Singer Gender Classification using Feature-based and Spectrograms with Deep Convolutional Neural Network

Mukkamala S.N.V. Jitendra[1], Dr. Y. Radhika[2]
Department of Computer Science and Engineering, GIT
GITAM (Deemed-to-be University), Visakhapatnam-530045, AP, India

***Abstract*—The task of music information retrieval (MIR) is gaining much importance since the digital cloud is growing sparklingly. An important attribute of MIR is the singer-id, which helps effectively during the recommendation process. It is highly difficult to identify a singer in the case of music as the number of signers available in the digital cloud is high. The process of identifying the gender of a singer may simplify the task of singer identification and also helps with the recommendation. Hence, an effort has been made to detect the gender information of a singer. Two different datasets have been considered. Of which, one is collected from Indian cine industries having 20 different singer details of four regional languages. The other dataset is standard Artist20. Various spectral, temporal, and pitch related features have been used to obtain better accuracy. The features considered for this task are Mel-frequency cepstral coefficients (MFCCs), pitch, velocity, and acceleration of MFCCs. The experimentation has been done on various combinations of the mentioned features with the support of artificial neural networks (ANNs) and random forest (RF). Further, the genetic algorithm-based feature selection (GAFS) has been used to select the suitable features out of the best combination obtained. Moreover, we have also utilized the recent popular convolutional neural networks (CNNs) with the support of spectrograms to obtain better accuracy over the traditional feature vector. Average accuracy of 91.70% is obtained for both the Indian and Western clips, which is an improved accuracy of 3% over hand engineering features.**

Keywords*—Gender identification; spectrogram; genetic algorithm-based feature selection (GAFS); music information retrieval (MIR); music recommendation; and singer's gender identification

I. Introduction

Technological advancements in the music industry have created an enormous number of music clips. It is difficult to categorize and organize such several clips if proper meta-information is not provided [1]. Hence, it is essential to provide the meta-information for the available clips of the digital cloud. Moreover, it is impractical to provide the meta-information for millions of tracks available in the digital cloud. The meta-information could be related to artists, instruments, genre, lyrics, etc. Of which, artist information is a much important factor where a majority of the listeners are usually listening to the songs of their favorite artist [2]. The artist is further characterized by a singer, an artist, or a composer. There is a small difference between an artist and a singer. A singer who contributes his vocals to the portion of a song during studio recordings. An artist who performs his skills on a stage in front of the audience. In general, a majority of the audience shows interest in the songs of a particular singer. For instance, Shreya Goshal is one such Bollywood singer who gets the attraction of most Indian listeners [3].

Since it is impractical to provide the meta-information manually due to the availability of million numbers of tracks, there should be an alternative approach for the provision of meta-information. The process of automatically extracting the information from music clips is called music information retrieval (MIR). Hence, there is a good amount of research happening to investigate an alternative approach or automatic approach which extracts the meta-information. The research on MIR is initiated during the initial years of the 21st century [4].

There are many works such as singer identification, genre classification, singer identification, lyrics transcription, instrument identification, mood estimation, and music annotation done on the aspects of MIR. However, the application which is designed for a particular regional song is not giving the same performance as the songs of other regions. Hence, a challenging event, called music information retrieval evaluation exchange (MIREX) has been initiated under the international music information retrieval systems evaluation laboratory (IMIRSEL) in the year 2000.

There are thousands of papers that have been published in ISMIR since 2000 on various MIR works mentioned [1]. This would give a clue to understanding the importance of automating the MIR tasks. From the above-mentioned tasks of MIR, we have identified one important problem named gender classification which is a sub-problem of singer identification. There are around 50,000 singers in the world. There are certain singers' datasets available with 3,000 and 48,800 singers. They are provided by the institute of computational perception of Johannes Kelper university, Linz, Austria, and are called c3ka and c49ka [5]. The process of singer identification gets complicated when there is an increase in the number of singers. Hence, it is essential to further categorize the singers based on their characteristics. One such important characteristic is gender. Based on its importance for MIR, the same has been considered for this work. In general, the task of singer gender identification is named automatic singer gender identification (ASGID).

In the case of speech recognition, the task of automatic gender identification (AGID) has been used in speaker recognition, biometric systems, security, and surveillance. Since the main objective of biometric systems is to identify a person, it was mentioned that the task of gender identification simplifies the task by segmenting a person into either male or female category [6]. Moreover, the implementation of AGID is also helpful in assigning a male customer care agent to the male customer in the case of transferring calls to an agent. And in some automatic regional language identification [7]. Similarly, some research has happened to categorize the genders of singers, which simplifies the process of singer identification. This task is further helpful for the music recommender system as well while recommending the songs to the listeners [8].

However, a majority of the research has happened in detecting the gender of a speaker. A less focus has been done on the case of the gender identity of a singer. However, the features related to speech processing are found to be sufficient to model the music as well. Hence, the features that can discriminate against gender have been considered and experimented within this work. The features such as Mel-frequency cepstral coefficients (MFCCs), variations of MFCCs such as velocity and acceleration, features related to pitch have been considered for this work. Artificial neural networks and random forests have been considered as classifiers to classify the category of feature dimension [47]. The dimensionality of the feature vector is 43. Feature selection methods such as principal component analysis (PCA), and cross-correlation analysis (CCA) are considered to select suitable features from the larger dimensionality. Further, the results have been compared with the popular convolutional neural networks (CNNs) by feeding the spectrogram images of the audio signals.

The rest of the paper is laid as follows: Section II gives detailed literature done for the speech and audio signals. Section III proposed methodology with a flow diagram and the details of features is described with different classification, models were elaborated along with the implementation of spectrogram-based CNN. Section IV presents the results and implications of work with the comparisons that are given for traditional feature-based approaches and recent popular convolutional neural networks. Section V concludes the work with future directions.

II. Background

The task of automatic singer gender identification (ASGID) is essential in simplifying the process of singer identification. It also helps in indexing and categorizing the audio clips into a class of male and female singers. Hence, it can be considered as one important factor while recommending songs to the listeners. However, the research which is done on gender identification is not up to the mark. The reason could be the similarities that can be identified among male and female singers. They may expose similar characteristics while singing a song. Whereas, one can observe the differences in terms of the pitch in the case of speech [9]. Based on this, we can conclude that the task of gender identification is a challenging issue when compared to gender identification for speech. Here, we have provided the literature on both the aspects of speech and music processing. It gives a clear understanding of the similarities and differences in both aspects.

The features considered for speech and music processing are divided into three classes namely low-level, mid-level, and high-level features. A signal of shorter length which is ranging from 10 to 100 milliseconds. It gives low-level inherent information to the researcher and also provides a way to map every portion of the signal with relevant information. Further, information from larger frames −frames is nothing but a portion of signal of the same length− has been considered to extract the mid-level information. To avoid loss of any information of the signal, a technique called overlapping has been introduced and generally, 50% of the frame will be considered to overlap [10]. However, the features extracted from low-level and mid-level are useful to provide high-level information i.e. gender, singer, artist, genre, raga, etc. This high-level information is useful to recommend or categorize the audio clips.

Fundamental frequency (F0) is one important feature in recognizing gender information [11]. The pitch range is around 100 Hz to 200 Hz in the case of males, and the same is around 120 Hz to 350 Hz for females [12]. Hence, the pitch has been used as a primary feature in most of the applications that are designed to categorize gender [13]. The accurate estimation of the fundamental frequency (F0) has been used to compute the set of acoustic features that are further used in various research works to estimate the gender of a speaker/singer [14]. However, the process of estimating the accurate F0 itself is a challenging task. An algorithm is yet to be designed to compute the accurate F0 value. The reason for poor performance obtained with the gender classification systems is due to the imprecise F0 obtained with the existing algorithms. Henceforth, various other spectral features obtained from the frequency spectrum have been used as supporting features for F0 to improve the performance of gender classification systems. Some notable features are including linear predictive coefficients (LPCs), Mel frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), pitch class profile (PCPs), perceptual linear predictive coefficients (PLPs), relative spectral MFCCs (RASTA-MFCCs), relative spectral PLP coefficients (RASTA-PLPs), etc. In contrast, some research works have concluded that the traditional features that are used for the tasks of speech recognition may not be suitable for the gender classification task. Further, a scope has been found to investigate the characteristics of the signal for different genders, which gives a clue to compute the relevant features for gender-specific tasks [15]. It could be possible to obtain better accuracy if the relevant features for gender are explored.

Humans have been categorized into male or female based on certain characteristics. Speech is one such important factor that helps in recognizing the male or female. The physical parameter of the glottis, vocal tract length, and thickness decide the category of a person to the male or female class. They are generally called acoustic parameters. Several works have been initiated to recognize gender with a variety of features related to acoustic parameters and popular classification models. A variety of performances were

observed with the features identified through acoustical parameters. Of which, pitch and first formant are the prominent features found in many research work with improved performance. Pitch is the feature which is related to the source of the voice and the first formant (F1) is related to vocal tract information. An approach of linear predictive analysis has been considered to compute the pitch and F1. Certain analysis has given the information that the pitch and F1 values of males are less when compared to those of females. Distance measure has been considered based on Euclidean distance to segregate the genders using the nearest neighbor classifier. Further, it is found that the features based on autocorrelation, cepstral analysis, linear prediction, vowels that are extracted from speech, reflection in voice, fricatives did well to identify the gender of a speaker. Pitch has been considered as a primary feature. In addition to pitch, MFCC features, and energy has been supplied to support vector machine (SVM) which is resulting in the performance with an accuracy of 95% from their dataset [16]. However, pitch alone could give an accuracy of 96% with the support of neural networks. The dataset also involves information that is phoneme and speaker-independent. Moreover, various vocal source parameters are extracted to detect the gender and an average accuracy of 95.1% is obtained in detecting the male and female classes.

In some works, only the portions of voiced segments have been detected to effectively estimate the category of gender. Various cepstral features such as PLPs, LPCCs, and MFCCs have been computed from those portions to classify the gender. Also, independent dimensions of the features mentioned above have been analyzed that helps in identifying the suitable feature vector for gender classification from the category of cepstral features.

However, a majority of the works mentioned above are designed to detect the gender for the speech that is recorded in the acoustically controlled environment. Since the biometric systems can be designed effectively to recognize the person's gender, the systems that are designed for studio-recorded speech are sufficient. They may not give the required performance in the real-time environment. Moreover, the singer's voice is always accompanied by background music. In such a case, the gender detection systems designed for the studio-recorded environment may not be useful. Hence, for the first time, the gender recognition system has been designed for a noisy environment. Besides, it is also important to note that the process of gender detection is to be done for various languages. There could be a variety of parameters related to the vocal tract that may affect the performance of gender detection. Hence, the system which is designed for one language may not give a similar performance with the other languages. The author has taken care of designing an effective system that could handle various languages [17]. An accuracy obtained with this approach is 95%, which shows the capability of the system. The features related to pitch and suitable spectral features have been used to obtain accuracy.

It is also an important aspect to know about the gender detection system, which could be affected by the age factor. It is stated in the literature that the person's voice gets changed every two years due to the change in vocal tract parameters. Moreover, the gender detection system is ineffective in the case of children when compared to elders. As there could not be many differences in the vocal tract parameters in the case of children, both male and female voices look similar. Research work has been done to verify the same by [18]. Various recordings have been collected from the age groups of 8-10 years and 16-20 years. Further, experimentation has been done to detect the genders of these two groups. An accuracy of 60% and 95% have been obtained from the children identification (CID) Dataset. Features that are based on acoustical cues, prominent peaks from cepstrum, pitch based on harmonic-to-noise ratio, and source spectral magnitude have been considered for this work. Similar experimentation has been done to compare the performance of the systems with two different databases [19]. A total set of features contains the vector size with 113 dimensions. A naive Bayes classifier has been used to classify the data into the male and female categories. However, the system is not giving an accurate performance in detecting the gender of a person.

Modified voice contour (MVC) is used to measure the intensity in voice in the speech sample, which further helps in discriminating against the genders [20]. The dataset has been collected, which is forming the signal with Arabic digits. As the standard dataset for the gender recognition system is the TIMIT dataset, the experimentation setup was compared with TIMIT using the supervised support vector machine (SVM) classifier [21].

A different approach to the classification of genders has been proposed in this work. Since the proposed work is mainly focusing on gender detection for singers, it is essential to develop a system that could handle the background music as well. It is not possible to find a song without background accompaniment. Moreover, 99% of the song portions are accompanied by instrumentals. Hence, we made an effort to suppress the noise up to some extent using Chebyshev infinite impulse response (IIR) filters. However, we are unable to neglect all the background support with this approach. Therefore, it is decided that the proposed system should be effective even if the background accompaniment is there. The database is collected from the Indian Bollywood, Tollywood, Kollywood, and Sandalwood cine industries that are involving four regional languages of India, namely Hindi, Telugu, Tamil, and Kannada. As there is no standard dataset for Indian language speaker's gender detection, we could not compare the proposed approach with any other work. However, we made an effort to compare the proposed work with the Western popular artist dataset called Artist20. Various features such as MFCCs, velocity, and acceleration of MFCCs called ΔMFCCs, and Δ ΔMFCCs have been computed. Besides, pitch related features are also computed to add strength to the above-mentioned spectral features. These features are fed to the two popular non-linear classifiers, such as random forest (RF) and artificial neural networks (ANNs). Further, we have used the genetic algorithm-based feature selection (GAFS) algorithm proposed by Murthy et al. to reduce the dimensionality and complexity issues. Moreover, the popular convolutional neural networks (CNNs) are also used by feeding the spectrograms as images. It is found that CNN's are more capable of discriminating against the gender of singers.

A. The main Focus of the Article

- Identifying suitable music data set for different regional languages in Indian and Western songs to identify singer gender.
- Implementing various feature extraction processes with GAFS (Genetic Algorithm based Feature Selection) by combining with traditional features like Mel-frequency cepstral coefficients (MFCCs), Pitch, and Temporal.
- Design a novel convolutional neural network model based on spectrogram images to Automatic gender identification of a singer in a given music track.

III. Proposed Methodology

The flow diagram of the proposed gender classification system has been depicted in Fig. 1. It has various blocks, namely dataset collection, the process of dividing the dataset into training and test sets, feature extraction, classification models, spectrogram generation, and convolutional neural networks (CNNs). This section describes the process of feature extraction, classification models, spectrogram generation, and CNNs.

B. Feature Extraction

Features represent the prominent information that is useful to discriminate different classes depending on the problem chosen. The features that are chosen for one task may not be suitable for any other task. Hence, we should perform a lot of analysis while choosing the feature vector. Since the features are to be extracted from the samples of the speech signal, we have used various signal processing approaches to identify the suitable features for the task of the singer's gender identification. Correlation is one important property that gives the similarity behavior of a particular feature over different classes. Hence, we have used the same metric to check the suitability of a particular feature. However, the evolutionary-based strategy has been used to select the relevant and effective feature vector for gender detection [22]. To select the suitable feature vector from the large dimensions to reduce the complexity we used a genetic algorithm-based feature selection (GAFS) algorithm and features related to voice source and spectral are more suitable for gender detection. Hence, Mel-frequency cepstral coefficients (MFCCs) and pitch related features such as minimum pitch (P_{min}), maximum pitch (P_{max}), average pitch (P_{avg}), and deviation in pitch (P_{std}) have been considered as base features for the task of the singer's gender identification.

Also, the variations in MFCCs such as velocity and acceleration called ΔMFCCs and ΔΔMFCCs have been computed by finding the first-order and second-order differentiation on MFCCs. MFCCs (13), ΔMFCCs (13), ΔΔMFCCs (13), and pitch (4) are together forming a feature vector of length 43. The process of computing features has been detailed below:

1) Mel-frequency cepstral coefficients: One of the prominent features in extracting relevant information from the speech signal is the MFCC feature vector. It can effectively model the music signal as well. Hence, it has been used as a base-line feature in many applications such as vocal and non-vocal segmentation [23], singer identification [24], genre recognition [25], etc., that are related to the music signal. It is a representation of the short-time power spectrum and computed using the non-linear Mel scale [26]. The log magnitude of the power spectrum has been computed to construct a cepstrum. The non-linear Mel scale will be applied to extract the prominent peaks after applying the triangular band filters on the cepstrum. However, MFCCs are extracted from the short-time frames and hence, come under the category of low-level features. The signal is divided into chunks of the length 25 ms with an overlap of 10 ms. The features have been computed from the 25 ms lengthened frames [27]. The steps to compute the MFCCs are given below:

- Divide the signal into the chunks of frames of length 25 ms with an overlap of 10 ms.
- Construct a spectrum using fast Fourier transformation (FFT).
- Construct a power spectrum from the FFT computed and compute the log magnitude of the spectrum which gives a cepstrum.
- Map the powers of the spectrum obtained above onto the Mel scale, using triangular bands.
- Identify the logs of powers at each Mel frequencies.
- Take the discrete cosine transform of the list of Mel log powers, which gives MFCCs.
- Consider the prominent 13 to 39 peaks and ignore the rest. Here, the first 13 MFCCs have been considered for experimentation.

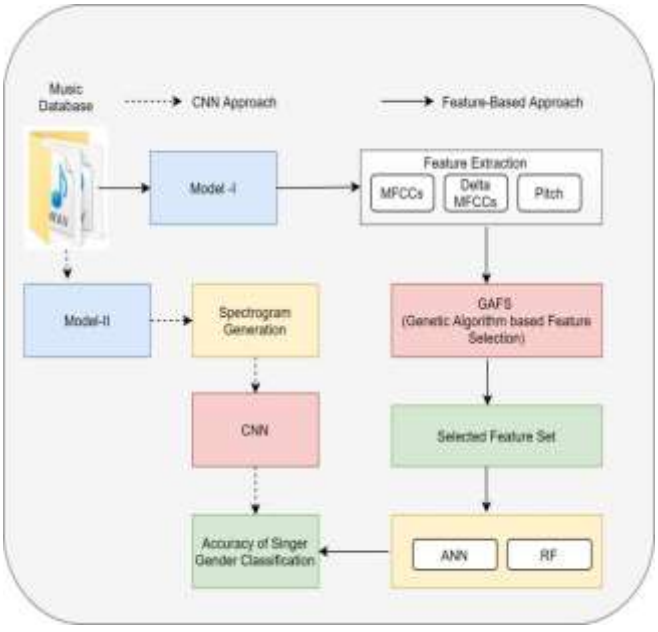

Fig. 1. Proposed Flow Work for Singer's Gender Detection and Comparison with Feature and CNN Approaches.

2) Velocity and Acceleration of MFCCs: The dynamic features that can extract the aggressive behavior of the music signal are temporal features. One such temporal feature is velocity and acceleration that are obtained from the static MFCCs. As MFCCs are static, they may not be able to detect the temporal information of the signal. The temporal information may be used in detecting the gender information of a singer. As the singers change their pitch information intentionally, the temporal information may give some clue in identifying the gender information. Hence, velocity and acceleration features have been used as supportive features for MFCCs. We found proper discrimination in the classification performance after adding velocity and acceleration features. The velocity features are extracted by computing first-order derivatives on the MFCCs and the acceleration features are the second-order derivatives [28-30]. Hence, they are generally called Delta (Δ) coefficients. Where the velocity features are represented as Δ , acceleration as Δ Δ of MFCCs. The general formula to compute the Δ coefficients are given in Eq. (1).

$$\Delta \mathrm{Cep}_{\mathrm{k}}(\mathrm{t}) = \frac{\sum_{\mathrm{i}=-\mathrm{n}}^{\mathrm{n}} \mathrm{m_i Cep_k(t+1)}}{\sum_{\mathrm{i}=-\mathrm{n}}^{\mathrm{n}} |\mathrm{i}|} \quad (1)$$

Where $Cep_k\ (t)$ is the MFCC feature that represents the k^{th} feature at time frame *t.* The total number of successive and predecessor frames are denoted with *n* and the weight mi has been added to the i^{th} frame. In general, the value of *n* considered for experimentation is 2. Further, the same process is applied to ΔMFCCs to compute accelerative features.

3) Pitch based Features: Since it is already mentioned that voice source parameters give much prominent information to recognize the gender of a singer effectively. Pitch is one such feature that is useful to detect gender. It has outperformed in recognizing gender from the speech. As there could be overlapping differences in the case of males and females, it would help efficiently in the case of the speaker's gender recognition. The same performance could not be observed with the pitch feature alone in the case of the singer's gender detection. Both the male and female singers can tune their pitch according to the tune of the song. This could be the main reason for an ineffective performance with the pitch feature alone. However, the temporal information observed from pitch may give some useful information to detect the gender effectively. Hence, some statistical methods are applied to pitch values for obtaining temporal information out of it. It results in obtaining the four different features such as minimum pitch (Pmin), maximum pitch (Pmax), average pitch (Pavg), and standard deviation of pitch (Pstd). We have used a harmonic-to-sub harmonic approach to obtain the pitch values as it performs well in the case of background accompaniment [31, 45].

4) Genetic algorithm based feature selection: Genetic Algorithms are designed to optimize the process to select the best solution and to discard the rest. The algorithm generates random values that are used to generate the population [32]. Initially, the random value of length 43 bits is generated. It is used to select the set of features from a total of 43 features. The population is the series of 0's and 1's, where '1' represents the feature consideration, and '0' represents its absence in the final set. The generation of the bits for the population is done through a random process [44, 46].

The selected features are then fed into the ANN and RF classifiers to get the accuracies. The accuracy obtained through this approach is highly efficient than that of the original feature set. The fitness of the population is calculated based on accuracy and the number of features selected. Higher the fitness, the greater the efficiency in the classification. This process is repeated for several epochs, and finally, an optimized set of the population is obtained. The mutation operation is performed by inverting or changing the bit values in the population. This process enriches the qualities of the child.

C. Spectrogram based CNNs

It is highly difficult to extract suitable features from the music signal as it is always accompanied by background music. It is also a known fact that the recent popular convolutional neural networks (CNNs) are doing well to classify the highly non-linear data. They already outperform in the field of image processing. Hence, we utilize the same for implementing the task of gender classification of Indian singers. As images are the possible input that we can feed to CNN to compute the suitable features automatically, spectrograms are constructed. Spectrograms represent the three-dimensional view of the signal having time, frequency, and intensity as the x, y, and z planes, respectively. The details of the spectrograms and the components of CNN are given in this section.

1) Spectrogram generation: Spectrograms help in analyzing the time-frequency information effectively. The frequency modulation can be observed in the case of spectrograms where it is not possible in the time domain. In general, the frequency domain gives information concerning single-frequency components. Time-frequency distribution (TFD) resolves the issue by providing both time and frequency information. Spectrograms give the information related to the moving sequence of the local spectra to any music signal [33, 34]. There are several ways of computing spectrograms. In this work, a short-time Fourier transformation has been utilized to construct a spectrogram. As the process isolates the distinguished components of two gender classes, the signal f*(t)* has been multiplied with the succeeding time windows which were shown in Eq. (2).

$$\mathrm{f(t)} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{f(t)}\, \omega(\mathrm{t} - \tau_{\mathrm{m}}) \quad (2)$$

The computed spectrograms have been used as RGB images having three dimensions for convolutional neural networks. All the images are scaled to the size of 128 * 128 pixels based on the normalization strategy.

D. Convolutional Neural Networks (CNNs)

One possible way to automatically extract the features from the images is by using the convolution operation. Various convolution layers are to be used to obtain the relevant features from the spectrogram images. With this nature, convolutional neural networks (CNNs) became popular in recent years. A majority of the applications that are based on image processing have been redesigned with the involvement of CNN. It is found that CNN's are outperforming in most of the cases when compared to traditional feature-based classification models. It is not possible to say that the CNNs are not based on a feature-based approach. However, they could extract suitable features automatically.

It is also a known fact that the speech signal is a one-dimensional time-invariant signal which gives no clue to estimate what it contains. Moreover, a music signal is highly complex when compared to the speech where instrumentals always accompany it. Hence, the task of the gender classification of a singer is harder than a speech signal. The features that are computed from time-domain and frequency-domain may give some base-line performance to discriminate against the gender of a singer. Hence, a three-dimensional spectrogram has been utilized as an image. It contains information related to frequency components and their intensity [35, 36]. It may further help in accurately classifying the gender of a singer. The components of CNNs have been detailed below, which gives fundamental information about the procedure.

A deep learning algorithm that could take an image as an input is the convolutional neural network (CNN), also called ConvNet. Various objects of the image are getting importance with the ConvNet and hence, able to discriminate each portion of the image using them. As it is known that the traditional feature-based approaches involve complicated preprocessing before extracting certain features from the image. However, CNN's can reduce the difficulty involved with the traditional approach. We have to manually change the filters to obtain the relevant features in the case of a feature-based approach. However, CNN has the inbuilt ability to apply the possible number of filters automatically.

The basic architecture of CNN for the problem chosen is given in Fig. 2. The organization of the human brain is the inspiration for designing such a connecting pattern of ConvNets. One more important piece of information that has to be noted for CNNs. The CNN might give an average performance with the grayscale images when compared to the RGB images. Hence, color spectrograms have been utilized for this task instead of monochrome images. Features obtained based on hand engineering may not be able to capture spatial and temporal variations from the spectrogram. However, ConvNets can estimate them effectively to extract suitable features. Moreover, the number of features selected is less and relevant when compared to the traditional method. The parts of CNN include the convolution layer(s) (CONV-RELU), pooling layer(s) (POOLING), fully connected network (FCN), and softmax layer (SOFT).

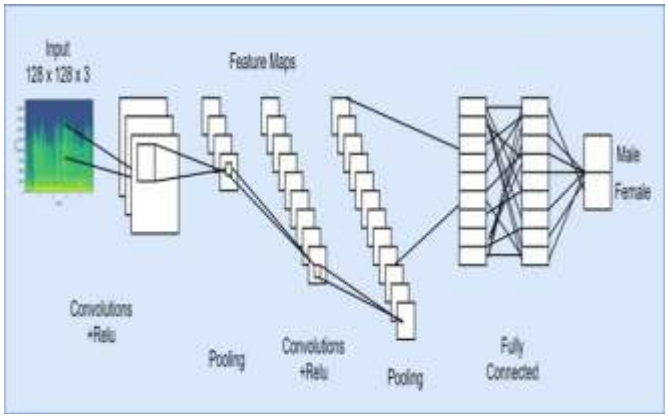

Fig. 2. The Components of the Convolutional Neural Network from Feeding Images to the Identification of Class Labels.

Algorithm for Automatic Singer Gender Identification (ASGID)

Input: *Spectrogram Images (Indian and Western Songs)*
Output: *Classification of Singer gender (Male or Female)*

1. *Start*
2. *Take* ***n * n * 3*** *images, the filter of size* ***f * f****. Were, n = 128, and f = 8 for the CNN0.*
3. *Padding* ***p = 2*** *to get effective results with the CNNs.*
4. *Generate Spectrogram image with the size of* ***(n+ 2p−f+1) * (n+2p−f+1) * f.***
5. *Consider max pool with a filter size of* ***2 * 2*** *and labeled the max filter as* ***w * w****.* ***Stride (s) = 2****.*
6. *The outcome of the pooling layer with* ***n1, w, and s*** *is* ***(n1 - f)/s + 1 * (n1 − f)/s + 1.***
7. ***Repeat*** *from Step 2* ***Until k!=4***
8. *Generate optimal features from the input spectrogram.*
9. *Flattening the output vectors of* ***Three channels into 2 classes*** *(Male or Female)*
10. *Stop*

IV. Result and Discussions

This section mainly focuses on two aspects. One is the datasets that are used for the task of gender classification, and the other part gives the detailed observations of the results obtained using the proposed approach.

A. Dataset Collection

Two datasets have been used for this work. One is the Indian popular songs dataset (IPSD), which has been designed with 20 singers. The other dataset is the standard Artist20 dataset. The IPSD has been designed with 250 audio clips. The average length of the audio clip is five seconds. The dataset includes song clips of various Indian cine industries, including Bollywood, Kollywood, Sandalwood, and Tollywood. The clips are based on regional languages, namely Hindi, Tamil, Kannada, and Telugu, respectively.

Further, each clip is segmented into 25 ms segments and an overlap of 10 ms. All the clips are recorded at the sampling frequency of 44100 Hz. The dataset has been collected based on the study done in [22]. Care has been taken to involve various background accompaniment instruments while collecting the dataset. The reason for selecting the songs of various languages is to make the system language independent. An Artist20 database is internationally accepted and acknowledged [37]. It comprises 20 songs of 20 singers of various genres. Since this database consists of only three female singers, we have considered only three male and female singers information for the task of gender classification. There are 100 clips for each male and female gender. The clips are with a sampling frequency of 44100 Hz. A sample data set of spectrograms are shown in Fig. 3. The length of each clip ranges from 2 to 5 seconds. The singers considered for the experiment have a different accent which makes the data-set very versatile and covers almost all the traits of the singers. This dataset is also divided into two parts for training 70% and testing 30%, respectively.

B. Results and Observations

The process of gender identification is quite easy with speech processing as pitch related features are sufficient to detect the same. Hence, the accuracy is around 97.40% with the suitable features that are extracted from the speech signal [38]. However, the above-mentioned accuracy is obtained for the studio-recorded speech where background noise is not considered. The system developed based on the traditional feature engineering approach gets failed if real-time noise and other speech gets involved. Hence, the modern popular convolutional neural networks have been used in the recent article and stated that an average of around 99% accuracy is obtained [39]. However, the speech considered for this work is recorded in environmentally controlled situations. It is very difficult to obtain that much level of accuracy in the case of gender identity for a singer. As the singer's vocals are always accompanied by instrumental sounds, the complexity of gender identification gets proportionately increased. Moreover, the singer intentionally changes his/her pitch since they trained their vocal cords accordingly. Based on this, one can say that the pitch related features alone could not perform well in the case of the singer's gender identification [14].

However, the supporting spectral and temporal features may support the pitch features to improve the performance of gender identification in the cast of singers. Hence, the Mel frequency cepstral coefficients (MFCCs) have been used as base-line supporting features as they have proven their capability in the modeling music signal. However, speech researchers know its importance as it has outperformed many speech-related tasks [40, 41]. Also, the first and second-order differentiations on MFCCs provide a new set of features called velocity and acceleration features of MFCCs. They are popularly called Δ and ΔΔMFCC features [42]. In many of the research works, it is mentioned that the Δ and ΔΔMFCCs carry temporal information of the signal. As the temporal information is much useful in discriminating the gender information, they have been added to support the spectral and pitch features [43].

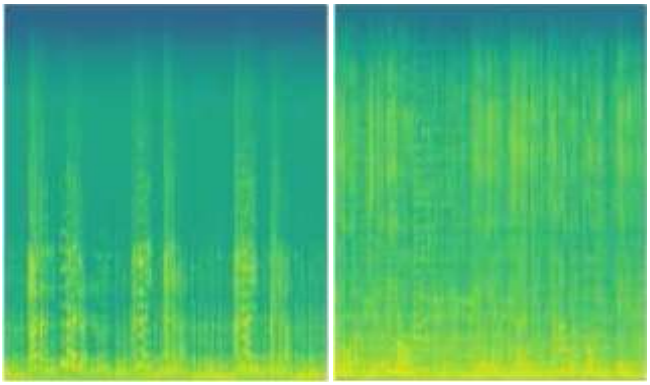

Fig. 3. A Sample Spectrogram Image that gives Some useful Discriminative Information of Male and Female Singers.

The consolidated features that are considered for this work are including MFCCs (13), pitch (4), ΔMFCCs (13), ΔΔMFCCs (13) forming a length of a 43-dimensional feature vector. The details of the features, acronyms, and length of each feature category are given in Table I. The second column represents the feature name, and the third one is the acronym that has been considered to represent the feature hereafter. The fourth column is the size of the respective feature. We have considered the features individually and different combinations that help estimate the relevant features for the task of the singer's gender identification.

The results obtained for the selective feature combinations are given in Table II and Fig. 4. The table gives complete information about the accuracies obtained for Indian and Western data clips. Moreover, three different classification models, such as ANN, RF, and CNN, have been considered to represent the data for both the categories. Initially, we experimented with the baseline MFCC features, obtaining an average accuracy of 63.60% and 69.42% for the Indian and Western clips, respectively. The percentage improvement with the combinations of {M+P}, {M+Δ+ΔΔ}, and {M+P+Δ+ΔΔ} is 16%, 3%, & 10% for Indian clips dataset, and 9%, 7%, & 6% for Western clips, respectively. However, the combination of {M+P+Δ+ΔΔ} is giving the best accuracy with the feature engineering process. The accuracies with the mentioned combination are 85.27% and 86.62% with Indian and Western clips, respectively. The values are average accuracies obtained for the ANN and RF. However, ANN and RF are found to be similar in their performances while classifying the genders for the above-mentioned feature combinations.

TABLE I. FEATURES, ACRONYMS, AND THEIR DIMENSIONAL SIZE THAT ARE CONSIDERED HEREAFTER

Sl. No.	Feature Name	Acronym	Size
1	MFCCs	M	13
2	Pitch	P	4
3	Δ MFCCs	Δ	13
4	Δ Δ MFCCs	Δ Δ	13

TABLE II. THE ACCURACY VALUES ARE OBTAINED USING THE VARIOUS COMBINATIONS OF FEATURE SETS FOR INDIAN AND WESTERN DATASETS

Features and Spectrograms	Accuracy (in %)					
	Indian			Western		
	ANN	*RF*	*CNN*	*ANN*	*RF*	*CNN*
M	62.34	64.86	-	69.29	69.54	-
M+P	74.82	73.24	-	76.34	75.25	-
M+Δ+ΔΔ	77.36	76.34	-	81.25	82.04	-
M+P+Δ+ΔΔ	83.72	86.82	-	85.19	88.05	-
GAFS<M+P+Δ+ΔΔ>	85.45	87.16	-	90.84	92.55	-
Spectrogram	-	-	***89.16***	-	-	***94.25***

There could be a chance of having some worthless feature dimensions in the selected combinational feature vector. Feature selection is one suitable approach to select the supporting feature dimensions and ignoring the rest, which may lead to an increase in the final accuracy. Hence, we applied a feature selection algorithm called genetic algorithm based feature selection (GAFS) to select the suitable features. The genetic algorithm comes under the category of evolutionary algorithms, which is purely based on randomness in the approach. The use of GAFS has given better performance on top of the best-combined feature vector. We have used the best combination {M+P+Δ+ΔΔ} to apply the GAFS algorithm mentioned in the 5th row of Table II, labeled GAFS<M+P+Δ+ΔΔ>. An increase of 1% and 5% have been obtained over the best-combined feature vector with the support of feature selection using GAFS.

Further, CNNs have been used to do experimentation to detect the accurate gender of the signer's voice. Spectrograms have been considered as they can discriminate the information related to male and female singers. Based on the visual differences observed in the spectrogram, an effort has been made to classify the gender with the support of CNN's. Table III gives detailed information about the hyperparameters and their values that are considered for the task of the singer's gender identification.

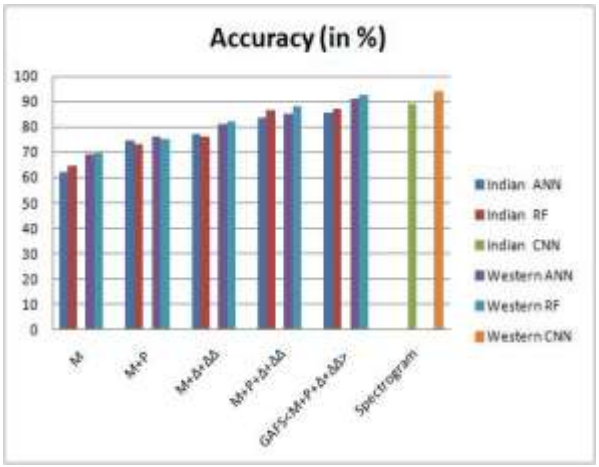

Fig. 4. The Graphical Representation of Accuracy Values that are Obtained using the Various Combinations of Feature sets for Indian and Western Datasets.

TABLE III. HYPERPARAMETERS ARE CONSIDERED FOR DESIGNING CNN FOR THE TASK OF THE SINGER'S GENDER CLASSIFICATION

Sl. No.	Parameter	Value
1	Batch size	8
2	#Channels	3 channels (RGB)
3	Filter size	3*3
4	Image size	128*128
5	#Convolution Layers	4
6	#Hidden layers	4
7	#Flatten layers	2
8	Softmax layer	1
9	#Output classes	2 (M & F)
10	Activation function(s)	ReLu
11	#Epochs	Around 250

A better accuracy has been obtained with the specified hyperparameters. CNN's outperform in many image processing tasks. Similarly, better performance has been observed in this case, as well. However, there is not much higher spike, which has been observed with the support of CNN's. A nominal improvement of 3% and 2.7% for the Indian and Western clips, respectively. However, CNN's can classify gender information with an accuracy of 89.16% and 94.25%, though there is complex background support by instrumentals. Hence, CNNs can be effectively utilized hereafter with various spectrogram models, which further could classify the singer's gender efficiently.

V. CONCLUSION

The process of the singer's gender identification will surely help the task of music information retrieval (MIR) and music recommender systems as well. The pitch alone features may not suffice to get better accuracy in the case of the singer's gender identification. An accuracy of 20% has been obtained using pitch features alone. The reason could be the support of complex background instrumentals involved in the case of music clips over speech signal processing. However, features that have been used in speech processing are effectively used in many music processing tasks as well. For instance, MFCCs are effectively utilized to model speech data and music also. Hence, spectral features MFCCs are suitable to give their support for pitch features to get better accuracy. Moreover, the temporal features obtained by applying first, and second-order differential equations on MFCCs resulting in velocity, and acceleration features. They are also useful to estimate the temporal variations in the signal effectively. It could be the reason for obtaining a considerable accuracy while using the combinational feature vector. It is also more important to omit worthless feature dimensions to avoid performance degradation. The support of recent CNNs is always effective in getting better accuracy over the traditional feature engineering process.

VI. FUTURE SCOPE

For future work, it is very important to establish a standard dataset for Indian singers. It may help in many of the MIR

tasks. The use of recurrent neural networks (RNNs) by feeding a one-dimensional signal could help in improving the accuracy as they outperform in many speech-related tasks. Hence, our future work focuses on the use of RNNs for gender identification. Moreover, it focuses on constructing an efficient dataset for Indian audio clips. Since the structure of Indian songs is completely different from Western clips, it is highly essential to construct the same. Further, we may focus on the task of singer identification using gender classification as a fundamental step. It means the task of singer identification can be effectively done if a two-level classification model is proposed.

AUTHORSHIP CONTRIBUTION STATEMENT

Mukkamala S N V Jitendra: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft & editing. Dr. Y. Radhika: Supervision, Conceptualization, Writing - review & editing.

ACKNOWLEDGMENT

We acknowledge support from the Department of Computer and Engineering, Gandhi Institute of Technology and Management (GITAM) Deemed to be University, Vishakhapatnam for guidance, reviews, valuable suggestions, and very useful discussions for all the support being extended to carry out this research work.

REFERENCES

[1] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi. "Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review." ACM Computing Surveys (CSUR), vol. 51(3), 2018, pp.1-46.

[2] Cai, Wei, Qiang Li, and Xin Guan. "Automatic singer identification based on auditory features." In 2011 seventh international conference on natural computation, vol. 3, IEEE 2011, pp. 1624-1628.

[3] Ter Bogt, Tom FM, Juul Mulder, Quinten AW Raaijmakers, and Saoirse Nic Gabhainn. "Moved by music: A typology of music listeners." Psychology of Music, vol.39(2), 2011, pp. 147-163.

[4] Downie, J. Stephen. "Music information retrieval." Annual review of information science and technology, vol.37 (1), 2003, pp. 295-340.

[5] Schedl, Markus, Peter Knees, and Gerhard Widmer. "Investigating web-based approaches to revealing prototypical music artists in genre taxonomies." In 2006 1st International Conference on Digital Information Management, IEEE 2006, pp. 519-524.

[6] Harb, Hadi, and Liming Chen. "Voice-based gender identification in multimedia applications." Journal of intelligent information systems, vol. 24(2), 2005, pp. 179-198.

[7] Alsharhan, E. and Ramsay, A., "Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions". Information Processing & Management, vol. 56(2), 2019, pp.343-353.

[8] Nakano, Tomoyasu, Kazuyoshi Yoshii, and Masataka Goto. "Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity." In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 5202-5206.

[9] Hu, Yakun, Dapeng Wu, and Antonio Nucci. "Pitch-based gender identification with two-stage classification." Security and Communication Networks, vol. 5(2), 2012, pp.211-225.

[10] Qian, Kun, Zixing Zhang, Fabien Ringeval, and Björn Schuller. "Bird sounds classification by large scale acoustic features and extreme learning machine." In 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2015, pp. 1317-1321.

[11] Huss, "Vocal Pitch Range and Habitual Pitch Level: The Study of Normal College Age Speakers" (1983). Master's Theses. 1590.

[12] Titze, I. R., & Martin, D. W. ,"Principles of voice production". The Journal of the Acoustical Society of America, vol. 104, 1998, pp. 1148.

[13] Barkana, Buket D., and Jingcheng Zhou. "A new pitch-range based feature set for a speaker's age and gender classification." Applied Acoustics, vol. 98, 2015, pp.52-61.

[14] Weninger, F., Wöllmer, M., & Schuller, B, "Automatic assessment of singer traits in popular music: Gender, age, height and race". Paper presented at the Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, pp. 37-42.

[15] Rao, K. Sreenivasa, and Shashidhar G. Koolagudi. "Identification of Hindi dialects and emotions using spectral and prosodic features of speech." IJSCI: International Journal of Systemics, Cybernetics and Informatics, vol. 9(4), 2011, pp. 24-33.

[16] Gaikwad, Santosh, Bharti Gawali, and S. C. Mehrotra. "Gender identification using SVM with combination of MFCC." Advances in Computational Research, vol. 4(1), 2012, pp. 69-73.

[17] Zeng, Yu-Min, Zhen-Yang Wu, Tiago Falk, and Wai-Yip 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.

[18] Chen, Gang, Xue Feng, Yen-Liang Shue, and Abeer Alwan. "On using voice source measures in automatic gender classification of children's speech." In Eleventh Annual Conference of the International Speech Communication Association, 2010, pp. 673-676.

[19] Sedaghi, M. "A comparative study of gender and age classification in speech signals". Iranian Journal of Electrical and Electronic Engineering, vol. 5(1), 2009, pp. 1-12.

[20] Alsulaiman, Mansour, Zulfiqar Ali, and Ghulam Muhammad. "Gender classification with voice intensity." In 2011 UKSim 5th European Symposium on Computer Modeling and Simulation, IEEE, 2011, pp. 205-209.

[21] Alsulaiman, Mansour, Zulfiqar Ali, and Ghulam Muhammad. "Voice intensity based gender classification by using Simpson's rule with SVM." In 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, 2012, pp. 552-555.

[22] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi. "Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS)." Expert Systems with Applications, vol. 106, 2018, pp. 77-91.

[23] Murthy, YV Srinivasa, and Shashidhar G. Koolagudi. "Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations." In 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2015, pp. 1271-1276.

[24] Murthy, Y. VS, T. K. R. Jeshventh, M. Zoeb, M. Saumyadip, and G. K. Shashidhar. "Singer identification from smaller snippets of audio clips using acoustic features and DNNs." In 2018 Eleventh International Conference on Contemporary Computing (IC3), . IEEE, 2018, pp. 1-6.

[25] Sharma, Rahul, YV Srinivasa Murthy, and Shashidhar G. Koolagudi. "Audio songs classification based on music patterns." In Proceedings of the second international conference on computer and communication technologies, Springer, New Delhi, 2016, vol. 381, pp. 157-166.

[26] Thomas, Matthew, YV Srinivasa Murthy, and Shashidhar G. Koolagudi. "Detection of largest possible repeated patterns in indian audio songs using spectral features." In 2016 IEEE Canadian conference on electrical and computer engineering (CCECE), IEEE, 2016. pp. 1-5.

[27] Jitendra, M. S. N. V., & Radhika, Y. "A review: Music feature extraction from an audio signal". International Journal of Advanced Trends in Computer Science and Engineering, vol. 9(2), 2020, pp. 973-980.

[28] Mason, J. S., and X. Zhang. "Velocity and acceleration features in speaker recognition." In Acoustics, Speech, and Signal Processing, IEEE International Conference on, pp. 3673-3674. IEEE Computer Society, 1991.

[29] Furui, Sadaoki. "Comparison of speaker recognition methods using statistical features and dynamic features." IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29(3), pp. 342-350, 1981.

[30] Qawaqneh, Zakariya, Arafat Abu Mallouh, and Buket D. Barkana. "Deep neural network framework and transformed MFCCs for speaker's

age and gender classification." Knowledge-Based Systems, vol. 115, pp. 5-14, 2017.

[31] Biswas, Roshni, YV Srinivasa Murthy, Shashidhar G. Koolagudi, and Swaroop G. Vishnu. "Objective Assessment of Pitch Accuracy in Equal-Tempered Vocal Music Using Signal Processing Approaches." In Smart Computing Paradigms: New Progresses and Challenges, vol. 766, pp. 161-168. Springer, Singapore, 2020.

[32] Murthy, YV Srinivasa, Shashidhar G. Koolagudi, and Vishnu G. Swaroop. "Vocal and Non-vocal Segmentation based on the Analysis of Formant Structure." In 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1-6. IEEE, 2017.

[33] Zeng, Yuni, Hua Mao, Dezhong Peng, and Zhang Yi. "Spectrogram based multi-task audio classification." Multimedia Tools and Applications, vol. 78(3), pp. 3705-3722, 2019.

[34] Russo, Mladen, Luka Kraljević, Maja Stella, and Marjan Sikora. "Cochleogram-based approach for detecting perceived emotions in music." Information Processing & Management, vol. 57(5), pp.102270. 2020.

[35] Costa, Yandre MG, Luiz S. Oliveira, and Carlos N. Silla Jr. "An evaluation of convolutional neural networks for music classification using spectrograms." Applied soft computing, vol. 52, pp. 28-38, 2017.

[36] Badshah, Abdul Malik, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. "Speech emotion recognition from spectrograms with deep convolutional neural network." In 2017 international conference on platform technology and service (PlatCon), pp. 1-5. IEEE, 2017.

[37] Ellis, Daniel PW. "Classifying music audio with timbral and chroma features". Paper presented at the Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, pp. 339-340.

[38] Grimaldi, Marco, and Fred Cummins. "Speaker identification using instantaneous frequencies." IEEE transactions on audio, speech, and language processing, vol. 16(6), pp. 1097-1111, 2008.

[39] Rami S. Alkhawaldeh, "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network", Scientific Programming, vol. 2019, Article ID 7213717, pp.1-12, 2019. https://doi.org/10.1155/2019/721371.

[40] Alam, Md Jahangir, Tomi Kinnunen, Patrick Kenny, Pierre Ouellet, and Douglas O'Shaughnessy. "Multitaper MFCC and PLP features for speaker verification using i-vectors." Speech communication, vol. 55(2), pp. 237-251, 2013.

[41] Gupta, Shruti, Md Shah Fahad, and Akshay Deepak. "Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition." Multimedia Tools and Applications, vol. 79, pp. 23347-23365, 2020.

[42] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. "A novel approach for MFCC feature extraction." In 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1-5. IEEE, 2010.

[43] Hu, Maodi, Yunhong Wang, Zhaoxiang Zhang, and Yiding Wang. "Combining spatial and temporal information for gait based gender classification." In 2010 20th International Conference on Pattern Recognition, pp. 3679-3682. IEEE, 2010.

[44] Kaluri, Rajesh, and P. Reddy. "Sign gesture recognition using modified region growing algorithm and adaptive genetic fuzzy classifier." Int J Intell Eng Syst 9 (2016): pp. 225-233.

[45] Kaluri, Rajesh, and C. H. Pradeep. "An enhanced framework for sign gesture recognition using hidden Markov model and adaptive histogram technique." Int J Intell Eng Syst 10 (2017):pp. 11-19.

[46] Kaluri, Rajesh, and Pradeep Reddy CH. "Optimized feature extraction for precise sign gesture recognition using self-improved genetic algorithm." Int. J. Eng. Technol. Innov 8, no. 1 (2018):pp. 25-37.

[47] Reddy, G. Thippa, M. Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. "Analysis of dimensionality reduction techniques on big data." IEEE Access 8 (2020): pp. 54776-54788.