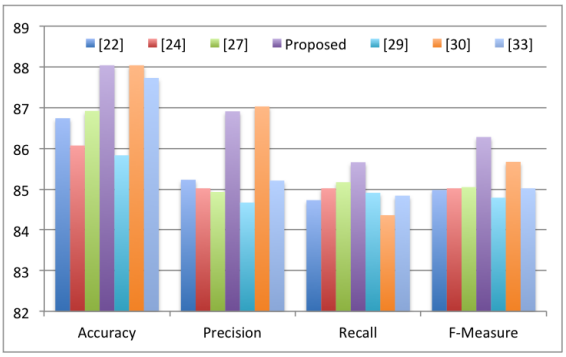
where only machine extracted features have been used for learning. So, we have considered the model $BFFSBR + CNN + GRUN$ as the final proposed gender recognition system in this work. To evaluate its effectiveness compare to other related state-of the art models, we have considered few deep neural network based gender recognition systems, such as Levi et al. [22], Kabil et al. [24], Rajeev et al. [27], Wang et al. [29], Markitantov et al. [30], and Ertam et al. [33], discussed in subsection 1.2. Here also, 10-fold cross validation and wilcoxon rank sum test are carried out for comparison purpose and the experimental results are listed in Table 3. For better visualization, the lists of values are represented by bar chart as shown in Figure 9. From the table and figure it is observed that, the proposed $BFFSBR + CNN + GRUN$ model outperforms others in terms of accuracy and F-Measure in all datasets except dataset DS_3 where [30] shows the best F-Measure. The proposed method provides the highest precision values for all datasets except DS_2, and highest recall values for all datasets except DS_3, and DS_4. Considering all four performance metrics, it is observed that the next two best models followed the proposed model are Markitantov et al.[30] and Ertam et al. [33], which is also true in terms of wilcoxon Ranksum test. It is also observed that, the proposed method provides the highest values of all four performance measure metrics in case of dataset DS_1, which contains the transgender speeches. As earlier mentioned that this dataset may contain more imprecise and uncertain information, so the proposed rough set and information theory based hybrid deep neural network model performs perfectly for gender recognition, which is the main objective of this paper.

The models are also Compared by analysing Receiver Operating Characteristic (ROC) curves generated for all datasets. It also helps us to measure the performance of the models by computing Area Under the Curve (AUC). More the AUC value implies better the model is and vice versa. To construct the ROC curve, we need to compute the True Positive Rate (TPR) and False Positive Rate (FPR) from the equation (22) and (23), respectively. The FPR is considered along X-axis and TPR is considered along the Y-axis, as shown in Fig. 10.

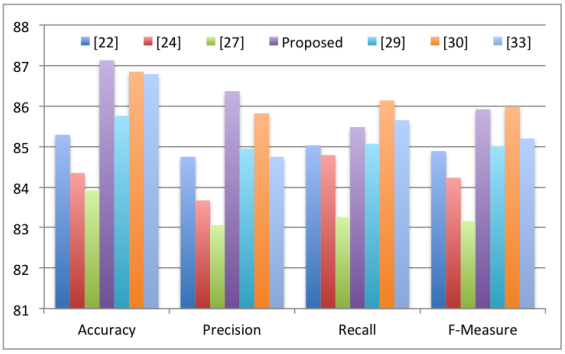
$$TPR = \frac{|TP|}{|FN| + |TP|} \tag{22}$$



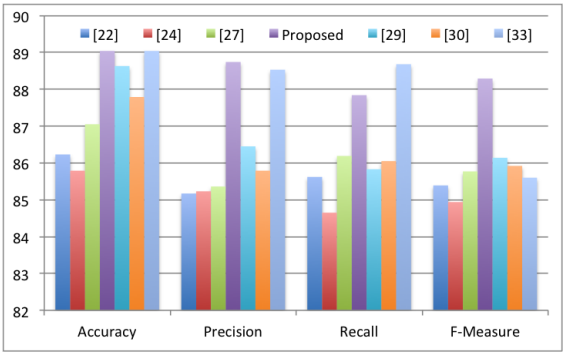
(a) Performance metrics for DS_1



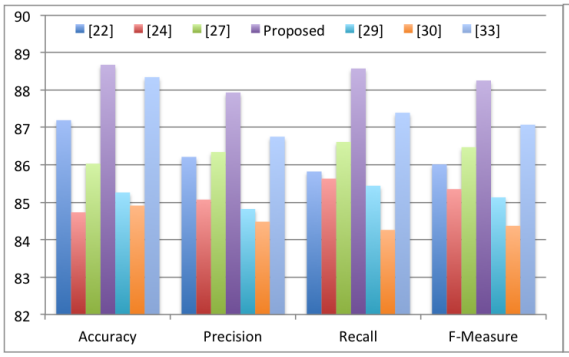
(b) Performance metrics for DS_2



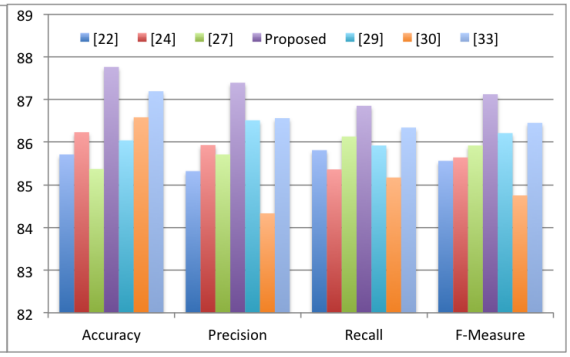
(c) Performance metrics for DS_3



(d) Performance metrics for DS_4



(e) Performance metrics for DS_5



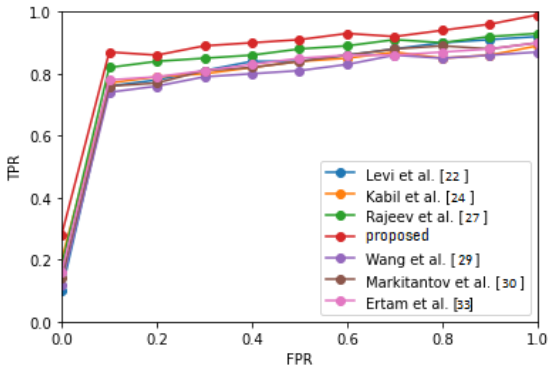
(f) Performance metrics for DS_6

Figure 9: Performance comparison of proposed model with related models

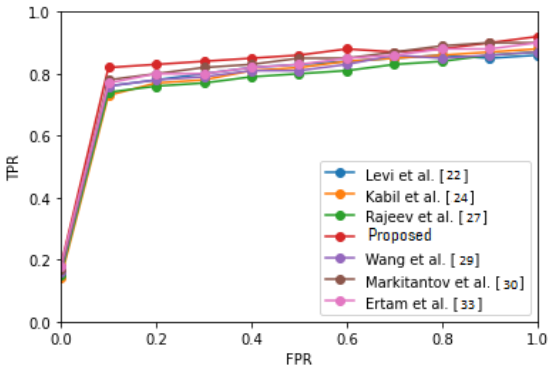
$$FPR = \frac{|FP|}{|TN| + |FP|} \tag{23}$$

From the equation (22) and (23), it is clear that both the TPR and FPR values ranges in between 0 and 1 and so the ROC curve lies within a square box of unit area. The ROC curve for the ideal or standard model should be the curve joining the points (0,0) to (0,1) and (0,1) to (1,1), which gives the AUC value as 1 square unit. The 45^o diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding to random chance. Thus the curve lies below the diagonal line is considered as the worse model and curve above the diagonal line is considered as the better model. It implies that larger the AUC value of a model implies better the model is. We have used all seven gender recognition models for all six datasets and the ROC curves are

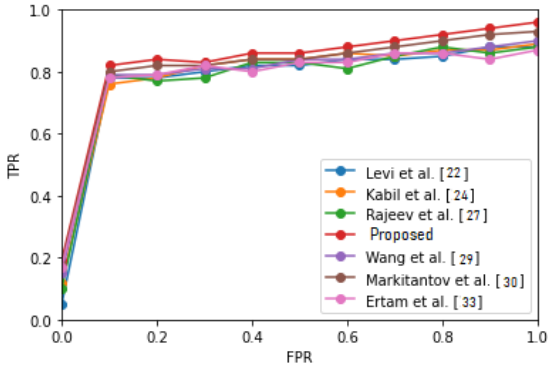
drawn as shown in Fig. 10. The Figure gives the ROC curves of all seven models for dataset DS_1 to DS_6, respectively. From the figure, it is observed that the curve obtained for proposed BFFSBR+CNN+GRUN model gives better result than other models with respect to the AUC value.



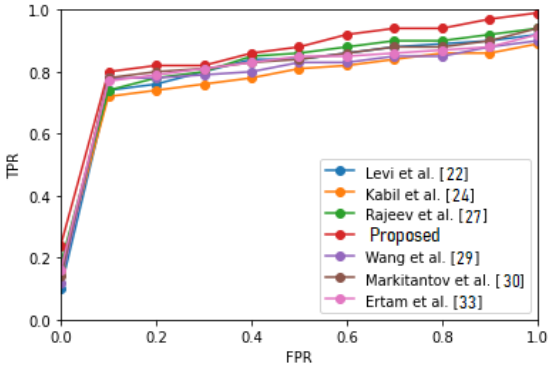
(a) ROC plotting for DS_1 dataset



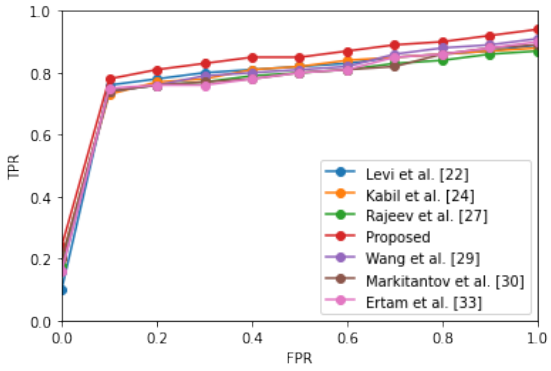
(b) ROC plotting for DS_2 dataset



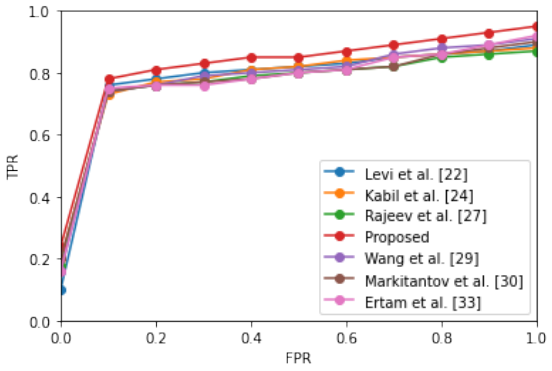
(c) ROC plotting for DS_3 dataset



(d) ROC plotting for DS_4 dataset



(e) ROC plotting for DS_5 dataset



(f) ROC plotting for DS_6 dataset

Figure 10: ROC curve based comparison of different gender recognition systems

6. Conclusion

The proposed gender recognition system is developed based on multi-view feature selection concepts. The human extracted features are evaluated using Rough set and information theory to select only the informative, precise and unambiguous features by removing uncertainty and ambiguity from the dataset. On the other hand, machine extracted features are generated using CNN and GRUN based deep neural networks. Finally, the features are combined and applied in the gender recognition system. Thus we extract features in different forms which are complementary to each other. The classification model is learned using this multi-view dataset to make full use of the hidden information. The method is applied for six different kinds of audio speech based datasets and obtained very promising results for gender recognition. The dataset DS_1 is the sampled dataset generated by us by collecting audio speeches of three different genders, where transgender is also considered. It is observed from the experimental results that, for DS_1, the proposed methods works more effectively, from which we may conclude that RST based human extracted features take important role in gender recognition. Generally, voice of transgender is different to distinguish from male or female voice accurately, and this uncertainty is tackled by the RST. We may apply different architectures of the deep neural networks together with fuzzy set theory and rough set theory for the same purpose, which is the future scope of this paper.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The propound work is seemingly supported by IIT kharagpur (IITKGP) for their contribution to provide the multilingual Indian speech data (DS_2). We convey our gratitude for contributing to our proposed research.

Compliance with ethical standards

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.
Funding details: No funding was received to assist with the preparation of this manuscript. **Conflict of Interest:** The authors declare that this manuscript has no conflict of interest with any other published source and has not been published previously (partly or in full). No data have been fabricated or manipulated to support our conclusions.
Informed Consent: The authors have no relevant financial or non-financial interests to disclose.

Authorship Contribution

Ghazaala Yasmin: Data curation, Methodology, Writing Original Draft
Asit Kumar Das: Conceptualization, Supervision
Janmenjoy Nayak: Suggestion and Editing
S. Vimal: Investigation and Validation
Soumi Dutta: Reviewing and Editing

References

[1] M. A. K. Halliday, J. J. Webster, Text linguistics: The how and why of meaning, Equinox Publishing Ltd., 2014.
[2] S. J. Arora, R. P. Singh, Automatic speech recognition: a review, International Journal of Computer Applications 60 (2012).
[3] H. K. Palo, M. N. Mohanty, M. Chandra, Emotion analysis from speech of different age groups., in: RICE, 2017, pp. 283–287.
[4] R. S. Sudhakar, M. C. Anil, Analysis of speech features for emotion detection: a review, in: 2015 International Conference on Computing Communication Control and Automation, IEEE, 2015, pp. 661–664.
[5] H. Erokyar, Age and gender recognition for speech applications based on support vector machines (2014).

[6] M. Yusnita, A. Hafiz, M. N. Fadzilah, A. Z. Zulhanip, M. Idris, Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal, in: 2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), IEEE, 2017, pp. 372–377.
[7] B. Jena, A. Mohanty, S. K. Mohanty, Gender recognition of speech signal using knn and svm, Available at SSRN 3769786 (2021).
[8] V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer, Efficient processing of deep neural networks: A tutorial and survey, Proceedings of the IEEE 105 (2017) 2295–2329.
[9] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F. E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.
[10] A. Khan, A. Sohail, U. Zahoora, A. S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, Artificial Intelligence Review 53 (2020) 5455–5516.
[11] R. Dey, F. M. Salem, Gate-variants of gated recurrent unit (gru) neural networks, in: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), IEEE, 2017, pp. 1597–1600.
[12] C. B. Ng, Y. H. Tay, B. M. Goi, Vision-based human gender recognition: A survey, arXiv preprint arXiv:1204.1611 (2012).
[13] R. R Mahajan, A. Ahuja, U. Mandawkar, A survey on automatic gender recognition using machine learning, IJRAR-International Journal of Research and Analytical Reviews (IJRAR), E-ISSN (2020) 2348–1269.
[14] C. B. Ng, Y. H. Tay, B.-M. Goi, Recognizing human gender in computer vision: a survey, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 2012, pp. 335–346.
[15] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, A. Sciarrone, Gender-driven emotion recognition through speech signals for ambient intelligence applications, IEEE transactions on Emerging topics in computing 1 (2013) 244–257.
[16] Y.-M. Zeng, Z.-Y. Wu, T. Falk, W.-Y. Chan, Robust gmm based gender classification using pitch and rasta-plp parameters of speech, in: 2006 International Conference on Machine Learning and Cybernetics, IEEE, 2006, pp. 3376–3379.
[17] M. Li, K. J. Han, S. Narayanan, Automatic speaker age and gender recognition using acoustic and prosodic level information fusion, Computer Speech & Language 27 (2013) 151–167.
[18] M. Li, C.-S. Jung, K. J. Han, Combining five acoustic level modeling methods for automatic speaker age and gender recognition, in: Eleventh Annual Conference of the International Speech Communication Association, 2010.
[19] G. Yasmin, O. Mullick, A. Ghosal, A. K. Das, Gender recognition inclusive with transgender from speech classification, in: Emerging Technologies in Data Mining and Information Security, Springer, 2019, pp. 89–98.
[20] J. Ahmad, M. Fiaz, S.-i. Kwon, M. Sodanil, B. Vo, S. W. Baik, Gender identification using mfcc for telephone applications-a comparative study, arXiv preprint arXiv:1601.01577 (2016).
[21] H. Harb, L. Chen, Gender identification using a general audio classifier, in: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), volume 2, IEEE, 2003, pp. II–733.
[22] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 34–42.
[23] R. S. Alkhawaldeh, Dgr: Gender recognition of human speech using one-dimensional conventional neural network, Scientific Programming 2019 (2019).
[24] S. H. Kabil, H. Muckenhirn, M. Magimai-Doss, On learning to identify genders from raw speech signal using cnns., in: INTERSPEECH, 2018, pp. 287–291.
[25] J. Mansanet, A. Albiol, R. Paredes, Local deep neural networks for gender recognition, Pattern Recognition Letters 70 (2016) 80–86.
[26] A. Dehghan, E. G. Ortiz, G. Shu, S. Z. Masood, Dager: Deep age, gender and emotion recognition using convolutional neural network, arXiv preprint arXiv:1702.04280 (2017).
[27] R. Ranjan, V. M. Patel, R. Chellappa, Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE transactions on pattern analysis and machine intelligence 41 (2017) 121–135.
[28] J. van de Wolfshaar, M. F. Karaaba, M. A. Wiering, Deep convolutional neural networks and support vector machines for gender recognition, in: 2015 IEEE Symposium Series on Computational Intelligence, IEEE, 2015, pp. 188–195.
[29] Z.-Q. Wang, I. Tashev, Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 5150–5154.
[30] M. Markitantov, O. Verkholyak, Automatic recognition of speaker age and gender based on deep neural networks, in: International Conference on Speech and Computer, Springer, 2019, pp. 327–336.
[31] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, M. Rosa-Zurera, Age and gender recognition from speech using deep neural networks, in: Workshop of Physical Agents, Springer, 2020, pp. 332–344.
[32] P. Gupta, S. Goel, A. Purwar, A stacked technique for gender recognition through voice, in: 2018 Eleventh International Conference on Contemporary Computing (IC3), IEEE, 2018, pp. 1–3.
[33] F. Ertam, An effective gender recognition approach using voice data via deeper lstm networks, Applied Acoustics 156 (2019) 351–358.
[34] P. Boersma, Praat, a system for doing phonetics by computer, Glot. Int. 5 (2001) 341–345.
[35] M. P. Gelfer, V. A. Mikos, The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels, Journal of Voice 19 (2005) 544–554.
[36] Y. Hu, D. Wu, A. Nucci, Pitch-based gender identification with two-stage classification, Security and Communication Networks 5 (2012) 211–225.
[37] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 835–838.

[38] J. Walters-Williams, Y. Li, Estimation of mutual information: A survey, in: International Conference on Rough Sets and Knowledge Technology, Springer, 2009, pp. 389–396.
[39] Q. Zhang, Q. Xie, G. Wang, A survey on rough set theory and its applications, CAAI Transactions on Intelligence Technology 1 (2016) 323–333.
[40] A. K. Das, S. Chakrabarty, S. Sengupta, Formation of a compact reduct set based on discernibility relation and attribute dependency of rough set theory, in: International Conference on Information Processing, Springer, 2012, pp. 253–261.
[41] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern recognition letters 15 (1994) 1119–1125.
[42] F. E. Tay, L. Shen, A modified chi2 algorithm for discretization, IEEE Transactions on knowledge and data engineering 14 (2002) 666–670.
[43] P. Das, A. K. Das, J. Nayak, Feature selection generating directed rough-spanning tree for crime pattern analysis, Neural Computing and Applications 32 (2020) 7623–7639.
[44] G. J. Woeginger, Exact algorithms for np-hard problems: A survey, in: Combinatorial optimization—eureka, you shrink!, Springer, 2003, pp. 185–207.
[45] D. Robinson, Z. Zhang, J. Tepper, Hate speech detection on twitter: Feature engineering vs feature selection, in: European Semantic Web Conference, Springer, 2018, pp. 46–49.
[46] Y. Qian, N. Chen, H. Dinkel, Z. Wu, Deep feature engineering for noise robust spoofing detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (2017) 1942–1955.
[47] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, D. S. Turaga, Learning feature engineering for classification., in: Ijcai, 2017, pp. 2529–2535.
[48] U. Khurana, D. Turaga, H. Samulowitz, S. Parthasrathy, Cognito: Automated feature engineering for supervised learning, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), IEEE, 2016, pp. 1304–1307.
[49] T. Stérin, N. Farrugia, V. Gripon, An intrinsic difference between vanilla rnns and gru models, COGNTIVE 2017 (2017) 84.
[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (2009) 10–18.
[51] V. R. Reddy, S. Maity, K. S. Rao, Identification of indian languages using multi-level spectral and prosodic features, International Journal of Speech Technology 16 (2013) 489–511.
[52] G. Yasmin, A. K. Das, J. Nayak, D. Pelusi, W. Ding, Graph based feature selection investigating boundary region of rough set for language identification, Expert Systems with Applications 158 (2020) 113575.
[53] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (2018) e0196391.
[54] J. Wilkins, P. Seetharaman, A. Wahl, B. Pardo, Vocalset: A singing voice dataset., in: ISMIR, 2018, pp. 468–474.
[55] Y. S. Murthy, S. G. Koolagudi, Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (gafs), Expert Systems with Applications 106 (2018) 77–91.
[56] A. K. Das, S. K. Pati, A. Ghosh, Relevant feature selection and ensemble classifier design using bi-objective genetic algorithm, Knowledge and Information Systems 62 (2020) 423–455.
[57] J. Rong, G. Li, Y.-P. P. Chen, Acoustic feature selection for automatic emotion recognition from speech, Information processing & management 45 (2009) 315–328.
[58] M. Sidorov, C. Brester, W. Minker, E. Semenkin, Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm., in: LREC, 2014, pp. 3481–3485.
[59] C. Wild, G. Seber, The wilcoxon rank-sum test, 2011.

Dataset	Feature Selection Method(#Features)	Classifiers							
		SVM	KNN	DT	NN	RF	NB	BST	SMO
DS_1	RMST (31)	79.06‡	73.37‡	77.29‡	**80.17**	79.83≈	74.72‡	73.37‡	75.17‡
	GAFS (33)	73.23‡	71.17‡	74.81‡	79.97≈	78.36‡	76.93‡	79.99≈	72.57‡
	RFSA (33)	71.47‡	73.39‡	71.28‡	77.39‡	74.83‡	76.24‡	72.64‡	79.53≈
	BFFSBR (27)	**81.24**	**82.79**	78.24‡	**80.17**	**80.03**	**80.95**	**80.49**	**79.79**
	AFSS (29)	73.29‡	73.39‡	74.74‡	71.24‡	75.79‡	78.39‡	76.05‡	77.24‡
	RSFS (29)	79.13‡	80.24‡	**83.06**	78.94‡	79.77≈	**80.95**	80.02≈	78.01‡
	SFGA (35)	78.13‡	78.12‡	76.76‡	79.86≈	77.14‡	80.63≈	78.09‡	73.71‡
DS_2	RMST **(29)**	74.11‡	71.23‡	74.01‡	73.36‡	76.93‡	75.38‡	74.19‡	77.49‡
	GAFS (34)	72.01‡	73.55‡	74.39‡	80.13≈	74.14‡	75.40‡	72.20‡	74.19‡
	RFSA (33)	74.85‡	71.90‡	74.84‡	72.35‡	72.37‡	73.49‡	72.58‡	73.49‡
	BFFSBR (29)	77.19‡	**78.57**	**81.83**	**80.47**	**81.70**	80.23≈	**81.15**	78.19‡
	AFSS (35)	77.54‡	76.92‡	75.37‡	73.76‡	81.37≈	79.29‡	78.74‡	77.43‡
	RSFS (33)	**78.93**‡	78.24≈	81.06	80.04≈	79.97‡	**80.36**	78.53‡	79.49≈
	SFGA (31)	75.76‡	73.79‡	77.28‡	75.12‡	74.19‡	78.84‡	80.93≈	**79.97**
DS_3	RMST (32)	74.03‡	76.73≈	**80.33**	75.08‡	76.47‡	73.01‡	77.18	76.45‡
	GAFS (35)	74.09‡	73.34‡	78.05‡	77.95‡	73.26‡	72.91‡	78.04‡	77.15‡
	RFSA **(30)**	75.28‡	76.23‡	78.79‡	76.29‡	74.04‡	76.08‡	78.28‡	73.79‡
	BFFSBR (30)	**78.25**	**80.06**	79.86≈	**78.61**	**82.02**	**81.40**	**80.10**	**82.10**
	AFSS (34)	76.69‡	79.67≈	74.39‡	77.23‡	74.34‡	76.05‡	75.39‡	76.43‡
	RSFS (31)	75.13‡	74.24‡	77.06‡	73.94‡	78.73‡	78.17‡	80.02≈	**82.10**
	SFGA (34)	74.93‡	76.25‡	76.72‡	77.34‡	75.69‡	76.23‡	74.38‡	81.79≈
DS_4	RMST (35)	74.01‡	79.73≈	75.03‡	74.98‡	**77.44**	73.01‡	76.10	76.45‡
	GAFS (34)	**78.75**	75.34‡	76.52‡	73.03‡	77.26≈	72.19‡	73.04‡	72.15‡
	RFSA (36)	74.28‡	72.23‡	79.31≈	78.39≈	74.07‡	76.08‡	73.28‡	74.29‡
	BFFSBR (30)	**78.75**	**80.06**	**79.62**	**78.61**	77.02≈	**81.40**	**81.19**	78.11≈
	AFSS (32)	73.69‡	72.39‡	73.37‡	71.93‡	72.84‡	73.75‡	74.89‡	**78.43**
	RSFS (34)	78.37≈	76.24‡	73.06‡	77.14‡	75.73‡	73.96‡	77.52‡	76.39‡
	SFGA (35)	72.43‡	74.75‡	76.02‡	78.27≈	74.39‡	73.73‡	74.68‡	73.95‡
DS_5	RMST (41)	82.17‡	81.89‡	77.13‡	75.69‡	77.15‡	81.18‡	80.25‡	77.64‡
	GAFS (39)	79.37	76.23‡	78.27‡	75.93‡	77.81‡	76.33‡	75.74‡	77.24‡
	RFSA (40)	77.32‡	74.37‡	81.61‡	81.97‡	77.87‡	79.83‡	76.98‡	78.49‡
	BFFSBR (32)	**84.87**	**85.37**	**83.76**	**84.16**	**85.33**	**86.34**	**86.79**	**84.31**
	AFSS (37)	76.36‡	77.19‡	75.77‡	75.47‡	76.38‡	77.67‡	79.09‡	83.97≈
	RSFS (38)	81.07‡	82.52‡	76.76‡	78.34‡	77.97‡	75.75‡	78.42‡	78.36‡
	SFGA (36)	75.93‡	76.85‡	78.09‡	80.32‡	77.19‡	76.77‡	77.69‡	76.91‡
DS_6	RMST (40)	78.21‡	**83.76**	79.01‡	80.29‡	81.74‡	82.39‡	80.11‡	81.29‡
	GAFS (39)	79.04	78.57‡	79.78‡	79.13‡	80.23‡	81.04‡	79.27‡	80.74‡
	RFSA (38)	77.83‡	77.09‡	**82.61**	80.76‡	79.92‡	80.61‡	78.54‡	79.56‡
	BFFSBR (36)	**81.36**	**83.76**	82.36≈	**82.06**	**82.87**	**83.56**	81.63≈	**84.75**
	AFSS (41)	78.53‡	79.92‡	77.18‡	76.35‡	80.73‡	78.75‡	80.26‡	82.61‡
	RSFS (40)	81.07≈	83.52≈	79.97‡	77.81‡	81.02‡	82.52‡	**81.79**	80.53‡
	SFGA (42)	79.74‡	78.91‡	80.37‡	81.78≈	80.36‡	81.78‡	79.66‡	83.77‡

Table 1: Evaluation of proposed feature selection method based on number of features selected and classification accuracy

Dataset	Proposed Model	Performance Metrics			
		A	P	R	F
DS_1	CNN	82.03‡	80.07	81.06	80.56 ‡
	CNN+GRUN	83.17‡	82.03	80.34	81.18‡
	BFFSBR+CNN+GRUN	**85.79**	**83.36**	**83.73**	**83.04**
DS_2	CNN	87.11‡	86.29	84.71	85.49 ‡
	CNN+GRUN	87.29‡	85.73	**85.83**	85.78≈
	BFFSBR+CNN+GRUN	**88.04**	**86.91**	85.66	**86.28**
DS_3	CNN	86.59‡	83.62	83.62	83.62 ‡
	CNN+GRUN	86.91≈	84.09	85.13	84.61‡
	BFFSBR+CNN+GRUN	**87.13**	**86.37**	**85.48**	**85.92**
DS_4	CNN	87.92‡	85.43	86.63	86.03 ‡
	CNN+GRUN	**89.04**	88.01	**87.92**	87.96≈
	BFFSBR+CNN+GRUN	**89.04**	**88.74**	87.84	**88.29**
DS_5	CNN	86.29‡	86.01	86.54	86.27 ‡
	CNN+GRUN	88.73≈	86.27	86.61	86.44‡
	BFFSBR+CNN+GRUN	**88.94**	**87.41**	**86.85**	**87.13**
DS_6	CNN	86.42‡	86.63	85.91	86.27 ‡
	CNN+GRUN	88.03‡	87.53	87.34	87.43‡
	BFFSBR+CNN+GRUN	**90.36**	**88.51**	**87.86**	**88.18**

Table 2: Evaluation of proposed models based on some performance metrics

Dataset	Proposed Model	Performance Metrics			
		A	P	R	F
DS_1	Levi et al. [22]	80.13 ‡	81.07	78.79	79.91 ‡
	Kabil et al. [24]	80.37‡	82.13	80.25	81.18‡
	Rajeev et al. [27]	82.15‡	81.97	81.53	81.75‡
	BFFSBR+CNN+GRUN	**85.79**	**83.36**	**82.73**	**83.04**
	Wang et al. [29]	84.35‡	82.29	82.29	82.29 ‡
	Markitantov et al. [30]	85.39≈	83.17	82.35	82.76≈
	Ertam et al. [33]	82.72‡	81.02	81.92	81.47‡
DS_2	Levi et al. [22]	86.74‡	85.23	84.73	84.98 ‡
	Kabil et al. [24]	86.07‡	85.02	85.02	85.02‡
	Rajeev et al. [27]	86.92‡	84.93	85.17	85.05‡
	BFFSBR+CNN+GRUN	**88.04**	86.91	**85.66**	**86.28**
	Wang et al. [29]	85.83‡	84.67	84.91	84.79 ‡
	Markitantov et al. [30]	**88.04**	**87.03**	84.36	85.67‡
	Ertam et al. [33]	87.73≈	85.21	84.84	85.02‡
DS_3	Levi et al. [22]	85.29‡	84.75	85.03	84.89 ‡
	Kabil et al. [24]	84.35‡	83.67	84.79	84.23‡
	Rajeev et al. [27]	83.92‡	83.06	83.26	83.16‡
	BFFSBR+CNN+GRUN	**87.13**	**86.37**	85.48	85.92 ≈
	Wang et al. [29]	85.76‡	84.95	85.07	85.01 ‡
	Markitantov et al. [30]	86.85≈	85.82	**86.14**	**85.98**
	Ertam et al. [33]	86.79≈	84.75	85.65	85.20‡
DS_4	Levi et al. [22]	86.23‡	85.17	85.62	85.39 ‡
	Kabil et al. [24]	85.79 ‡	85.23	84.65	84.94 ‡
	Rajeev et al. [27]	87.05 ‡	85.36	86.19	85.77 ‡
	BFFSBR+CNN+GRUN	**89.04**	**88.74**	87.84	**88.29**
	Wang et al. [29]	88.63≈	86.45	85.83	86.14 ‡
	Markitantov et al. [30]	87.79‡	85.79	86.05	85.92 ‡
	Ertam et al. [33]	**89.04**	88.53	**88.68**	85.60 ‡
DS_5	Levi et al. [22]	87.19‡	86.21	85.82	86.01 ‡
	Kabil et al. [24]	84.73 ‡	85.07	85.63	85.35 ‡
	Rajeev et al. [27]	86.03 ‡	86.34	86.61	86.47 ‡
	BFFSBR+CNN+GRUN	**88.67**	**87.93**	**88.57**	**88.25**
	Wang et al. [29]	85.26‡	84.82	85.44	85.13 ‡
	Markitantov et al. [30]	84.91‡	84.48	84.26	84.37 ‡
	Ertam et al. [33]	88.34≈	86.75	87.39	87.07 ‡
DS_6	Levi et al. [22]	85.71‡	85.32	85.81	85.56 ‡
	Kabil et al. [24]	86.23 ‡	85.93	85.36	85.64 ‡
	Rajeev et al. [27]	85.37 ‡	85.71	86.13	85.92 ‡
	BFFSBR+CNN+GRUN	**87.76**	**87.39**	**86.85**	**87.12**
	Wang et al. [29]	86.04‡	86.51	85.92	86.21 ‡
	Markitantov et al. [30]	86.58‡	84.33	85.17	84.75 ‡
	Ertam et al. [33]	87.19‡	86.56	86.34	86.45 ‡

Table 3: Comparison of the proposed model with other related models