Received: September 20, 2021. Revised: October 29, 2021. 262

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Bird Voice Classification Based on Combination Feature Extraction and Reduction Dimension with the K-Nearest Neighbor

Pulung Nurtantio Andono[1]* **Guruh Fajar Shidik[1]** **Dwi Puji Prabowo[1]**
Dewi Pergiwati[1] **Ricardus Anggi Pramunendar[1]**

[1]*Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia*
* Corresponding author's Email: pulung@dsn.dinus.ac.id

Abstract: Indonesia is rich in flora and fauna diversity, but it is experiencing an increasing number of endangered populations. The public can prevent the expanding population threatened with extinction by knowing the types of fauna, especially birds threatened with extinction. This study proposes an automatic grouping method to recognize bird sound patterns in an open environment. This experiment used bird data from a public dataset and obtained bird sound patterns by recording from a distance far enough to make sounds without feeling disturbed freely. However, the sound of birds may contain noise and need data processing using YAMME to be carried out the noise, and birds' voices can be separated. After that, the data was extracted using the combination of Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Cepstral Coefficients (GTCC) methods also reduced the dimensions of the feature before completing the identification. The bird identification obtained provides an accuracy performance that reaches 78.11%, and these results are higher than other feature extraction methods that also apply dimensional reduction.

Keywords: Bird sound, Feature segmentation, Extraction, Pattern recognition, Automatic recognition.

1. Introduction

Birds are experiencing an increasing threat of extinction. The increase is due to many illegal hunting and habitat changes [1–4]. Data from the World Bank in 2018 states that out of 4,584 bird species globally, 160 bird species are threatened with extinction. According to data obtained in the Red List of the World Conservation Agency (IUCN Red List) states that in 2018 there were 1,771 species of birds that are endemic to Indonesia. Based on the data that more than 9% of Indonesian bird species are threatened with extinction.

In reducing the number of endangered bird species and saving bird population habitats, the government has made conservation efforts in various areas. Conservation efforts must be carried out jointly by the community and researchers [5, 6]. The role of the community is to understand the species, classification, morphological characteristics, habitat and distribution, and protection of the threat animal status.

The difference knowledge becomes an obstacle in recognizing all types of birds, especially if done manually in the open based on the abilities possessed by the limited human sense of hearing. Limitations cause differences in performance in grouping [7]. The role of the researcher is to find a way out of rules to make it easier for people to understand the animals around them [8].

Many researchers are active in saving some endemics around them [9–11] because endangered animals are not cases from some countries and affect other countries. These efforts give fewer performance results when applied in classifying various bird species in the open. Therefore, this study proposes a new method to identify various bird species in different countries based on sound pattern data to address the gaps from previous studies. The pattern of bird sound is obtained from a combination of multiple features extracted according to the state-of-

the-art [9, 10, 12–14] of the original sound. The combination is used because noise resistance is a severe problem of signal recognition, and dimension reduction is used to reduce every signal data by utilizing the distribution of similarity of the characteristics of each data. The classification method used in this study uses the K-Nearest Network (KNN) method because it is very rarely used in classifying bird sounds.

This study contribution is validated by comparing the results of signal enhancement based on the classification method so that it can: (1) show the effect of a combination of feature extraction methods on classification performance; (2) analyze the relationship between feature extraction method and bird signal classification performance, and (3) determine the parameters needed to achieve the bird signal classification performance.

This paper is organized as follows: In Section 2, previous research on this research is highlighted. Section 3 presents our proposed model. Section 4 describes our experimental design. Next, Section 5 describes the results of the experiment and discussion. Finally, we conclude Section 6.

2. Related research

Several researchers from ornithologists have worked on classifying birds in the open. Automatic classification of bird species has been carried out based on image data [11, 15], and sound [9, 10, 12–14, 16–22], which is an essential computational tool in the field of ornithology, as well as conservation monitoring. Researchers are working to improve performance in classifying and categorizing bird species.

The recognition of birds based on sound patterns has been carried out by Briggs et al. [12] by proposing the nearest neighbor classification method using the Kullback Leibler Divergence and Hellinger matric. The combination of the spectral density and Mel-frequency cepstral coefficient (MFCC) feature extraction method is applied to the nearest neighbor classification method based on Hellinger Matric. It produces an accuracy rate of 92.10%. Stowell and Plumbey [13] compared twelve feature representations derived from the Mel spectrum and classified them using the random forest method. In this study, MFCC performed worse than the original Mel spectrum data, but the actual data impacted time. Raghuram et al. [9] also proposed a new framework for bird voice recognition, using 27 features, consisting of 4 pitch features, four energy features, four duration features, 13 MFCC features, and two tempo features. This study compared four classification methods as Naïve Bayes (NB), Neural Network (NN), Random Forest (RF), and Support Vector Machines (SVM). The results showed that the RF method obtained a maximum accuracy of 96.13%, while NB, NN, and SVM produced 91.16%, 93.93%, and 87.29%. Qian et al. [17] proposed a new framework by implementing reduced feature space dimensionality using refiefF. The classification method uses an extreme learning machine (ELM) and gives an average performance result of more than 80% based on the total species observed. Qian et al. [10] continued their research by proposing sparse-instance-based and least-confidence-score-based active learning methods to select informative data features automatically. The proposal obtained a performance of more than 85% by using the unweighted average recall (UAR) measurement method, which was applied to 3,483 audio data with a total of 60 bird species. Kahl et al. [16] used 36,496 audios consisting of 1500 species with the convolutional neural network (CNN) classification method. CNN produces features from a visual representation of audio recordings. The results in this study reached the mAP performance of 0.605. Bird recognition was also carried out by Supriya et al. [14] using the MFCC feature with Gaussian Mixture Model (GMM) and SVM classification method. The results show that the GMM classification method performance is higher than the SVM, 95% for GMM, while the SVM is 86%.

Research conducted by P. Jancovic and M. Kokuer [21] also did this study, which used a 48-species decomposition based on the sinusoid and expressed by frequency and magnitude values. The classification method is obtained by using the hybrid deep learning method with the Markov model. The performance received reached 98.7%. Using five local bird species, Ramashini et al. [20] performed a classification based on the Nearest Centroid (NC) classification method and compared them with SVM and KNN. This study also uses a reduction method based on Linear Discriminant Analysis (LDA) to get the best performance results. Sukri et al. [18] also used four species of birds. They utilized the power spectral density (PSD) as a feature extraction method to obtain power per unit of frequency and the neural network for classification. Chandu et al. [19] classified 400 samples of bird recording data using the CNN method based on the Alexnet architecture. In this study, feature extraction was carried out using spectrogram generation, where before the extraction process, the initial processing was carried out to make the sound better.

Based on that, many researchers used various

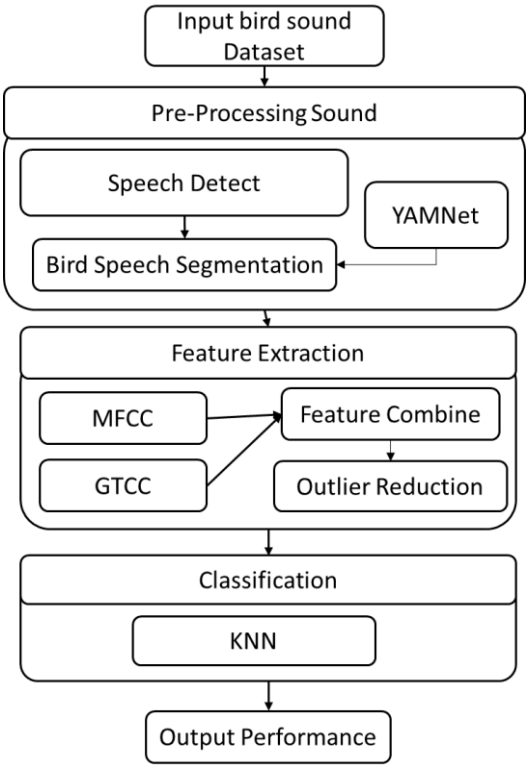

Figure. 1 Research proposed

datasets and methods to classify the bird voices. Different methods are carried out differently between researchers to improve the classification performance, from the feature extraction method to the classification method. On the other hand, many researchers use the same extraction and classification methods but use different data. Not all of these studies make use of techniques to reduce noise from their original audio quality. Furthermore, no research has been done on the use of feature extraction techniques to bird sound data.

3. The proposed approach

According to several studies, data, characteristics, and classification methods all influence the performance of classification methods. The classification method's performance demonstrates that essential characteristics do not always imply low accuracy. Proposed can take advantage of these properties to display accurate voice characteristics based on the purpose. The data source, however, has an impact on these aspects. This study presents a novel workflow for bird sound identification that combines the original bird voice attributes (see Fig. 1).

3.1 Data acquisition

The voice of birds is believed to have distinct qualities in this study, but the environment that impacts it has the necessary properties. As a result, to determine the optimal model, this study uses datasets collected from public sources. The original bird voice dataset, which included 21,375 bird sounds from 264 bird species, was obtained from the Cornell Lab of Ornithology's Center [23]. Due to the enormous amount of data and labels, this data into 18 experiments with 14 to 15 classes for each experiment.

3.2 Data processing

Data processing is carried out in several stages: preparing data and classifying bird voices based on the parts considered bird voice. The supplied audio files are of different duration and quality and have different sample rates, bit depths, and channels. In the first step, a data set is formed with file properties with homogeneous types. The experiment performed the training by extracting audio using time-coded annotations of the newly provided proper space validation metadata set. All files included in the training set are further processed to create additional data sets with different content.

At this stage, the speech boundary is detected in the audio signal, and then the voice is grouped or segmented, which is considered the voice of birds and eliminates other voices. This step uses the YAMNet Neural Network AudioSet ontology to get the part of the audio signal regarded as a bird voice [24], [25]. The YAMNet has 632 separate classes with 670 connections to accurately describe the existing voice characteristics in ontology visualization.

3.3 Feature extraction

Bird speech features are extracted via feature extraction. The birds voice must have certain qualities. As a result, at this point, a combination of Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Cepstral Coefficients (GTCC) characteristics should be used. This experiment decreases the quantity of data deemed an outlier from the resultant features and merging features. Outliers are chosen based on the feature group with the fewest groups provided in the k-Means Clustering-based clustering technique.

3.3.1. Mel frequency cepstral coefficients

Mel Frequency Cepstral Coefficients (MFCC) have been used in the field of speech recognition since their calculations are based on a perception-based frequency scale in the first stage (the Mel-scale-inspired model of human hearing). The steps of the MFCC method are shown as follows:

- Pre-processing

This stage is carried out to improve the quality of the incoming voice signal by eliminating noise from the voice signal. The search for endpoint detection is carried out and destroys the voice signal noise. Determination of the starting and ending points of the voice signal is carried out for noise search. The amplitude value of the high and low frequencies is balanced to minimize the noise that may still be present in the voice signal and obtain a better quality with a noise-free voice signal.

- Frame Blocking

The voice signal obtained will be separated into several frames. The length of the frame is separated from the voice signal by N. The M value represents the value that separates the frames or the number of overlaps to maintain the voice signal value, where $M < N$. The overlap parameter is provided to preserve the value in the frame; thus, the value is not lost during the following step. The number of frames in each voice signal is obtained based on Eq. (1), where M is the number of overlaps, N is the frame size, I is the sample rate, and $J(f)$ is the number of frames.

$$\mathrm{J(f)} = \frac{\mathrm{I} - \mathrm{N}}{\mathrm{M}} + 1 \tag{1}$$

- Windowing

Windowing is done to produce distortion between frames and the signal in the frame to avoid breaking continuity. This step is done by duplicating each sample frame from the starting point to the endpoint of the frame. It is used to improve the continuity of the voice signal at the start and endpoints of the frame. Hamming windows are used for the windowing process in the hope of being able to produce a precise and value extraction process from the voice signal and can be shown in Eq. (2). This value of an is commonly approximated α as 0.54, with N is the number of samples to be processed and n is number of samples.

$$h(n) = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N}\right), 0 \le n \le N-1$$
$$h(n) = 0, otherwise \tag{1}$$

- Discrete Fourier Transform (DFT)

DFT converts a signal from the time domain to the frequency domain and is carried out on each frame that has maintained continuity and calculates each power spectrum, which is needed to find out the frequency that appears in the frame. Therefore, we get the frequency and power spectrum in each frame. DFT is applied to a windowed signal representing the magnitude and phase of the signal using Eq. (3). N is the maximum number of samples to be processed, h_n represents the sample signal value, k represents the discrete frequency variable and i always $\sqrt{-1}$.

$$\mathrm{H_k} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{h_n e^{-i\frac{2\pi kn}{N}}} \tag{3}$$

- Mel Filtering

The range of human perception can be determined using the Mel Scale and modeled in the human auditory system using a 24-band filter bank. The results from the previous stage, the DFT spectrum, do not consider that human hearing is less sensitive at frequencies above 1000 Hz. The DFT calculation is only related to the linear frequency scale, so a warping frequency process is needed, or the frequency spectrum is converted to a smaller number using a logarithmic Mel scale. Mel Filtering filters the voice signal processed in the previous stage and creates a pattern called the Mel-spectrum. Steps before filtering, it is necessary to determine the value of the filter bank. This filter bank effectively maps the center of the DFT frequency bin. After the value is specified, the filtering process can be applied to the processed voice signal to produce a Mel-spectrum. The equation of mel-filtering can be shown in Eq. (4).

$$\mathrm{mel(f)} = 1127 \ln\left(1 + \frac{\mathrm{f}}{700}\right) \tag{4}$$

- Inverse Discrete Fourier Transform (IDFT)

The IDFT of the Mel-Spectrum was calculated to produce the Mel-Frequency Cepstral Coefficients. Signal analysis cepstral domain proved helpful given its invariance of invariance concerning linear spectral distortion. The value of the Cepstrum contains meaningful information to give the unique characteristics of the waveform. The resulting MFCC features are 39 features consisting of 13 MFCC values and 26 deltas MFCC.

3.3.2. Gammatone frequency cepstral coefficients

Gammatone Cepstral Coefficients (GTCC) is a feature based on a set of Gammatone Filter banks. The cochleagram is a time-frequency representation of the signal obtained from the output of the Gammatone filter bank. A cochleagram is used to calculate GTCC features and the calculation stages are similar to the MFCC.

- Gammatone Filter

Gammatone filters are designed to simulate the

processes of the human auditory system. A Gammatone filter $g(t)$ with a center frequency (f_c) can be defined in Eq. (5). The Gammatone filter $g(t)$ is described by the parameters order n (integer), ringing frequency c (rad/s), beginning phase φ (rad), and one-sided pole bandwidth b (rad/s) is obtained by Eq. (6). The phase φ is close to zero, the variable a controls the gain value, the filter order is determined by the weight of n set to a value less than four. In the filter response, a is an arbitrary factor commonly used to make the peak gain equal unity.

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c + \varphi) \quad (5)$$

$$b = 25.17\left(\frac{4.37F_c}{1000} + 1\right) \quad (6)$$

In obtaining a representation similar to the FFT-based spectrogram, a set of Gammatone filters, are considered channels with different center frequencies for the Gammatone filter bank.

- Windowing

The GTCC, similar to MFCC, requires a window to cover K points and shift each L point in each frame. Each frame is defined by $x(t; f_c(m))$ with the center of the frequency (f_c) in m filter. The resulting Cochleagram representation for each frame is calculated on average across the t window and is defined in Eq. (7). Where γ is the dependent factor in frequency, and the other represents the magnitude of the complex number. M is the number of filter bank channels with K values of 400, L of 160, and M of 32 for 16 kHz signals that produce 100 frames per second.

$$\bar{x}(t; f_c(m)) = \frac{1}{K}\sum_{L=0}^{K-1} \gamma|x(tL + i; f_c(m)) \quad (7)$$

- Discrete Continue Transform (DCT)

DCT was applied to obtain uncorrelated cepstral coefficients. Similar to the MFCC operation in Eq. (8) and the range u starts from 0 to 31.

$$h_k = 2\sum_{n=0}^{N-1} x(n)\cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \quad (8)$$

N is the maximum number of samples to be processed, h_n represents the sample signal value, k are represents the continue frequency variable. The features provided by the GTCC method produce 39 features consisting of 13 GTCC values and 26 GTCC deltas.

3.3.3. Outlier reduction using k-means clustering

The resulting data is sometimes redundant, and if it is assumed that the feature sets are correlated, it can reduce the feature vector without losing much information. k-Means clustering can reduce the amount of redundant data for a label without losing much information. Outlier reduction is made by grouping the resulting data and selecting the data that has the most significant members. The data is generated every time the voice signal is performed feature extraction. It does not detract from the data and produces the best quality information because it is gathered in the same group.

The k-Means method groups the existing data into several groups. The data in one group have the same characteristics and different characteristics from the data in other groups. They minimized the objective function by minimizing variations between data in a cluster and maximizing variation with data in other sets. The objective function used is based on Eq. (9). The frequency of the signal employed is f, and the centroid is ce, which would be selected at random.

$$dist = \sqrt{(f - ce)^2} \quad (9)$$

3.4 Classification step

K-nearest neighbors (KNN) is a simple algorithm that stores all available cases and classifies new topics based on similarity measures. There are training data sets, and each data set is labeled. When entering new data without labels, compare features according to the training set to find the closest k-similar. KNN has high accuracy, is not sensitive to outliers, and does not require data input assumptions. However, K-Nearest neighbors have high computational complexity and high spatial complexity.

3.5 Performance evaluation step

Performance evaluation of the classification method is done by calculating the accuracy. Accuracy is defined as the correct classification of all data obtained. The accuracy value is obtained using Eq. (10) with t and n as the number of correctly classified sample data and n as the total sample data.

$$accuracy = \frac{t}{n} \times 100 \quad (10)$$

4. Experiment design

The investigation was carried out to determine the performance of the suggested technique using accurate performance data from the Cornell Lab of Ornithology Center for Conservation Bioacoustics (CCB). This study suggests combining the Mel Frequency Cepstral Coefficients (MFCC) technique and the Gammatone Cepstral Coefficients (GTCC) method for feature extraction. The k-Means technique lowers the quantity of data in a signal by grouping it according to the most crucial cluster and treating the least number of clusters as an outlier.

The KNN approach is used to classify data and obtain an accuracy number that is compared to standard learning algorithms such as naïve Bayes (NB), neural networks (NN), and decision trees (DT). Other methods of feature extraction were also used to make comparisons. The parameters' default values restrict the use of parameters in this study. The MFCC and GTCC feature extraction methods, employ windows in real vectors in the form of a hamming function with a size of 0.03 times the frequency value, which is repeated regularly. Size An integer with a value of 0.02 times the frequency value is used to specify the overlapping length of adjacent windows. The window limit in detecting voice is likewise calculated with a value of 0.02. The MFCC and GTCC feature each have 39 features, resulting in a total of 78 features. Simultaneously, the quantity of data acquired through feature extraction and data reduction was 625,381 data from 21,375 original data. Each learning algorithm's parameter settings are also done by default. MATLAB R2021a version 9.10.0 was utilized in this study to improve the signal, feature extraction, and classification.

5. Results and discussion

This discussion is divided into three stages: initial processing, feature extraction, and classification. The results are shown for each step in detail, accompanied by a discussion.

5.1 Pre-processing voice

The data shown in Fig. 2 shows that not all data is the voice of birds. Therefore, in this study, a search for good boundaries was carried out to obtain features that only consisted of bird voices. Fig. 3 has shown as many as nine parts that do not contain voice signals. When these restrictions are eliminated and the voice signals are merged, the result is a complete signal comprising the voices of the birds seen in Fig. 4. As a result, the signal is intended to offer visible characteristics without any voice other than birds'. In addition, to recognize bird sounds, this study used YAMNet, which is based on a convolution neural network (CNN). Table 1 shows the findings obtained from the original voice data or without identification and deletion of non-speech data. These findings demonstrate that most of the voices extracted from the voice data are bird sounds; nevertheless, as shown in Fig. 5 and Table 1, some aspects of the voice are non-bird. The voice that used cleaning noise to minimize noise is shown in Table 2 and Fig. 6. The result demonstrates that the data has been cleansed of non-bird voices. Audio signals with identified labels in voice areas are shown in Fig. 5 and Fig. 6.

5.2 Feature extraction

The signal quality is enhanced by extracting characteristics from the signal data and converting

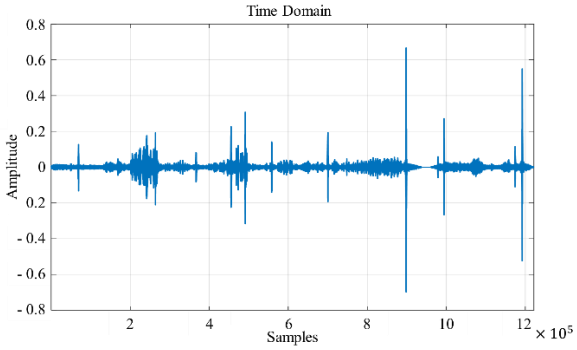

Figure. 2 Original bird audio

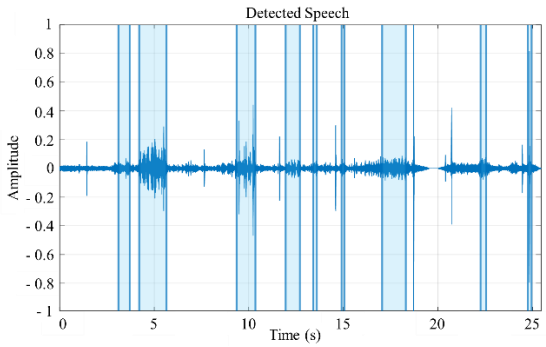

Figure. 3 Bird voice and other

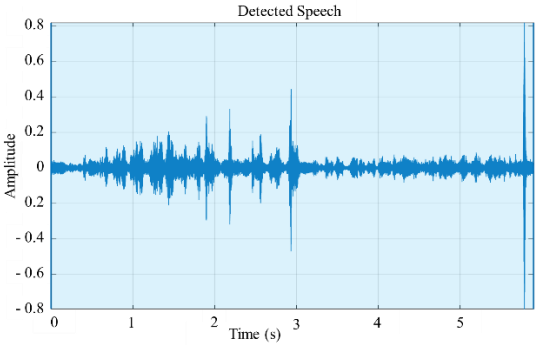

Figure. 4 Clear bird voice

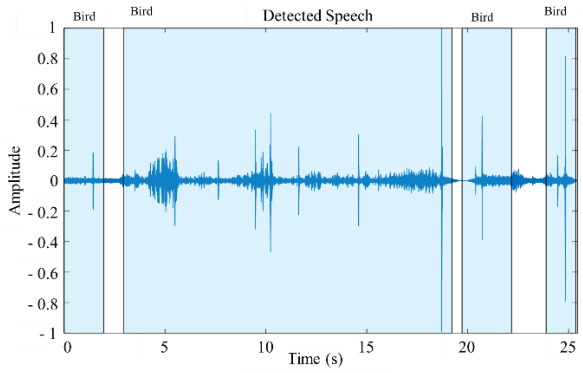

Figure. 5 Spectral entropy of audio signal

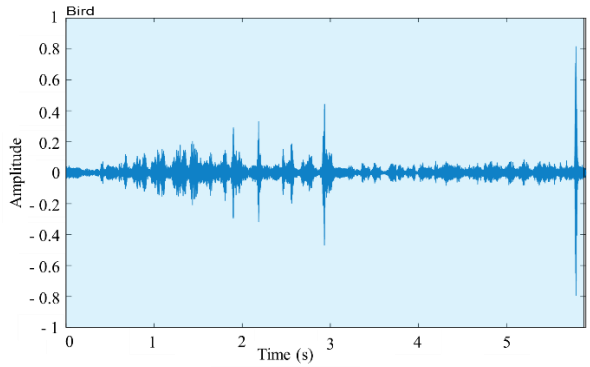

Figure. 6 Clear noise audio signal

Table 1. Detected region

Time Stamp (s)		Label	Mean	Max Score
0.01	1.96	Bird	0.56	0.75
		Bird vocalization, bird calls, bird song	0.51	0.67
		Chirp, tweet	0.42	0.58
2.94	19.23	Bird	0.78	0.95
		Bird vocalization, bird calls, bird song	0.75	0.93
		Chirp, tweet	0.60	0.83
19.72	22.17	Bird vocalization, bird calls, bird song	0.58	0.82
		Bird	0.57	0.84
		Chirp, tweet	0.47	0.67
23.89	25.36	Bird	0.57	0.65
		Bird vocalization, bird calls, bird song	0.37	0.49

them into numbers that machine learning algorithms can identify. Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Cepstral Coefficients (GTCC) were used to extract features from a total of 21,375 bird sounds. Because each data extraction result from signal feature extraction offers actual data,

Table 2. Detected region

Time Stamp (s)		Label	Mean	Max Score
0.00	5.88	Bird	0.94	0.98
		Bird vocalization, bird calls, bird song	0.92	0.97
		Chirp, tweet	0.84	0.90

Table 3. Feature extraction applied using data reduction

No	mfcc1	mfcc2	.	gtcc1	.	gtcc78
1	0.47	-1.50	.	-0.67	.	-0.21
2	0.81	-0.84	.	0.09	.	-0.56
3	-1.10	-1.79	.	0.37	.	-0.74
4	1.22	-0.78	.	0.18	.	-0.79
5	-0.93	-1.48	.	-0.32	.	-0.64
6	-0.83	-1.19	.	-0.20	.	-0.28
7	0.47	-1.50	.	-0.10	.	0.10
.	.	.	.	.	.	.
.	.	.	.	.	.	.
625,381	-0.69	-0.78		-0.84		0.04

which is 16,924,040 data extracted with 78 data features, the two extractions are merged, and data reduction is performed. The k-Means clustering method is used to reduce data by identifying the most prominent total cluster member from the number of data that have been clustered on each data. The outcome of k-Means clustering data reduction is 625,831 data with 78 characteristics. k-Means helps reduce training time by decreasing the number of data records generated by feature extraction.

5.3 Classification

The MFCC and GTCC feature outputs are combined and submitted to k-Means clustering to reduce the amount of data. The next step is to determine the accuracy of the data. We propose the k-Nearest Neighbor (KNN) method to get the best identification results. The findings reveal that the proposed method's combination of characteristics produces the best accuracy, averaging 78.11% when feature techniques are integrated with the KNN algorithm. As demonstrated in Fig. (7), the combination method of GTCC and MFCC delivers a slight performance improvement over the MFCC method proposed by [9, 13]. The proposed method incorporates attributes that disclose more information about each bird's voice characteristics than the GTCC features. When compared to our suggested spectrum, the essential Mel and Bark spectrums perform poorly.

The study comprised 18 experiments, each

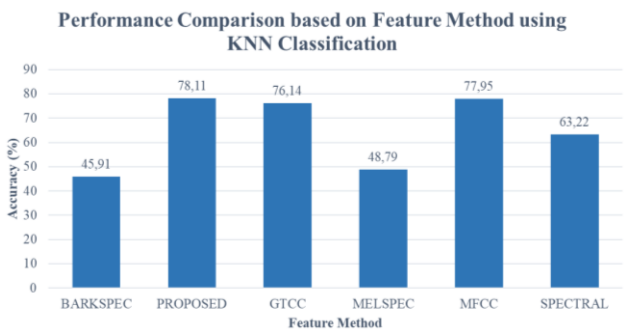

Figure. 7 Performance comparison based on feature method using KNN classification

Table 4. Performance comparison based on number of experiments

Experiment	Proposed	MFCC	GTCC
1	78.66	77.44	75.78
2	77.20	76.59	75.15
3	79.77	79.73	77.51
4	78.37	78.41	75.35
5	78.68	78.34	76.68
6	76.82	76.39	74.35
7	77.89	77.73	76.30
8	79.84	79.70	77.96
9	78.98	78.61	76.61
10	76.71	76.01	74.23
11	77.56	77.85	75.45
12	80.39	80.51	79.59
13	75.28	75.46	73.62
14	77.42	75.92	75.28
15	79.59	79.82	78.14
16	78.50	78.91	77.82
17	77.20	76.61	74.25
18	77.12	78.71	76.51

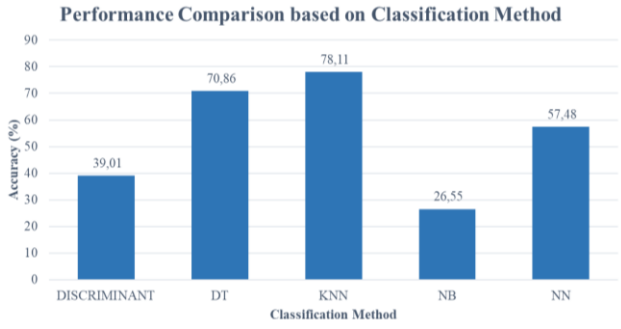

Figure. 8 Performance comparison based on the classification method

conducted out on 14 to 15 separate labels each experiment. Fig. 7 shows the experimental results using the MFCC and GFCC feature extraction approaches, and the data demonstrate that the suggested method performs the best. However, as seen in Table 4, not all of the proposed configurations follow the same trend. The presented technique outperforms the MFCC feature extraction method in 11 experiments. Furthermore, in terms of performance, the proposed scheme surpasses the GFCC feature extraction method. The above is possible because the recommended approach, which includes MFCC, corrects the flaws in the GFCC method.

The experiment's findings indicated that the GFCC and MFCC performance outcomes are not substantially different. However, the results of the Statistics Friedman test show that asymptotic significance value of 0.000 < 0.05. Then HO is rejected, and Ha is accepted, or in other words, there is a difference in the average recognition performance of the three voice feature extraction methods. Thus, the proposed method by combining both MFCC and GFCC features can provide high recognition performance.

GFCC uses the equivalent rectangular bandwidth (ERB) scale, whereas MFCC employs the mel-scale. In the process, GFCC employs cubic root operations, whereas MFCC uses log operations. The technique changes the convolution between the excitation source and the vocal tract to add in the spectrum domain. As a result, GFCC is less susceptible to noise than MFCC. When it comes to accuracy, however, MFCC makes greater use of the noise that already exists and produces somewhat better outcomes than GFCC. Consequently, the proposed scheme may be considered to have generated good results with the excellent accuracy.

5.4 Evaluation using several classification methods

This study uses several classification methods to assess the extracted features and compare the original unprocessed signal data. Naive Bayes (NB), Neural Network (NN), k-Nearest Neighbor (KNN), and decision tree (DT) are some of the classification methods employed [9]. When several classification methods were evaluated, the KNN classification method produced the best results, with an accuracy value of 7.25% greater than the DT method and considerably higher than the other classification methods (see Fig. 8).

The KNN algorithm does not require any training time, whereas neural network training takes an extended period. However, if you have many data points and do not utilize forecast lookups, KNN may take longer to evaluate. Compared to DT, adding updated information to an existing model In KNN,

adding that point to a current data set is possible; however, updated information cannot modify the model in DT; therefore, DT must create a new tree from scratch. The KNN method is relatively simple and only requires the setting of one hyperparameter (the k-value). In contrast, neural network training requires a set of several hyperparameters that determine the network's size and structure, as well as optimization techniques. The aforementioned is what causes KNN to get better performance than NN. In its implementation, NB performs poorly since it is a parametric classification technique and a generative model that creates new instances. Despite this, KNN is a real-time technique with a high computational value.

6. Conclusion

Our goal is to protect endangered bird populations, arguably one of the most challenging issues many countries face. Some automatic bird voice recognition systems, on the other hand, continue to provide high feature findings, resulting in the usage of computing devices. This study can finish the identification stage by decreasing the number of records from the extraction results and using limited computing resources during the identification process.

Furthermore, when the MFCC and GTCC characteristics are combined with the k-Means technique to minimize the number of records before identification, the accuracy performance gained improves by 0.16% compared to the MFCC and 1.97% compared to the GTCC, respectively. This combination proposed yields a better outcome than the previous feature approaches. These findings suggest that our suggested technique may increase the accuracy of bird sound recognition with our restricted resources. Future studies will hopefully be able to find and test the ability of additional reduction strategies and enhance performance.

Conflicts of interest

Following the International Journal of Intelligent Engineering and Systems, policy, and my ethical obligation as a researcher, I am reporting that this paper has not been published previously, has not been copyrighted, has not been submitted elsewhere. I have disclosed those entirely to the International Journal of Intelligent Engineering and Systems and have gotten approval from all authors to manage any potential conflicts from that research.

Author contributions

Conceptualisation, Pulung Nurtantio Andono; methodology, Pulung Nurtantio Andono and Ricardus Anggi Pramunendar; software, Ricardus Anggi Pramunendar; investigation, Dwi Puji Prabowo; formal analysis, Ricardus Anggi Pramunendar, and Dwi Puji Prabowo; resources, Ricardus Anggi Pramunendar and Dwi Puji Prabowo; data curation, Dwi Puji Prabowo and Dewi Pergiwati; writing—original draft preparation, Ricardus Anggi Pramunendar; writing—review and editing, Pulung Nurtantio Andono and Guruh Fajar Shidik; visualisation, Ricardus Anggi Pramunendar; supervision, Pulung Nurtantio Andono and Guruh Fajar Shidik; project administration, Dwi Puji Prabowo and Dewi Pergiwati; funding acquisition, Ricardus Anggi Pramunendar.

Acknowledgments

This work is funded by the Indonesian Ministry of Research and Higher Learning (DPRM-DIKTI) and supported by the Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang.

References

[1] S. Chng and J. Eaton, "In the market for extinction : Eastern and Central Java Eastern and Central Java", *Traffic Report*, No. 2016.

[2] V. Nijman, S. L. Sari, P. Siriwat, M. Sigaud, and K. A. I. Nekaris, "Records of 4 CE songbirds on the markets of Java suggests domestic trade is a major impediment of their conservation", *BirdingASIA*, Vol. 27, No. 2016, pp. 20–25, 2017.

[3] S. H. M. Butchart, J. P. W. Scharlemann, M. I. Evans, S. Quader, S. Arico, J. Arinaitwe, M. Balman, L. A. Bennun, B. Bertzky, C. Besancon, T. M. Boucher, T. M. Brooks, I. J. Burfield, N. D. Burgess, S. Chan, R. P. Clay, M. J. Crosby, N. C. Davidson, N. D. Silva, C. Devenish, G. C. L. Dutson, D. Fernandez, L. D. C. Fishpool, C. Fitzgerald, M. Foster, M. F. Heath, M. Hockings, M. Hoffmann, D. Knox, F. W. Larsen, J. F. Lamoreux, C. Loucks, I. May, J. Millett, D. Molloy, P. Morling, M. Parr, T. H. Ricketts, N. Seddon, B. Skolnik, S. N. Stuart, A. Upgren, and S. Woodley, "Protecting Important Sites for Biodiversity Contributes to Meeting Global Conservation Targets", *PLoS ONE*, Vol. 7, No. 3, p. e32529, 2012.

[4] E. Buechley and C. H. Sekercioglu, "Endangered species", *Grzimek's Animal Life Encyclopedia*, No. 2013, pp. 159–176, 2018.

[5] V. Capmourteres and M. Anand, "Conservation value: a review of the concept and its quantification", *Ecosphere*, Vol. 7, No. 10, pp. 1–19, 2016.

[6] N. J. Bennett, R. Roth, S. C. Klain, K. Chan, P. Christie, D. A. Clark, G. Cullman, D. Curran, T. J. Durbin, G. Epstein, A. Greenberg, M. P. Nelson, J. Sandlos, R. Stedman, T. L. Teel, R. Thomas, D. Veríssimo, and C. Wyborn, "Conservation social science: Understanding and integrating human dimensions to improve conservation", *Biological Conservation*, Vol. 205, pp. 93–108, 2017.

[7] R. A. Pramunendar, S. Wibirama, P. I. Santosa, P. N. Andono, and M. A. Soeleman, "A Robust Image Enhancement Techniques for Underwater Fish Classification in Marine Environment", *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 5, pp. 116–129, 2019.

[8] N. A. Rose and E. C. M. Parsons, "'Back off, man, I'm a scientist!' When marine conservation science meets policy", *Ocean & Coastal Management*, Vol. 115, pp. 71–76, 2015.

[9] M. A. Raghuram, N. R. Chavan, R. Belur, and S. G. Koolagudi, "Bird classification based on their sound patterns", *International Journal of Speech Technology*, Vol. 19, No. 4, pp. 791–804, 2016.

[10] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active Learning for Bird Sounds Classification", *Acta Acustica United with Acustica*, Vol. 103, No. 3, pp. 361–364, 2017.

[11] J. Atanbori, W. Duan, J. Murray, K. Appiah, and P. Dickinson, "Automatic classification of flying bird species using computer vision techniques", *Pattern Recognition Letters*, Vol. 81, pp. 53–62, 2016.

[12] F. Briggs, R. Raich, and X. Z. Fern, "Audio classification of bird species: A statistical manifold approach", In: *Proc. of IEEE International Conference on Data Mining, ICDM*, pp. 51–60, 2009.

[13] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning", *PeerJ*, Vol. 2, No. 1, p. e488, 2014.

[14] Supriya, R. Prakrithi, S. Bhat, Shivani, and S. Santhosh, "Classification of birds based on their sound patterns using GMM and SVM classifiers", *International Research Journal of Engineering and Technology*, Vol. 5, No. 2004, pp. 4708–4711, 2018.

[15] A. Marini, J. Facon, and A. L. Koerich, "Bird species classification based on color features", *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp. 4336–4341, 2013.

[16] S. Kahl, T. W. Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks", *CEUR Workshop Proceedings*, Vol. 1866, 2017.

[17] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine", *The Journal of the Acoustical Society of America*, Vol. 142, No. 4, pp. 1796–1804, 2017.

[18] M. M. M. Sukri, U. Fadlilah, S. Saon, A. K. Mahamad, M. M. Som, and A. Sidek, "Bird Sound Identification based on Artificial Neural Network", In: *Proc. of 2020 IEEE Student Conference on Research and Development (SCOReD)*, 2020, pp. 342–345.

[19] B. Chandu, A. Munikoti, K. S. Murthy, G. Murthy V, and C. Nagaraj, "Automated Bird Species Identification using Audio Signal Processing and Neural Networks", In: *Proc. of 2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, pp. 1–5, 2020.

[20] M. Ramashini, P. E. Abas, U. Grafe, and L. C. D. Silva, "Bird Sounds Classification Using Linear Discriminant Analysis", In: *Proc. of 2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1–6, 2019.

[21] P. Jancovic and M. Kokuer, "Bird species recognition using unsupervised modeling of individual vocalization elements", *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 27, No. 5, pp. 932–947, 2019.

[22] L. Müller and M. Marti, "Bird sound classification using a bidirectional LSTM", In: *Proc. of CEUR Workshop Proceedings*, Vol. 2125, 2018.

[23] J. Stastny, M. Munk, and L. Juranek, "Automatic bird species recognition based on birds vocalization", *Eurasip Journal on Audio, Speech, and Music Processing*, Vol. 2018, No. 1, 2018.

[24] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 776–780, 2017.

[25] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures

for large-scale audio classification", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings,* pp. 131–135, 2017.