Improving quality of the GTZAN dataset for SVM genre classifiers.

Leonard in't Veen Supervisor(s): Cynthia Liem , Jaehun Kim
EEMCS, Delft University of Technology, The Netherlands
l.j.intveen@student.tudelft.nl, c.c.s.liem@tudelft.nl j.h.Kim@tudelft.nl

Abstract

The GTZAN dataset, a collection of 1000 songs spanning 10 genres, proposed by Tzanetakis has been around for 20 years. In this time hundreds of researches and applications have included this database. However, there seem to be some serious limitations to this dataset. There are duplicates, mislabellings, low audio recordings and narrow representations of genres. This paper aims to research the effects of both audio quality and the content of this dataset on genre classification. A Support Vector Machine (SVM) has been used to retrain and compare different versions of the dataset. Two experiments have been proposed in the paper. In the first experiment, a comparison between a lossless dataset of high audio quality and an mp3 version of that same dataset of a lower audio quality have been investigated. The lower quality dataset performed worse on the SVM classifier of this size. The second experiment proposed a new metal dataset, based on a wider and more balanced range of metal sub-genres. This metal dataset has replaced the original metal part of the GTZAN dataset. Some retrainings done this way had a higher accuracy than the original, giving confidence that representing a well-balanced genre might improve classification performance. Finally, it has been found that the original GTZAN classifier is inaccurate on audio samples outside of its dataset, where the new retrainings done on lossless datasets without much preprocessing seem to perform substantially better. This last finding has not been verified systematically and asks for more verification.

1 Introduction

In the field of Music Information Retrieval (MIR), music descriptors play a large role. These can be subdivided into high-level descriptors and low-level descriptors. High-level descriptors aim to determine properties like mood, danceability, and genre. In doing so, often, low-level descriptors are being used to determine features more closely related to the audio signal such as loudness, silence rate and spectral energy. Instead of using databases with songs, research often uses databases consisting of pre-extracted audio features instead. This circumvents copyright issues and reduces run time considerably.

Models based on similar audio features can display different results. Varying results on seemingly similar applications and models put into question the validity and performance of such systems. This raises questions about what causes these differences. Are audio features maybe not robust across different codecs? The discussion following these findings asks for more validation and testing. The importance of these issues is highlighted by the widespread utilization of these models and descriptors. An example of these utilizations is Essentia (Bogdanov et al., 2013), a widely adopted audio analysis library, which will also be used as a basis for this research.

One of the datasets used for Essentia's applications is the GTZAN dataset (Tzanetakis and Cook, 2002). The GTZAN dataset is used in hundreds of researches over the last two decades in various algorithms. While algorithms grew from statistical analysis to Support Vector Machines (SVM) and machine learning models, music genres evolved to include a wider variety of music, but the dataset remained the same. Despite being heavily criticized, no better alternative was found. This raises the question of how this widely adopted dataset could be improved. Errors in the dataset and the impact of version differences should be investigated to determine where best the dataset could find improvement.

The main question this paper tries to answer is: **How does the audio quality and content of the GTZAN dataset impact Essentia's SVM genre classifier?** This is addressed through 2 sub-questions. This paper will describe the experimental set-up and corresponding results for these subquestions:

- What is the effect of degrading the database's audio quality on Essentia's SVM classifiers accuracy?
- What is the effect of changing GTZAN's song selection on Essentia's SVM classifiers accuracy?

The first experiment compares a classifier based on a lossless version of the GTZAN dataset to a classifier based on a lower quality lossy dataset. The performance of both versions represented through labellings in a confusion matrix will be discussed.

The second experiment compares the same lossless classifier against versions with an augmented metal partition. 5 different versions have been created based on a random song selection from a metal dataset. This metal dataset has been created to represent a balanced and complete portion of metal sub-genres.

Comparing against the original GTZAN dataset and corresponding classifier introduced a lot of unknowns. Classifiers from both experiments have been compared against a new classifier based on the GTZAN dataset. This dataset was created to address some of the issues with GTZAN while providing a known baseline from which can be compared. However, this dataset brings some issues of it's own which will be discussed in the discussion section of the paper. Even though no formal comparison between the new version and the GTZAN classifier has been done, informal comparisons did provide some interesting observations, which will be briefly touched upon as a preliminary study.

The paper follows the following structure: More in-depth motivation and related work will be discussed in section 2. The methodology will be explained in section 3. Section 4 and 5 will present the results and discuss these results. A reflection on reproducibility and replicability will be given in section 6. Afterwards, the paper will be concluded and suggestions for future work will be given in section 7.

2 Related work

The motivation for this research topic rose from recent studies which challenge established systems and found unexpected results. Studies indicate that audio features can not be taken as a ground truth. Urbano et al. (2014) examined the robustness of audio descriptors to varying audio qualities and codecs and proposes a systematic way of performing this comparison. In doing so he found that although most results showed to be robust, some combinations of codec and audio quality impact the performance of algorithms. This requires more testing and tweaking of parameters on algorithms. Another example is the research of Liem and Mostert (2020) who found that certain genre classifiers based on similar audio features have surprisingly low correlations between each other or even negative correlations.

Algorithms to improve genre classification have been proposed over the last years. Many of these algorithms are based on more and more advanced algorithms using SVM's, such as the classifier proposed by Xu et al. (2003), and more advanced machine-learning models, such as Kour and Mehan's (2015) classifier who combines and SVM with a Neural Network . In this research, an SVM classifier was chosen as this seemed a relatively up to date solution while not requiring immense retraining times. Essentia provides an SVM model trainer in it's library that trains an SVM classifier based on provided ground truth and audio feature files (Bogdanov et al., 2013). This SVM is based on the LIBSVM library (Chang and Lin, 2011), a library providing tools and extensions for SVM's.

Many people have looked at ways to improve classifiers and algorithms, but few have looked at ways to improve the datasets used or have critically assessed the datasets' quality. As of today, no standardized way to assess the quality of a dataset has been created. Someone who did an in-depth analysis on a dataset is Tzanetakis and Cook (2002). He analysed the GTZAN dataset which has been used in hundreds of researches over the last 2 decades. This is a 1000 song dataset consisting of 30-second samples from 10 different genres. Each genre is represented by 100 songs. However, the exact content and samples used in this dataset are not known. Sturm has identified all but 23 songs from the dataset (Sturm, 2012). Sturm identified some problems with this dataset. There are multiple duplicates (5%), mislabelling's (10.8%) and low-quality samples in the dataset (Sturm, 2014). Taking this into account Sturm has estimated that a perfect classifier would only be able to reach an accuracy of 94.5%. This motivates this research to improve the dataset instead of classifiers.

Rodriguez-Algarra et al. (2019) addressed the impact of confounding factors in the design of music classification experiments, the inability to distinguish the effects of multiple potential influencing variables in the measurements. For Algarra's research performed together with Sturm the GTZAN dataset was also a subject. To avoid these confounding factors, a lossless version of GTZAN has been created in this research to function as a baseline.

3 Methodology

To draw meaningful conclusions from comparisons, an effort was made to minimize the number of confounding variables. To achieve this, first, a new version of GTZAN based on a lossless format has been made, namely LGTZAN. This dataset could then be used as a baseline to test the 2 subquestions of this research: What is the effect of degrading the database's audio quality on Essentia's SVM classifiers accuracy and what is the effect of changing GTZAN's song selection on Essentia's SVM classifiers accuracy. Both of these will be done by altering LGTZAN and comparing it to the original using 5 fold cross-validation to obtain accuracies. This section will first describe the data collection. Following will be a description of the experimental setups of the 2 experiments done.

3.1 GTZAN

As mentioned earlier GTZAN has some issues. Several are addressed in this research based on Sturms work (Sturm, 2013). Mislabellings have been corrected or removed and low-quality files have been replaced by lossless files. Duplicates have been replaced by random song samples from the same genre. Another problem with GTZAN is that it hasn not been updated, seeing as new music is constantly being made some of the genres in GTZAN might not be representative anymore. Especially the metal portion of GTZAN seems like a problem, as it already has a large overlap with rock and consist of a small partition of the metal genre as a whole. The first issue will be addressed in the LGTZAN subsection while the last issue will be addressed in the MLGTZAN subsection.

LGTZAN

The LGTZAN dataset is a musical dataset based on the GTZAN dataset. Initially, an effort was made to create an exact replica in the lossless format. However, in the data collec-

tion phase of the research, a couple of issues were found with this approach. The current GTZAN list of songs is not completely known, there are as of this moment 23 songs missing. The known list of songs has been queried against the Muziekweb database. Since the people of Muziekweb[1] are providing the FLAC (lossless) files. However, not all the albums were available in their database. In addition, after receiving the data from Muziekweb, it turned out some albums were sent without content, both metadata and audio were missing. This resulted in a dataset that was around 80% complete. With this in mind instead, a choice was made to not make an exact replica, but a modified version of GTZAN. For the resulting database for all songs one album containing this song has been added to the database. Resulting in a much larger version of the GTZAN database consisting of 470 albums 14000 audio files. Unfortunately, this database can not be shared due to copyright constraints.

MLGTZAN

A metal dataset has been created consisting of a balanced set of metal sub-genres. The dataset contains 200 albums provided by Muziekweb in FLAC format. The division of the metal dataset has been made to find a representative distribution of current metal sub-genres. The sub-genres were chosen based on recent research into metal sub-genres (Hillier, 2020). From the proposed sub-genres, a larger focus was put on the sub-genres which are closer to other genres in the dataset, such as power metal (classical + metal). For that reason, the extreme metal genres selected (death, black, thrash and doom) are represented with 10 albums each. The sub-genres heavy, folk, power, symphonic, prog and nu are represented with 20 albums each. Hybrid genres with already participating sub-genres were not considered as the overlap would be too large. The ground truth of the metal genres is verified using Discogs tags (Hartnett, 2015). Albums were hand-selected based on collection rate. Only albums with the target sub-genre as their main tag were considered. A focus was put on albums having as few as possible alternative tags. To create the MLGTZAN dataset the LGTZAN dataset was taken without the metal portion of this set. To this set 100 songs from 100 different albums were added from the metal dataset.

3.2 Experimental design

In the experiments described in this section, the aim was to answer to questions: What is the effect of compression a dataset on Essentia's SVM classifiers accuracy and what is the effect of changing the song distribution on an SVM classifier based on GTZAN? But before the methods to answer these questions are described, we must first dive deeper into how sought to control the experiment.

Creating LGTZAN classifier

With the use of some python scripts [2], the lists and ids of songs and albums have been combined. From every album corresponding to a song on the GTZAN list, a random song for the LGTZAN list was selected. After this process around 200 songs were missing. The remaining songs were selected from random albums corresponding to the correct genre. Since there is a random element in the LGTZAN dataset the retraining of the SVM classifier has been done 5 times. This is to get the best results and make sure results are consistent. The SVM model used was available in Essentia's libraries, this model uses LIBSVM, an SVM library(Chang and Lin, 2011). For the retraining, a ground truth list and file location list have been provided to Essentia's built-in SVM trainer. Instead of audio files, audio features files have been provided to the training module. These audio features were precomputed by Muziekweb using the same low-level descriptors as Essentia. Several hyperparameters for the SVM model are considered: regularization hyperparameter C, kernel coefficient Gamma and Kernel type, which is polynomial or RBF. Optimal hyperparameters have been found using a grid search. The final SVM selected is the one selected with the best accuracy.

The low-level descriptor to compute tonal_key_key and tonal_key_scale values used to indicate the key and scale of the song have been replaced by three newer algorithms in the audio features: Krumhansl, Temperley and Edma. Edma has been primarily trained on dance, Krumhansl has been primarily trained on pop and Temperley has been primarily trained on Euro-classical music. All of them have been run on the same dataset. The accuracies for Krumhansl were the highest, so Krumhansl was used for the remaining retrainings. No additional tweaking on parameters and low-level descriptors has been done.

The results from the LGTZAN classifier after 5 runs turned out to be close to the results posted by Essentia (Essentia, 2020). Which gave confidence in the working of the LGTZAN classifier.

comparing audio quality

The LGTZAN dataset has been compressed to match the original audio quality of GTZAN,22050Hz 16-bit mono audio (MP3), using pydub (Robert et al., 2018). On this set of MP3 data, Essentia's low-level descriptors have been run. After which the resulting audio features could be used to train the compressed LGTZAN classifier. The bias and performance towards encodings have been compared using the accuracies of this result to the LGTZAN accuracies. Degrading the audio quality means there should be a lower amount of information for the classifier to train on. This could affect the classifier's accuracy.

comparing musical content

Since the current representation of metal in LGTZAN is mostly heavy rock and heavy metal, a retraining has been done using the MLGTZAN dataset. Since there is a random component to the selection of songs again the training has been done 5 times. As in the creation of the LGTZAN classifier, these results have been checked for variance. The given accuracies from the MLGTZAN have been compared to the LGTZAN classifier. The effect of adding the wider range of metal to the dataset means metal songs could be easier classified as other genres. For example, power metal is closer to classical than heavy metal. Less overlap between rock and

[1] https://www.muziekweb.nl/

[2] https://gitlab.ewi.tudelft.nl/cse3k-21q2-music-faithfulness/gtzan_augmentations/-/tree/main/parsers(windows)

Accuracy: 74.5

Predicted (%)												Actual (%)
	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock		Proportion
blues	77.00	1.00	9.00	3.00	0.00	5.00	2.00	1.00	2.00	0.00	blues	10.00 %
classical	2.00	94.00	2.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	classical	10.00 %
country	1.00	1.00	75.00	3.00	3.00	6.00	2.00	5.00	1.00	3.00	country	10.00 %
disco	6.00	1.00	9.00	58.00	7.00	1.00	2.00	7.00	3.00	6.00	disco	10.00 %
hiphop	1.00	0.00	1.00	11.00	73.00	0.00	2.00	7.00	3.00	2.00	hiphop	10.00 %
jazz	2.00	0.00	4.00	2.00	2.00	86.00	1.00	0.00	1.00	2.00	jazz	10.00 %
metal	1.00	0.00	3.00	0.00	1.00	0.00	89.00	2.00	0.00	4.00	metal	10.00 %
pop	1.00	2.00	4.00	11.00	10.00	1.00	0.00	62.00	7.00	2.00	pop	10.00 %
reggae	4.00	0.00	3.00	7.00	6.00	0.00	0.00	6.00	72.00	2.00	reggae	10.00 %
rock	3.00	1.00	8.00	9.00	5.00	1.00	11.00	3.00	0.00	59.00	rock	10.00 %

Figure 1: Resulting confusion matrix LGTZAN4

metal could also mean that the accuracies for rock songs improve.

4 Results

Due to the random song selection, the LGTZAN set has been retrained 5 times with an average accuracy of 72.46. The standard deviation was 1.15, meaning the classifiers were relatively consistent. Since the goal is to improve the dataset, the highest accuracy of 74.5 was chosen to continue with the remaining research questions. Since it was the fourth retraining this will be referred to as LGTZAN 4. Figure 1 shows the confusion matrix corresponding to this classifier, with genre accuracies in green and mislabellings per genre in the rows. This value is close to the original value from the GTZAN classifier(75.5) which gives confidence in the quality of the LGTZAN classifier.

The parameters with the best results from the 5 retrainings were all different. The best parameters for LGTZAN 4 were: c:1, gamma: -3 and kernel: poly. A full overview of the results can be found on the GitLab page[3].

LGTZAN 4 has also been tested with different key descriptors. Of the three descriptors used Edma and Temperley performed worse with an accuracy of respectively 74.1 and 73.1. As expected Temperley performed better on classical, Edma performed better on blues classical and jazz and Krumhansl performed better on all other genres. However, one unexpected result is that Krumhansl performed by far the worst on pop. As the low-level descriptors were not the focus of this research, no further testing was done on the key descriptors. Between the different LGTZAN versions, some differences can be seen between genre mislabellings. Not a single genre is consistent across all retrainings. With a maximum difference in genre mislabellings of 11 for blues and a minimum difference of 4 for rock. The worst performing genres were mostly disco and rock, followed by reggae and pop. The GTZAN confusion matrix(Essentia, 2020) is by comparison much more balanced. With the largest differences being in disco and pop. Rock is still at the bottom of performance here, this issue is addressed in the ”dataset content” chapter.

4.1 Audio quality

Converting the audio from LGTZAN 4 to mp3, and reducing the audio quality and amount of channels to match GTZAN's audio format gave a worse result as can be seen in figure 2. On average genres decreased, with hiphop and rock performing worst. From this can be seen that degrading the audio quality of the LGTZAN set might decrease SVM classifier performance. However, we cannot say improving the GTZAN audio quality will also improve classifiers trained on that dataset. More factors play a role in this comparison as will be discussed in the following chapter.

4.2 Dataset content

For the MLGTZAN classifier, the LGTZAN 4 classifier dataset was augmented by swapping out the metal part of the dataset with a partition of the varied metal dataset as described in section 3.1. Expectations for the MLGTZAN classifier were that by addressing the possible overlap between metal and rock the accuracy for rock might increase by mislabelling fewer rock songs as metal. Metal songs accuracy would probably be spread more over the other genres as the metal component of the dataset was spread out more. Metal songs being mislabeled as rock indeed decreased in all cases however, rock didn not improve in all cases. Metal mislabeled as rock decreased by an average of 2.6, while rock only increased by an average of 1.6. Metal did indeed decrease in all cases, but interestingly in almost all cases most of the

[3]https://gitlab.ewi.tudelft.nl/cse3k-21q2-music-faithfulness/gtzan_augmentations

Accuracy: 72.0

Predicted (%)												Actual (%)
	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock		Proportion
blues	74.00	1.00	10.00	3.00	0.00	3.00	1.00	0.00	4.00	4.00	blues	10.00 %
classical	0.00	97.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	classical	10.00 %
country	7.00	0.00	78.00	2.00	1.00	2.00	1.00	1.00	0.00	8.00	country	10.00 %
disco	3.00	1.00	10.00	57.00	5.00	1.00	4.00	11.00	4.00	4.00	disco	10.00 %
hiphop	1.00	0.00	0.00	10.00	63.00	2.00	3.00	5.00	13.00	3.00	hiphop	10.00 %
jazz	7.00	0.00	5.00	1.00	0.00	82.00	0.00	1.00	1.00	3.00	jazz	10.00 %
metal	1.00	0.00	0.00	1.00	3.00	1.00	87.00	2.00	0.00	5.00	metal	10.00 %
pop	1.00	1.00	6.00	9.00	10.00	0.00	1.00	65.00	3.00	4.00	pop	10.00 %
reggae	4.00	0.00	3.00	2.00	14.00	0.00	0.00	8.00	67.00	2.00	reggae	10.00 %
rock	6.00	1.00	12.00	9.00	8.00	0.00	8.00	5.00	1.00	50.00	rock	10.00 %

Figure 2: Resulting confusion matrix after mp3 conversion

added mislabelling were from metal to rock. With metal going down an average of 6.2 and metal to rock mislabellings going up an average of 5.

Overall the MLGTZAN classifiers had similar accuracies to LGTZAN 4 classifier. With an average of 74.28 and a standard deviation of 0.39. MLGTZAN 4 and 5 outperformed LGTZAN 4. MLGTZAN 5, shown in figure 3, performed the best with an accuracy of 74.8. Although similar results were seen a lot of differences can be found in the confusion matrixes. Unexpectedly, the other genres fluctuated between different versions of the metal subset, even though they seem invariable to a change in metal. In all cases, overall fewer songs were mislabeled as metal, opposite from what was expected. Another interesting phenomenon is that classical improved in all cases, even though through the addition of power metal and symphonic metal this was expected to be decreased.

Since the metal songs were selected at random from the metal sub-genres the difference between MLGTZAN version might be explained by the clustering of metal songs being closer to certain genres in certain retrainings. To draw more meaningful conclusions an analysis of the mislabellings and corresponding audio features should be done.

4.3 Everything is jazz?

One question that has, since the start of the research, been found infeasible to answer within the scope of this project is the comparison of GTZAN to LGTZAN through the use of non-standard music samples. Although no systematic way was used to test such samples, some private and varying music recordings from my private collection were used to investigate this issue. An interesting result that was found here is that most non standard 30 second samples run through the GTZAN classifier gives a default answer of jazz. The exact probabilities are shown in figure 4. On all of the samples used, including the LGTZAN and metal dataset, this was the case. However, testing the same full-length samples on LGTZAN 4, excluding the LGTZAN and metal dataset, other values were given. Of which a large amount was correctly classified. It seems like the lossless version trainings perform better on non-standard data however, this has not been researched systematically and thoroughly enough to draw that conclusion.


```
"genre_tzanetakis": {
    "all": {
        "blu": 0.0514126457274,
        "cla": 0.189598113298,
        "cou": 0.0246246308088,
        "dis": 0.02359331958,
        "hip": 0.0690673515201,
        "jaz": 0.463441282511,
        "met": 0.0488544665277,
        "pop": 0.0266393795609,
        "reg": 0.0361540466547,
        "roc": 0.0666147470474
```


Figure 4: Default result GTZAN classifier

5 Discussion

This research aimed to test the impact of audio quality and content changes to the musical dataset GTZAN on an SVM classifier. As mentioned earlier a new LGTZAN dataset was chosen as a basis to test the impact of changes, the impact of this new LGTZAN 4 dataset and it's implications and limitations will be discussed here. As well as the results of the two experiments.

5.1 GTZAN comparison

A number of faults were removed from the GTZAN dataset by incorporating the work of Sturm (2014). Known mislabelling were relabeled or removed, duplicates were elim-

Accuracy: 74.8

Predicted (%)												Actual (%)
	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock		Proportion
blues	81.00	2.00	6.00	1.00	1.00	4.00	1.00	0.00	1.00	3.00	blues	10.00 %
classical	2.00	96.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	classical	10.00 %
country	5.00	1.00	76.00	3.00	2.00	3.00	0.00	5.00	1.00	4.00	country	10.00 %
disco	4.00	0.00	11.00	59.00	6.00	0.00	2.00	9.00	2.00	7.00	disco	10.00 %
hiphop	1.00	0.00	1.00	5.00	64.00	0.00	2.00	6.00	14.00	7.00	hiphop	10.00 %
jazz	4.00	0.00	3.00	4.00	1.00	87.00	0.00	0.00	0.00	1.00	jazz	10.00 %
metal	0.00	0.00	0.00	1.00	4.00	0.00	82.00	3.00	0.00	10.00	metal	10.00 %
pop	0.00	2.00	4.00	8.00	6.00	2.00	2.00	69.00	5.00	2.00	pop	10.00 %
reggae	4.00	0.00	1.00	6.00	10.00	0.00	0.00	5.00	73.00	1.00	reggae	10.00 %
rock	2.00	0.00	9.00	11.00	7.00	1.00	7.00	2.00	0.00	61.00	rock	10.00 %

Figure 3: Resulting matrix confusion MLGTZAN5

inated and bad recordings were fixed through the use of a lossless dataset. In principle, this means that comparing this new dataset to the GTZAN dataset should give improved results however, a direct comparison turned out to be unfeasible and rather meaningless. The first and most important issue with this direct comparison is that the lossless dataset used did not contain all the audio files needed. This combined with the fact that not all GTZAN entries are identified meant that LGTZAN could have at most around 800/1000 of GTZAN's original songs. GTZAN was also trained using 30 second fragments of which the exact timings are unknown. And lastly, these files most likely went through some kind of prepossessing we can not recreate. The number of unknowns here means that any kind of result could come from several factors. So, to maximize control of the experiment LGTZAN was created as a baseline.

Sturm commented on his paper (Rodriguez-Algarra et al., 2019): "We have found sub-20 Hz information that greatly inflates the performance of particular models in GTZAN. I wonder if creating a "lossless" version will cure that". Although not an exact lossless version of GTZAN was created in this research and examining this inflation fell outside the scope of this research, the lossless version created was created to minimize the confounding factors. Perhaps a further analysis or improvement on this dataset could explain these performance inflations.

5.2 LGTZAN and data collection

The creation of LGTZAN brings with it some flaws of its own. By basing it on GTZAN the number of duplicate artists remained in the LGTZAN dataset, meaning it would be biased towards certain artists especially in problematic genres mentioned by Sturm such as blues. However, by adding randomness to the selection more diversity was added to the LGTZAN dataset. Another important characteristic of LGTZAN to consider is that LGTZAN takes full song lengths into account, whereas the previous GTZAN classifier used 30-second fragments. This might increase the accuracy, but it might also introduce a bias based on song length. Compared to the GTZAN classifier (accuracies shown in appendix A) disco stands out as a worse classified genre, this probably has to do with preprocessing of the audio or some changes in low-level descriptors.

In the selection process of the lossless dataset, songs were mapped to albums which were then added to the dataset. To minimize time spend, the first available album was selected. This means that in some cases certain artists might have many of their albums in the dataset while others have 1 album with all their songs. Additionally, some albums might be collections from different artists, adding new artists to the dataset. The training sets for the classifiers were selected by mapping every song with a known album id to a random song from that album. This increases the randomness and possible distribution of the classifier however, it might also introduce more duplicates. In some rare cases, albums might also contain songs from different genres.

The selection of the metal dataset was done with a filter on most owned albums per sub-genre. This introduces a bias towards more popular metal music in each genre. Albums with multiple metal sub-genres were not considered, which might create a deficit for certain merged genres for example alternative metal, combining alt-rock and metal.

5.3 Experiment outcomes

Converting LGTZAN to mp3 with worse audio quality gave a clear worse outcome. These results suggest that for a dataset of this size the SVM classifier perform better with more data points. However classical improved. Classical already having the highest accuracy is the most distinct genre in the list. A possible explanation for the improved accuracy might be

that the obfuscation of other genres made classical even more distinct.

The accuracies of the MLGTZAN classifiers increased over LGTZAN in some cases. Considering this was a random selection this gives a strong indication an improved classifier can be made by selecting an optimal dataset for metal and possibly other genres. However, using the same test method biases towards the dataset itself. Due to the large variance, it is uncertain at this point if adding more cases towards the edge of a genre spectrum can improve or decrease accuracy.

5.4 non-standard samples and optimization

In order to further optimize the tested classifiers, two things need to be considered. The first thing is how to measure optimization. Sturm mentions in his paper the "perfect results" of the GTZAN set, however, this will not mean the classifier performs well on songs outside of it's dataset. As was shortly seen the GTZAN classifier classifies a majority of songs as jazz. LGTZAN or MLGTZAN could be optimized by tweaking parameters from low-level classifiers. Changing this can be untactful as illustrated by changing the tonal_descriptor. However, completely optimizing might lead to overfitting on the dataset, which might be one of the reasons the GTZAN classifier performs badly on non-standard samples. Besides mentioned suggestions to optimize and research low- and high-level classifiers. I believe the best way to improve the dataset is to test it against a well balanced non-standard test set consisting of various genres.

6 Responsible Research

In this section, the ethical implications and reproducibility of the paper will be discussed. Reproducibility is an important problem to consider inherent to this field of study. As highlighted by Liem and Mostert (2020), large scale musical datasets cannot be shared because of this licensing. Which is one of the premier reasons for the need of audio feature databases such as Acoustic Brainz[4].

With regards to this research, reproducibility is also an interesting topic to address. As discussed earlier an exact replica of GTZAN was not reproducaible on account of missing information and undocumented timeslices and preprocessing. In a way, this research tries to tackle the problem of reproduacibility with the GTZAN dataset by creating a new version. This new version has been well documented. All the selection procedures have been described in the methodology section. Seeds and code have been saved to the GitLab repository. Titles of most songs have also been saved in gitlab[5]. However, some of the metadata is missing and the actual data used has been provided under contract by Muziekweb. The contract specifies no reuse of the data after the project ends is allowed. This means one would have to make a contract with Muziekweb to reproduce the results. To make this process easier the needed list of albums has also been saved to the repository. Replicating the research can be done without making a contract with Muziekweb, by following all the steps provided in the research with an alternative database. Further Evaluation and testing have been made possible by sharing the pre-trained classifiers on GitLab. These classifiers are free to use under the license the CC BY-NC-ND 4.0 license[6].

Ethical issues from this research stem from the classifier behaviour and database contents. One should always consider who is responsible when the classifier performs in an ethically questionable way. However, there is no interaction with sensitive data or with people directly as the classifier only interacts with audio features.

In the process of tagging the database for genres Hartnett (2015) was used. Discogs uses expert opinions of multiple experts for their tagging, reducing the likelihood of biases in their tagging. The GTZAN database has been modified to be more inclusive for the metal sub-genre by including more nationalities. One issue however is that Muziekweb mostly represents European and American bands, resulting in Asian metal not being represented in the dataset.

7 Conclusions and Future Work

This paper tried to answer the research question: How does the audio quality and content of the GTZAN dataset impact Essentia's SVM classifier. Comparing effects on the exact GTZAN dataset turned out to be unreachable and uninformative due to a lot of unknowns and uncontrollable variables. Instead, a retraining with full-length lossless files has been performed and used as a base case for several other retrainings. This base case called LGTZAN 4 had an accuracy of 74.5% close to the original GTZAN classifier, without preprocessing and optimization of the SVM model.

Experiment one tried to answer if decreasing audio quality would impact the classifiers accuracy. After retraining on lower audio quality the accuracy dropped by 2.5% showing that audio quality indeed decreases accuracy for this dataset. Decreasing the audio quality leads to fewer data points for the SVM to work with. Which in turn leads to worse fittings.

The second experiment tried to determine the impact the content of the dataset has on the accuracy. Changing the metal part of the dataset with a more balanced representation of the metal sub-genre turned out to be able to increase the accuracy. The exact reason is still unclear and needs more research as there seems to be a lot of variety between random metal songs selected for the dataset. A handpicked selection might give more insight. This experiment also gave confidence that improving other parts of the dataset in the same way as proposed by the paper might improve the quality of that classifier.

The quality of the classifier has only been measured in terms of 5-fold cross-validation accuracy. However, a stronger accuracy does not necessarily indicate better performance. Some testing with music samples of varying formats not included in both datasets showed that the original GTZAN classifier is not able to classify songs outside of it's dataset well, giving almost all samples a default value of jazz. Where LGTZAN 4 was accurate on the tested samples. However, these samples were not tested structurally.

[4] https://acousticbrainz.org/

[5] https://gitlab.ewi.tudelft.nl/cse3k-21q2-music-faithfulness/gtzan_augmentations

[6] https://creativecommons.org/licenses/by-nc-nd/4.0/

Optimizing the SVM model might be able to further improve the classifier in the future and possibly explain bad accuracies of certain genres like disco compared to the GTZAN classifier. Additionally, mislabellings should be analysed with their corresponding audio features to draw more meaningful conclusions on why certain mislabellings took place. Another way to improve as mentioned by other researchers is improving the underlying low-level descriptors used to extract audio features. The last way to improve on this is by tackling the issue of duplicate artists in the datasets. This is shown to at least partially work by the changing of the metal partition. Before being able to confidently use the improved dataset a further analysis of its potential flaws as described earlier should be done.

Further research also asks for a structural way to test the effect of non standard music samples in varying formats to test the performance of these classifiers "in the wild". Lossless retrainings seemed to identify genre more accurately in this setting. The reason for this is unclear from current studies.

References

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., Serra, X., et al. (2013). Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR).

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Essentia (2020). Svm model accuracies. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm at 21-01-2022.

Hartnett, J. (2015). Discogs. com. *The Charleston Advisor*, 16(4):26–33.

Hillier, B. (2020). Considering genre in metal music. *Metal Music Studies*, 6(1):5–26.

Kour, G. and Mehan, N. (2015). Music genre classification using mfcc, svm and bpnn. *International Journal of Computer Applications*, 112(6).

Liem, C. C. and Mostert, C. (2020). Can't trust the feeling? how open data reveals unexpected behavior of high-level music descriptors. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*.

Robert, J., Webbie, M., et al. (2018). Pydub.

Rodriguez-Algarra, F., Sturm, B. L., and Dixon, S. (2019). Characterising confounding effects in music classification experiments through interventions. *Transactions of the International Society for Music Information Retrieval*.

Sturm, B. L. (2012). An analysis of the gtzan music genre dataset. In *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, MIRUM '12, page 7–12, New York, NY, USA. Association for Computing Machinery.

Sturm, B. L. (2013). The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*.

Sturm, B. L. (2014). The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2):147–172.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Urbano, J., Bogdanov, D., Boyer, H., Gómez Gutiérrez, E., Serra, X., et al. (2014). What is the effect of audio quality on the robustness of mfccs and chroma features? In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014); 2014 Oct 27-31; Taipei, Taiwan.[place unknown]: International Society for Music Information Retrieval; 2014. p. 573-578.* International Society for Music Information Retrieval (ISMIR).

Xu, C., Maddage, N., Shao, X., Cao, F., and Tian, Q. (2003). Musical genre classification using support vector machines. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 5, pages V–429.

A GTZAN SVM accuracies

Accuracy: 75.528701%

Predicted (%)												
	blu	cla	cou	dis	hip	jaz	met	pop	reg	roc		Proportion
blu	78.00	1.00	8.00	3.00	1.00	1.00	3.00	0.00	2.00	3.00	blu	10.07 %
cla	2.15	92.47	1.08	2.15	0.00	0.00	0.00	1.08	0.00	1.08	cla	9.37 %
cou	1.00	1.00	78.00	7.00	0.00	2.00	0.00	4.00	3.00	4.00	cou	10.07 %
dis	0.00	1.00	5.00	71.00	3.00	1.00	2.00	7.00	5.00	5.00	dis	10.07 %
hip	2.00	1.00	0.00	6.00	73.00	0.00	3.00	3.00	11.00	1.00	hip	10.07 %
jaz	7.00	4.00	4.00	3.00	1.00	79.00	0.00	1.00	1.00	0.00	jaz	10.07 %
met	2.00	0.00	0.00	1.00	3.00	2.00	86.00	1.00	0.00	5.00	met	10.07 %
pop	0.00	1.00	6.00	6.00	5.00	0.00	0.00	75.00	4.00	3.00	pop	10.07 %
reg	3.00	2.00	4.00	4.00	11.00	2.00	0.00	5.00	64.00	5.00	reg	10.07 %
roc	7.00	2.00	6.00	10.00	3.00	2.00	4.00	2.00	4.00	60.00	roc	10.07 %

Actual (%)