e-ISSN: 2614-8404
p-ISSN:2776-3234

Music Genre Recommendations Based on Spectrogram Analysis Using Convolutional Neural Network Algorithm with RESNET-50 and VGG-16 Architecture

I Nyoman Purnama[1]
[1]Sistem Informasi, STMIK PRIMAKARA
Email: [1]purnama@primakara.ac.id

Abstract – Recommendations are a very useful tool in many industries. Recommendations provide the best selection of what the user wants and provide satisfaction compared to ordinary searches. In the music industry, recommendations are used to provide songs that have similarities in terms of genre or theme. There are various kinds of genres in the world of music, including pop, classic, reggae and others. With genre, the difference between one song and another can be heard clearly. This genre can be analyzed by spectrogram analysis. Convolutional Neural Network(CNN) is a neural network algorithm that is commonly used to recognize and classify image data. In this study, an image spectrogram analysis was developed which will be the input feature for the Convolutional Neural Network. CNN will classify and provide song recommendations according to what the user wants. In addition, testing was carried out with two different architectures from CCN, namely VGG-16 and RESNET-50. From the results of the study obtained, the best accuracy results were obtained by the VGG-16 model with 20 epochs with accuracy 60%, compared to the RESNET-50 model with more than 20 epochs. The results of the recommendations generated on the test data obtained a good similarity value for VGG-16 compared to RESNET-50.

Keywords – Recommendation, VGG16, Resnet50, CNN, Spectogram, Music

I. INTRODUCTION

Music is an inseparable part of people's lives. Music often accompanies someone in their activities. Sometimes listening to music can also affect the mood of the listeners. Usually someone listens to music according to his feelings at that time. So that the role of music becomes important in managing people psychology[1].

Correct song listened by someone can affect listener's feelings. Due to the large amount of music available either through the internet or other music service applications, it will be difficult for people to choose the songs they want. Music is also distinguished by a variety of genres, speed, tempo and themes that vary and vary widely[2]. For western songs, the genres are distinguished by Hip-Hop, International, Electronic, Folk, Experimental, Rock, Pop, and Instrumental. This makes it difficult for music lovers to choose the right song.

Music lovers usually choose songs using manual method in finding the desired music. Like asking for recommendations from friends or listening to music shows to choose music[3]. Often the song that is listened to, does not match his mood or is not a fan of the genre of the song. Recommendations are implemented in various music player platforms on the internet, to provide more experience in listening to the music. The recommendation system is able to predict the favorite music desired by the user. Besides for users, recommendations are also useful for music service providers, because they can increase user satisfaction for using the music service.

Deep learning is a part of Artificial Neural Network-based Machine Learning. With deep learning, a computer can classify and recommend data in the form of images or sounds[4]. One of the methods commonly used for the classification and recommendation process is the Convolutional Neural Network (CNN). CNN is an extension of Multilayer Perceptron. CNN is able to learn from an image by using supervised learning techniques. This technique will provide a target for output by comparing past learning experiences. There are several architectures that can be used to optimize CNN so that they can have optimal classification results. There are the VGG architecture, mobileNet, ResNet etc. ResNet, short for Residual Networks is a classic neural network[5]. This model is also the winner of the ImageNet challenge in 2015. This model is also easier to optimize, and can get accuracy from great depths increases. ResNet 50 is the best CNN architecture, it is proved on the research by Talo was to conduct research on the classification of brain diseases with MRI images[6]. The spectrogram is a visual representation of the frequency spectrum of the signal. The spectrogram is formed using the Fourier transform. Making a spectrogram with FFT (Fast Fourier Transform) is done by first taking the data in the time domain, and breaking the data into several parts, and doing a Fourier Transform to calculate the magnitude of the frequency spectrum for each part.

The spectrogram is very useful for analyzing sound, where the spectrogram forms a ratio of magnitude to frequency at a given time. Music recommendations can also be made based on the mood of the user. Where in the research that has been done by Amala George et al. The system developed is able to analyze the mood of the user based on his face, then analyze it using the CNN algorithm[2]. From the results of this mood classification, recommendations are then given using the recommendation module. From the research that has been done, the accuracy is 98%.

Research of music types Classification using the CNN algorithm has also been carried out by analyzing spectrogram images. The spectrogram image that has been generated from the music, then deep learning process will be carried out by using the CNN model. Based on the research that has been done, it is found that the use of 35 epochs has an optimal accuracy of 81.33%. When compared with the KNN method, CNN produces a better level of accuracy[1]. Other research on spectrogram analysis for music genre classification using CNN and Mel-spectograms has been carried out and the test results depend on the number of datasets, training iterations and computer specifications greatly affect the level of accuracy and duration of modeling. The resulting accuracy is very optimal in classifying music genres, which is 99% for the RELU activation function and 95% for ELU[7].

Music recommendations based on genre have also been carried out using the Convolutional Recurrent Neural Network. Where in this study also uses a spectrogram and analyzes it using CRNN. This study also compared the use of CRNN and CNN methods to classify music genres. From the research results, it is found that CRNN which takes into account the frequency and time sequence features has better performance than CNN[8]. Research on next-song recommendation has also been carried out, where Neural network has performed well in all types of tests. In this study it was concluded that the NN-based next-song recommenders, CNN-rec, NN-rec and Word2Vec, outperform the non-NN based ones[9]. In this research demonstrate that the NN-based next-song recommenders, which combine users' general preference and sequential listening patterns, have the highest performance.

Music recommendation using deep content also done by A¨aron van den Oord dkk[10]. In their research showed that recent advances in deep learning translate very well to the music recommendation setting in combination with approach used in this study, with deep convolutional neural networks significantly outperforming a more traditional approach using bag-of-words representations of audio signals. Also other research on music recommendation done by using user behaviour [11]. The approach considered genre, recording year, freshness, favor and time pattern as factors to recommend songs. The evaluation results demonstrate that the approach is effective.

Research on music recommendations by genre is carried out by comparing several machine learning algorithms such as KNN, RF, NB, DT dan SVM[12]. According to the results summarized in this research, SVM achieved better classification results than other methods. In addition, changing the window size and window type caused very small performance changes. Research about music recommendation using similarity between using decided genre value and using feature vector distance also have been done by Jonseol Dee et al. In their paper, proposed a recommendation system based on a preference classification using real-time user brainwaves and genre feature classification. Proposed user's preference clasifier achieved an overall accuracy of 81.07%[13].

Based on the research that has been done previously, this study will carry out a music genre recommendation process using the GZTAN dataset which is composed of 10 types of genres, where the music data is first processed using a spectrogram. The results will be classified using the CNN algorithm with RESNET50 and VGG16 architecture. The results of the recommendations generated will be tested whether they are in accordance with the song desired by the user.

II. Research Methodology

The method used in this research is the dataset preparation process, pre-processing, spectrogram, classification process and calculating similarity using cosine similiarity.

A. Dataset

This research uses a dataset in the form of spectrogram images taken from the GTZAN dataset. To simplify the classification of music data using a neural network, it is necessary to change the music data data into a mel-spectrogram to be processed by the Neural Network. GTZAN consists of music data and Mel spectrogram results from that music file. Where this dataset is a public dataset that is widely used for evaluating the introduction of music genres (Music Genre Recognition / MGR). GTZAN is a collection of music collected from 2000-2001, which comes from various sources such as CDs, radio and microphone recordings. This dataset consists of 10 genres, namely blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. The duration of each of these music is 30 seconds. Each genre contains 100 music files. The number of datasets used in this study is divided into 3 parts : training data, validation data and test data. With details in each section as follows:

Table 1. Number dataset used on each class

Genre	Training data	Validation data	Testing data
Blues	80	10	10
Classical	80	10	10
Country	80	10	10
Disco	80	10	10
Hiphop	80	10	10
Jazz	80	10	10
Metal	80	10	10
Pop	80	10	10
Regae	80	10	10
Rock	80	10	10

B. Spectogram

The spectrogram is a visual representation of the frequency spectrum of the signal[14]. In the GTZAN dataset, spectrograms have been generated and stored in their respective classes. Before being entered into the CNN network, this data is further divided into training data, validation data and test data. Each of these spectrogram images will be included in the array, then labeled according to their respective index folders. Then after being given a label, the data will be appended into an array to make it easier to pass the data. From the spectrogram image there are many values and features of the music file that can be displayed. The following is an example of an illustration of the spectrogram of each class in this study.

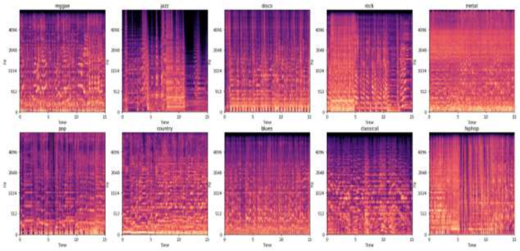


Figure 2 Example of Spectogram on each classes

Based on the picture above, it can be seen that there are differences in the spectrograms of each genre. The image on the right shows a spectrogram for the hip hop and rock genres, here you can see the frequency density compared to the spectrogram on the left. The spectrogram image is a wave generated as an audio representation in the time, frequency and magnitude domains. To generate spectrograms from each music genre and use it on an artificial neural network, in this study, the Librosa library was used. With librosa, we can retrieve important features in a music file, such as tempo, chroma and spectrogram.

C. Convolutional Neural Network

Convolutional Neural Network is an artificial neural network that is widely used in the field of image classification. In this study, the audio/music signal is represented as a spectrogram which has a 2D image. CNN is used to classify music genres with the help of spectrograms. Based on the spectrogram images of each music genre, the pattern of the audio signal can be seen. So that each of these genres can be input of the CNN artificial neural network.

In this study, two CNN architectures were used, namely Resnet and VGG16. Resnet is a residual network, which is in charge of image recognition. RESNET-50 is an improved version of VGG-16. Where the last number of this architecture represents the number of layers in the network. RESNET stands for Residual Network which is an artificial neural network innovation that won the 2015 ILSVRC classification competition with an error rate of only 3.15%[15]. While VGG-16 stands for Visual Geometry Group and 16 is the number of layers. The VGG-16 is also a well-known model that participated in the 2014 ILSVRC and obtained an accuracy rate of 92.7%. VGG-16 is also used in image classification and is very popular because of its ease of implementation. The following in Figure 3 is a comparison of the RESNET and VGG architectures.

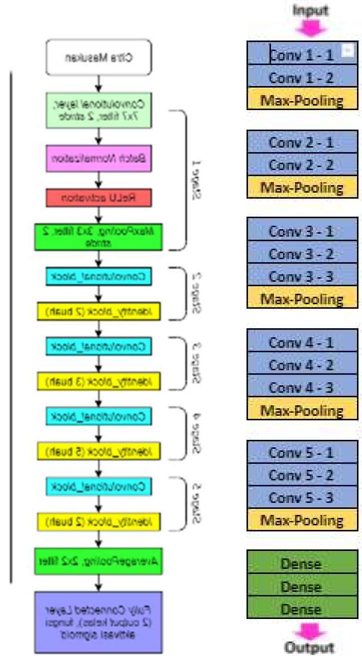


Figure 3. Comparison of Resnet-50 and VGG-16 . architectures

D. Research flow

The research flow used in this study can be described in outline as follows:

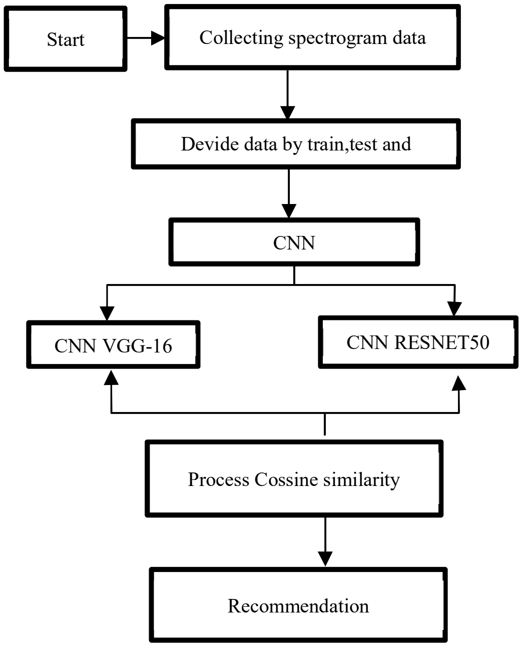

Start
Collecting spectrogram data
Devide data by train,test and
CNN
CNN VGG-16
CNN RESNET50
Process Cossine similarity
Recommendation

Figure 4. Research flow

The first step in this research is to collect the dataset, in the form of a spectrogram image from the GTZAN dataset. After that by using the required libraries such as *imageDataGenerator* in the Keras library to manage training, test and validation data. After that the MFCC image data from 10 music genres have been grouped by category. The next step is to build a CNN model with Keras.

There are 2 different processes will be carried out using the VGG-16 and Resnet 50 architectures.

After the model is obtained from the training process using two different architectures. Then the process of finding similarities between feature vectors is carried out using cosine similarity. The application will display a recommendation of 5 songs that match those in the validation data. Where the recommendation process is carried out by calculating the value of the similarity of features between one music and another. The first process is to choose music from each genre that will be used as the basis for the recommendation system. Then the forecast from the music base is calculated based on an artificial neural network. The cosine similarity value is calculated from the 2 featured vector being compared. To calculate the similarity of 2 pieces of music with the number of features N, where the first music has a feature vector x=[x1,x2,x3….xn] and the second music has a feature vector y=[y1,y2,y3,…yn] then the formula which is used as follows:

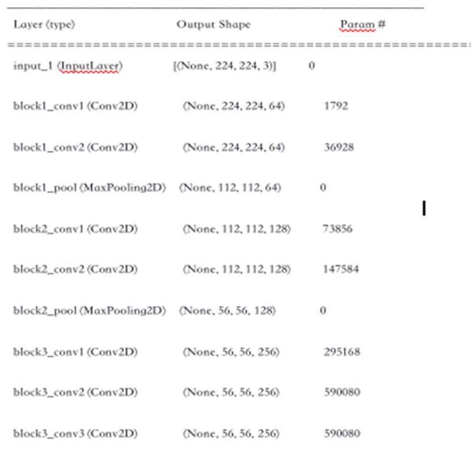
$$\cos\theta = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Figure 5. Cossine similarity formula

III. Results And Discussion

System implementation is done using Google Colabs. The libraries used in the making of this research are numpy, pandas, librosa, Keras and Scikit learns. This research uses a spectrogram image dataset obtained from the GTZAN dataset with a total of 1000 music data and is divided into 10 categories namely Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, Rock. This spectrogram image will be the input for the Convolutional Neural Network. Where the image in the form of a mel spectrogram representation of this audio file is saved in ".jpg" format.

To build the image dataset in this study, the *ImageDataGenerator* library was used. As a parameter of this library, we must divide the GTZAN dataset into 3 folder namely training, test and validation data. The number of classes used is 10 classes which are divided into their respective folders. Some parameters that must be initialized are batch_size=64 and the number of initial epochs used is 20. An important parameter that we need to initialize is Input_shape for all images, in this study we set the input shape at (224,224.3) / RGB channel and normalize image with a scale of 1./255. To implement CNN in this research, using Python programming language with Keras library and tensorflow. CNN modeling is done by initializing the CNN network layer parameters, namely the number of conv2d layers in the model=2, the number of conv2d layers=32, filter size=(3,3), initializer=glorot_uniform, activation function=relu, layer dropout=0.2 and the optimized optimizer. "Adam" is used. The CNN model makes several layers, namely convolution layers, pooling layers, dropout layers, flatten layers, dense layers and the RELU activation function. The result of the convolution process is a feature map that is used in the convolution process repeatedly. The resulting model output is shown as shown below.



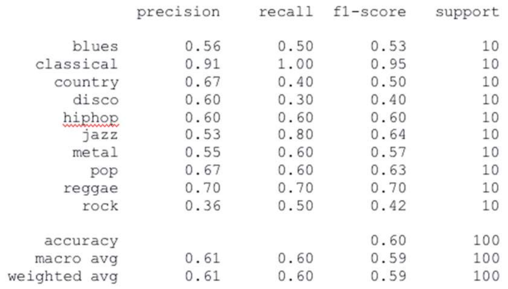
```
Layer (type)                  Output Shape              Param #
==================================================================
input_1 (InputLayer)          [(None, 224, 224, 3)]     0

block1_conv1 (Conv2D)         (None, 224, 224, 64)      1792

block1_conv2 (Conv2D)         (None, 224, 224, 64)      36928

block1_pool (MaxPooling2D)    (None, 112, 112, 64)      0

block2_conv1 (Conv2D)         (None, 112, 112, 128)     73856

block2_conv2 (Conv2D)         (None, 112, 112, 128)     147584

block2_pool (MaxPooling2D)    (None, 56, 56, 128)       0

block3_conv1 (Conv2D)         (None, 56, 56, 256)       295168

block3_conv2 (Conv2D)         (None, 56, 56, 256)       590080

block3_conv3 (Conv2D)         (None, 56, 56, 256)       590080
```

Figure 6. CNN output model

The next process is to carry out the transfer learning process with 2 different architectures, namely VGG16 and RESNET50. Transfer learning is the process of using an existing model for different problems. By using transfer learning, it is hoped that the results of the training will be better. The parameters needed in this transfer learning process are *lastfourtrainable*, if the value of this parameter is false then the last fully connected layer will be trained. If true then the last 4 layer models that have parameters will be trained. For these two architectural models, "adam" optimization is used. The training process was carried out on each model of 20 epochs. This training will produce a model that will be used in the testing process.

A. Result analysis

After the training process was carried out, the precision, recall, and f1-score values were obtained from each music class. Following are the values of Precision, recall, f1-score and accuracy on the VGG16 model.



```
              precision    recall  f1-score   support

       blues       0.56      0.50      0.53        10
   classical       0.91      1.00      0.95        10
     country       0.67      0.40      0.50        10
       disco       0.60      0.30      0.40        10
      hiphop       0.60      0.60      0.60        10
        jazz       0.53      0.80      0.64        10
       metal       0.55      0.60      0.57        10
         pop       0.67      0.60      0.63        10
      reggae       0.70      0.70      0.70        10
        rock       0.36      0.50      0.42        10

    accuracy                           0.60       100
   macro avg       0.61      0.60      0.59       100
weighted avg       0.61      0.60      0.59       100
```

Figure 7. VGG16 Accuracy value

The results of the confusion matrix for the CNN-VGG16 model are shown in Fig. where the results obtained are quite good in classifying the class of the music dataset used. The best classification process was obtained from classical, jazz and reggae classes. While the lowest classification was obtained in the disco class.

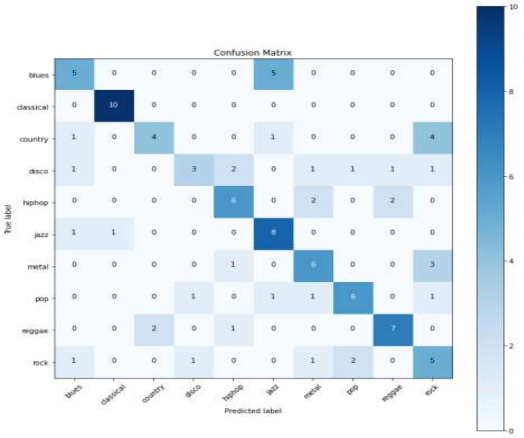


Figure 8. VGG16 Confusion matrix

After the process of model formation with transfer learning VGG16 and RESNET50. Then proceed with making feature extraction. In the VGG16 model, for example, in this study, we will take a model that has been previously stored in the training process. After that, the output weight will be obtained before the classification layer of this model. From this model we will derive the feature vectors for the training and validation datasets. The result of this feature vector is then searched for its similarity with cosine similarity.

In Figure 9, 5 music recommendations are obtained based on the spectrogram image of the music file desired by the user. Where "test image" is the music spectrogram testing data. As Seen with VGG16, the recommended spectrogram has almost the same shape as the test spectrogram. With the test image from the Blues class, the recommendation results are also obtained from the Blues class with a similarity level of 1.

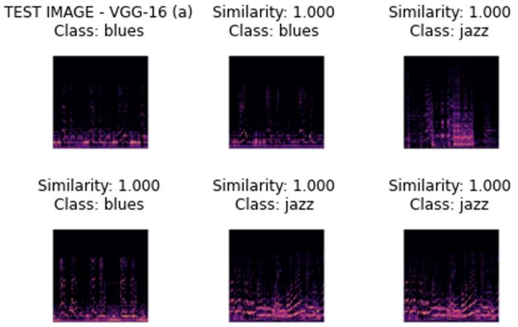
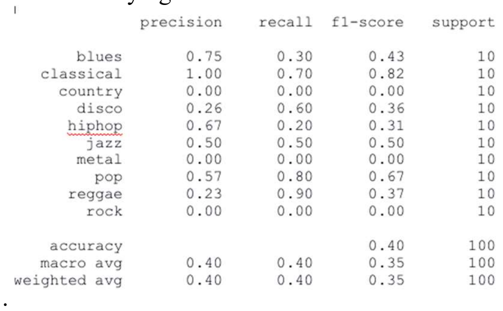


Figure 9. Recommendation output for VGG16 model

While on RESNET50, it has the same testing process as VGG16. After the experiment, the training process with RESNET50 requires a larger epoch to get better accuracy results. In this study, quite good accuracy results were obtained at epochs of 30 for RESNET50. The picture below shows the calculation results of Precision, recall, f1-score and accuracy on the RESNET50 model. The resulting Accuracy value is slightly lower than the VGG16 model with a larger number of epochs. The results of the confusion matrix for the CNN-RESNET50 model are shown in Fig. Where the results obtained are still lower than the VGG16 model in classifying classes from the music dataset used.



```
              precision    recall  f1-score   support

       blues       0.75      0.30      0.43        10
   classical       1.00      0.70      0.82        10
     country       0.00      0.00      0.00        10
       disco       0.26      0.60      0.36        10
      hiphop       0.67      0.20      0.31        10
        jazz       0.50      0.50      0.50        10
       metal       0.00      0.00      0.00        10
         pop       0.57      0.80      0.67        10
      reggae       0.23      0.90      0.37        10
        rock       0.00      0.00      0.00        10

    accuracy                           0.40       100
   macro avg       0.40      0.40      0.35       100
weighted avg       0.40      0.40      0.35       100
```

Figure 10. RESNET50 Accuracy value

The results of the confusion matrix for the CNN-RESNET50 model are shown in Fig. Where the results obtained are still lower than the VGG16 model in classifying classes from the music dataset used. The best classification is obtained from the reggeae and pop classes. In rock class, the RESNET50 model is not able to give good classification results.

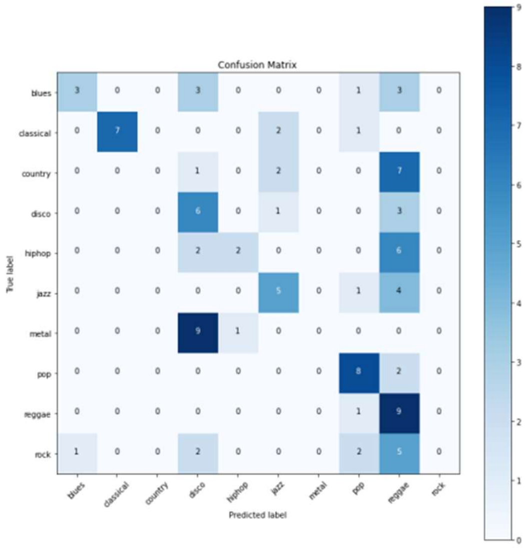


Figure 11 RESNET50 Confusion matrix

For the testing process, the steps taken are the same as the process in the VGG16 model, that is looking for feature extraction from the test image and looking for its cosine similarity with feature extraction from the training dataset. So that the results of the music spectrogram recommendations are obtained in accordance with the testing dataset used. The following is in Figure 12 the results of 5 image similarities from the tested test data. It can be seen that the results of the spectrogram recommendation are quite good, only the level of similarity is lower than the VGG16 model.

e-ISSN: 2614-8404
p-ISSN:2776-3234

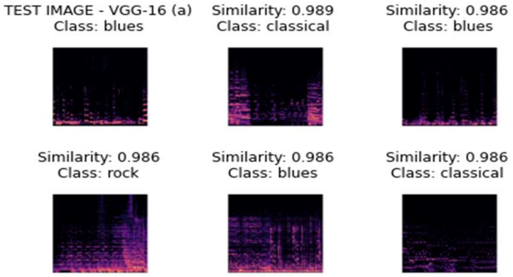


Figure 12. Recommendation output for RESNET50 model

IV. Conclusion

In this research it is implemented using Python, Google Colab, and TensorFlow and hard libraries. Input shape on CNN model in This research is 224x224 pixels, the filter size is 3x3, the number of epochs is 20 and 30, and the training data is 799 and the validation is 100 data. CNN is the most widely used method in image data. For research with audio data, this data is first processed by spectrogram analysis in the form of Cartesian coordinates with the amplitude of the music as the y-axis. In this study, the spectrogram results become input for CNN with VGG16 and RESNET50 architectures.

Based on the results of the design of prediction system using the CNN method, the accuracy value for the VGG16 training data model is 0.8408, the training data loss is 0.4827, the test data accuracy is is 0.6094 and the test data loss is 1.2762. Meanwhile, for the RESNET50 model, the training data accuracy value is 0.6286, the training data loss is 1.0383, the test data accuracy is is 0.3438 and the test data loss is 1.8529. So, from these results it can be concluded that the results in both the data is still underfitting. This is because there are still many datasets that are more numerous in number and variants that have characteristics that are similar to each class.

The best accuracy results were obtained by the VGG16 model with 20 epochs compared to the RESNET50 model with more than 20 epochs. The results of the recommendations generated on the test data obtained a good similarity value for VGG16 compared to RESNET50. The suggestion for this research is that in the future it can increase the dataset so that the accuracy obtained is even better, because in this study the songs in the dataset do not have clear boundaries between one genre and another. In addition, the epoch value during the training process is also further improved so that the accuracy level is even better for each CNN model.

References

[1] Y. M. G. Costa, L. S. Oliveira, A. L. Koericb, and F. Gouyon, “Music genre recognition using spectrograms,” *Int. Conf. Syst. Signals, Image Process.*, pp. 151–154, 2011.

[2] A. George, S. Suneesh, S. Sreelakshmi, and T. E. Paul, “Music Recommendation System Using CNN,” vol. 9, no. 6, pp. 4197–4200, 2020.

[3] C. R. Wairata, E. R. Swedia, and M. Cahyanti, “Pengklasifikasian Genre Musik Indonesia Menggunakan Convolutional Neural Network,” *Sebatik*, vol. 25, no. 1, pp. 255–261, 2021, doi: 10.46984/sebatik.v25i1.1286.

[4] J. Dias, “Music genre classification from Spectrogram using CNN,” [Online]. Available: http://cs230.stanford.edu/files_winter_2018/projects/6936608.pdf.

[5] Faiz Nashrullah, Suryo Adhi Wibowo, and Gelar Budiman, “The Investigation of Epoch Parameters in ResNet-50 Architecture for Pornographic Classification,” *J. Comput. Electron. Telecommun.*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.52435/complete.v1i1.51.

[6] M. Talo, O. Yildirim, U. B. Baloglu, G. Aydin, and U. R. Acharya, “Convolutional neural networks for multi-class brain disease detection using MRI images,” *Comput. Med. Imaging Graph.*, vol. 78, p. 101673, Dec. 2019, doi: 10.1016/J.COMPMEDIMAG.2019.101673.

[7] D. Lionel, R. Adipranata, and E. Setyati, “Klasifikasi Genre Musik Menggunakan Metode Deep Learning Convolutional Neural Network dan Mel- Spektrogram,” *J. Infra Petra*, vol. 7, no. 1, pp. 51–55, 2019, [Online]. Available: http://publication.petra.ac.id/index.php/teknik-informatika/article/view/8044.

[8] Adiyansjah, A. A. S. Gunawan, and D. Suhartono, “Music recommender system based on genre using convolutional recurrent neural networks,” *Procedia Comput. Sci.*, vol. 157, pp. 99–109, 2019, doi: 10.1016/j.procs.2019.08.146.

[9] K.-C. Hsu, S.-Y. Chou, Y.-H. Yang, and T.-S. Chi, “Neural Network Based Next-Song Recommendation,” 2016, [Online]. Available: http://arxiv.org/abs/1606.07722.

[10] D. G. W. Ingram *et al.*, “Computer Aided Design, International Conference, University of Southampton, Engl, Apr 24-28 1972.,” *Inst Electr Eng, Conf Publ*, no. 86 . IEE, pp. 1–9, 1972.

[11] Y. Hu, “12th International Society for Music Information Retrieval Conference (ISMIR 2011) NEXTONE PLAYER : A MUSIC RECOMMENDATION SYSTEM BASED ON USER BEHAVIOR,” no. Ismir, pp. 103–108, 2011.

[12] A. Elbir, H. Bilal Çam, M. Emre Iyican, B. Öztürk, and N. Aydin, “Music Genre Classification and Recommendation by Using Machine Learning Techniques,” *Proc. - 2018 Innov. Intell. Syst. Appl. Conf. ASYU 2018*, no. October 2018, 2018, doi: 10.1109/ASYU.2018.8554016.

[13] J. Lee, K. Yoon, D. Jang, S. J. Jang, S. Shin, and J. H. Kim, “Music recommendation system based on genre distance and user preference classification,” *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 5, pp. 1285–1292, 2018.

[14] M. H. Ashshiddieqy, Jondri, and A. Rizal, “Klasifikasi Suara Paru Dengan Convolutional Neural Network (CNN),” *eProceedings Eng.*, vol. 07, no. 02, pp. 8506–8512, 2020.

[15] W. Setiawan, “Perbandingan Arsitektur Convolutional Neural Network Untuk Klasifikasi Fundus,” *J. Simantec*, vol. 7, no. 2, pp. 48–53, 2020, doi: 10.21107/simantec.v7i2.6551.