INACCURATE PREDICTION OR GENRE EVOLUTION? RETHINKING GENRE CLASSIFICATION

Ke Nie
University of California, San Diego
knie@ucsd.edu

ABSTRACT

The existing MIR research on genre classification primarily focuses on how to classify a song into the "correct" genre while downplaying the fact that genres mutate over time and in response to social change in terms of their musical properties. Songs claiming the same genre can sound very different if they are released years apart, and genres may revive musical traditions from the past. In this paper, I show that the performance of genre classifiers fluctuates as genres evolve. Unsatisfactory performance of the classifiers may not indicate algorithmic flaws but rather the change of genre characteristics. I demonstrate this by studying the case of Chinese Hip-Hop music. Specifically, I collected and analyzed 69,427 songs from four genres (Hip-Hop, Pop, Rock, and Folk) released on a Chinese music platform between 2009 and 2019. Using classifiers trained from the songs in different year cohorts to predict the genre of all the songs, I show how genre classifiers can be used to detect the stylistic shift in Hip-Hop that happened during this period. The paper thus offers a novel, sociological perspective on contending with the much-challenged idea of improving genre classification accuracy for its own sake. However, instead of questioning the effort, I argue that MIR research on genre classification can be helpful for studying genre as a social construct and cultural phenomenon if the pursuit of prediction performance and the cultural meaning of inaccurate prediction are carefully balanced.

1. INTRODUCTION

Past MIR research on genre classification has primarily strived to find ways for algorithms to correctly detect music genres. Numerous studies have experimented with various data sources, algorithmic approaches, and evaluation metrics to improve algorithmic attempts to grasp the characteristics of music genres [1-7]. Most of the existing studies rely on fixed datasets and metrics to evaluate the performance of classifiers for the convenience of comparison.

One major concern in this area, however, is that genre is an ambiguous concept [8]. After all, what distinguishes one genre from another is *subjective*, since it is "based upon subjective responses with little inter-participant consensus" [3]. This is because the boundaries according to which genres are distinguished are not entirely rooted in the musical or sonic elements; rather, they are based on people's understanding of the difference between genres, which depends on their cognitive perception and cultural knowledge [3,9]. In fact, research has found that human evaluators may also be ambivalent about the classification of genres, which problematizes the very idea of genre and challenges the purpose of classification tasks [10, 11].

Adding to this line of questioning, this paper underscores the social dimension of genres that further complicates the concept by exploring critical implications for the design and implementation of genre classifiers. I argue that genres are social constructs that constantly change over time and in response to social, economic, and political pressure [12-15]; therefore, if genres evolve, the performance of genre classifiers trained on songs from precedented cohorts will fluctuate in predicting future cohorts. Given this feature, I argue that the "inaccurate" prediction of genre classifiers can be used to detect and map the evolution of genres. I demonstrate this point by presenting a case study of Chinese Hip-Hop music. Using songs from different year cohorts as the training set, I show that the performance of the classifiers mutates over time as the genre evolves. The findings demonstrate that prediction accuracy may not be the most valuable feature of algorithmic classifiers and that inaccurate predictions may suggest genre evolution.

This paper is organized as follows. In Section 2, I review the current MIR research on genre classification tasks and identify the common concern shared in this research area regarding the ambiguity of the genre concept. I propose taking a deeper look at the social dimension of genres, a generally understudied subject in the MIR field, as a way to advance our understanding of the evaluation of genre classifiers. In Section 3, I present the case of Hip-Hop music in China, a genre that has enjoyed a dramatic turn of events in the past decade. Drawing from different pieces of evidence, in Section 4 I show how the evolution of genre could complicate the performance of genre classifiers and how classifiers can be used to help understand the development of a genre. I then conclude the paper by discussing the impacts of genre evolution on the design and performance of genre classifiers.

2. RELATED WORK

The existing MIR literature on genre classification is primarily aimed at exploring, devising, and experimenting with algorithmic designs that could be used to understand and predict music genres automatically. These studies can be roughly categorized into three main themes.

First and foremost, researchers have attempted to identify different kinds of *data sources* that are useful for earmarking the characteristics of music genres. Two types of sources that are deemed significant are the music content — e.g., music and sonic features extracted from the digital signals of the songs [1] — and the music context — e.g., information on the performers, musicians, communities, and other social-cultural features associated with the genre [2]. Thus, discovering relevant data sources is the foundational work for performing genre classification tasks.

Second, contingent upon the data sources mentioned above, researchers have been trying to improve on the *algorithmic approaches* to achieve better performance in genre classification tasks. For example, researchers who use music content for the task have been experimenting with various types of acoustic representations (MFCCs, spectrograms, etc.) and machine learning architectures (SVM, CNN, etc.) [4,5,7,8,16]. There are also attempts to create a multimodal framework that juxtaposes both music content and music context data, such as the lyrics, to enhance the performance classification task [6]. The primary goal here is to improve the accuracy of the classification algorithms, which are usually tested on an existing dataset, such as GTZAN [1].

Studies aligned with the above two themes comprise much of the MIR research on genre classification. Their common goal is to find ways to enhance genre classification performance, which is usually evaluated based on metrics such as prediction accuracy. While these studies have explored pathbreaking methods through which genre classification tasks are performed at higher accuracy, some scholars have also reflected on the *evaluation metrics* of genre classifiers, arguing that "accuracy is not enough," as genre classification is essentially a task of musical recognition rather than classification accuracy [17,18]. Furthermore, researchers highlight the ambiguity of the concept of "genre" itself, pointing out that the notion is essentially "subjective" [3]; therefore, human evaluation needs to be included in the process [19]. While there is suspicion about the validity of genre classification tasks due to the ambiguity of the genre concept, most scholars still acknowledge the relevance of the research while looking for instruments to mitigate the problem of subjectivity [11]. Previous studies from this third, reflective approach have shared incisive thoughts on how to refine algorithmic frameworks for better classification performance.

Following past reflections on this issue, I want to point out another dimension that could further complicate but also help the task. As many contend, the concept of genre is subjective in the sense that the exact boundaries between genres are subject to human evaluation, which diverges across different human evaluators, leading to ambivalent classification outcomes in the first place [3,19]. Genre is, therefore, a human construct that depends on mutual agreement among evaluators or the community as a whole. This view is close to that of the social scientists and business scholars of the music industry, who have long argued that genres are social constructs that are classificatory apparatuses used by producers, consumers, and intermediaries to make sense of a distinct type of music [12-14]. This sociological view of genre claims that while music content may matter in the identification of genre, genres are, above all, divided by communities. This helps explain cases where different genres sound similar to each other or where the same genre may sound very different in different regions. For example, sociologist William Roy pointed out that "hillbilly music" and "race records" in the 1920s America were two genres that were associated with the same type of music, although the former was used as a market label for African American audience while the latter was for the "rural whites" due to segregations at the time [20]. They are, nevertheless, considered two disparate "genres" as they are tied to different communities.

A more important component of the sociological view of genre that is usually underplayed in the MIR research is that genres can evolve. In other words, the way in which genre classification systems upon which producers, consumers, and intermediaries agree to distinguish different types of music can change across time and localities [12,13]. For example, the "Rock" genre has experienced significant changes in terms of its associated sound since its inception in the 1950s, so music labeled as "Rock" today sounds very different from music given the same label 70 years ago. Moreover, genres may also influence each other in terms of how they sound. For example, "Nu-Metal," which prevailed in the late 1990s and early 2000s, arguably sounds more similar to Hip-Hop contemporaries than their Metal predecessors from the 1970s.

Additionally, genre categories may also evolve in response to external shocks to the industry or the social environment in general. One example is the case of censorship, in which censored genres have to change their content in response to political pressure [15]. Similarly, advancements in technology or sound devices also help to define or shift music genres, such as the use of the Roland TR-808 as the defining sound of early-1980s Hip-Hop music. The idea here is that genres are constantly shaped and reshaped by social, economic, and political forces; therefore, they cannot be seen as fixed, immutable entities.

The mutability of genre adds a further question to the design of genre classification algorithms in terms of their data sources, algorithmic architecture, and evaluation metrics. This paper attempts to address the issue of evaluation specifically. If genres evolve over time, one should then expect that the classification accuracy of a genre classifier will drop when it is used to predict the songs that are released a long time apart from the songs that were used to train the classifier. Specifically, we may expect that a classifier that was trained on a set of songs released in the 1980s will have a harder time predicting the correct genre of songs from the 2010s than songs from the 1990s. In this case, the accuracy metric is not useful for understanding the performance of the classifier.

In this case, evaluation metrics may have a richer meaning than simply as a criterion for making the genre distinction: they may be used to understand how genre evolves or how genres influence each other. For example, inaccurate predictions may not necessarily indicate algorithmic flaws; they may imply that the actual genre is moving toward the predicted genre in terms of its sound. The trickiest part of making such a claim, however, is how to distinguish genre evolution from the drawbacks of the algorithmic design when prediction accuracy is low. This will be discussed further in the analysis and discussion sections below.

3. DATA & METHOD

To demonstrate how genre evolution may confound the performance of genre classifiers, I study the case of Chinese Hip-Hop music from the past decade. Hip-Hop has long been a relatively unpopular genre in China since it was introduced to the country in the 1990s [21]. In recent years, however, Chinese Hip-Hop experienced a series of dramatic external shocks, which elevated the genre's prominence extraordinarily. In summer 2017, the first season of *The Rap of China* started to stream on iQiyi, one of the most popular streaming platforms in China. The reality-show-style music competition attracted an unexpectedly expansive audience, securing 2.7 billion views in the course of three months and becoming the most successful online program ever in the history of the sector. The contestants featured in this program, as well as their music, soon became the iconic representatives of an unprecedented fad. Most significantly, many of the high-profile contestants, including the co-winners of the competition, GAI and PG One, were known for their involvement in Trap music, a style that originated in the Southern United States and is characterized by its distinct use of synthesized drums (above all, Roland 808 drums). Moreover, it is remarkably different from the jazzy, pop-oriented Hip-Hop style prevalent in the genre before. In January 2018, Hip-Hop's rise was met with state censorship aimed at blocking Hip-Hop musicians from mainstream media, as the state was concerned about past Hip-Hop figures' problematic practices of including subversive and delinquent content in their songs. Although many musicians had to remove some of their songs from the internet under political pressure, Hip-Hop songs, in general, were still allowed to circulate online via music streaming platforms, and the censorship loosened up six months later without any official announcement [15].

The story of Chinese Hip-Hop music thus provides an interesting case for investigating the issue of genre evolution, as the genre became remarkably popular and potentially turned in a new stylistic direction due to the success of *The Rap of China*. To understand the impact of this stylistic change on machine learning classifiers, I collected songs from one of the most popular music streaming platforms in China. I collected the songs' audio files, as well as their metadata, including their genre and tags. The platform's setup requires each song to have only one identified genre, while the song can have multiple tags that are considered its "subgenres." Because of copyright issues, only a portion of the available songs could be downloaded at the time of the data collection in September 2019. Eventually, I collected 69,427 songs released between 2009 and September 2019 from four genres (Hip-Hop, Pop, Rock, and Folk), constituting the dataset on which I conducted my analysis. The other three genres are also well established in China in that they have distinct fanbases, although the Chinese Hip-Hop community was more connected to the Rock community particularly in the early 2000s for sharing resources and venues [22].

The structure of the dataset is given in Figure 1 below. As shown in the graphic, the number of total song releases consistently increases over the years, with the exception of the year 2019, when the data collection was suspended. The data and the code for the following analysis are publicly available,[1] although the metadata of the songs is left out due to copyright issues.

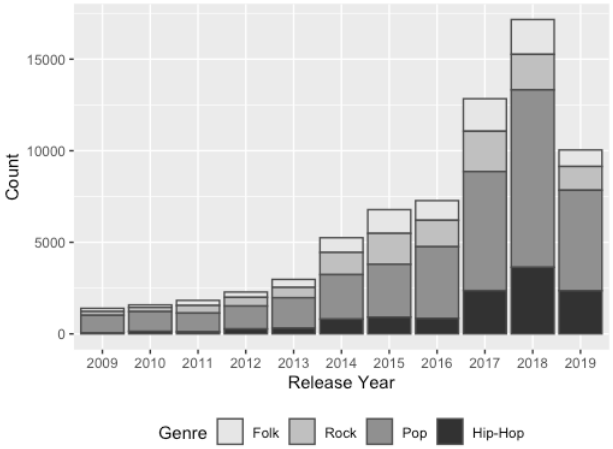

Figure 1. The structure of the dataset. The height of the bar represents the number of releases.

I used librosa [23] to extract sonic features from the audio files to prepare for training the classifiers. As my goal was not to improve classification accuracy, I followed Tzanetakis & Cook's [1] classic approach for computational efficiency. Specifically, I sliced a 1-minute long excerpt of audio signals from each song starting at 30 seconds into the song, which supposedly captures a substantive part of the song, and extracted 11 timbral texture features from this slice, including the chroma features (chroma_stft), root-mean-square of the spectrogram (rms), spectral centroid (spectral_centroid), spectral bandwidth (spectral_bandwidth), spectral roll-off (spectral_rolloff), zero-crossing rate (zero_crossing_rate), and the first five Mel-frequency cepstral coefficients (MFCC). I normalized the value of the features using a 0-1 scale and took the means of the features for each song, resulting in an 11-dimensional feature vector.

I used the songs released in 2009 and 2018, respectively, to train the genre classifiers. They represent the first and the last full year in the dataset, a fact that can be conveniently used to track changes over the year. I also used the widely studied GTZAN dataset [1] — its filtered version [24] due to the errors in the original collection [25] — as a supplementary check of the classifiers' performance trained on a dataset from a different time and region. For each of these three cohorts, I trained four classifiers: a Gaussian Naïve Bayes classifier (hereafter, GNB), a K-Nearest Neighbor classifier (hereafter, KNN), a Random Forest classifier (hereafter, RF), and a classifier trained by Extreme Gradient Boosting, a more recently prevailing tree-based algorithm. They were used in parallel to triangulate the results. For the 2009 cohort, I picked the top 50 most streamed songs of each genre, for a total of 200 songs, to compile the training set, whereas for the 2018 cohort, I picked the top 1900 songs of each genre, totaling 7600 songs. My reason for choosing these two numbers is

[1] The data can be found at https://github.com/norzvic.

that they are the closest number that can be divided by ten to the total number of releases of a genre with the least releases in that year (54 Hip-Hop songs in 2009; 1902 Folk songs in 2018), making it convenient for a 4:1 split for training and testing the classifiers. For the GTZAN dataset, I only used the data for three genres — Pop, Rock, and Hip-Hop — to train the classifiers, as there is no "Folk" in the GTZAN dataset. I extracted the data on the 11 features, normalized them, and used all 300 songs from the three genres in the dataset to train four classifiers similarly. The performance of the classifiers is shown in Table 1. I then used all four sets of classifiers to predict the genre of all the songs in the dataset, the results of which are reported in the next section.

		2009	2018	GTZAN
GNB	Acc	0.575	0.472	0.750
	AUC	0.832	0.755	0.867
KNN	Acc	0.625	0.584	0.800
	AUC	0.646	0.578	0.608
RF	Acc	0.800	0.580	0.833
	AUC	0.908	0.786	0.927
XGB	Acc	0.700	0.629	0.732
	AUC	0.905	0.864	0.916

Table 1. The performance of the classifiers across cohorts. Acc refers to the accuracy of predicting the test set, whereas AUC refers to the corresponding area under the ROC line for multi-classes.

4. FINDINGS

4.1 Overall Accuracy and Recall

As the focus of this study is Hip-Hop, I will primarily heed two metrics: Hip-Hop accuracy and Hip-Hop recall. Here, Hip-Hop accuracy refers to the rate at which the classifier correctly identifies Hip-Hop versus non-Hip-Hop, whereas Hip-Hop recall refers to the rate at which the classifier correctly identifies Hip-Hop among true Hip-Hop songs. The reason to incorporate recall is that the metric can be useful to detect potential genre evolution: the lower the recall is, the more probable that a Hip-Hop song is classified to other genres, which could mean that the genre is straying away from the sound captured by the training set.

The overall performance of the classifiers is shown in Figure 2. The "Average" column on the right of the graphic takes the average value of the corresponding metrics of the left three classifiers, showing an average trend of the metrics. The graphic demonstrates that the performance of most classifiers follows a fuzzy U-shape trend, with higher values at the start and the end than in the middle years of the dataset. The U-shape trend is statistically tested by a series of polynomial regressions of the metric on year and its quadratic term. In particular, the U-shape is demonstrable in the performance of the averaged classifier. All three recalls from the averaged classifier (bottom right of Figure 2) fit a polynomial regression on year and its quadratic term, where their coefficients are all statistically significant ($p < 0.05$). On the other hand, only the averaged classifier trained from the 2018 cohort songs yields significant results in the regression of accuracy on year and its quadratic term, while the other two averaged classifiers did not.

The results from the recalls, which imply how the true Hip-Hop songs of each year are similar to those of the benchmark cohort, indicate that Hip-Hop deviated from its late 2000s conventions in the middle of the 2010s but bounced back later, in particular after 2017, when the first season of *The Rap of China* aired. The downward trend of the prediction accuracy of the 2009 cohort classifiers may suggest more nuanced changes in the genre: given that the recall of the 2009 cohort is also U-shaped, it implicates that the 2009 cohort classifiers are able to detect the bouncing back of the genre conventions but confused other genres with Hip-Hop when the bounce took place in the late 2010s. This may be primarily because of the relatively small size of the 2009 cohort classifiers' training set, but it also suggests the possibility of other genres' cross-over to Hip-Hip, which I will discuss in 4.3.

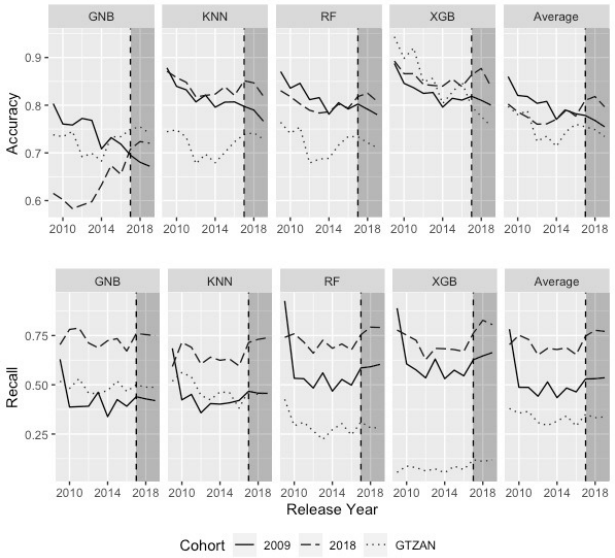

Figure 2. Hip-Hop accuracy and recall across cohorts and classifiers. The darker section on the right of each grid indicates the years since the first season of *The Rap of China* in 2017.

4.2 Metrics and Subgenre Salience

Genres are comprised of "subgenres," which are subsets of their parent genre but can be distinctly different from each other in terms of their sound [1,26]. The idea of distinct subgenres under the same genre category thus suggests that genre evolution may be understood as changes in the most salient subgenres, i.e., subgenres that take up the highest proportion of the genre. It is intuitive, then, to speculate that the classifier's performance is better when it predicts a set of songs whose most salient subgenres are more similar to those of the training set.

Testing this speculation requires, above all, identifying the most salient subgenres of the training set, or the subgenre to which most songs of that genre are affiliated. I thereby designed an original approach to measuring subgenre salience, and I used the approach for measuring the salience of Hip-Hop subgenres as follows. First, I extracted the subgenres from all the Hip-Hop songs in the

		GNB			KNN			RF			XGB			Average		
		Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec	Subg	Acc	Rec
2009	Top1	OS	0.236* (0.103)	0.625*** (0.095)	OS	0.188* (0.066)	0.636** (0.149)	OS	0.165* (0.067)	1.010*** (0.182)	OS	0.180** (0.050)	0.773** (0.165)	OS	0.154* (0.062)	0.761*** (0.139)
	Top5	OS; I; Con; A; CR	0.143* (0.045)	0.265** (0.075)	OS; I; Con; A; CR	0.123*** (0.021)	0.329** (0.072)	OS; A; U; Pop; I	0.090 (0.065)	0.001 (0.314)	OS; A; U; Pop; I	0.082 (0.058)	-0.049 (0.251)	OS; I; Con; A; CR	0.138*** (0.029)	0.301 (0.168)
2018	Top1	HH	0.211** (0.045)	0.065 (0.264)	HH	-0.001 (0.029)	0.208** (0.055)	HH	0.033 (0.023)	0.149* (0.048)	HH	0.003 (0.028)	0.206* (0.071)	HH	0.071* (0.022)	0.157** (0.047)
	Top5	HH; Pop; T; OS; CU	0.245** (0.058)	0.016 (0.070)	HH; Pop; T; OS; CU	-0.022 (0.034)	0.150 (0.093)	HH; Pop; T; OS; CU	0.245** (0.058)	0.106 (0.074)	HH; Pop; T; OS; CU	-0.013 (0.033)	0.136 (0.109)	HH; Pop; T; OS; CU	0.082** (0.028)	0.102 (0.077)

Note: Standard errors are in parentheses. Among the subtitles, Subg refers to subgenre, Acc refers to accuracy, and Rec refers to recall. Abbreviations in the Subg column: OS for Old-School Hip-Hop; I for Instrumental Hip-Hop; Con for Conscious Hip-Hop; A for Alternative Hip-Hop; CR for Cloud Rap; U for Underground Hip-Hop; Pop for Pop Rap; HH for Hip-Hop; T for Trap; CU for Chinese Underground Hip-Hop.
*p < .05; **p <.01; ***p<.001

Table 2. Performance of genre classifiers on songs released 2010-2019.

dataset and counted their appearances each year. In many cases where a song has multiple subgenres, one more appearance was added for each of them. To account for the fact that prediction performance on each subgenre can significantly impact the prediction performance on the genre as a whole, I used the 2009 cohort classifiers and the 2018 cohort classifiers, respectively, to predict the genre of the Hip-Hop songs released in the year of the same cohort. After that, I calculated the prediction accuracy of each subgenre (which is also identical to recall, as only true Hip-Hop songs are examined here) by computing the proportion of correct predictions among all the songs with the subgenre in question. In this case, the prediction accuracy of each subgenre also indicates how the classifiers were trained to favor certain subgenres over others. I multiplied the prediction accuracy of each subgenre by their number of appearances. The product is the measurement of subgenre salience, which nicely captures the size of the impact that a subgenre can make on the pooled metrics of the genre. I identified the top 1 subgenre and top 5 subgenres with the highest salience score, which are presented in the note in Table 2. Then, I calculated their proportion in the total number of releases each year, respectively. Finally, I employed linear regression of Hip-Hop accuracy and recall on the proportion of the most salient subgenre(s) each year to understand the association between them. The objective of the linear regression is to see if the classifier performs better when the salient subgenres, which are favored by the classifiers by design, are more dominant among all subgenres of a year. The GTZAN cohort was not included in the analysis in this part, as it does not have a corresponding year in the dataset.

The coefficient estimates of the proportions of salient subgenres according to the regression models are presented in Table 2. The overall pattern generally substantiates a positive relationship between performance metrics and the size of the salient subgenres' proportions. In other words, the classifiers perform worse in years where there are fewer salient subgenres among the total releases of the year in question. Specifically, the average classifier trained on 2009 songs performs better when there are more songs in that year that are Old School Hip-Hop, Instrumental Hip-Hop, Conscious Hip-Hop, Alternative Hip-Hop, or Cloud Rap, whereas the average classifier trained on 2018 songs performs better when there are more Pop Rap, Trap, Old School Hip-Hop, or Chinese Underground Hip-Hop present in that year. The results imply that it is possible to detect genre evolution from the performance of the classifiers, which is demonstrable in the metrics when the genre deviates from the salient subgenres of the classifier.

4.3 Metrics and Genre-Crossing

As implied in Figure 2, the evolution of a genre may involve other genres. This could either mean "assimilation," where the genre in question is absorbing musical elements from other genres, or "dispersion," where the musical elements of the genre in question penetrate other genres. In either case, it is reasonable to speculate that the interaction between genres will likely do harm to the performance of the classifiers that are trained on songs where genre-crossing is not prominent.

To investigate this issue, I first identified the number of releases in non-Hip-Hop genres (i.e., Folk, Rock, and Pop) with at least one subgenre that is explicitly associated with Hip-Hop. I did so by searching all the songs whose subgenres incorporate strings such as "Hip-Hop," "Hip Hop," "Rap," "Trap," and their Chinese equivalent. Table 3 presents the number of Hip-Hop-crossing non-Hip-Hop songs and their percentage among the total releases over the years studied. The table has some important implications for understanding classifiers' performance, as illustrated in Figure 2. First, there are no Hip-Hop-crossing non-Hip-Hop songs from 2009-2011. This means that the 2009 cohort classifiers were not well-trained to correctly detect

	Hip-Hop crossing non-Hip-Hop	Total Non-Hip-Hop	Percentage (%)
2009	0	1344	0
2010	0	1446	0
2011	0	1717	0
2012	22	1991	1.1
2013	59	2603	2.2
2014	98	4347	2.2
2015	184	5682	3.1
2016	163	6264	2.5
2017	94	10381	0.9
2018	135	13394	1.0
2019	118	7565	1.5

Table 3. The number of Hip-Hop-crossing non-Hip-Hop songs and their percentage in the total non-Hip-Hop releases over the years studied.

	GNB		KNN		RF		XGB		Average	
	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN	Diff. Acc	Diff. FN
2009	-0.063*** (0.015)	0.087*** (0.016)	-0.020 (0.017)	0.163*** (0.015)	-0.089*** (0.015)	0.162*** (0.016)	-0.076*** (0.016)	0.181*** (0.016)	-0.062*** (0.008)	0.148*** (0.008)
2018	-0.090*** (0.012)	0.217*** (0.017)	-0.125*** (0.016)	0.270*** (0.017)	-0.131*** (0.016)	0.245*** (0.017)	-0.161*** (0.016)	0.261*** (0.017)	-0.127*** (0.008)	0.249*** (0.008)
GTZAN	-0.079*** (0.011)	0.182*** (0.017)	-0.120*** (0.016)	0.134*** (0.017)	-0.086*** (0.015)	0.263*** (0.015)	-0.081*** (0.014)	0.073*** (0.010)	-0.092*** (0.007)	0.120*** (0.008)

Note: Standard errors are in parentheses.
*p < .05; **p <.01; ***p<.001

Table 4. Two-sample T-tests of key metrics between Hip-Hop-crossing non-Hip-Hop songs and those that do not cross the Hip-Hop boundary. Among the subtitles,

Hip-Hop-crossing non-Hip-Hop songs, which occur more frequently in the subsequent years. This possibly explains why 2009 cohort classifiers have a downward trend in their accuracy, as they are underperformed when identifying crossover songs. Second, the fact that there are more Hip-Hop-crossing non-Hip-Hop songs in terms of their percentage in the total non-Hip-Hop releases in the middle years than in the early and the late 2010s coincides, inversely, with the U-shaped trend of the metrics in Figure 2. This makes sense, as the prominence of genre-crossing is supposed to curb classification performance.

To further understand the relationship between genre-crossing and the performance of the classifiers, I also analyzed the difference in performance between genre-crossing non-Hip-Hop songs and those that did not cross the boundary of Hip-Hop. To do so, I conducted two-sample t-tests on two metrics between the two groups. The first metric was prediction accuracy, indicating the rate at which the non-Hip-Hop genre of the song was correctly identified. The second metric was the False Negative rate, indicating the rate at which the non-Hip-Hop song was incorrectly identified as Hip-Hop. Table 4 shows the results of the test, which unanimously demonstrate how genre-crossing discounts classification performance. Specifically, the consistent and statistically significant negative values under the columns of the difference in prediction accuracy suggest that the classifiers' ability to predict non-Hip-Hop songs that do not cross the boundary of Hip-Hop is always better than their ability to predict those that do so. Similarly, all the positive values under the False Negative columns provide strong evidence that classifiers are better at identifying non-Hip-Hop songs when they do not claim Hip-Hop as one of their subgenres. In short, the performance of genre classifiers may be an indicator of genre evolution, as bad performance points to genre-crossing.

5. DISCUSSION

Genres will evolve, and this will affect the performance of genre classifiers regardless of the data sources or algorithmic approaches they use. By analyzing the case of Chinese Hip-Hop, I demonstrate that the performance of genre classifiers is significantly impacted by genre evolution, particularly when the training set has less songs of the salient subgenres and when genres start to cross boundaries.

The goal of the present study is to highlight genre as a mutable cultural construct and to examine what this means for MIR research on genre classification. This study has at least three implications. First, similar to what has been argued in some of the previous MIR studies of genre, I contend that musical genre is a cultural construct and that the musical elements associated with a genre can change over time. My findings also suggest that the prevalence of a particular music style in a genre may be revived after a period of relative silence, which echoes what music industry scholars call a "revival" of music styles [13].

Second, because genres change and evolve, one possible application of genre classifiers is to understand the patterns of such change. I show it is possible to use genre classifiers, if trained on a particularly designed dataset, to detect the way genres evolve. It is difficult, though, to separate genre evolution from flaws in algorithmic design; therefore, it is important to supplement the analysis with a more detailed investigation of the soundscape of the songs across subgenres and songs that cross genres.

Third, the findings suggest that using accuracy or accuracy-related metrics to train genre classifiers might not be the only best application in practice. Regardless of their algorithmic framework, genre classifiers are almost always better at predicting the genre of songs released within a similar timeframe as those in the training set. As a result, if genre classifiers are meant to help classify newly released songs, it might be futile to aim at improving prediction accuracy, as the genre might evolve. This implies that we need to find an optimal balance between raising the prediction accuracy of the algorithms and understanding the developing trend of genres, which might require data exploration or even qualitative data beyond the design and optimization of machine learning algorithms.

There are obvious limitations in the data analysis; however, I argue that they will not severely affect the findings. Above all, a significant number of songs were inaccessible, especially in very distant years, which may twist the conclusion that Hip-Hop has been "revived" from its convention 10 years ago. It is certainly likely that the actual dominant subgenre in 2009 may not be "Old-School Hip-Hop." However, the finding that the way classifiers are trained can affect their performance still holds. It remains true that, if a genre classifier is trained on the songs of a particular composition of subgenres, the classifier might better predict the genre of a song when it shares commonalities with that subgenre's composition. In that case, classifiers can still be used to understand the evolution of genres, although they would be better "detectors" if more data were available.

6. ACKNOWLEDGEMENT

The author would like to thank Jin Ha Lee, Oriol Nieto, and the anonymous reviewers for their insightful comments and suggestions. The author would also like to acknowledge Leo Y. Yang for his assistance with data collection.

7. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] P. Knees, E. Pampalk, and G. Widmer, "Artist Classification with Web-Based Data.," 2004.

[3] A. J. Craft, G. A. Wiggins, and T. Crawford, "How Many Beans Make Five? The Consensus Problem in Music-Genre Classification and a New Evaluation Method for Single-Genre Categorisation Systems.," in *ISMIR*, 2007, pp. 73–76.

[4] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification," *Applied Sciences*, vol. 8, no. 1, p. 150, 2018.

[5] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE signal processing magazine*, vol. 36, no. 1, pp. 41–51, 2018.

[6] S. Oramas, F. Barbieri, O. Nieto Caballero, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[7] J. Pons and X. Serra, "Randomly weighted cnns for (music) audio classification," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019, pp. 336–340.

[8] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[9] M. Schedl, E. Gómez Gutiérrez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval. 2014 Sept 12; 8 (2-3): 127-261.*, 2014.

[10] M. Sordo, O. Celma, M. Blech, and E. Guaus, "The quest for musical genres: Do the experts and the wisdom of crowds agree?," in *ISMIR*, 2008, pp. 255–260.

[11] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?," in *ISMIR*, 2006, pp. 101–106.

[12] J. C. Lena and R. A. Peterson, "Classification as culture: Types and trajectories of music genres," *American Sociological Review*, vol. 73, no. 5, pp. 697–718, 2008.

[13] J. C. Lena, *Banding together: how communities create genres in popular music*. Princeton, N.J: Princeton University Press, 2012.

[14] P. DiMaggio, "Classification in Art," *American Sociological Review*, vol. 52, no. 4, pp. 440–455, 1987.

[15] K. Nie, "Disperse and preserve the perverse: computing how hip-hop censorship changed popular music genres in China," *Poetics*, p. 101590, 2021.

[16] D. Diakopoulos, O. Vallis, J. Hochenbaum, J. W. Murphy, and A. Kapur, "21st Century Electronica: MIR Techniques for Classification and Performance.," in *ISMIR*, 2009, pp. 465–470.

[17] B. L. Sturm, "Two systems for automatic music genre recognition: What are they really recognizing?," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2012, pp. 69–74.

[18] B. L. Sturm, "Classification accuracy is not enough," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 371–406, 2013.

[19] S. Lippens, J.-P. Martens, and T. De Mulder, "A comparison of human and automatic musical genre classification," in *2004 IEEE international conference on acoustics, speech, and signal processing*, 2004, vol. 4, pp. iv–iv.

[20] W. G. Roy, "'Race records' and 'hillbilly music': institutional origins of racial categories in the American commercial recording industry," *Poetics*, vol. 32, no. 3–4, pp. 265–279, Jun. 2004.

[21] J. Sullivan and Y. Zhao, "Rappers as knights-errant: classic allusions in the mainstreaming of Chinese rap," *Popular Music and Society*, pp. 1–18, 2019.

[22] J. De Kloet, *China with a cut: globalisation, urban youth and popular music*, vol. 3. Amsterdam University Press, 2010

[23] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.

[24] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015

[25] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *Journal of New Music Research*, vol. 43, no. 2, pp. 147–172, Apr. 2014

[26] A. Caparrini, J. Arroyo, L. Pérez-Molina, and J. Sánchez-Hernández, "Automatic subgenre classifi-

cation in an electronic dance music taxonomy," *Journal of New Music Research*, vol. 49, no. 3, pp. 269–284, 2020.