National College of Ireland

Impact of Missing Data on Audio Genre Classification using Convolutional Neural Network

MSc Research Project
MSc in Data Analytics

Raunak Milind Sathe
Student ID: x20118350

School of Computing
National College of Ireland

Supervisor: Michael Bradford

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Raunak Milind Sathe
Student ID:	x20118350
Programme:	MSc in Data Analytics **Year:** 2021-2022
Module:	Research Project
Supervisor:	Michael Bradford
Submission Due Date:	31/01/2022
Project Title:	Impact of Missing Data on Audio Genre Classification using Convolutional Neural Network
Word Count:	6985 **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Raunak Milind Sathe

Date: 31/01/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Impact of Missing Data on Audio Genre Classification using Convolutional Neural Network

Raunak Milind Sathe
x20118350

Abstract

Genre classification of audio files has many applications such as recommendation systems, Digital Signal Processing. Missing data in audio files is very common in real world applications. In this project a novel approach has been implemented to study the impact of missing data on genre classification using deep learning. Convolutional Neural Network is a very powerful deep learning algorithm which is very popular in classification tasks. The dataset used is the GTZAN dataset containing 1000 audio samples from 10 genres. In this project, three experiments were run, first the CNN model using the VGG16 architecture was trained on the GTZAN dataset and then two models were implemented on the two processed GTZAN dataset where segments were removed from the dataset to make new, shorter datasets. Surprisingly, the model showed no downgrade in performance on the processed dataset and thus, the conclusion was reached that missing data need not be very crucial in achieving genre classification.

Keywords— CNN, VGG16, Genres, Recommendation, Missing Data, GTZAN

1. Introduction

Genre classification is a very popular area of research due to its increasing applications in the fields of recommendation systems, digital signal processing etc. Over the last few years with the advent of modern technologies, internet and modern electronic devices the audio data has exploded making genre classification ever so important. (Song, Dixon, and Pearce 2012) have studied the explosion of audio data. They concluded that for this data to be usable, the biggest hurdle remains in getting this audio data into a usable and organized format. One of the most prominent problems associated with Digital Signal Processing is the presence of noise of distortions in the audio signal. (Bako 2015) have studied the presence of distortions in audio signals. They studied the different factors for distortions and that it is very difficult to completely eliminate the presence of distortions or noise in audio signals. Some of these distortions maybe intentional or unintentional but are likely to persist unless a perfect ecosystem is provided which is very unlikely everywhere.

Convolutional Neural Networks (CNN) have become very popular over the last few years. (Sharma, Jain, and Mishra 2018) have studied the popularity of CNN in image processing. They studied the efficacy of the popular network GoogLeNet and concluded that this architecture showed high precision and accuracy for image classification problems. Another such image classification work was carried out by (Mo et al. 2019) who were able to generate a model with over 99% accuracy. Many such research works have been carried out over the last few years which all point towards the use of CNN for image classification.

However, since we are dealing with audio signals in this research, the use of CNN becomes insignificant unless the audio files are converted into image data. A lot of research work has been carried out in this domain over the last few years. (Wyse 2017) in their research work have discussed the possibility of using audio files as an input for CNN. There are other research works in this domain that all used CNN for audio classification by converting them into spectrograms.

In this project segments from the original audio have been removed to make shorter audio files where each audio file will be 15 seconds and 20 seconds long respectively. A CNN will be initially implemented on the original dataset which is first converted into an image as an input for the model. After fine tuning the model to get the best result, the same model will be trained on the processed datasets to see the results. The goal of this research paper will be to answer the following research paper -

How effective is the deep learning technique such as Convolutional Neural Network in classifying audio signals into genres when there is a significant amount of missing data present in the audio signals?

The following are the research objectives of the project -

- Study Audio Genre Classification using Convolutional Neural Network in the presence of significant missing data.

- Understand how the model responds to missing data in each individual genre.
- study whether model trained on a data with significant missing data can generate respectable results or is there a need for data imputation.

The findings from this research will have several applications not just in recommendation systems but also in applications of digital signal processing such as voice recognition, radar applications etc. Since the goal of this project is to study the impact of missing data, this work can be a reference for identifying the impact of information loss on future applications.

1.1. Report Structure

The report is divided into sections for the ease of understanding. The first section is the Introduction section which introduces the project topic and the motivation along with the research question. This is followed by the Related Work section which studies some of the work carried out in the domain, its outcomes and how this project will provide a novelty to this work. The next sections are the Research Methodology and the Design Specification of the project. This section is followed by the Implementation. This section includes all the details carried out at each stage along with the experiments and the models used. The next section is the results and evaluation where the project will be evaluated. This section is followed by the discussion and finally the Conclusion and future work. At the end of the report are the references.

2. Related Work

2.1. Area of Research

Before carrying out the literature review, it is critical to identify the domains of this project. The project is focused on the use of Audio to Image conversion to carry out Audio Genre classification using Convolutional Neural Network. The following diagram shows the areas of research -

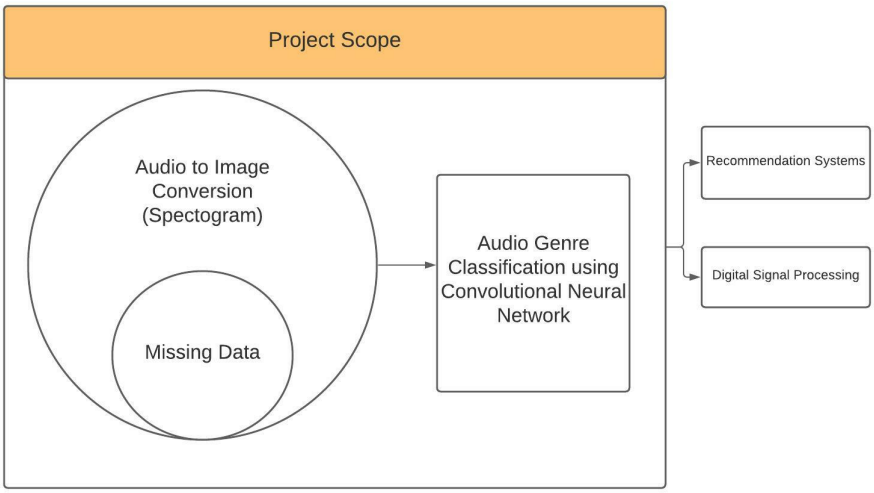

Figure 1: Areas of Research

2.2. Audio Genre Classification using CNN

The work carried out by (Elbir et al. 2018) is the closest to the research work that has been carried out in this paper. They have done a comparison of machine learning techniques on the GTZAN dataset which is also used for this study. They were able to conclude that SVM is the most suitable algorithm for this dataset and achieved higher accuracy than CNN. The research conducted by (Xu and Zhou 2020) also used the GTZAN dataset and also built a CNN model similar to this project. The research part of this paper included using the Squeeze and Excitation block (SE-Block). The SE-Block learns the weight depending on the loss. This is then included in the final layer of the model. Bayesian optimization technique was used to select the best features of the SE-Block. They originally built a 5 layer model and then in the experiment embedded the SE-Block by layer. The 2 experiments showed that accuracy improved from 83.2% to 91%. Another such research work was carried out by (J. Liu, C. Wang, and Zha 2021) where they used deep learning to classify music genres by extracting features from the audio files. Their research revolves around the fact there is very little interaction between the multi-level features and hence, there is a lack of learning features. They were able to achieve a very high accuracy of 93.2% which is better than any state-of-art models on the GTZAN dataset. The researchers (Agrawal and Nandy 2020) have proposed a novel approach of genre classification where they used 2 stages. The first stage is the standard CNN model using spectrograms after converting the audio files. In the second stage metadata is fetched which contains the title, artist and the lyrics. The result is then combined with the first stage to get the output. They were able to build a model with 76.2% accuracy which may not be as high as other research works but still a good accuracy given that this was a novel approach. The research paper (Nanni et al. 2021) conducted a similar experiment of using CNN on audio signals. They used 3 different dataets which included sounds of cats, birds and the ESC50 dataset. The standard process of converting them in spectrograms and then using pre-trained models on them

worked very well achieving a minimum accuracy of 88% and a very high accuracy of 97% on the ESC dataset. These two works have shown contrasting results where one research showed high accuracy on one dataset and other showed low accuracy. Another similar work has been carried out by (Jaiswal and Kalpeshbhai Patel 2018) where they compared CNN and DNN on the ESC dataset. A total of 5776 spectrograms were obtained from this dataset. The model was built which had the following features - 1 input layer and 1 output layer, 2 hidden layers, 1 activation function using the Relu function, 1 Optimizer which was the Adam optimizer and the model was run for 15 epochs. They were able to achieve a very impressive accuracy of 85% on the test data which exceeds the state-of-the-art models which produced an accuracy of 78%.

Since this project has an aspect of using transfer learning, a research into transfer learning has been carried out. A lot of research work has been carried out into transfer learning. (Weiss, Khoshgoftaar, and D. Wang 2016) carried out a survey of transfer learning approaches taken up by other researchers. They investigated different transfer learning approaches such as Homogeneous transfer learning which uses instances, features and parameters as a base for information transfer technique. Symmetric and asymmetric approaches are also widely used which form the Heterogeneous transfer learning approach which is relatively new. Another approach is the domain adaption process. The focus is on correcting the conditional distribution differences which is used when the dataset is not properly labelled. The research presented by (Oquab et al. 2014) is focused on the problems that arise when dealing with limited data. CNN is a model that requires large data. Their work focuses on using CNN trained on large datasets on smaller training datasets. The work focuses on reusing layers on smaller dataset. They were able to conclude that their model showed better result on the PASCAL VOC datasets despite the obvious limitations in the data. Their work was compared with the work carried out by (Marszalek et al. 2007), which was considered the state-of-art model on the PASCAL datasetand, and showed improvement in the performance.

The research carried out by (Yosinski et al. 2014) have studied the efficiency of transfer learning in deep learning. They observed that many times the CNN layers behave similarly. The first layer detects colour bubbles in the images as well Gabor filters. These features are common and present in other datasets. They first analyzed the negatives of transfer learning and scenarios of inefficacy and incompatibility. The high layer neurons designed for the particular task if vastly different from the original task and optimization challenges between neurons were the two major difficulties they noticed while using transfer learning. After running experiments on ImageNet, they concluded that even if the target task is vastly different from the original task, it can still improve the performance of the model much more than randomly selecting and tuning features.

The VGG architecture is of particular interest to us as it will be used in this research project. The research carried out by (Tammina 2019) has used the VGG-16 architecture for image classification. The dataset used contained 25000 images of cats and dogs. 3 experiments were conducted to study if pre-trained state-of-art VGG16 model will show any better results. The first model was a basic CNN model, the second a CNN model using image augmentation and lastly using VGG16 and image augmentation. The end

result showed that the VGG16 architecture showed a very high validation accuracy of 95.40%. Another such work was carried out (Guan et al. 2019) where they used the VGG16 architecture to identify cancer images from the benign ones. The dataset contained 279 images of thyroid nodules. They acheived a very high accuracy of 97.66% as compared to 92% using the Inception-v3. The results from both the above research papers show the high

A lot of work has been carried in this domain, particularly towards classifying animal or environmental sounds using CNN and pre-trained CNN models such as VGG19, GooglNet models that have shown good accuracy. A lot of this work is focused on comparison of models and thus defining the best performing model for a particular dataset. There is thus, a lack of novelty in this domain with many researchers focusing solely on paramterization, changing models and using different datasets. This situation is not ideal in a real world scenario where noise and information loss are common.

2.3. Effect of Missing Data in Audio Signals for Genre Classification

As we saw in the above review, the work was related to applying machine learning and deep learning techniques to audio datasets and classifying genres. In this section, we will take a look at research works where missing information has been accounted for while building audio classification models. This is a scenario that resembles real life applications closely. The research carried out by (Jiang, Chen, and Yuan 2005) is similar to the work carried out in this project where the impact of missing values has been studied. The approach that they have taken is dividing the dataset in subgroups which will then act as individual datasets as an input. Their approach is based on training different models on different subsections of the data and by putting them together, the full dataset can be utilized. They used AdaBoost and Bagging techniques and were able to generate very good results with accuracy in the 90% range. Another similar research work carried out in this domain was by (P. Liu et al. 2005) who focused more on using different techniques to fill the missing data values. Some of the methods used were Case deletion, using mean and mode for filling missing values, using a probability based model and finally using kNN. The model that they used was the Naive Bayes approach giving an average accuracy in the 50% range for 3 different datasets.

The researchers (Pikrakis et al. 2015) in their work have also focused on audio genre classification with missing labels. Their research work varies slightly from others where they focused more on missing labels rather than missing values. They used the Restricted Boltzmann Machines and and Dictionary Learning Algorithms. With an accuracy of 82.3% this was lower than the Deep Learning approach used by (Pikrakis 2013) where the accuracy was 84%. However, with the increased complexity of the deep learning architectures, the accuracy did not demonstrate a significant improvement. Thus, they concluded that the Restricted Boltzmann Machines is a sufficiently good algorithm for genre classification with missing data. The research work carried out by (Smaragdis, Raj, and Shashanka 2009) was in the similar domain of missing data imputation for image classification. However, their approach differed in the methodology part. The imputation of missing information was

carried out on the spectrograms after they were obtained from transforming the audio file. This was referred to as imputation in the time-frequency domain. The results were more the form of qualitative rather than quantitative and they were able to conclude that their proposed algorithm provided better results than SVD and KNN imputation. The research carried out by (Jang et al. 2019) was in a similar domain of imputing missing values using the zero-inflated denoising convolutional autoencoder. They scope of application was in the actigraphy devices worn on wrists. They used the parameters RMSE and MAE as the evaluation metrics. The objective of this research was to impute missing values using deep learning instead of statistical approaches. Their approach produced an RMSE of 839.3 counts which was better than the Bayesian Regression approach which showed a RMSE of 924.5 counts. Thus they were able to conclude that the deep learning approach generated better results than statistical approaches.

Image augmentation has gained popularity in deep learning as it increases the dataset without needing additional data input. The researchers (Shorten and Khoshgoftaar 2019) have conducted a survey to understand whether image augmentation should be used and if used what are the considerations. They concluded that image augmentation has many benefits in deep learning. One of the benefits was removing the bias from the data. The 2 most popular techniques data warping and oversampling have shown high potential in many research works. However, there seems to be no general agreement regarding image augmentation and how to calculate the new data size. There is still a risk of bias depending on the type of filters that will be used.

Recommendation systems can be divided in 2 types the content based and collaborative based. The researchers (Oord, Dieleman, and Schrauwen 2013) have discussed the drawbacks of these types of systems and also analyzed how deep learning has helped in achieving better performance than the existing state-of-art traditional approaches. Their research revolved around the use of neural nets to discover latent factors when there is a significant lack of user information. Genre based recommendation systems are very popular as with increasing data availability, service providers can recommend music based on what the users listen. The researchers (Lee et al. 2015) have focused on usage history and genre classification for music recommendation. They used 5 and 10 feature vectors with audio features to build a recommendation system. The distance metric learning algorithm was used which will reduce the dimensionality of the features. The data gathering process was a very complex process where they gathered usage history for a total of 4 days. The conclusion was that a genre and usage history based recommendation system is a viable solution.

2.4. Literature Review Summary

A brief literature review was conducted to study the existing works in the field of recommendation. The initial discovery was that neural networks have produced better results compared to the traditional content and collaborative based traditional approaches. Next the literature review looked into capturing the power of the CNN over audio files. A

lot of existing material was found in this section, especially on the popular GTZAN dataset which has been used for this project. Different approaches were taken by researchers to introduce novelty into the field of audio genre classification using CNN. This section also looked at the use of transfer learning in image classification as transfer learning has been used in this project. From this section, it was clear that these research papers did not have any missing data considerations. As missing data values and information is common in real world scenario, the next section looked into the work carried out in the domain of genre classification using deep learning with missing data values. Again, a lot of research work was carried out in this domain. All the research works were towards imputing or filling the missing data values, be in the audio files or spectrograms after converted from audio signals. This research focuses on not filling the missing values and seeing if satisfactory results can still be achieved. Finally, image augmentation its effects have been discussed as this is another part of the project. Another element of this project is finding how robust are the genres to missing data values while using deep learning for genre classification, a research domain which has not been explored. Thus, taking all the elements of the project into account and the research work that has been carried out, we can conclude that this project is novel and explores unexplored fields in the domain of genre based recommendation systems using deep learning.

3. Research Methodology

The Knowledge Discovery in Databases (KDD) methodology has been adopted for this project. All the stages of the process have been rigorously applied to acquire the best set of results. The following diagram shows the research methodology that has been undertaken in this project -

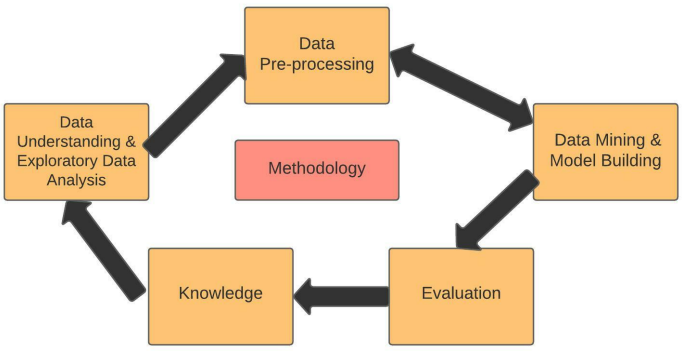

Figure 2: Research Methodology

The methodology starts with data acquisition and understanding, followed by pre-processing, modelling and evaluating which are feedback based processes, followed finally by gaining knowledge from the process

4. Design Specification

This stage is a more detailed description of the project stages with a brief description of the activities that will be carried in the main sections from the above diagram. The above diagram shows the Design Specification of this project -

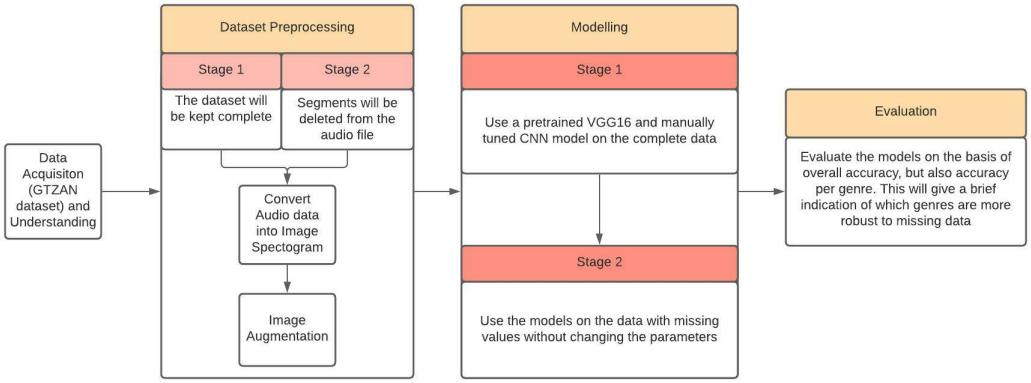

Figure 3: Design Specification

The first stage of the project is acquiring the GTZAN dataset and understand the features of the dataset. There are three case studies in total in this project. case study 1 is building the model on the original dataset. Case study 2 and 3 are similar and this where pre-processing comes into picture. The dataset will be processed to make 2 new datasets where case study 2 dataset will have audio files of only 15 seconds and case study 3 dataset will have audio files of 20 seconds. The VGG16 model will be trained on these datasets and the results will be compared.

5. implementation

5.1. Data Acquisition & Understanding

The dataset has been downloaded from the Marsyas website and the name of the data is the GTZAN dataset. Alhtough the dataset is available, the permission of the creator Mr. George Tzanetakis has been taken for the use of the dataset. This is a very popular dataset due to its easy availability and low complexity of the data, as seen in the research work where multiple researchers have used this dataset. The dataset is a very clean and properly labelled. The dataset contains a total of 1000 songs from 10 genres which are rock, reggae, disco, blues, classical, country, hiphop, pop, jazz and Metal. Each genre has 100 songs each, with a length of 30 seconds per song.

5.2. Data Pre-processing

This section is very critical to the project, as the novelty factor is implemented in this section. The purpose of this project is to see how the models behave in the presence of missing data. In order to replicate this, segments have been removed from the audio file, to make a smaller audio file which will be missing segments from the original file. The process has been explained below -

- The audio file will be split into segments using the pydub library in python.
- Each audio segment will of 5 seconds.
- Case Study 1 - This is not a part of the pre-processing but will contain the original dataset with the complete audio segments and the model will be trained on this data.
- Case Study 2 - Three audio segments will be merged together, to form a new audio file of 15 seconds. The audio segments will be selected from random for eg. 0-5 seconds, 20-25 seconds, 25-30 seconds. The difference in the original audio clip and the processed audio clip will now be 15 seconds, which will be considered as the missing data.
- Case Study 3 - Four audio segments will merged together to form a new audio file of 20 seconds. The The difference in the original audio clip and the processed audio clip will now be 10 seconds, which will be considered as the missing data.

The following image explains the pre-processing stage using a visualization -

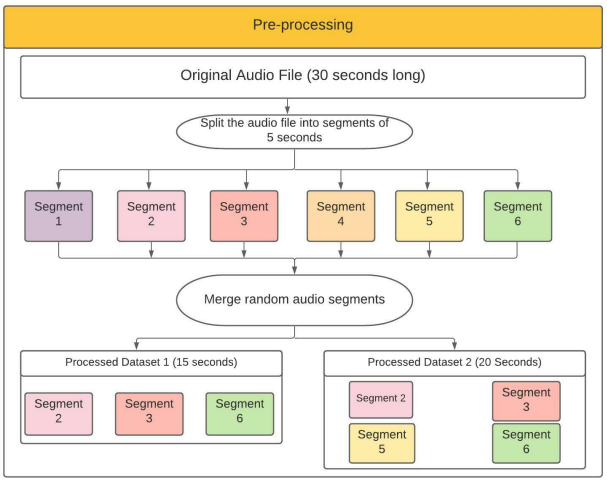

Figure 4: Pre-processing Methodology

In order to further demonstrate how the above process worked, we can take a look at the following waveform diagrams showing the original audio file and the processed audio files -

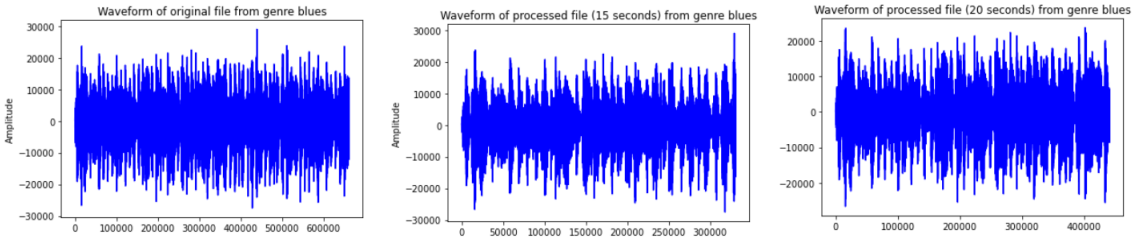

Figure 5: Waveforms of original, 15 seconds and 20 seconds file

As expected, there is a notable difference between the waveforms of the original and the processed audio files.

Convolutional Neural Network is best suited for image classification problems. However, the dataset used is an audio file and hence, the audio file has been converted into image spectrograms as was carried out by all the researchers who used the audio dataset for genre classification using CNN. The Librosa library in python has been used for this process of converting the audio files into spectrograms. The following figure shows the spectrograms of the original and processed audio files from the blues genre -

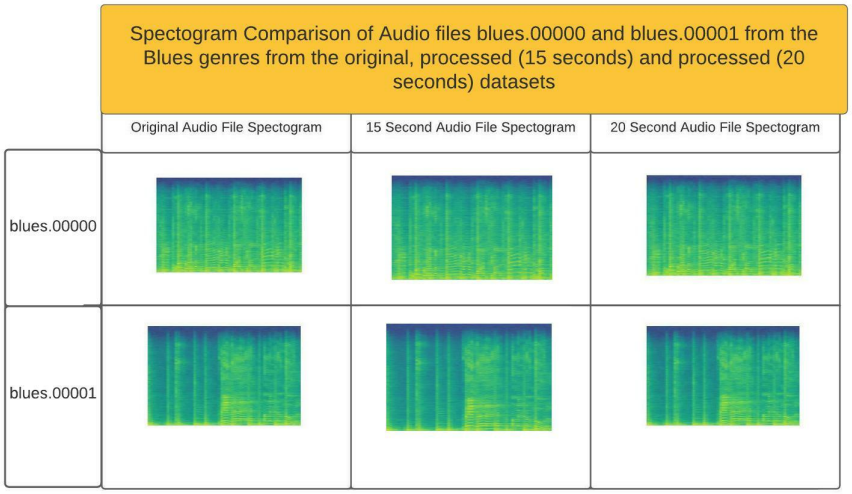

Figure 6: A comparison of the Spectrograms of the Original and Processed Audio Files (15 Seconds and 20 Seconds) from Genre Blues

The changes may not be significant and only the model results will show clearly how these changes affected the performance. These spectrograms are nothing but a description of the loudness of the song and by removing segments from the dataset the 'loudness' of the images will also change and would potentially affect the model performance.

5.3. Image Augmentation

This is a process that has been used by most of the researchers in their projects and will be implemented for this project as well. First, the image will be scaled to make all the images in the range (0,1). Next, random transformations are applied followed by zoom function. The final step is to apply the horizontal flip. Although classically image augmentation is used to bulk up the data this has not been done here and the output from the ImageDataGenerator is essentially the same number of images just they have been processed. This also means that the spectrograms from the original dataset have now been replaced.

5.4. Model

Three models were built and trained on the original and processed datasets (15 seconds and 20 seconds). The dataset was split into 80% train data and 20% test data. The VGG16 architecture model was used. The architecture of the CNN model has been described below

5.4.1 VGG16

The project uses a transfer learning approach where a pre-trained state-of-art model known as VGG16 architecture was used. The advantages of using transfer learning have been discussed in the literature review section. The following diagram describes the VGG16 architecture -

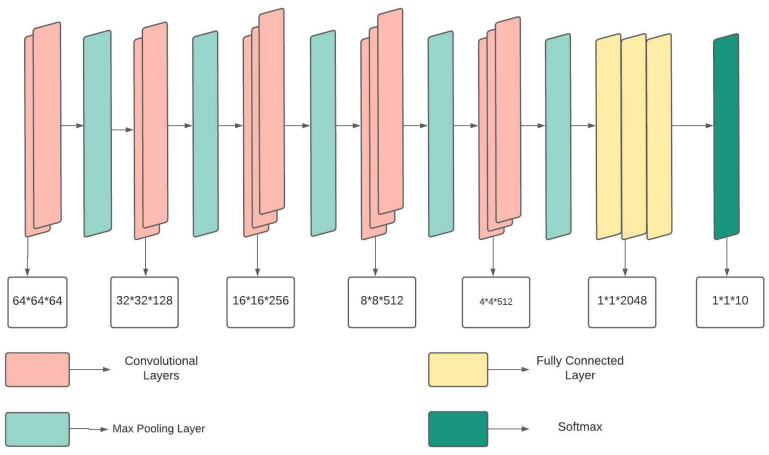

Figure 7: VGG16 Architecture diagram

The above figure is a summary of the model architecture used in this project with the input size being 64*64*3 and the output being a 10 channel layer for the 10 genres. The architecture is as follows -

- Two layer convolution layer of size 64*64*3, where 64*64 is the image size and 3 is the number of channels and the number of filters is 64, which means that the new image dimensions will be 64, 64, 64.
- This is followed by a Max Pooling layer, of filter size 2*2 and stride 2. The following equation explains the output that will be received from the Max Pooling layer every time

$$Output = (((n + 2p - f)/2) + 1) \tag{1}$$

  The output size that will be received will be of size (32, 32, 64).
- This is followed by 2 more convolution layers, where the size changes to 32*32*128, thus the size of the filters changes from 64 to 128 and the image sizes reduce to half as per the formula.
- This is followed by another maxpooling layer, where the dimensions are 16,16, 64.
- There is a change in the convolution layers which changes from 2 to 3, making it a 3 layer convolution layer. This process is repeated 2 more times, thus giving the final dimensions as (4*4*512).
- This is followed by 3 fully connected layers, where the first 2 layers give an output as (1*1*2048), and then the final fully connected layer with the output size (1*1*10), where 10 is the output classes which are the 10 genres.

The model cost and optimization techniques used are categorical cross entropy as the loss function and optimizer used as the optimizer. The metrics used is accuracy. The model was run for a number of epochs and the final number of epochs that was chosen was 1000. The model case studies were as follows -

- The VGG16 model will be first trained on the original dataset. The model will be judged on the basis of the overall accuracy parameter. The training and validation losses will be noted, along with the Precision, Recall and F1 score per genre.
- The same model will be run on the processed datasets where the audio clips are 15 seconds and 20 seconds long. The model will be also judged on the accuracy metrics along with the Precision, Recall and F1 score per genre.

6. Results & Evaluation

In this section, the models will be evaluated and the model best suited to the case study will be identified. A detailed discussion along with visualizations has been provided for the ease of understanding. The following table summarizes the results of the findings.

Model Used	Dataset Used	Testing Accuracy	Loss
VGG16	Original Dataset	46.07%	1.472
	Processed Dataset (15 seconds audio file)	49.23%	1.412
	Processed Datset (20 seconds audio file)	55.12%	1.272

Figure 8: Result Summary

As we can see from the above table, the experiments showed very surprising results. Against expectations, the performance of the model on the processed datasets showed higher accuracy than the model on the original dataset. The accuracy increases slighlty by using 15 seconds of the audio file as the dataset but increases drastically after using the dataset with 20 seconds audio files. This shows that the dataset did not suffer any degradation but in fact became better after processing. The results have been further evaluated in detail in the following sections -

6.1. Case Study 1 - Original Dataset

The following image shows the graphs from the training part of the model -

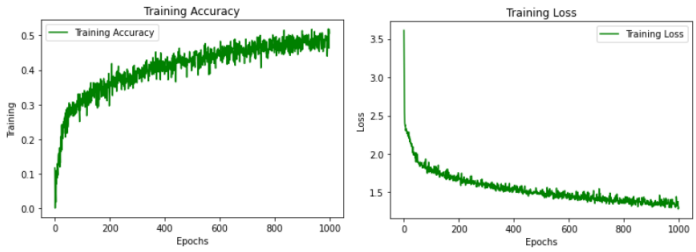

Figure 9: Training Accuracy, Training Loss

As can be seen from the above images, the training accuracy remains fairly stable and also on occasions crosses the 0.5 mark. The training loss starts above 3.5 but drops drastically and dropping below 1.5 mark on occasions. The following diagram shows the results from the testing part of the model starting with the confusion matrix and the results per genre.

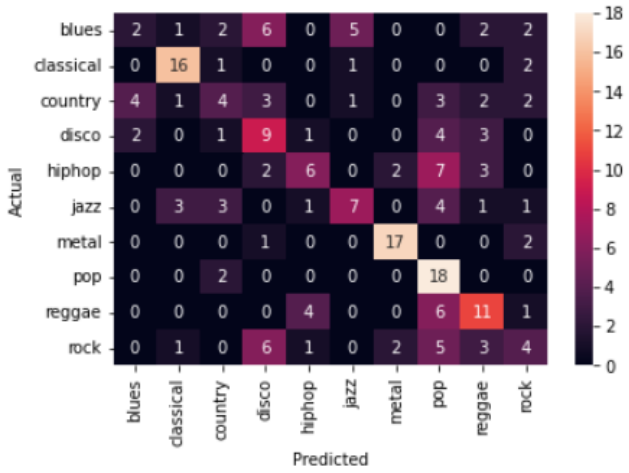

Figure 10: Confusion Matrix (Test Results)

One way to look at the confusion matrix is taking a look at the x and y axis. For example, in the genre Classical 16 songs were correctly predicted. However, 6 songs were predicted as classical but were not classical, hence making the score 16/22.

Genre	Precision	Recall	F1-score
Blues	0.25	0.10	0.14
Classical	0.73	0.80	0.76
Country	0.31	0.20	0.24
Disco	0.33	0.45	0.38
hiphop	0.46	0.30	0.36
Jazz	0.50	0.35	0.41
Metal	0.81	0.85	0.83
Pop	0.38	0.90	0.54
Reggae	0.44	0.50	0.47
Rock	0.29	0.18	0.22

Table 1: Precision, Recall and F1-score Test results per genre (Original Dataset)

From the above table, the genres Classical and Metal show very good results. It is very important to take a look at the precision-recall balance as the genre Pop shows very bad

performance. The difference in the precision recall values is so high that it clearly states the dominant presence the irrelevant results meaning many songs from other genres were classified as pop.

6.2. Case Study 2 - Processed Dataset (15 seconds audio file)

The graphs for the training accuracy as well as training loss are shown below -

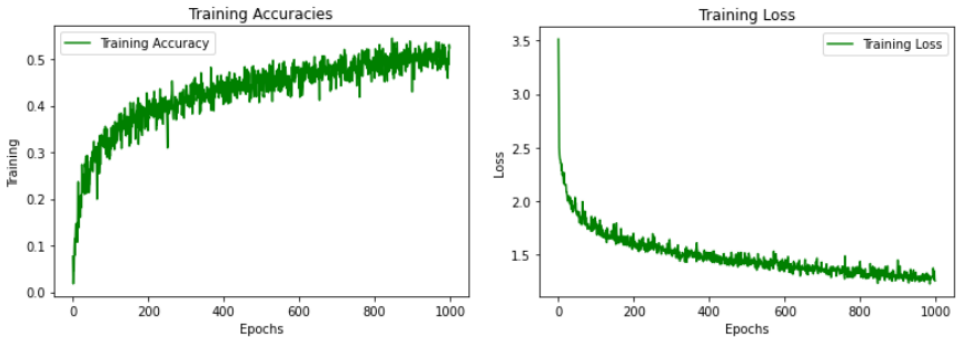

Figure 11: Training Accuracy, Training Loss

As we can see from the images the training accuracy starts at 0 and on occasions crosses the 0.5 mark. The graph however, shows more instability than case study 1. The training loss graph is similar to the case study 1.

The following shows the confusion matrix and the test results -

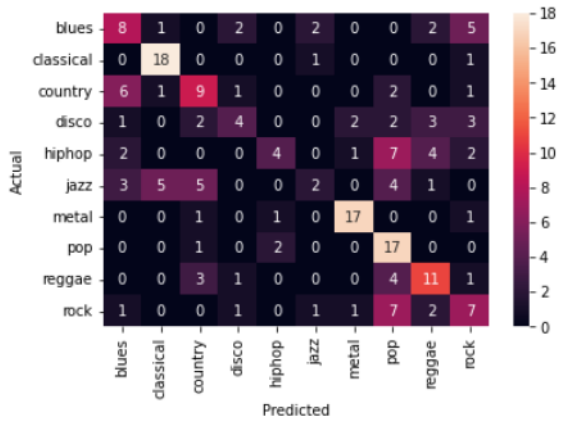

Figure 12: Confusion matrix (Test Results)

As we saw in the case study 1, the confusion matrix can be read in terms of how many songs were correctly calculated. Again taking the case of genre Classical, the model correctly predicted 18 but incorrectly predicted 7 songs as classical.

Genre	Precision	Recall	F1-score
Blues	0.38	0.40	0.39
Classical	0.72	0.90	0.80
Country	0.43	0.45	0.44
Disco	0.44	0.24	0.31
hiphop	0.57	0.20	0.30
Jazz	0.33	0.10	0.15
Metal	0.81	0.85	0.83
Pop	0.40	0.85	0.54
Reggae	0.48	0.55	0.51
Rock	0.33	0.35	0.34

Table 2: Precision, Recall, F1-score Test results per genre (Processed Dataset)

The results from case study are very similar to that of case study 1. The Precision-Recall balance and the F1-score show that the genres Classical and Metal are the best performing genres. There is also a vast difference between the precision recall values in the genres Hiphop and Pop thus making the predictions in these genres very unreliable.

6.3. Case Study 3 - Processed Dataset (20 seconds audio file)

The training accuracy and training loss graphs are shown below -

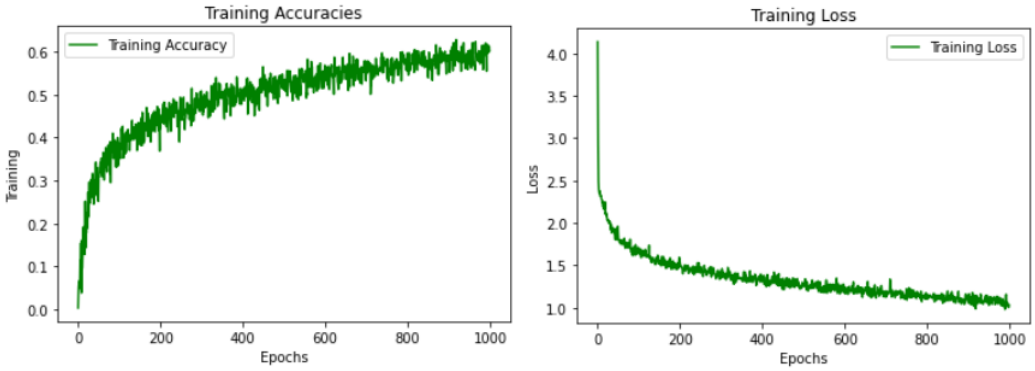

Figure 13: Training accuracy and Training loss

The training accuracy graph shows more stability than the earlier case studies although very minimum. The accuracy also crossed the 0.6 mark at times and the overall accuracy was a very good 55%. The training loss characteristics are fairly similar in all the case studies, except here it starts very high at around 4 but drops drastically and gives a final

reading of 1.2. The following show the confusion matrix and the results summary per genre of the model -

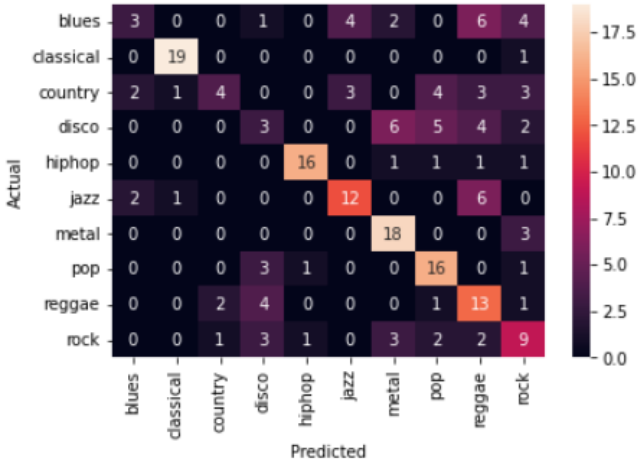

Figure 14: Confusion matrix

Genre	Precision	Recall	F1-score
Blues	0.43	0.15	0.22
Classical	0.90	0.95	0.93
Country	0.57	0.20	0.30
Disco	0.21	0.15	0.18
hiphop	0.89	0.80	0.84
Jazz	0.63	0.57	0.60
Metal	0.60	0.86	0.71
Pop	0.55	0.76	0.64
Reggae	0.37	0.62	0.46
Rock	0.36	0.43	0.39

Table 3: Precision, Recall and F1-score Test results per genre (Processed Dataset (20 second audio file)

The importance of using different evaluation metrics is seen in this case study again, especially in a multi-class classification problem. The genre with the best performance is the Classical with a very high precision, recall and F1-score. A total of 19/20 songs were correctly classified in this genre. The high precision score also suggests that the model didn't just simply classify majority songs as classical songs, but correctly classified only classical songs as classical songs. No other genres showed such a high accuracy however, the performance of the genre Hiphop comes very close. It becomes increasingly evident to

consider the precision-recall balance as the most accurate metrics after taking a look at the high difference in the precision-recall differences in the Blues, Country, Metal and Reggae genres. This suggest the high presence of false positives and false negatives values in the model.

7. Discussion

This section discusses the results and the interpretation of those results. The following list describes the findings of the project -

- The VGG16 CNN Model performed better on the processed datasets with shorter audio clips than the original dataset. This may be associated with the following 2 factors and requires further research -
  - The difference in the duration of the audio clips in the original and processed data did not contribute towards assisting the model in making predictions. This may also be due to the fact that the spectrograms are a measure of loudness, but the audio files had duration which were silent filler music and not really associated with any genre.
  - The original dataset was more susceptible to overfitting resulting in lower accuracy. However, in the processed data, the robustness increased, variance decreased and the accuracy increased.
- All the case studies showed good performance on certain genres such as Classical. There were other common genres which showed very low performance. Thus, in industry applications if the data is associated with genres or if there are target genres such as Classical, this model can be used to make predictions.
- The overall accuracy is not a good evaluation metric and precision-recall balance should be given more importance.
- The papers reviewed in section 2.2 showed very high accuracy using CNN on the GTZAN dataset, something that this project was not able to achieve despite the use of VGG16, however they adopted different processes and the experiment design was not the same.
- The section 2.3 focused on the data imputation methodologies used by researchers. This project shows that the model actually performed better on the processed data. Thus, we can say that data imputation may not be necessary after all.
- The models trained on processed dataset also take less processing time due to less data which is another positive point other than the better accuracy.

8. Conclusion & Future Work

A Convolutional Neural Network model using the VGG16 architecture was built on the original GTZAN dataset and two processed versions where only 15 seconds and 20 seconds of the original data were considered. Contrary to expectations, the model performed better on the processed data rather than the original data with the highest accuracy of 55.12%. The models gave the best results on the genre Classical. The precision-recall balance was more indicative of the true performance than the accuracy parameter. We can successfully conclude that the missing data had no significant degrading effect on the model and in fact the model performed better on the processed datasets.

The future work associated with this model can be to increase the accuracy of the model as this project failed to match the accuracy percentage of the state-of-art models. One of the ways to do this is to add a metadata file containing the extracted features of the audio files. Further experiments can be carried out to understand at what stage does the missing data become a significant deteriorating factor and the accuracy drops exponentially making it no longer viable for real world application. Another approach would be to add noise to the data, which means adding random audio clips to the dataset and seeing the results for this experiment.

9. Acknowledgement

I would like to thank my supervisor Michael Bradford for guiding me through the thesis work. I would also like to thank all the faculty members at NCI. Finally I would like to thank my parents, friends and family for their continuous support.

References

Agrawal, Manish and Abhilash Nandy (2020). "A Novel Multimodal Music Genre Classifier using Hierarchical Attention and Convolutional Neural Network". In: *CoRR* abs/2011.11970. arXiv: 2011.11970. URL: https://arxiv.org/abs/2011.11970.

Bako, Tamas (Jan. 2015). "Nonlinear distortions in audio devices". In.

Elbir, Ahmet et al. (Oct. 2018). "Music Genre Classification and Recommendation by Using Machine Learning Techniques". In: pp. 1–5. DOI: 10.1109/ASYU.2018.8554016.

Guan, Qing et al. (2019). "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study". In: *J Cancer* 10, pp. 4876–4882. DOI: 10.7150/jca.28769. URL: https://www.jcancer.org/v10p4876.htm.

Jaiswal, Kaustumbh and Dhairya Kalpeshbhai Patel (2018). "Sound Classification Using Convolutional Neural Networks". In: *2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 81–84. DOI: 10.1109/CCEM.2018.00021.

Jang, Jong-Hwan et al. (Sept. 2019). "Deep Learning Approach for Imputation of Missing Values in Actigraphy Data: Algorithms Development Study (Preprint)". In: *JMIR mHealth and uHealth* 8. DOI: 10.2196/16113.

Jiang, Kai, Haixia Chen, and Senmiao Yuan (2005). "Classification for Incomplete Data Using Classifier Ensembles". In: *2005 International Conference on Neural Networks and Brain* 1, pp. 559–563.

Lee, Jongseol et al. (2015). "Music recommendation system based on usage history and automatic genre classification". In: *2015 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 134–135. DOI: 10.1109/ICCE.2015.7066352.

Liu, Jinliang, Changhui Wang, and Lijuan Zha (2021). "A Middle-Level Learning Feature Interaction Method with Deep Learning for Multi-Feature Music Genre Classification". In: *Electronics* 10.18. ISSN: 2079-9292. DOI: 10.3390/electronics10182206. URL: https://www.mdpi.com/2079-9292/10/18/2206.

Liu, Peng et al. (2005). "An Analysis of Missing Data Treatment Methods and Their Application to Health Care Dataset". In: *Advanced Data Mining and Applications*. Ed. by Xue Li, Shuliang Wang, and Zhao Yang Dong. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 583–590.

Marszalek, Marcin et al. (Oct. 2007). *Learning Object Representations for Visual Object Class Recognition*. Visual Recognition Challange workshop, in conjunction with ICCV. URL: https://hal.inria.fr/inria-00548669.

Mo, Weilong et al. (June 2019). "Image recognition using convolutional neural network combined with ensemble learning algorithm". In: *Journal of Physics: Conference Series* 1237.2, p. 022026. DOI: 10.1088/1742-6596/1237/2/022026. URL: https://doi.org/10.1088/1742-6596/1237/2/022026.

Nanni, Loris et al. (2021). "An Ensemble of Convolutional Neural Networks for Audio Classification". In: *Applied Sciences* 11.13. ISSN: 2076-3417. DOI: 10.3390/app11135796. URL: https://www.mdpi.com/2076-3417/11/13/5796.

Oord, Aäron van den, Sander Dieleman, and Benjamin Schrauwen (2013). "Deep Content-Based Music Recommendation". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., pp. 2643–2651.

Oquab, Maxime et al. (2014). "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724. DOI: 10.1109/CVPR.2014.222.

Pikrakis, Aggelos (2013). "A deep learning approach to rhythm modeling with applications". In: *Proc. Int. Workshop Machine Learning and Music*.

Pikrakis, Aggelos et al. (2015). "Pattern classification formulated as a missing data task: The audio genre classification case". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2026–2030. DOI: 10.1109/ICASSP.2015.7178326.

Sharma, Neha, Vibhor Jain, and Anju Mishra (2018). "An Analysis Of Convolutional Neural Networks For Image Classification". In: *Procedia Computer Science* 132. International Conference on Computational Intelligence and Data Science, pp. 377–384. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.05.198. URL: https://www.sciencedirect.com/science/article/pii/S1877050918309335.

Shorten, Connor and Taghi M. Khoshgoftaar (2019). "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6, pp. 1–48.

Smaragdis, Paris, Bhiksha Raj, and Madhusudana Shashanka (2009). "Missing data imputation for spectral audio signals". In: *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. DOI: 10.1109/MLSP.2009.5306194.

Song, Yading, Simon Dixon, and Marcus Pearce (June 2012). "A Survey of Music Recommendation Systems and Future Perspectives". In.

Tammina, Srikanth (Oct. 2019). "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images". In: *International Journal of Scientific and Research Publications (IJSRP)* 9, p9420. DOI: 10.29322/IJSRP.9.10.2019.p9420.

"The impact of the Lombard effect on audio and visual speech recognition systems" (2018). In: *Speech Communication* 100, pp. 58–68. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2018.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0167639317302674.

Weiss, Karl R., Taghi M. Khoshgoftaar, and Dingding Wang (2016). "A survey of transfer learning". In: *Journal of Big Data* 3, pp. 1–40.

Wyse, Lonce (June 2017). "Audio Spectrogram Representations for Processing with Convolutional Neural Networks". In.

Xu, Yijie and Wuneng Zhou (2020). "A deep music genres classification model based on CNN with Squeeze amp; Excitation Block". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 332–338.

Yosinski, Jason et al. (2014). "How Transferable Are Features in Deep Neural Networks?" In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, pp. 3320–3328.