Building Indonesian Music Dataset: Collection and Analysis

1st M. Octaviano Pratama
Faculty of Computer Science
UPN Veteran Jakarta
South Jakarta, Indonesia
octav@upnvj.ac.id

2nd Pamela Kareen
BISA AI Academy
Bisa Artifisial Indonesia
South Jakarta, Indonesia
pamela@ptbisaai.com

3rd Ermatita
Faculty of Computer Science
Universitas Sriwijaya
Sumatera Selatan, Indonesia
ermatita@unsri.ac.id

2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) | 978-1-6654-2733-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIMCIS53775.2021.9699332

***Abstract*—We introduce The Indonesian Music Dataset (IMD), a collection of audio features and text lyrics features for thousand Indonesian popular songs which has been developed for automatic music era classification and other classification tasks. Dataset collection consists of audio features represented by Spectrogram, Chroma Feature and Low-level audio features. The dataset also consists of lyric features in order to support multimodal tasks. Dataset is equipped with eras (year of publication) labels starting from '70 until the current era, mood labels from Valence-Arousal (Anger, Sadness, Happiness and Relax), and genre labels (Rock, Pop, Jazz). In this paper, we also present era, mood and genre prediction as an example of a dataset experiment for each modality (audio features and text lyrics features) that shows positive results using benchmarking models.**

Keywords—Indonesian Music Dataset, Music Classification, audio features, text lyrics features

I. INTRODUCTION

Music Information Retrieval (MIR) topics like lyric classification [1], audio classification [2], mood classification from sound [3] and music era [4] become interesting research topics. Several music trend applications in 2021 like playlist generation in Spotify and query by humming in Sound hound [5] used MIR as its background process. MIR components like music content (rhythm, timbre and melody), music context (lyrics and music poster) and user context (genre, mood and era) hold important parts in music application implementation [5]. Based on MIR components, surprisingly few researchers focus on music era classification tasks because of the rareness of music era dataset. Similar with mainstream music classification tasks like genre and mood, music era has its own characteristics that distinguish music in an era with other music in other eras. Music era is a year of music publication that connects various similar characteristics of songs from similar decades like 70', 80' and current decade [6]. Research from Choi [4] tried to recognize the music era, genre and mood using audio features. Music era, genre and mood classification can be implemented in a music player in order to recognize a given audio file automatically without annotation.

In order to conduct research in music classification, several music representation files such as audios [4], lyrics [1] and posters [7] can be extracted as feature sets. From overall music representation files, the use of audio representation has become one of essential parts in music classification tasks. Research from Choi [4], Irvin [2], and Jeong [8] used Spectrogram as audio features. On the other side, [9] used Chroma features as audio features. Both of these features offer appropriate image audio features. The appropriate selection of models also affects classification accuracy results. Choi et al [4] used Convolutional Recurrent Neural Networks (CRNN) to do genre and mood classification while Irvin et al [2] used Recurrent Neural Networks with Attention to do genre classification. Based on model selection above, several researchers used deep learning model [10] in order to produce higher classification accuracy. Current music dataset like Million Song Dataset (MSD) [11], and Free Music Archive (FMA) [12] offer large-scale music dataset, but not specific to extracted audio feature representation and eras label annotation. Issue in music dataset creation is also becoming an interesting research topic. Music dataset is also taken from songs in country or region like Latin Music Dataset [13]. Surprisingly none of the existing music dataset taken from Indonesian songs.

Based on above explanation, in this research we build Indonesian Music Era Dataset (IMD) as an alternative of existing music dataset that consists of Audio Features (Spectrogram, Chroma features and Low-level Audio Features) and Text Lyrics Features from Indonesia songs with 5 era labels (70', 80', 90', 00' and current era), 3 genres (Rock, Pop and Jazz) and 4 Moods (Anger, Sadness, Happiness, and Relax). In the next section, the process of dataset creation from data gathering, audio preprocessing and feature generation is explained clearly. In order to ensure appropriate dataset feasible to be used in music classification tasks, Convolutional Neural Networks (CNN) model is performed to classify Spectrogram and Chroma audio features, LSTM is also performed to classify Text lyrics features. Our contributions in this research can be separately divided into: (1) we are the first researcher that created the complete Indonesian Music Dataset that consists of a thousand audio features, text lyrics features, and also equipped with era, genre and mood annotation, (2) Indonesian Music Dataset (IMD) is equipped with combination of multimodal features, hence it can be used in multimodal tasks such as fusion between text and audio and others.

II. RELATED WORKS

Music Information Retrieval (MIR) is a process that focuses on music feature extraction and the using of music features for music tasks like classification, recommendation and generation [5]. Several researchers have created music dataset like Million Song Dataset [11], Free Music Archive [12], and others. Most existing dataset provide large-scale raw audio or audio features and are equipped with music labels like genre and mood. Music dataset also refers to collection music in a country or region like Latin Music Dataset [13]. Features in music dataset also vary starting from low-level features like bandwidth, and zero crossing rate [14] to high-level features like Spectrogram and Chroma Feature [14].

High-level features in music can be obtained from extracted raw audio, music symbol representation, and lyrics [5]. One of the most used high-level features in MIR is audio representation like Spectrogram and Chroma Feature. Several researches that used Spectrograms like Irvin [15], Choi [16] and Costa [17] have proved that this feature is fit to be used in MIR. On the other hand, Chroma Feature is also used by

researchers as a music feature in several tasks like cover song detection [9]. Both of these features further can be fit with deep learning models for classification or recommendation tasks based on previous research [9][15][16][17]. Deep learning uses multiple layers of neural network architecture [10]. Various well-known deep learning architectures like Convolutional Neural Networks (CNN) [18], and Recurrent Neural Networks (RNN) [19] have been implemented in the MIR domain. Detail past datasets and experiments can be seen in Table I

TABLE I. RELATED WORKS

Dataset	Dataset Properties			Modal
	Ref	*Feature*	*Tasks*	
GTZAN	[2], [8], [9]	MFCC, Low-level Feature	*genre*	Audio
Million Song Dataset	[6], [20]	*Mel-spectrogram*, Low-level Feature	*Genre, era, mood*	Audio
Lyric Find	[1]	Text Lyrics	*genre*	Text Lyrics
MDL	[3]	Lyrics and MFCC	*Mood*	Text and Lyrics

Based on table I, we can see for each dataset, the dataset only provides one modal whether audio or text lyrics, hence it does not support the multimodal classification task. Based on task, not all dataset support multi task such as genre, era and mood classification. The idea of Indonesian Music Dataset was formed in 2017, when we observed music genre and mood recognition, we used benchmark dataset from Latin Music Dataset, Million Song Dataset and others. we never found the music dataset from Indonesia, hence we started to collect the dataset from audio recording and transform it into features in order to keep the dataset not break copyright. Another reason, we found that existing music datasets have only few features such as single modal audio features, or single modal lyrics features. In this work, we will include not only audio features, but also lyrics features. There are several advantages in order to create Indonesian Music Dataset as follow: (1) large scale Indonesian Music Dataset assists local researchers to conduct research about music recognition in context to Indonesian music, (2) Indonesia Music Dataset consists of audio features and lyrics which can be used in multimodal tasks, and single modality tasks in texts and audios, (3) Indonesia Music Dataset can be used as benchmarking of dataset come from Indonesia and will be updated the features as soon as possible.

III. METHODOLOGY

A. Audio Features Creation

A thousand songs were collected from several sources, then raw audio files were cut into 30 second audio length. Frequencies of audio are also changed into 44100 Hz. Every single audio file is annotated manually with era labels among 70', 80', 90', 00' or current era and also annotated manually with genre labels: Rock, Pop and Jazz. The last step, audio files are changed into high-level audio features (such as Spectrogram and Chroma Feature) and low-level audio features (such as Spectral Centroid, Energy and Root Mean Square Energy, Zero Crossing Rate, and other low-level features).

Process of Spectrogram creation adopted from [6] as follows: (1) divide for each 3 second audio track into overlapped audio frame with 5 millisecond shifting, (2) given Fourier Transform to each frame and stack together in frequency and time axis, (3) given triangular filter bank in order to obtain response for each frame, (4) given logarithmic for each spectral intensity, (5) previous step will produce 600 frame that represented 3 second audio song, so the last step is to create tensor input 600 x 128 in order to make input support CNN model.

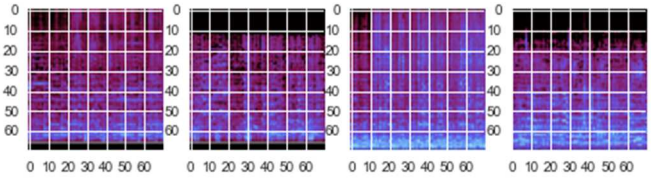
On the other side, in order to create Chroma feature representation, the step is explained as follows: (1) raw audio is changed into sequence Short-Time Chroma frame, (2) Ifgram and Ifpack from Librosa package will be used to create Chroma resolution, (3) 3 second audio file is represented as 600 x 128 tensor size. The IMD dataset contains a thousand Spectrogram and Chroma features derived from 100 Indonesian artists. 3 most popular Indonesian genres: Rock, Pop and Jazz contributes most of this dataset. All of the extracted features are bundled in a large-scale data storage file: HDF5 (Hierarchical Data Format Version 5) format.

The creation of low-level features such as Spectral Centroid (SC), Energy and Root Mean Square Energy (RMSE), Zero Crossing Rate (ZCR), and other low-level features using Python programming library: Librosa which is known as audio feature extraction library. The energy is the total magnitude of signal that can be defined as $\sum_{i=1}^{n} |X(n)|^2$ and we can count RMSE from Energy with formula $\sqrt{\frac{1}{n}\sum_{i=1}^{n} |X(n)|^2}$. On the other hand, ZCR indicates the number of times the signal crosses the horizontal axis. All of the low-level features are stored in data frame format as follow:

	label	ZCR	SC	RMSE
0	1	0.082031	5336.356749	6.270399e-07
1	2	0.029785	1812.679530	8.130488e-02
2	3	0.043457	1871.356553	5.096431e-02
3	4	0.041504	1777.395888	6.642678e-02
4	3	0.033203	1884.911718	8.011928e-02

Fig. 1. Low-level audio features represented by ZCR, SC, and RMSE

The completed audio features contains spectrogram audio features, Chroma audio features, low-level audio features and era labels for a thousand Indonesian songs from 1970 until 2017 packaged into 5 classes: '70, '80, '90, '00, and current era. Dataset is also equipped with 3 well-known labels: Rock, Pop and Jazz, and mood labels. The dataset is stored in an HDF5 file and can be obtained by requesting the corresponding author. Spectrogram and Chroma file example can be seen as follows:



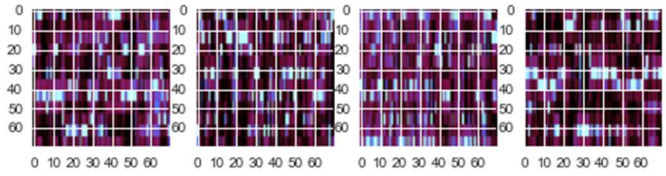


Fig. 2. Spectrogram (Top) and Chroma (Below) feature visualization

The classification will be performed on the dataset in order to ensure the dataset is feasible to be used. Convolutional Neural Networks [17] is performed to each high-level audio feature (Spectrogram and Chroma Features), meanwhile Deep Neural Networks is performed to each low-level audio feature (ZCR, SC, RMSE and others). The task is limited to era classification.

B. Text Lyrics Features Creation

We collect Lyrics Features from Indonesian popular songs from 1970 until 2010. The lyrics are adjusted with audio features. We use Python Library for lyrics collection from online websites. The lyrics are then annotated in three main tasks: genre, mood aend era label using two annotators. The agreement between annotators are measured by using Kappa Score in order to evaluate validity of the dataset. Here Lyrics Features flowchart:

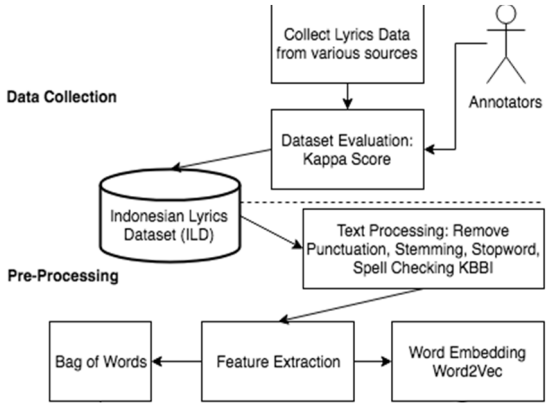


Fig. 3. Flow of Text Lyrics Collection

Based on flow, after collecting and evaluating lyrics, then text processing is performed to obtain clean text lyrics. The first step is to remove punctuation like @#$%^&* then change each sentence into words using tokenization. The next step, Indonesian Stemming techniques [22] and Stopword is also performed to remove inappropriate words. The final result, checking each word in Indonesian Dictionary or Kamus Besar Bahasa Indonesia (KBBI) for spell correction. If the basic word is not included in KBBI, then it would be eliminated. We created Bag of Words (BoW) and Word Embedding Word2Vec [23] for each cleaned text, hence all of the Lyrics format will be on BoW Features with numeric format.

The classification will be performed to text lyrics dataset in order to ensure the dataset is feasible to be used. Long Short Term Memory (LSTM) [24] is performed to classify text lyrics.

IV. Implementation

A. Audio Classification

After creating a music era dataset, the dataset is evaluated using Kapp score in order to ensure validity of data. We obtained a 0.769 evaluation score that indicated agreement between annotators is valid. The complete result can be seen as follow:

$$p_A = \frac{180+13}{200} = 0.965$$

$$p_{relevan} = 180 + 180 + 3 + 4 = 0.9175$$

$$p_{no} = 13 + 13 + 3 + 4 = 0.0825$$

$$p_e = 0.9175^2 + 0.0825^2 = 0.842 + 0.006 = 0.848$$

$$Kappa = \frac{(0.965-0.848)}{(1-0.848)} = \frac{0.117}{0.152} = 0.769$$

Music Era classification based on CNN model is performed to spectrogram and Chroma features in order to prove and analyze the feasibility of the IMD dataset that was created before. The process of dataset analysis is explained as follows: First, both a thousand spectrogram and Chroma Feature dataset is divided into three separate parts: (1) training data, (2) validation data and (3) testing data. Second, model is created based on CNN architecture as follows:

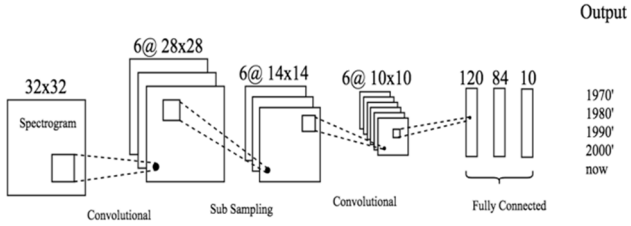


Fig. 4. Convolutional Neural Networks Architecture [18]

Based on above architecture, input features $\{x_1, x_2, \ldots, x_n\}$ such as spectrogram and Chroma Feature is put in input layer that would be feed forwarded into convolutional layer $S(x, W)$ in Eq. 1 [18]. For each convolutional layer, it will be equipped with Rectified Linear Unit (Eq. 2) activation function [18] to produce hidden state h_i. Then output from ReLU will be forwarded into the Pooling layer for subsampling each receptive field with 2x2 kernel. This pipeline process will be repetitively done depending on architecture. In the last layer of convolution, the layer will be flattened in order to be fit with a fully connected layer. The last step is to put output from a fully connected layer into the Softmax classifier (eq.3) [18] in order to determine Era class.

$$S(x, W) = \sum_{i=1}^{n} \quad \sum_{j=1}^{m} \quad x_{ij} W_{(i-m, j-n)} \tag{1}$$

$$h_i = \{s_i > 0, \ then \ s_i \ otherwise, then \ 0 \tag{2}$$

$$Softmax(h_i) = \frac{exp(h_i)}{\sum_{i=1}^{n} \ exp(h_j)} \tag{3}$$

This training process used Mini-batch Gradient Descent to determine the best hyperparameter θ given spectrogram or Chroma Feature $x^{i:i+n}$ and era label $y^{i:i+n}$ [25]. ADAM optimizer [25] is also used in training for optimization of neural networks. The last step, after creating an appropriate model, measurement evaluation is performed using accuracy, precision, recall and F1-Score based on confusion matrix for each Era label [5]. Precision describes how many retrieved items *relevant* to the query (Eq.4). Recall measures how

much relevant items are retrieved (Eq. 5). F1-Score defined as an average from precision and recall (Eq.6) [5].

$$Precision\ (P) = \frac{|Relevant \cap Retrieved|}{|Relevant|} \quad (4)$$

$$Recall\ (R) = \frac{|Relevant \cap Retrieved|}{|Retrieved|} \quad (5)$$

$$\text{F-Score}\ (\boldsymbol{F}) = \boldsymbol{2}\frac{\boldsymbol{P}\,.\,\boldsymbol{R}}{\boldsymbol{P}+\boldsymbol{R}} \quad (6)$$

Training is performed in GPU computer GTX 1050 with Cuda Architecture in order to decrease training time. 30 epoch training was performed for each Spectrogram and Chroma Feature Training data, then Validation data was used to measure accuracy and loss compared with Training data in order to prevent overfitting. Based on figure 3, both training and validation accuracy increase as a time reverse loss function decreases as a time that indicates our model is not overfitting.

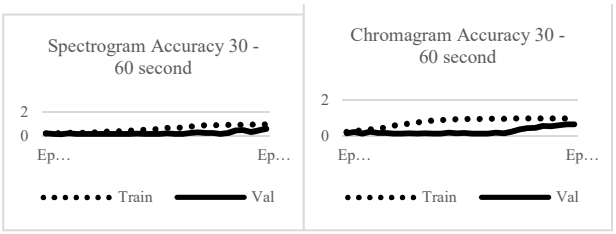


Fig. 5. Training Accuracy and Validation in Spectrogram and Chroma Feature

After performing approximately 3 hours training for each design experiment in order to produce appropriate models, testing data is used to test overall accuracy of the model. Complete prediction results for each design experiment can be seen in Table 2.

TABLE II. AUDIO CLASSIFICATION RESULT

Model	Feature	Task	Overall Accuracy
CNN	Spectrogram	Era	74%
CNN	Chroma	Era	72%
DNN	ZCR, SC, RMSE, etc	Era	75%

Based on the prediction result in table 2, CNN model produces positive result in predicting both spectrogram and Chroma and also DNN model produces positive result in predicting combination of low-level features such as ZCR, SC and RMSE. Beside overall accuracy, Confusion Matrix is also used to show prediction results of each era label as can be seen in figure 4.

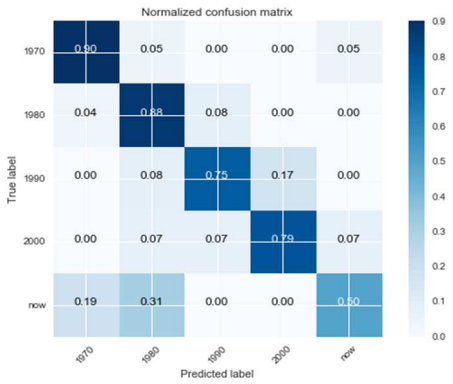


Fig. 6. Confusion Matrix of CNN to Spectrogram Feature

B. *Text Lyrics Classification*

After creating a music era dataset, the text dataset is evaluated using Kappa score in order to ensure validity of data. We obtained a 0.812 evaluation score that indicates agreement between annotators is valid. The complete result can be seen as follow:

$$p_A = \frac{330+3}{337} = 0.981$$

$$p_{relevan} = \frac{330+2}{330+3+2+2} \cdot \frac{330+2}{330+3+2+2} = 0.970$$

$$p_{no} = \frac{3+2}{330+3+2+2} \cdot \frac{3+2}{330+3+2+2} = 0.0002$$

$$p_e = 0.970^2 + 0.02^2 = 0.9764$$

$$Kappa = \frac{(0.981-0.9764)}{(1-0.9764)} = 0.812$$

Music Era classification based on the LSTM model is performed by BoW and Word2Vec in order to prove and analyze the feasibility of the IMD dataset that was created before. BoW and Word2Vec feature x_i will be forwarded into LSTM architecture which consists of 4 gates: input i_t in Equation 6, output o_t in Equation 7, forget f_t in Equation 8 and Candidate C_t in Equation 9. For each gate, the formula can be computed as follow:

$$i_t = \sigma(W_i * [\,C_{t-1}, h_{t-1}, x_t] + b_i) \quad (7)$$

$$o_t = \sigma(W_o * [\,C_t, h_{t-1}, x_t] + b_o) \quad (8)$$

$$f_t = \sigma(W_f * [\,C_{t-1}, h_{t-1}, x_t] + b_f) \quad (9)$$

$$C_t = f_t * C_{t-1} + (1 - f_t) * {\sim}C_t \quad (10)$$

Input gate layer decides what value will be updated, forget gate layer decides previous state will be kept or thrown. Output gate layer decides the output and the candidate is the value that will be added into state. Detail LSTM architecture as follow:

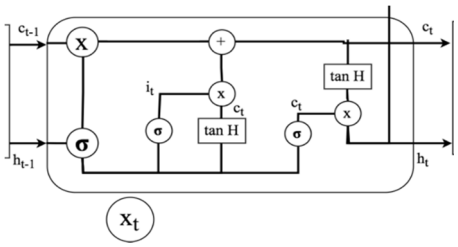


Fig. 7. Long Short Term Memory

Training is performed in GPU computer GTX 1050 with Cuda Architecture in order to decrease training time. 30 epoch training was performed for each BoW and Word2Vec data, then Validation data was used to measure accuracy and loss compared with Training data in order to prevent overfitting. After performing training for each design experiment in order to produce appropriate models, testing data is used to test overall accuracy of the model. Complete prediction results for each design experiment can be seen in Table 3.

TABLE III. TEXT CLASSIFICATION RESULT

Class	*Genre*	*Precision*	*Recall*	*F-score*	*Accuracy*
Genre	Rock	0.94	0.92	0.93	95%
	Pop	0.95	0.98	0.96	
	Jazz	0.94	0.90	0.92	
Mood	Sad	0.85	0.83	0.84	85%
	Happy	0.84	0.85	0.83	
	angry	0.84	0.80	0.82	

	relax	0.84	0.80	0.82	
Era	1970	0.95	0.93	0.94	94%
	1980	0.94	0.95	0.93	
	1990	0.94	0.90	0.92	
	2000	0.94	0.90	0.92	
	2010	0.95	0.93	0.94	

V. Discussion

IMD can be used by researchers to try out models or algorithms in MIR tasks specifically in genre, mood and era classification. We encourage researchers to use this dataset by sending e-mail to us for obtaining a complete dataset. This dataset can be a benchmark comparing other dataset like Million Song Dataset (MSD) [11], Free Music Archive (FMA) [12] and others. The advantages of using this dataset are: (1) this dataset contains audio features (Spectrogram, Chroma features and low-level features) extracted from a thousand Indonesian songs that have file size less than 20 Mb. This dataset is also contain lyric features (BoW and Word2Vec), (2) this dataset does not contain raw audio files like FMA, so researcher can merely focus on models not preprocessing dataset, (3) this dataset is equipped with Era labels, Mood labels and Genre labels that focus on Indonesia Music Dataset, (4) this dataset has already proved by using benchmarking model such as CNN and DNN in audio features and LSTM in text features that can produce promising results.

Based on experiments using CNN model to Audio Features, we can produce overall testing accuracy 74% and 72% for Spectrogram and Chroma features respectively and 75% for Low-level Audio Features that indicated our dataset is feasible and positively used by other researchers in MIR tasks. Based on the confusion matrix and histogram for each Era prediction, we can produce positive results of precision, recall and F1 score that indicate our IMD dataset was correctly collected.

Based on experiments using the LSTM model to Text Features, we can produce overall testing accuracy 95%, 85% and 94% for Genre, Mood and Era respectively, indicating that our dataset is feasible and positively used by other researchers in MIR tasks. Based on the confusion matrix and histogram for each Era prediction, we can produce positive results of precision, recall and F1 score that indicate our IMD dataset was correctly collected.

VI. Conclusion

In this paper, we have shown Indonesian Music Dataset (IMD) that contains Audio Feature as well as Lyrics Features is feasible to be used in Music Information Retrieval tasks like Genre, Mood and Era classification. This dataset contains a thousand of high-level audio features like Spectrogram and Chroma Feature, low-level audio features like ZCR, SC and RMSE and also contains text audio features such as BoW and Word2Vec. This dataset is also equipped by Mood, Era and Genre labels that have already been validated using Kappa Score. To prove our IMD dataset can be correctly used in MIR tasks, we try Music Era Classification using Convolutional Neural Networks (CNN) to audio features. Based on experiment design, we show both 74% and 72% of overall accuracy in predicting Spectrogram and Chroma features respectively, meanwhile in low-level we obtained promising results with 75% of accuracy. We also show, average Precision, Recall and F1 Score for each era classification to Spectrogram feature reach 0.78, 0.77 and 0.77. On the other hand, in text lyrics classification, we obtained accuracy 95%, 85% and 94% for Genre, Mood and Era respectively that indicated our dataset feasible and positively used by other researcher in MIR tasks

In the future, we plan to collect more and more datasets in order to augment existing IMD dataset. We also plan to complete other class labels of MIR.

Acknowledgment

We are pleased to dedicate this work to Mirna Adriani from the Faculty of Computer Science, Universitas Indonesia who passed away in 2020 who gave suggestions and deep discussion.

References

[1] A. Tsaptsinos, “Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network,” 2017 Int. Soc. Music Inf. Retr., pp. 694–701, 2017.

[2] J. Irvin, E. Chartock, and N. Hollander, “Recurrent Neural Networks with Attention for Genre Classification,” 2016.

[3] X. Hu and J. S. Downie, “Improving mood classification in music digital libraries by combining lyrics and audio,” in Proceedings of the 10th annual joint conference on Digital libraries - JCDL ’10, 2010, p. 159.

[4] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional Recurrent Neural Networks for Music Classification,” 2017 ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 1–5, Sep. 2017.

[5] M. Schedl, E. Gómez, and J. Urbano, Music Information Retrieval: Recent Developments and Applications, vol. 8, no. 2–3. 2014.

[6] A. Ramachandran, S. Vasudevan, and V. Naganathan, “Deep Learning for Music Era Classification,” pp. 1–8, 2016.

[7] W. Chu, “Movie Genre Classification based on Poster Images with Deep Neural Networks,” 2017 Proc. Work. Multimodal Underst. Soc. Affect. Subj. Attrib., pp. 39–45, 2017.

[8] I.-Y. Jeong and K. Lee, “Learning Temporal Features Using a Deep Neural Network and its Application to Music Genres Classification,” Ismir, pp. 434–440, 2016.

[9] G. Gwardys and D. Grzywczak, “Deep image features in music information retrieval,” Int. J. Electron. Telecommun., vol. 60, no. 4, pp. 187–199, 2014.

[10] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[11] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, “The Million Song Dataset,” in Proceedings of the 12th International Conference on Music Information Retrieval ({ISMIR} 2011), 2011.

[12] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A Dataset For Music Analysis,” 2017 Int. Soc. Music Inf. Retr., Dec. 2017.

[13] C. N. Silla, A. L. Koerich, and C. A. A. Kaestner, “THE LATIN MUSIC DATABASE.”

[14] P. Knees and M. Schedl, Music Similarity and Retrieval, vol. 36. 2016.

[15] J. Irvin, E. Chartock, and N. Hollander, “Recurrent Neural Networks with Attention for Genre Classification,” 2016.

[16] G. Keunwoo Choi and K. C. orgy Fazekas, Mark Sandler, “Convolutional Recurrent Neural Networks for Music Classification,” arXiv Prepr. arXiv1609.04243v3, pp. 1–5, 2016.

[17] Y. M. G. Costa, L. S. Oliveira, and C. N. Silla, “An evaluation of Convolutional Neural Networks for music classification using spectrograms,” Appl. Soft Comput., vol. 52, pp. 28–38, 2017.

[18] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., “ImageNet Classification with Deep Convolutional Neural Networks,” Adv. Neural Inf. Process. Syst. 25, pp. 1–9, 2012.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” NIPS, p. 9, Sep. 2014.

[20] K. Choi, G. Fazekas, K. Cho, and M. Sandler, “A Tutorial on Deep Learning for Music Information Retrieval,” arXiv:1709.04396, Sep. 2017.

[21] C. D. Manning, P. Raghavan, and H. Schutze, “An Introduction to Information Retrieval,” Online, no. c, p. 569, 2009.

[22] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, “Stemming Indonesian,” ACM Trans. Asian Lang. Inf. Process., vol. 6, no. 4, pp. 1–33, Dec. 2007.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” 2013 NIPS, Oct. 2013.

[24] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[25] S. Ruder, “An overview of gradient descent optimization algorithms,” Jun. 2016.