Natural Language Processing (2024), 1–19
doi:10.1017/nlp.2024.23

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Resource building and classification of Mizo folk songs

Esther Ramdinmawii and Sanghamitra Nath

Department of of Computer Science & Engineering. Tezpur University, Napaam, Sonitpur, Assam, India
Corresponding author: Esther Ramdinmawii; Email: esther.rdmchhakchhuak@gmail.com

(Received 24 December 2022; revised 5 November 2023; accepted 4 January 2024)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

Folk culture represents the social, ethnic, and traditional livelihood of people belonging to a certain tribe or community and is important in keeping their culture and tradition alive. The Mizo people are a Tibeto-Burmese ethnic group, native to the Indian state of Mizoram and neighboring regions of Northeast India. Mizo folk culture is an amalgamation of festivity, celebration, liveliness, kinship, brotherhood, and merriment, and above all, preserves the ethnicity of this tribal community that is fundamentally entrenched. Unfortunately, the Mizos are fast giving up their old customs and adopting the new mode of life that is greatly influenced by the western culture. This makes it all the more crucial to preserve the intangible cultural heritage of this ethnic tribe whose folk cultures are vanishing day by day. To the best of our knowledge, this work is the first attempt at preservation and classification of Mizo folk songs. The first part of this paper presents a literature survey on preservation, analysis, and classification of folk songs. The second part presents the methodology for preliminary classification of Mizo folk songs. Three categories of Mizo folk songs—Hunting chants (*Hlado*), Children's songs (*Pawnto hla*), and Elderly songs (*Pi pu zai*)—are used in this study. A total of 29 acoustic features are used. A long short-term memory network using custom attention layer has been proposed for classification, whose results are compared with four supervised models (Support Vector Machines, K-Nearest Neighbor, Naive Baye's, and Ensemble). Experimental results from the proposed model are promising, with an implication of scope for future research in acoustic analysis and classification of Mizo folk songs using recent unsupervised methods.

Keywords: low-resource regional songs; acoustic analysis; Mizo; folk song resources; folk song classification

1. Introduction

Folk music embodies a profound legacy and diversity while also transcending cultural boundaries with its universal language. It serves as a means to safeguard our cultural heritage and pass down our history to future generations, ensuring the preservation of our rich legacy. Folk music in India encompasses a vast array of traditions, each reflecting the unique cultural heritage of its respective region. The Northeastern states of India exhibit a particularly diverse and rich folk music tradition. These cultures have a vibrant tapestry of folk songs that are deeply rooted in their history, customs, and rituals. These songs provide insights into the local customs, beliefs, and values of the communities they belong to.

One such community is the Mizo community from Mizoram, which is the southernmost state among the Northeastern states of India, as shown in Figure 1. The Mizo tribe has a significant population, with over 8,40,000 speakers (as per 2011 census) within Mizoram as well as its neighboring states such as Manipur, Tripura, and Assam. There are also Mizo-speaking populations

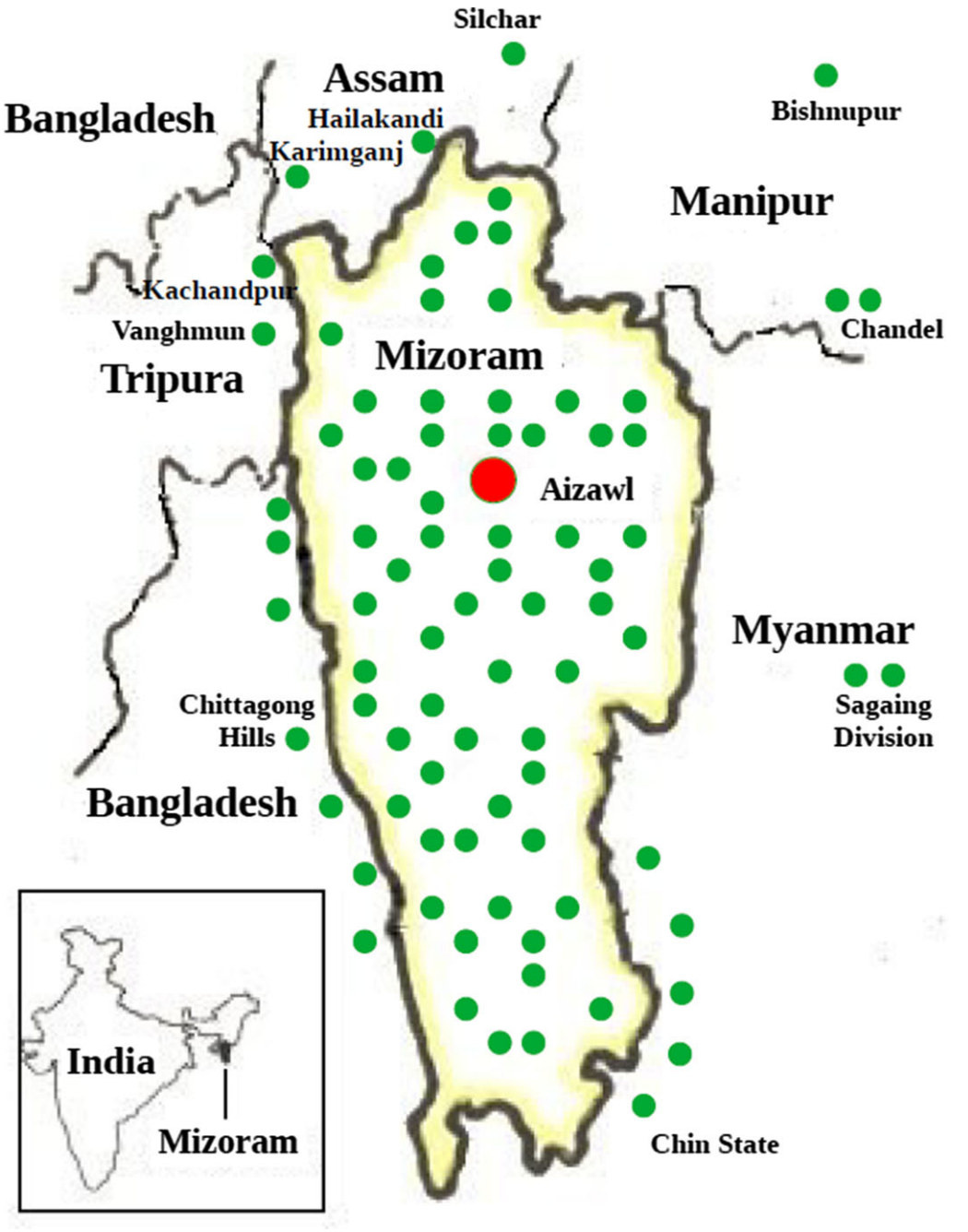

Figure 1. Speaker population of Mizo language[a] (marked in green dots).

in certain parts of Bangladesh, toward the western border of Mizoram, as well as in Myanmar, on the eastern border of Mizoram. Various sub-tribes like Thado, Paite, Lusei, Pawi, and others reside within the Mizo community, with the Lusei dialect adopted as the lingua franca of modern-day Mizoram. The Mizo language belongs to the Tibeto-Burman language family (Weidert 1975),

[a]https://www.kamat.com/kalranga/nindia/mizoram/map.htm

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

specifically the Kuki-Chin subgroup. Mizo tribe and their language, with their vibrant cultural traditions and linguistic heritage, contribute to the diverse tapestry of North-East India's rich ethnic and linguistic mosaic.

The Mizos have a rich and diverse cultural heritage. Their folklore is naturally passed down from previous generations orally. Mizo folk songs depict stories of the Mizo society, tradition, and culture, at a certain time in history. They reflect the glorious past of the Mizos, including their way of farming and harvesting, hunting, war, natural disasters, romance and nuptials, place of females in social strata, place of males in society, etc.

According to the works of Thanmawia (1998), Lalthangliana (2005), and Lalremruati (2012, 2019), Mizo folk songs can be categorized under five main themes depending on their purpose and use—*hunting and war chants, lamentations, satire, love, and nature* themes. Songs are also categorized depending on their types of tune, called *thlûk* (Lalthangliana 1993; Khiangte 2001, 2002; Lalzarzova, 2016). This is termed as 'Hlabu'. Those having the same tune or melody are called 'hlabu khat'. Different categories from these earlier works are discussed in brief as follows:

1. **War chants:** *Bawh hla* (war chants) are chanted solo by warriors after a successful war or raid where they have taken the head of an enemy. These personal and subjective chants are spontaneous and convey the singer's emotions and mood, reflecting a sense of pride and ego (Lalremruati 2019).
2. **Hunting chants:** *Hlado* (hunting chants) share the same melody as *Bawh hla* and are spontaneously composed and chanted after a triumphant hunt. They typically emphasize the singer's supremacy over the common man, employing words that express the singer's ego and pride (Lalremruati 2019).
3. **Lamentations:** Songs of this nature are traditionally chanted during times of adversity. The Mizos experienced famines during their settlement in the Than ranges, leading to significant loss of life (Lalremruati 2019). These songs, known as *'ṭhuthmun zai'* (songs sung while sitting), emerged as a way for people to offer condolences and gather together, sitting and singing these songs in solidarity (Thanmawia 1998; 1998, Lalremruati 2012, 2019).
4. **Satire:** These songs serve the purpose of ridicule and can be both aggressive and offensive, but they also encompass cheerful and humorous elements. Known as *'intuk hla'*, they are utilized to lighten the mood in gatherings that are otherwise heavy and tense (Thanmawia 1998; Lalthangliana 2005; Lalremruati 2019).
5. **Nature themed:** Mizo folk songs frequently depict the beauty of nature and its influence on society, emphasizing the Mizo people's reliance on and connection with the natural world, including its flora and fauna. These songs often draw parallels between elements of nature and the affection shared between couples, intertwining themes of nature and love (Lalremruati 2019).
6. **Couplet and triplet:** Mizo folk songs can be categorized based on the number of lines they contain. The first form of folk song, known as a couplet (*tlar hnih zai*), is believed to consist of two lines (Lalthangliana 1993). It is further believed that earlier songs primarily comprised couplets and triplets (Khiangte 2001).
7. **Songs named after individuals:** Mizo folk songs are also categorized according to the names of their original composers. Subsequently, other composers utilize the same tune to create different lyrics, a practice often done as a tribute to honor the original composer (Lalremruati 2012; Lalzarzova 2016).
8. **Songs named after merry and festive occasions:** The Mizos celebrate various festivals, many of which are connected to the agricultural season. Among the most common ones are Chapchar Kût, Mim Kût, and Ṭhalfavang Kût. Chapchar Kût marks the joyous completion

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

of rice plantation, Thalfavang Kût celebrates the harvest, while Mim Kût is a solemn festival dedicated to the souls of the deceased, featuring rituals, feasting, and mournful singing and dancing (Khiangte 2002).

The Indian Government has implemented heritage preservation schemes that aim to preserve and promote oral traditions, performing arts, social and ritual events, etc., from various states. Projects by the All India Radio and the Indira Gandhi National Centre for the Arts[b] focus on preserving dying folk songs and classical Indian music; they have not yet included Mizo folk songs or the Mizo language in the Technology Development for Indian Languages[c] repository. It is crucial to protect the language, culture, and traditions of the Mizo community, as Mizo is classified as 'vulnerable' on the 2010 UNESCO list of endangered languages (Moseley 2010).

The cultural development of Mizo society has been significantly influenced by the impact of globalization, which has gradually diminished the significance of traditional folk songs due to linguistic changes in the Mizo language. The vocabulary of these songs differs from spoken Mizo and includes borrowings from the Paite dialect, making it challenging to sing or comprehend the lyrics. As a result, passing down this cultural heritage to the younger generation has become increasingly difficult amidst the rapid social changes, depriving them of access to and practice folk tales, which are vital for maintaining cultural roots. Thus, preserving folk songs has become even more crucial.

Hence, this work aims to address this issue by proposing a framework for preservation and classification of Mizo folk songs. The main contributions of this work are listed below:

- Creation of Mizo folk song database for research in music processing.
- Utilization of the database toward identification of unique acoustic characteristics of the Mizo folk songs from a speech processing point of view.
- Acoustic classification of Mizo folk song categories using a long short-term memory (LSTM) network with custom attention layer (LSTM-attn).

The paper is structured as follows. Section 2 presents a brief summary of existing literature on Music Information Retrieval (MIR), existing analysis methods, features, and classification methods of folk songs in other languages. Section 3 discusses the methodology including data used, acoustic features employed, as well as detailed discussion of the proposed LSTM-attn model. Section 4 discusses the experiments and results. Section 5 concludes with a summarization of the work, its limitations, and future scope.

2. Literature survey

2.1 Music information retrieval and recent techniques

MIR deals with problems of music access, filtering, tool development, and retrieval (Orio, 2006). According to (Orio, 2006), the applications of MIR are intended to help users find specific music in a large collection by a particular similarity matching technique and criteria. Major tasks in MIR include (i) audio fingerprinting, (ii) audio-textual alignment, (iii) cover song identification, (iv) music genre identification and classification, and (v) music recommendation, among others (Srinivasa Murthy and Koolagudi 2018; Blaß and Bader, 2019). Our proposed framework focuses on the tasks of music identification and classification. The basic framework of an MIR system is shown in Figure 2.

In Deruty *et al.* (2022), music production of contemporary pop music is carried out using AI tools. Different musical instrument sound generation tools are utilized with automatic music

[b] https://ignca.gov.in/regional-centers/southern-regional-centre/
[c] https://tdil-dc.in/index.php?option=com_vertical&parentid=58&lang=en

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

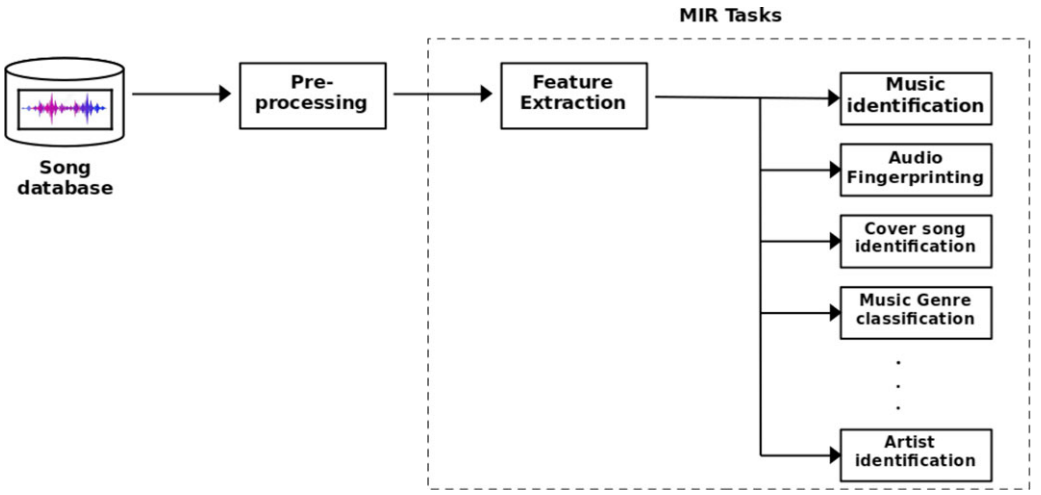

Figure 2. Block diagram of a typical MIR system.

labeling features in the form of symbolic representations and the coupling of composition with sound editing and mixing. In Shah *et al.* (2022), music genre classification is carried out using machine learning models such as Support Vector Machines (SVM), Random Forest, Extreme Gradient Boosting, and Convolutional Neural Networks (CNN). They used the popular GTZAN dataset for training and testing. It is seen that CNN performs the best compared to the traditional models. Deep learning models are also used in Mersy (2021), wherein depth-wise separable CNN is trained on electronic dance music and validated the performance with a CNN that is tested on a source-separated spectrogram and a normal spectrogram. The source-separated spectrogram proves better in terms of classification performance for limited dataset. Genre classification on the GTZAN dataset and Free Music Archive dataset is also undertaken in by Ashraf *et al.* (2020), using deep learning models such as CNN, Recurrent Neural Network (RNN), and CNN-RNN models with global layer regularization (GLR) using Mel-spectrograms to evaluate the performances. In the GLR technique, every hidden unit of a layer shares the same normalization terms. It is seen that CNN-RNN networks performed better on the two datasets due to the utilization of deep features.

In recent years, retrieval of music information from real-time embeddings has been seen in Stefani and Turchet (2022). Twelve acoustic guitar techniques are compiled, and the onset of such musical instances is detected. Cepstral features are used to train and test Deep Neural Network (DNN) models, and deployed to a Raspberry Pi-based embedded computer, with accuracy of 99.1%. Image embedding and acoustic embeddings are used in Dogan *et al.* (2022) for zero-shot audio classification. Similarly, in Lazzari, Poltronieri, and Presutti (2023), the pitch class from music structures is embedded into continuous vectors using existing methods and custom encodings using LSTM neural networks. This performs better than those techniques that use chord symbolic annotations.

2.2 Classification methods for identification of folk song and music

In music and singing processing, songs that share the same tune or melodies and have similar acoustic components are grouped together. It is usually done for efficient storage and retrieval, where ordering is necessary for ease of access. Classification of songs is also important in finding out the geographical origin of folk songs/music, for music recommendation, etc.

A few decades back, classification methods based on the ending notes, number of lines in the song, number of syllables in a line, etc., were adopted in Elschekova (1966), Keller (1984),

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Bohlman (1988), Umapathy, Krishnan, and Jimaa (2005), Van Kranenburg *et al.* (2007), but not without problems and limitations (Keller, 1984). In recent years, machine learning models have been heavily employed for musical classification, mainly based on the genre. Music genre classification techniques are found in Jiang *et al.* (2002), Aucouturier and Pachet (2003), Umapathy *et al.* (2005), Orio (2006); Meng *et al.* (2007), Lee *et al.* (2009), and Fu *et al.* (2010), which are based on different temporal and spectral methods. Several approaches and models have emerged over the years. Conditional Random Fields (CRFs) have been employed by (Liu *et al.* 2007; Li *et al.* 2019) along with GMM and Restricted Boltzmann Machine. In Li, Ding, and Yang (2017), it is seen that CRF-GMM outperformed traditional classification models (approx. 4.6 %–18.13 %).

Attention neural network-based architecture for folk song classification is also explored (Arronte-Alvarez and Gomez-Martin 2019). Musical motif embedding is also introduced to represent folk songs in different languages. For motif embedding, Word2Vec model (Mikolov *et al.* 2013) has been used and then later on employed in the ANN architecture. They classified folk songs of Chinese, Swedish, and German origins. Their results are comparable to those in existing studies (Cuthbert, Ariza, and Friedland 2011; Le and Mikolov 2014). In Loh and Emmanuel (2006), the extreme learning machine (ELM) for music genre classification is used, wherein features from 160 songs of four different genres in classical, pop music, rock music, and dance music are extracted. ELM and SVM have been used to classify these folk music.

In the Indian context, classification of Punjabi folk musical instruments from audio segments is carried out by Singh and Koolagudi (2017). Although vocal singing is not considered in their paper, the classification methods they used for polyphonic musical signals could be employed for the classification of vocal singing with multiple singers. Classification accuracy of 71% is achieved using the J48 classifier, which is increased to 91% by further improvement of input data samples. Feature selection is performed based on the performance of the J48 classifier. The selected features are then supplied to eight additional classifiers, where the highest classification rate is achieved by logistics classifier (95%). In Das, Bhattacharyya, and Debbarma (2021), a classification system for Kokborok music using traditional machine learning techniques is developed. A computational method to minimize the errors for each class is developed, with an alpha (α) value defined to estimate better accuracy, which successfully improved the original classification accuracy. In Das *et al.* (2023), music source separation is carried out for Hunting chants of Mizo folk songs using techniques like REpeating Pattern Extraction Technique (REPET), Robust Principal Component Analysis (RPCA), and Non-negative Matrix Factorization (NMF). It is seen that RPCA obtained the best signal-to-distortion ratio and signal-to-noise ratio for separation of vocals and musical accompaniments, followed by REPET, and NMF.

Based on the literature survey, the following research gaps are noted:

- Folk songs have received relatively less attention compared to mainstream or commercial music genres. Consequently, there is a limited pool of research studies and resources dedicated to under-resourced folk songs. This limitation affects the depth and breadth of research findings and the development of specialized tools and techniques for analysis.
- Folk songs are often part of an oral tradition, passed down through generations without written documentation. This poses challenges in preserving and documenting these songs, leading to the risk of songs being lost or forgotten over time.
- The oral nature of folk songs and the lack of standardized annotation and metadata for these songs make it difficult to compare and analyze them systematically.

3. Methodology

In Figure 3, the methodology followed for Mizo folk song classification is depicted. Firstly, the dataset is built by collecting folk songs from different sources. Then pre-processing is carried out

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

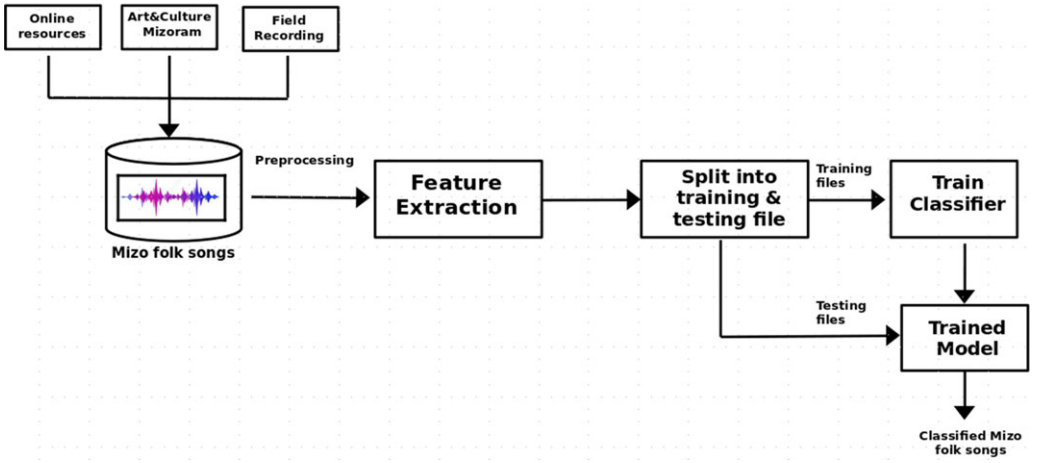

Figure 3. Methodology for classification of Mizo folk songs.

for extraction of acoustic features. Next, the extracted features are used for training and testing of the classification models. These different stages of the framework are detailed in the subsequent sections.

3.1 Mizo folk song dataset

The dataset for this study is collected from three sources, namely, publicly available Mizo folk songs of performances in cultural events and competitions, which are sourced from the internet, songs provided by the Art & Culture Department (A&C), Mizoram, and songs collected in field recordings. The internet data included audio from YouTube videos, mainly from documentaries, competition performances, and recordings made by educational institutes. The sampling frequency is 48,000 samples per second in mono channel.

Songs obtained from A&C Department have vocals accompanied by cow-hide drums. Recording has been done using the Shure SM58 dynamic vocal microphone, at 44,100 samples per second as the sampling rate and 1,411 kbps bit rate, with stereo channel. The songs were originally recorded for use in a folk song competition by the technical staff at A&C Department, and later shared with the authors of this study. Field data is recorded from a male singer who performed several categories of folk songs. Recording is carried out in a quiet room by the authors of this work. Zoom H1n portable recorder has been used, with sampling frequency of 44,100 samples per second, bit rate of 1,411 kbps, with stereo mode. The recording is placed approximately 1 ft. from the singer and mounted on a tripod.

Data imbalance is observed mainly due to singers being more familiar with certain song categories than others. From all the collected songs, *Hunting chants (Hlado), Children's songs (Pawnto hla), and Elderly songs (Pi Pu zai)* have the most number of song samples and longer duration. Here, the *Hlado* songs obtained from the internet have been chanted in an open field, and the ones from field data are recorded in a quiet room. *Pawnto hla* has been sung by kids in an open playground. *Pi pu zai* has been recorded in a room full of people who sang the songs together in a group. The data distribution can be seen in Table 1. Except for these three chosen categories, other categories contain varying song samples ranging between 1 and 20. This huge imbalance in data makes it infeasible for classification using all available categories. Hence, for the purpose of this paper, the said three categories have been chosen.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Table 1. Categories of Mizo folk song dataset used in this study

Category	No. of songs	Performance type	No. singers	Duration (apprx. mins)
Hunting chants (*Hlado*)	56	solo chanting	5	40
Children's songs (*Pawnto hla*)	19	group	≈ 10	41
Elderly songs (*Pi Pu zai*)	18	group	≈ 20	49
			Total duration	≈ 130

Data pre-processing tasks such as cleaning, segmentation, noise removal, and normalization are carried out on the dataset. Unwanted segments like background noise, coughing sounds, swallowing sounds, and tongue clicking in the recordings are removed. However, in order to avoid aliasing and windowing effects, about 0.5–1 sec regions of silence are left uncut at the start and end of each audio clip. Amplitude normalization is performed by taking the absolute maximum amplitude of the song signal, in order to keep the amplitudes in the range of -1 and $+1$.

A consistent naming structure is maintained for each category of the data: `songcategory_source_genderSpeakerNo_speechNo`. So, for a folk song type *Hlado*, the first song, performed by the second male singer obtained from A&C Department, can be written as: `hlado_src2_m2_0001.wav`. This dataset will be made available upon request to the authors or through the Natural Language Processing Laboratory, Department of Computer Science & Engineering, Tezpur University.

3.2 Acoustic feature extraction

For the experiments, Matlab (MATLAB 2022) and Praat (Boersma 2001) tools are used. For the purpose of feature extraction, the songs are sampled at 48,000 samples per second, and frame-wise extraction is carried out. Frame size of 25 msec and frameshift of 10 msec are used (Paliwal, Lyons, and Wójcicki 2010; O'Shaugnessy 1987, p. 179). The following acoustic features are used:

1. **Fundamental Frequency (F0):** F0 is the frequency at which the vocal folds vibrate during voice production. In this work, F0 is extracted using the autocorrelation function, which is computed as:

$$R(i) = \sum_{n=i}^{N-1} x(n)x(n-i) \tag{1}$$

where $1 \leq i \leq p$ for a finite duration of $x(n)$ and p is a range of lag values ((O'Shaugnessy 1987, p. 196); (Huang *et al.* 2001, p. 321)). Six parameters of F0, minimum, maximum, mean, range, standard deviation, and median, are extracted.

2. **Signal energy:** The energy of a continuous-time signal $x(t)$ can be calculated by taking the square of amplitude of each time instance of x ((Haykin and Van Veen, 2007, p.20); (O'Shaugnessy 1987, p. 180)). It is computed as

$$E_x = \int_{-\infty}^{\infty} x^2(t)dt \tag{2}$$

3. **Zero crossing rate:** The amount of time a signal crosses the x-axis is known as zero-crossing rate (ZCR). For a signal $x(t)$, ZCR is computed as

$$ZCR(x(t)) = \frac{1}{2M}|sgn(x(t)) - sgn(x(t-1))|W(i-j) \tag{3}$$

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

where *W(i)* represents a window of size *M* samples, and the signum function returns output of ZCR in the range of [0, 1] (O'Shaughnessy 1987, p.182). Higher ZCR value implies higher frequency content in the signal (Lerch 2012, p. 62)

4. **Strength of excitation (SoE):** SoE is the relative strength of impulse-like excitation at the Glottal Closure Instants. In this work, SoE is extracted using zero frequency filtering (ZFF) method (Yegnanarayana and Murty 2009). The slope of the ZFF signal at each epoch is the SoE (Mittal 2016; Kadiri and Alku 2020).
5. **Cepstral peak prominence (CPP):** It is a commonly used method for acoustic measure of voice quality in different applications of speech analysis like singing voice studies (Baker *et al.* 2022) and speech dysphonia (Fraile and Godino-Llorente 2014). We have extracted CPP with voice detection and without voice detection, as found in (Murton, Hillman, and Mehta 2020).
6. **Mel frequency cepstral coefficients (MFCC):** MFCCs describe the overall spectral envelope of a signal (Lerch 2012; O'Shaughnessy, 1987). The i^{th} coefficient, as in Lerch (2012, p. 51), is computed as

$$MFCC_i(n) == \sum_{k'=1}^{K'} log|X'(k', n)|.cos\left(i.\left(k' - \frac{1}{2}\right)\frac{\pi}{K'}\right) \tag{4}$$

where $|X'(k', n)|$ is the Mel spectrum at that frame block. In this work, 13 MFCC coefficients are used.

7. **Formant frequencies:** Acoustic resonances in the vocal tract are called formants (O'Shaughnessy, 1987). They are crucial in examining the articulatory response of the vocal tract (Ladefoged and Johnson 2014). A 10^{th} order linear prediction is used for generating the first four formants, and the songs are resampled to 10,000 samples per second.

3.3 Proposed models

At present, it is still difficult to implement a fully unsupervised learning model for audio, since singing signal is an exceedingly non-linear data. Moreover, sufficient data to implement an unsupervised model is currently unavailable for Mizo folk songs. Hence, an approach using a supervised deep learning model, LSTM, is proposed in the following subsection.

3.3.1 LSTM with attention mechanism (LSTM-attn)

A LSTM model with attention mechanism has been proposed. This attention mechanism enhances the ability of LSTM to focus on specific regions of the input acoustic feature vector at each time step. It computes attention scores based on the similarity between the data points in the feature vector, and assigns weights to different regions of the input vector giving 'attention' to the most relevant data points. LSTM-attn in this work is computed as in Algorithm 1, using the following parameters:

- One-hot encoded input sequence, x, with dimension $2102 \times 29 \times 1$ (for the 29 selected acoustic features)
- Two weight parameters, Q_w and K_w, are defined as learnable weight matrices for Query projection and Key projection, respectively.
- Output sequence, y, which is an attention-weighted sequence with the same dimension as x.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Algorithm 1. Attention mechanism for LSTM

1: Input sequence x

2: Perform linear transformation of the input sequence, x, using weights of Query and Key projections: $q\_proj = x \cdot Q_w$ and $k\_proj = x \cdot K_w$. Here, q_proj and k_proj have dimensions defined by (batchSize, sequenceLength, inputDim)

3: Compute the attention scores by taking dot product of q_proj and k_proj. Transpose k_proj to avoid dimension mismatch: $attn\_scores = q\_proj \cdot k\_proj^T$

4: Apply softmax activation function along the last axis (axis $= -1$) to normalize the attention scores: $attn\_scores = softmax(attn\_scores, axis = -1)$

5: Calculate the weighted sum, v_w, for the values in x using the computed attention scores: $y = attn\_scores \cdot v_w$

6: Return the output, y.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 29, 64)            16896

 attention_layer (Attention  (None, 29, 64)            8192
 Layer)

 lstm_1 (LSTM)               (None, 29, 128)           98816

 flatten (Flatten)           (None, 3712)              0

 dense (Dense)               (None, 128)               475264

 dense_1 (Dense)             (None, 3)                 387

=================================================================
Total params: 599555 (2.29 MB)
Trainable params: 599555 (2.29 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 4. Summary of the proposed LSTM-attn model with custom attention layer.

The model summary of LSTM-attn is shown in Figure 4. The first LSTM layer in this figure is the input layer, which takes the input having a shape of $2102 \times 29 \times 1$. This layer uses ReLU (Rectified Linear Unit) with 64 units to introduce non-linearity in the input vector. This layer allows to capture the temporal differences in the input feature sequence. The output produced has a shape of $32 \times 29 \times 64$, as batchSize $= 32$.

This is then passed to the attention layer, where *attn_scores* are computed based on the importance of the data points, as per the algorithm mentioned above. The dimension of the weights for the matrices Q_w and K_w are customized as 64×29, taking the size of both the time axis and

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

the feature axis from the input vector, rather than weighing on the time axis alone as done in conventional attention mechanisms. This projection of input data to a higher dimensionality for query vector allows the model to concentrate on the most relevant features in the batch, while the key weight is set to retain the dimension of the input vector. Moreover, this customization allows for pairwise relationships between features within the input sequence while preserving all input features. The shape of the output is maintained from the previous layer. This setting was seen to improve the model performance than allowing the model to assign random weights.

Next is another LSTM layer with 128 ReLU units, whose output sequence has $32 \times 29 \times 128$ shapes. Then, the output is flattened to get a vector of shape 32×3712. A fully connected dense layer using ReLU activation with 128 units is again added, which reduces the shape of the vector to 32×128. Subsequently, the softmax layer follows with 3 units (i.e. the number of classes, which in our case is the number of song categories) to produce the 32×3 output as class probabilities.

3.3.2 Machine learning models

In addition to the proposed LSTM-attn model above, four commonly used supervised machine learning models, SVM, K-Nearest Neighbor (KNN), Naive Bayes, and Ensemble learning, are employed for comparing the results obtained from the LSTM-attn model. These models have been found to have the highest classification rates as compared to other models for shorter segments of speech (Grimaldi, Cunningham, and Kokaram 2003; Huang *et al.* 2014).

4. Experiments and discussion of results

4.1 Experiments

In this work, a total of 29 acoustic features and parameters have been extracted and divided into four different combination sets. This is done to find out which group of features are relevant for the classification task based on their acoustic properties. The features are grouped as follows: ***set-1:*** *Temporal features (F0, Energy, ZCR);* ***set-2:*** *Source feature (SoE);* ***set-3:*** *Source-system features (CPP, MFCC);* ***set-4:*** *System features (Formants);* ***set-1 + 2:*** *Temporal + source features (F0, Energy, ZCR, SoE);* ***set-1 + 2 + 3:*** *Temporal + source + source-system features ((F0, Energy, ZCR, SoE, CPP, MFFC); and* ***set-1 + 2 + 3 + 4:*** *all sets of feature (F0, Energy, ZCR, SoE, CPP, MFFC, Formants).* Class labels 1, 2, and 3 are assigned to the hunting chants, children's songs, and elderly songs, respectively.

In total, there are seven feature combinations used for classification of the folk songs. Performing the classification with such combinations will help to identify what acoustic features are relevant for the classification of Mizo folk songs. The three categories of songs, whose typical sample length is 1–5 mins, are divided into manageable chunks of 3 sec. So, from the original 93 song files, a total of 2948 samples are generated.

After removal of NaN and zero values, the feature vector is one-hot encoded to reshape and make it compatible with the model. The shape of the vector becomes $2102 \times 29 \times 1$. Using the seven feature set combinations, experiments are carried out wherein the shape of the input vector changes depending on the number of features considered. For these experiments, the *'adam'* optimizer is used, along with a constant batch size of 32 for different epochs—10, 20, 30, 40, and 50. Only the epoch with the best result, i.e., 10, is reported in this study. With the small size of the feature vector, it is deemed sufficient to choose 10 epochs for this work. For the four ML models, after eliminating NaN values, the dimension of the final feature vector becomes 2183×29. The models are trained using k-fold cross validation ($k = 5$) on 80% of the data, and 20% is set aside for testing.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Table 2. Macro-averages of long short-term memory with attention layer model for classification of three categories of Mizo folk songs, with 20% testing data

Feature Set	Train (%)	Test (%)	Precision	Recall	f1-score
Set-1	96.61	95.01	0.94	0.94	0.94
Set-2	72.46	71.73	0.74	0.73	0.74
Set-3	93.58	91.69	0.89	0.88	0.89
Set-4	73.47	68.88	0.77	0.79	0.78
Set-1 + 2	95.84	93.82	0.93	0.93	0.93
Set-1 + 2 + 3	95.54	93.11	0.92	0.93	0.92
Set-1 + 2 + 3 + 4	95.30	91.21	0.90	0.90	0.90

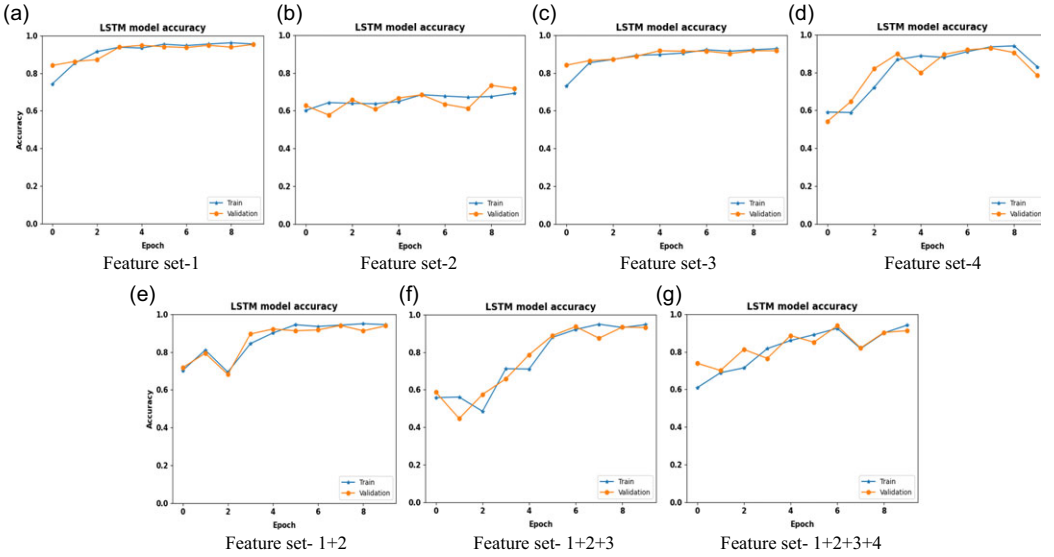

Figure 5. Accuracy plots for LSTM-attn with 10 epochs for different feature sets (Accuracy × 100).

4.2 Discussions

In Table 2, the accuracy results of the proposed LSTM-attn model are shown. The training and testing results are split 80:20 from the input sequence. Although slightly lesser, the performance of the model for each feature set is comparable to the existing ML models used in this study. As the classes are rather distinctive from one another, it has been observed that class-2 and class-3 exhibit minimal misclassification between them, and higher number of misclassifications are seen between class-1 and class-3.

Interestingly, it is observed that LSTM-attn performance deteriorates as the feature set combines more acoustic features. The accuracy plots of the feature sets are shown in Figure 5. The ***set-1*** provides the best accuracy (95.01%) among the individual feature sets. However, as the combinations are increased, the model appears to gradually reduce in performance (91.21%) in case of ***set-1*** + ***2*** + ***3*** + ***4*** for all the features. This is attributed to the fact that LSTMs are sequential models and work well in capturing patterns and dependencies in sequential data. However, as the

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Table 3. Classifier performances for three categories of Mizo folk songs, with different combinations of acoustic features (5-fold cross validation with 20% data for testing)

Feature Set	Classifier	C.V. (%)	Test (%)	Correctly Classified	Incorrectly Classified
Set-1	KNN	96.51	96.79	1685	61
	SVM	95.19	96.79	1662	84
	Naive Baye's	93.81	96.79	1638	108
	Ensemble	96.22	97.71	1680	66
Set-2	KNN	56.99	61.01	1693	53
	SVM	53.32	48.85	1679	67
	Naive Baye's	62.94	57.57	1665	81
	Ensemble	65.18	64.91	1689	57
Set-3	KNN	89.98	92.43	1571	175
	SVM	88.89	89.68	1552	194
	Naive Baye's	91.07	92.20	1590	156
	Ensemble	90.89	92.66	1587	159
Set-4	KNN	61.57	62.84	1075	671
	SVM	54.93	58.49	959	787
	Naive Baye's	48.68	50.69	850	896
	Ensemble	64.83	65.60	1132	614
Set-1 + 2	KNN	96.96	97.02	1693	53
	SVM	96.16	95.41	1679	67
	Naive Baye's	95.36	94.72	1664	81
	Ensemble	96.74	95.41	1689	57
Set-1 + 2 + 3	KNN	97.19	97.25	1697	49
	SVM	96.91	97.02	1692	54
	Naive Baye's	96.28	96.79	1681	65
	Ensemble	96.33	97.48	1682	64
Set-1 + 2 + 3 + 4	KNN	95.88	96.10	1674	72
	SVM	95.42	96.33	1666	80
	Naive Baye's	96.22	96.33	1680	66
	Ensemble	96.51	97.71	1685	61

feature sets combined are not inherently sequential by nature, the performance of the LSTM-attn is seen to deteriorate.

Out of the four different supervised classifiers employed, it can be seen from Table 3 that Ensemble method achieves the best accuracy of 97.71% for temporal features in ***set-1*** and all features in ***set-1*** + ***2*** + ***3*** + ***4***, with 66 incorrectly classified data points. It can also be observed from Figure 6 that there is hardly any misclassification between class-3 (elderly songs) and class-2 (Children's songs). This is because children's voice and adult's voice have clear distinction and lack similarity, so the models are able to train and predict well. Misclassification is highest in case of hunting chants and the elderly songs. Although the rhythm and tempo of the songs are not similar, there is still the fact that both are sung and performed by adults. As such, the characteristics of these two categories of songs may show some similarity in terms of excitation features.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

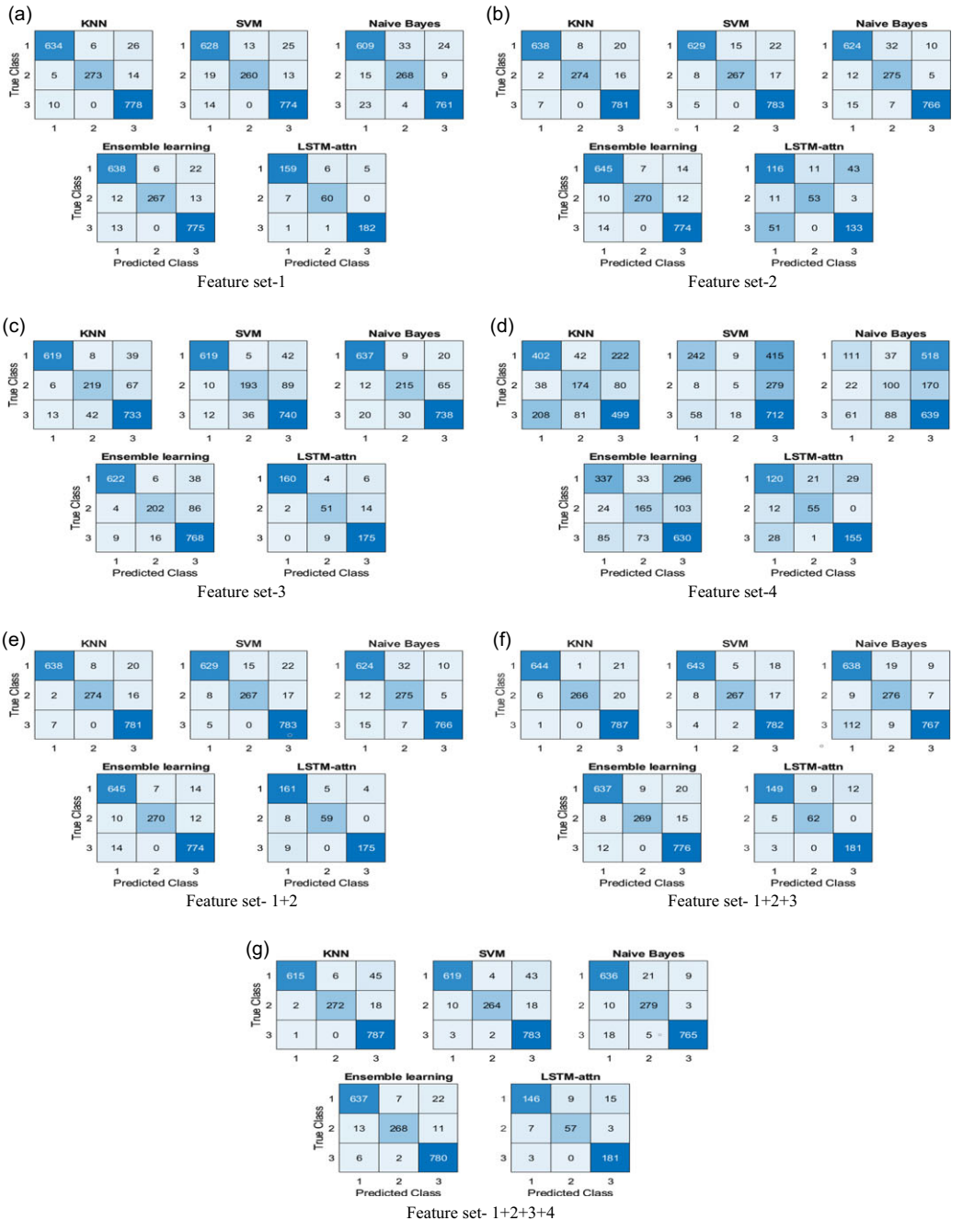

Figure 6. Confusion matrices of the four ML models and LSTM-attn with different feature sets.

4.2.1 A comparative analysis of classification using different feature sets

In ***set-1***, the LSTM-attn model yields a testing accuracy of 95.01%, which, despite a slight decrease, is considered a fairly good performance given the limited data size. The accuracy plot in Figure 5(a) shows slight improvements. The precision, recall, and f1-score are also the highest

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

among all the feature sets, as shown in Table 2. Ensemble model also performs well, achieving an accuracy rate of 97.71%. KNN, SVM, and Naive Bayes classifiers obtained testing accuracies of 96.79%. Overall, this feature set with temporal features demonstrates the best accuracy when considering both machine learning models and the proposed LSTM-attn model.

In ***set-2***, LSTM-attn achieves the highest performance (71.73%) despite challenges with reduced features and its f1-score dipping to 0.74. There is no misclassification of class-3 as class-2 for the models except Naive Bayes seen in Figure 6(b). Ensemble model also performs the best among (64.91%) the ML models. The use of a single acoustic feature (SoE) in this set affects the performance. It can also be due to the fact that estimation of excitation strength is obtained using a ZFF model (Yegnanarayana and Gangashetty 2011), which uses a fixed window length in the trend removal of zero-frequency resonators. This fixed windowing does not work well for singing voice and expressive voice due to high source-filter system interaction (Kadiri and Yegnanarayana 2015; Kadiri, Alku, and Yegnanarayana 2020).

In ***set-3***, source-system features produced better classification accuracy than ***set-2***. LSTM-attn obtained 91.69% accuracy with its f1-score at 0.89. Ensemble method performs the best (92.66%) while SVM has the lowest accuracy (89.68%) and the highest classification error. The accuracy plot in Figure 5(c) shows higher accuracies with little improvement over 10 epochs.

In ***set-4***, LSTM-attn obtains the best performance (69.88%). The challenges with feature set-4 become evident as the models struggle to effectively categorize folk songs. As can be seen in Figure 6(d), the misclassification ratio closely mirrors the classification accuracy, which can be attributed to the influence of musical notes in vocal singing, causing shifts in formant frequencies (Heaton, 2010). Although relatively steady, the accuracy plot also shows a slight dip toward the last epoch in Figure 5(d).

With ***set-1*** + ***2***, there is an improvement to the classification accuracy when the temporal features are combined with source features, as it has been observed that ***set-2*** does not perform well on its own. The LSTM-attn obtains 93.11% accuracy, although the ML models achieved better accuracy.

With feature ***set-1*** + ***2*** + ***3***, improvements in the accuracy of the classification are observed with fewer classification errors as shown in Figure 6(f). The f1-score is at 0.92 and the accuracy plot in Figure 5(f) shows a steady curve between training and testing data. KNN model does a good job of classification with the least amount of misclassified data points. The LSTM-attn achieves the lowest accuracy of 93.11% with f1 score of 0.93.

Finally, for ***set-1*** + ***2*** + ***3*** + ***4***, the incorporation of system features (formants) to the previous three types of features does not improve the classification accuracy for LSTM-attn, KNN, SVM, and Naive Bayes. However, in Ensemble model, the accuracy is improved and misclassification is reduced. It is observed that the performance of LSTM-attn does not necessarily improve with more diversified feature combinations. It is seen to have lesser accuracy rate than those whose features are of the same feature type.

4.2.2 Comparison with existing works

Given the absence of prior research on the acoustic analysis and classification of Mizo folk songs, this study draws comparisons with similar research on folk songs in other low-resource Indian languages. Currently, the existing works do not typically employ more recent techniques using deep learning methods due to being under-resourced.

As depicted in Table 4, different works on classification of Indian folk songs are shown, out of which Kokborok (Das *et al.* 2021) seems to be closest to Mizo in terms of language family (Tibeto-Burman language family) and geographical location. Although the experimental setup is dissimilar, in case of Kokborok, 63% classification accuracy has been achieved by using statistical computational methods to improve the classification error of the feature sets.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Table 4. Classifier performance compared with existing studies of under-resourced folk songs

Slno.	Classifiers	Database and Experimental Setup	Accuracy	Ref.
1.	SVM	116 folk songs consisting of Gais, Rais, and Phag genres of Central and North India, Kernel Density Estimates are generated for n-grams of the 3 genres, 80:20 training-testing ratio and 5-fold cross validation	91.30%	(Pandey and Dutta 2014)
2.	SVM	103 devotional songs and 113 folk songs in Tamil, spectral and cepstral features are used, 30 sec excerpts of songs with 20 msec frame and 10 msec shift, 60:40 training-testing ratio	84.21%	(Betsy and Bhalke 2015)
3.	J48	150 Punjabi folk music, 13 features from musical instruments algoza, dhol, dilruba, sarangi, tumbi, are extracted using 50 msec frame and 25 msec shift	91.00%	(Singh and Koolagudi 2017)
4.	Neural network	160 songs of Garba, Lavani, Ghoomar, and Bhangra dance songs for classification, 13 MFCC and 13 LPCC features ($p = 12$) extracted with 15 msec frame and 10 msec shift	90.80%	(Bhatt and Patalia 2017)
5.	LibSVM	300 Kokborok songs, extracted timbre, rhythm, and intensity based features with window size of 512 samples, proposed statistical method based on Henver's mood taxonomy for musical mood analysis, reduction of classification error in each feature type	63.00%	(Das *et al*. 2021)
6.	CNN	80 mins of Hindustani vocal and music, 13 MFCC features extracted for classification of steady notes and transition regions	70.00%	(Rege and Sindal 2021)
7.	LSTM-attn	approx. 2 hrs of Mizo folk songs, consisting 3 categories, input vector of size 2102×29, 80:20 training-testing ratio, LSTM with custom attention layer and a fully-connected dense layer with activation function = *'ReLU'*, batchSize = 32, optimizer function = *'adam'*, loss function = *'categorical_crossentropy'*,	95.01%	**our work**

In case of existing work with similar experimental setup, the folk songs of different categories of Gais, Rais, and Phag are classified by Pandey and Dutta (2014). A 5-fold cross validation and 80:20 training-testing ratio has been employed, which achieved 91.3% using SVM classifier. Despite the performance of our proposed LSTM-attn model being lower than the four existing ML models used in this study, there is a slight improvement than the existing works.

5. Summary and conclusion

Mizo is a low-resource language that lacks tools and technology required for the archival of its folk music. It has been observed that very few acoustic studies exist for Indian folk songs and music in spite of its richness in cultural and regional diversity. A survey of literature on Mizo folk songs as well as on recent methods of folk song and music classification have been carried out.

This work proposes an LSTM-attn consisting of an LSTM layer, a custom attention layer, a fully connected dense layer, and a softmax layer. Its performance is compared with those of existing machine learning models like SVM, KNN, Naive Bayes, and Ensemble models. Three categories of Mizo folk songs are used as dataset for classification, *Hunting chants (Hlado), Children's songs (Pawnto hla), and Elderly songs (Pi pu zai),* with the total duration of the songs being approximately 2 hrs. A total of 29 acoustic features grouped into temporal features, source features, source-system features, and vocal tract filter features are extracted from the Mizo folk songs.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Classification is carried out with 20% of the data segregated for testing. The highest accuracy achieved for the LSTM-attn is 95.01% (for temporal features), while it achieved 91.21% for all features combined. The results are comparable to existing studies of folk song classification in other Indian languages.

Our work is constrained by the relatively small dataset, which necessitates the segmentation of song samples into 3-sec segments. This approach may result in the loss of contextual information and potential discontinuities. Consequently, important audio events or transitions that span longer duration could be divided across different segments, posing challenges in capturing the complete audio content. Moreover, this employed frame-wise analysis might have overlooked important tonal characteristics of the Mizo language present in these folk songs. Performance of the proposed LSTM-attn model could be improved with larger sample size in each class. A comprehensive evaluation of the model will be undertaken in future. Additionally, analysis will be conducted to address the issue of lower accuracy in diverse acoustic feature sets. Exploration of tone-tune relationship in the Mizo language will also be undertaken, building upon previous studies in by Ramdinmawii and Nath (2022) and Gogoi and Nath (2023).

This work would significantly contribute to India's efforts in preserving intangible cultural heritage, benefiting Mizoram's Art & Culture Department, currently engaged in archiving the state's heritage. Additionally, this method can have broader applications in MIR, not only in Mizo but also in other Tibeto-Burman languages like Tani (Arunachal Pradesh), Meitei (Manipur), and Garo (Meghalaya).

Acknowledgement. Authors thank the Director and Technician, Department of Art & Culture, Mizoram, for their contribution in sharing their prerecorded songs. Authors are also grateful to the late Pu Lalkhuma (Sialsuk) for his valuable contribution to the dataset. Lastly, a great appreciation goes to the owners of YouTube channels who permitted us to use their content for our dataset in this work.

References

Arronte-Alvarez A. and Gomez-Martin F. (2019). An attentional neural network architecture for folk song classification, arXiv preprint arXiv: 1904.

Ashraf M., Geng G., Wang X., Ahmad F. and Abid F. (2020). A globally regularized joint neural architecture for music classification. *IEEE Access* **8**, 220980–220989.

Aucouturier J.-J. and Pachet F. (2003). Representing musical genre: A state of the art. *Journal of New Music Research* **32**(1), 83–93.

Baker C. P., Sundberg J., Purdy S. C., de SLeão, S. H., et al. (2022). CPPS and voice-source parameters: Objective analysis of the singing voice. *Journal of Voice* **38**(3), 549–560.

Betsy S. and Bhalke D. (2015). Genre classification of Indian Tamil music using mel-frequency cepstral coefficients. *International Journal of Engineering Research & Technology* **4**(12), 423–427.

Bhatt M. and Patalia T. (2017). Neural network based Indian folk dance song classification using MFCC and LPC. *International Journal of Intelligent Engineering and Systems* **10**(3), 173–183.

Blaß M. and Bader R. (2019). Content-based music retrieval and visualization system for ethnomusicological music archives. *Computational Phonogram Archiving*, **5**, 145–173.

Boersma P. (2001). Praat, a system for doing phonetics by computer. *Glot International* **5**(9), 341–345.

Bohlman P. V. (1988). *The Study of Folk Music in the Modern World*. Bloomington, Indiana, U.S.: Indiana University Press.

Cuthbert M. S., Ariza C. and Friedland L. (2011). Feature extraction and machine learning on symbolic music using the music21 toolkit. In *ISMIR*, pp. 387–392.

Das N., Ramdinmawii E., Kumar A. and Nath S. (2023). Vocal singing and music separation of mizo folk songs. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, IEEE, pp. 1–6.

Das S., Bhattacharyya B. K. and Debbarma S. (2021). Building a computational model for mood classification of music by integrating an asymptotic approach with the machine learning techniques. *Journal of Ambient Intelligence and Humanized Computing* **12**(6), 5955–5967.

Deruty E., Grachten M., Lattner S., Nistal J. and Aouameur C. (2022). On the development and practice of ai technology for contemporary popular music production. *Transactions of the International Society for Music Information Retrieval* **5**(1), 35.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Dogan D., Xie H., Heittola T. and Virtanen T. (2022). Zero-shot audio classification using image embeddings. In *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 1–5.

Elschekova A. (1966). Methods of classification of folk-tunes. *Journal of the International Folk Music Council* **18**, 56–76.

Fraile R. and Godino-Llorente J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control* **14**, 42–54.

Fu Z., Lu G., Ting K. M. and Zhang D. (2010). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia* **13**(2), 303–319.

Gogoi J. and Nath S. (2023). Analysing word stress and its effects on assamese and mizo using machine learning. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, IEEE, pp. 1–6.

Grimaldi M., Cunningham P. and Kokaram A. (2003). A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 102–108.

Haykin S. and Van Veen B. (2007). *Signals and Systems*. Daryaganj, New Delhi, India: Wiley India Pvt. Ltd .

Heaton E. M. (2010). *Formant Changes in Amateur Singers After Instruction in a Vowel Equalization Technique*. Ann Arbor, Michigan, U.S.: ProQuest LLC.

Huang X., Acero A., Hon H.-W. and Reddy R. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, vol. **95**. Hoboken, New Jersey, U.S.: Prentice Hall.

Huang Y.-F., Lin S.-M., Wu H.-Y. and Li Y.-S. (2014). Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. *Data & Knowledge Engineering* **92**, 60–76.

Jiang D.-N., Lu L., Zhang H.-J., Tao J.-H. and Cai L.-H. (2002). Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, IEEE, vol **1**, pp. 113–116.

Kadiri S. R. and Alku P. (2020). Excitation features of speech for speaker-specific emotion detection. *IEEE Access* **8**, 60382–60391.

Kadiri S. R., Alku P. and Yegnanarayana B. (2020). Analysis and classification of phonation types in speech and singing voice. *Speech Communication* **118**, 33–47.

Kadiri S. R. and Yegnanarayana B. (2015). Analysis of singing voice for epoch extraction using zero frequency filtering method. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4260–4264.

Keller M. S. (1984). The problem of classification in folksong research: A short history. *Folklore* **95**(1), 100–104.

Khiangte L. (2001). Mizo folk literature. *Indian Literature* **45**(1), 72–83.

Khiangte L. (2002). *Mizo Songs and Folk Tales*. New Delhi, India: Sahitya Akademi.

Ladefoged P. and Johnson K. (2014). *A Course in Phonetics*. Stamford, Connecticut, U.S.: Cengage Learning.

Lalremruati R. (2012). Oral literature: a study of Mizo folk songs, PhD thesis. Mizoram University.

Lalremruati R. (2019). Narratives of Mizo traditional songs : A thematic analysis. *International Journal of Research and Analytical Reviews* **6**(2), 422–425.

Lalthangliana B. (1993). *History of Mizo Literature*. Aizawl: R.T.M. Press.

Lalthangliana B. (2005). *Culture and Folklore of Mizoram*. Publications Division Ministry of Information & Broadcasting. New Delhi, India: Publications Division, Ministry of Information and Broadcasting, Govt. of India.

Lalzarzova, 2016). Thanglunghnemi zai bihchianna. *Mizo Studies* **VIII**(2), 27–35.

Lazzari N., Poltronieri A. and Presutti V. (2023). Pitchclass2vec: Symbolic music structure segmentation with chord embeddings, arXiv preprint arXiv: 2303.15306.

Le Q. and Mikolov T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, PMLR, pp. 1188–1196.

Lee C.-H., Shih J.-L., Yu K.-M. and Lin H.-S. (2009). Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Transactions on Multimedia* **11**(4), 670–682.

Lerch A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Hoboken, New Jersey, U.S.: John Wiley & Sons, Inc.

Li J., Ding J. and Yang X. (2017). The regional style classification of Chinese folk songs based on GMM-CRF model. In *Proceedings of the 9th International Conference on Computer and Automation Engineering*, pp. 66–72.

Li J., Luo J., Ding J., Zhao X. and Yang X. (2019). Regional classification of chinese folk songs based on crf model. *Multimedia Tools and Applications* **78**(9), 11563–11584.

Liu Y., Xu J., Wei L. and Tian Y. (2007). The study of the classification of Chinese folk songs by regional style. In *International Conference on Semantic Computing (ICSC 2007)*, IEEE, pp. 657–662.

Loh Q.-J. B. and Emmanuel S. (2006). ELM for the classification of music genres. In *2006 9th International Conference on Control, Automation, Robotics and Vision*, IEEE, pp. 1–6.

MATLAB (2022). *Version 9.12.0 (R2022b)*. Natick, Massachusetts: The MathWorks Inc.

Meng A., Ahrendt P., Larsen J. and Hansen L. K. (2007). Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing* **15**(5), 1654–1664.

https://doi.org/10.1017/nlp.2024.23 Published online by Cambridge University Press

Mersy G. (2021). Efficient robust music genre classification with depthwise separable convolutions and source separation. In*Proceedings of the AAAI Conference on Artificial Intelligence*, vol.**35**(18), pp. 15972–15973.

Mikolov T., **Chen K.**, **Corrado G. and Dean J.** (2013). Efficient estimation of word representations in vector space, arXiv preprint arXiv: 1301.3781.

Mittal V. K. (2016). Discriminating features of infant cry acoustic signal for automated detection of cause of crying. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, pp. 1–5.

Moseley C. (2010). *Atlas of the World's Languages in Danger*. place de Fontenoy, Paris, France: UNESCO Publishing.

Murton O., **Hillman R. and Mehta D.** (2020). Cepstral peak prominence values for clinical voice evaluation. *American Journal of Speech-Language Pathology* **29**(3), 1596–1607.

Orio N. (2006). Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval* **1**(1), 1–90.

O'Shaugnessy D. (1987). *Speech Communication: Human and Machine*. Boston, Massachusetts, U.S.: Addison-Wesley Publishing Company.

Paliwal K. K., **Lyons J. G. and Wójcicki K. K.** (2010). Preference for 20-40 MS window duration in speech analysis. In *2010 4th International Conference on Signal Processing and Communication Systems*, IEEE, pp. 1–4.

Pandey A. and Dutta I. (2014). Bundeli folk-song genre classification with KNN and SVM. In *Proceedings of the 11th International Conference on Natural Language Processing*, pp. 133–138.

Ramdinmawii E. and Nath S. (2022). A preliminary analysis on the correlates of stress and tones in Mizo. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**(2), 1–15.

Rege A. and Sindal R. (2021). Audio classification for music information retrieval of Hindustani vocal music. *Indonesian Journal of Electrical Engineering and Computer Science* **24**(3), 1481.

Shah M., **Pujara N.**, **Mangaroliya K.**, **Gohil L.**, **Vyas T. and Degadwala S.** (2022). Music genre classification using deep learning. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, pp. 974–978.

Singh I. and Koolagudi S. G. (2017). Classification of Punjabi folk musical instruments based on acoustic features. In *Proceedings of the International Conference on Data Engineering and Communication Technology*, Springer, pp. 445–454.

Srinivasa Murthy Y. V. and Koolagudi S. G. (2018). Content-based music information retrieval (CB-MIR) and its applications toward the music industry: A review. *ACM Computing Surveys* **51**(3), 1–46.

Stefani D. and Turchet L. (2022). On the challenges of embedded real-time music information retrieval. In *Proceedings of the 25-th International Conference on Digital Audio Effects (DAFx20in22)*, vol. **3**, pp. 177–184.

Thanmawia R. (1998). *Mizo Poetry*. Publications Division, Ministry of Information & Broadcasting, Government of India. Aizawl, Mizoram, India: Din Din Heaven.

Umapathy K., **Krishnan S. and Jimaa S.** (2005). Multigroup classification of audio signals using time-frequency parameters. *IEEE Transactions on Multimedia* **7**(2), 308–315.

Van Kranenburg P., **Garbers J.**, **Volk A.**, **Wiering F.**, **Grijp L. and Veltkamp R.** (2007). Towards integration of music information retrieval and folk song research. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pp. 505–508.

Weidert A. (1975). *Componential Analysis of Lushai Phonology*, vol. **2**. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Yegnanarayana B. and Gangashetty S. V. (2011). Epoch-based analysis of speech signals. *Sadhana* **36**(5), 651–697.

Yegnanarayana B. and Murty K. S. R. (2009). Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing* **17**(4), 614–624.

Cite this article: Ramdinmawii E. and Nath S. Resource building and classification of Mizo folk songs. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.23