International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Apr-24)

http://dx.doi.org/10.12785/ijcds/1501106

Rocking Across Borders : An Analysis of the Musical Differences between Bangladesh and West-Bengal Rock Songs Using Spotify Audio Features

Moshiur Rahman Autul[1], Durjoy Dey[1], Partha Protim Paul[1] and Mohammed Raihan Ullah[1]

[1]*Institute of Information and Communication Technology,Shahjalal University of Science and Technology, Sylhet, Bangladesh*

Received 17 Sep. 2023, Revised 14 Mar. 2024, Accepted 18 Mar. 2024, Published 1 Apr. 2024

Abstract: Over the last few decades, there has been a significant increase in the availability and utilization of large music collections. However, most studies of these collections have been limited to Western music, which hinders our ability to comprehend the diversity and commonality of music across all cultures. Based on popularity from Spotify, it has been discovered that Bangladeshi rock music is more popular than the rock music from West Bengal, India. Previous research suggests that listeners from diverse cultural backgrounds may have varying preferences when comes to music appreciation. This research aimed to explore the reasons behind the popularity of Bangladeshi rock music compared to West Bengal rock music. By extracting various features from songs, the study sought to identify what is the reason behind the popularity of Bangladeshi rock music and whether there are any differences in terms of musical features.

Keywords: Machine Learning, Random Forest, Support Vector Machine, Data Mining, Empirical Study, Statistical Analysis, Musicological Analysis, Cultural Diversity, Feature Extraction, Feature Scaling, Pearson Correlation, Correlation Coefficient, Chi-Square Testing, Hypothesis Testing

1. INTRODUCTION

Both Bangladesh and West Bengal share the Bengali language as their native tongue. But, from Spotify, we know that Bangladeshi rock songs are more popular than West Bengal's rock songs. The reason for this popularity remains unknown. Both regions share the same language and have somewhat similar cultures, so it's interesting to figure out what things make rock songs more popular in Bangladesh. In this paper, our primary motivation is to uncover the reasons behind this trend. Another important incentive is the potential impact on the rock music industry. Successful identification of these reasons could provide valuable insights for producers. They could gain a deeper understanding of the musical features in Bangladesh and West Bengal rock music, identify patterns, and use this knowledge to tailor their songs for greater popularity. For producers in West Bengal, understanding the aspects that make their songs less popular than those from Bangladesh could guide them in enhancing their songs' appeal. This knowledge could potentially elevate West Bengal rock songs to the same level of popularity as those from Bangladesh. Moreover, our research aims to shed light on the commonalities and differences between the rock music of these two regions.

Music is a universal language that has the power to transcend cultural boundaries and unite people across the globe. Despite its widespread appeal, music can vary greatly depending on the region and culture from which it originates. This is precisely why the classification of cross-cultural songs is so vital - it allows us to appreciate and comprehend the vast and diverse range of musical traditions from around the world. Cross-cultural song classification involves identifying and categorizing songs based on their cultural origin, style, and instruments used. This classification can be particularly challenging, given the increasing globalization of music and the blending of different genres and styles. Moreover, music audio features can be influenced by cultural differences [1]. Nevertheless, it is essential to recognize the distinct musical traditions and unique cultural elements that each song represents.

One approach to cross-cultural song classification is to use geographic regions as a starting point. For example, songs from Western Europe often feature classical music and orchestral arrangements, while African music is characterized by complex rhythms and percussion instruments. Meanwhile, music from the Middle East typically features stringed instruments such as the oud and qanun [2]. Another approach to cross-cultural song classification is to focus on the cultural elements present in the music. For example, many traditional Native American songs are characterized by the use of vocals, drums, and rattles, while Celtic music often features fiddles and bagpipes. These cultural elements

E-mail address:moshiur21@student.sust.edu, durjoy62@student.sust.edu, partha-iict@sust.edu, raihan-iict@sust.edu

help to distinguish each musical tradition and provide a deeper understanding of the culture that produced it.

Cross-cultural song classification is important not only for understanding the music itself but also for appreciating the cultural and historical contexts that gave rise to it. By exploring the diverse musical traditions from around the world, we can broaden our understanding of the human experience and celebrate the unique cultural contributions that each society has made to the world of music [3].

The cross-cultural song classification between Bangladesh and West Bengal likely stems from a desire to understand and appreciate the shared cultural heritage of the two regions. Bangladesh and West Bengal, which are adjacent regions located in the eastern part of the Indian subcontinent, share a long history of cultural and linguistic exchange. Both regions have a rich musical heritage, with a diverse range of genres and styles that have evolved over centuries. However, due to the historical and political boundaries that have separated these regions over time, the musical traditions of each region have developed unique characteristics. For this separation, there have been some cultural changes and Cultural changes can have an impact on music, influencing its diversity and prompting alterations [4]. The cross-cultural classification of songs between Bangladesh and West Bengal seeks to bridge this divide by identifying the commonalities and differences between the musical traditions of the two regions. This can lead to a deeper understanding of music and appreciation of the cultural heritage of each region, and foster greater cross-cultural exchange and collaboration in the future [5], [6].

By extracting key musical features from both regions' songs, we aimed to find out the relationships between these features. And identify which features differ significantly between the songs of both regions. Our ultimate goal is to determine if these distinctive features significantly contribute to the elevated popularity of Bangladeshi rock songs compared to those from West Bengal, or if psychological and regional cultural factors play a dominant role.

2. BACKGROUND AND RELATED STUDY

Musical classification has not received much attention in the first half of the 20th century. And musicology has never quite found its comfort zone in cross-cultural classification. The task of classifying music acoustically poses various challenges, such as the requirement for classification schemes to be universally applicable. However, these challenges do not necessarily invalidate the concept of cross-cultural classification. Lyrical content and style provide insights into cross-cultural similarities or diversities in distinct societies [7]. Furthermore, emotion influences music choice, East Asian cultures prefer high-energy songs, strongly associated with anger downregulation [8].

There are two primary methodological obstacles to the cross-cultural classification of music. The first pertains to instrumental music and involves ensuring that we are making fair comparisons between different culture's use of dissimilar instruments with distinct acoustic features, production methods, and tuning systems. The second challenge is related to vocal music and necessitates creating a classification system that is inclusive enough to encompass all musical cultures while still being able to distinguish between "song" and "speech." Furthermore, Among the key considerations in music selection, mood holds significant importance, and emotional feelings associated with music also play an important role [9], [10]. However, the interpretation of mood is subjective and can be shaped by various factors, with the listener's cultural background being a notable influence [11]. Music serves as a form of communication with a distinct ability to unite individuals and convey a diverse array of emotions [12].

A. *CantoCore: (a song-classification scheme)*

The "Cantometrics" system, which is considered the most well-established song-classification scheme, includes multi-dimensional and musilinguistic spectra as a significant design element. This system categorizes songs based on multiple acoustic characteristics related to their structure, performance style, and accompanying instruments. Each characteristic has several character states, ranging from individualized to groupy, which are arranged along a social continuum.

The main aim of this study is to provide a comprehensive examination of a novel song classification system that is universally applicable. The system is known as "CantoCore" and is focused on the fundamental structural features of a song [13]. The classification system places its attention solely on the structural attributes of songs, rather than their instrumentation or performance style. This is because the creators of the system hypothesize that structural traits are more dependable and consistent than other aspects of a song. The song classification system is intended to cover all musical forms, from basic sentence structures to intricate responsorial polyphony, across the entire musilinguistic spectrum. The study also includes an evaluation of the reliability of the song codes through two comparisons: CantoCore vs. Cantometrics, and the structural characteristics of Cantometrics vs. its performance and instrumental characteristics.

Both song classification systems utilize only acoustic data, without taking into account non-acoustic factors. While Cantometrics (represented by the green box) considers a song's performance, structure, and accompanying instruments, CantoCore (represented by the red box) concentrates solely on the structural attributes of the vocal component, excluding performance and instrumental features.

B. *Classification Scheme*

Music is composed of multiple levels of organization arranged hierarchically. The figure 2 below illustrates a representation of this musical hierarchy, which is used to

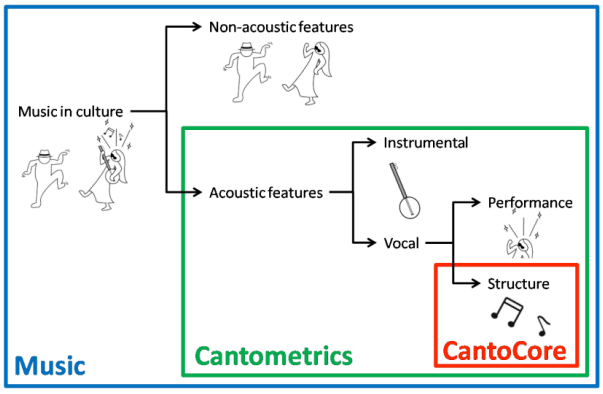

Figure 1. CantoCore vs Cantometrics [13].

categorize the attributes of the CantoCore classification system.

One way to understand the CantoCore classification system is to compare it to a biological organism. In this analogy, a song is akin to a complex arrangement of notes, much like an organism is made up of various cells. However, just as cells interact with each other and with their environment in multifaceted ways, notes in a song also have intricate relationships and interact with their surrounding musical and non-musical elements on various levels. While these interactions cannot be entirely measured, they can be modeled in a useful way.

The fundamental categorization in the CantoCore system is between the note level and the supra-note level. At the note level, the most basic building block of music is the note itself, which can be further categorized into three characteristics: 1) rhythm, which represents the duration of a note and is colored red. 2) pitch, which reflects the acoustic frequency of a note and is colored blue; and 3) syllable, which represents the articulation of a sung note and is colored green in the figure, exemplified by the syllable ”la” [14].

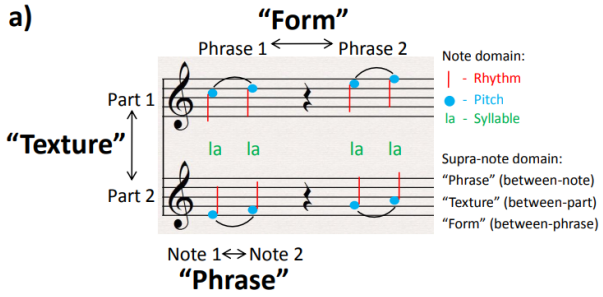

Figure 2. The musical hierarchy [13].

The supra-note level of the CantoCore system refers to the interactions between notes, which are organized into three overarching hierarchical domains. The first is the phrase domain, which represents the between-note level within individual vocal parts. The second is the texture domain, which represents the between-part level, where simultaneous phrases in different vocal parts overlap in time. The third is the form domain, which represents the between-phrase level, where successive phrases come together to form larger melodic units.

The figure 3 outlines the classification characteristics associated with each of the three supra-note domains mentioned earlier. Additionally, it demonstrates that the domain of ”phrase” encompasses the three note-level characteristics of rhythm, pitch, and syllable.

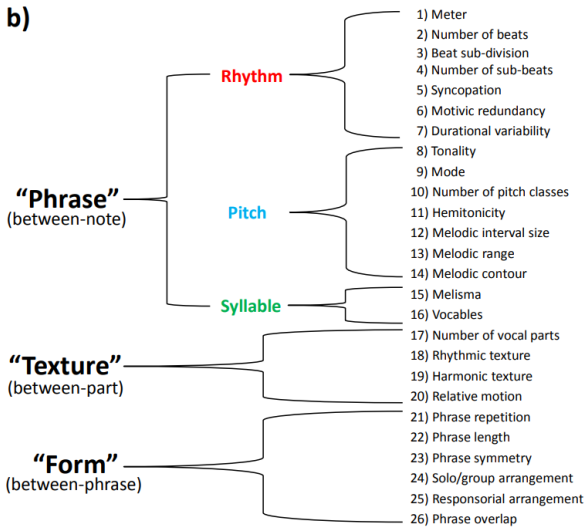

Figure 3. The 26 characters of CantoCore classification scheme[13].

The CantoCore song classification scheme categorizes 26 structural characteristics of songs, as presented in Figure 3. These are organized into categories associated with the note and supra-note domains previously mentioned. Among the 26 structural characteristics, 15 are more detailed versions of the structural characteristics already present in Cantometrics. Meanwhile, the other 11 characteristics, mostly related to rhythm and scale, are entirely new.

C. Quantitative vs. qualitative characters

Classification theory makes a fundamental distinction between two types of characters: quantitative (continuous) and qualitative (discrete). Quantitative traits can be classified based on their size, such as melodic intervals which can range from very small to very large, and can also be coded by their frequency of occurrence in a song. In the CantoCore system, vocables are coded based on their frequency, from being completely absent to being ubiquitous. In contrast, qualitative traits cannot be measured on a numerical spectrum and are instead organized into discrete states. For example, melodic contours are classified into different types, such as descending, ascending, or arched contours. Of the 26 characters in the CantoCore system,

15 are considered quantitative traits and 11 are qualitative traits based on classification theory.

D. Feature Dependence

In some cases, certain characters are dependent on others. For instance, songs without a beat cannot have a sub-beat, so a ”n/a” character state is added to indicate that a character is unclassifiable.

To this day, there is a considerable amount of research on music, providing valuable insights into musical features, cross-cultural songs, and genre classification [15], [16]. However, there is a notable gap in research regarding cross-cultural song classification in the Bengali language. Additionally, there is a lack of investigation into the relationship between Bengali musical features and their popularity. Therefore, our study focuses on Bengali rock music. By extracting key musical features and employing machine learning algorithms, we aim to understand the correlation between rock music characteristics and its popularity.

3. METHODOLOGY

In this study, we aim to find the musical differences between Bangladeshi Rock Songs and West Bengal Rock Songs. A systematic approach was undertaken to complete this study. Here are the approaches.

A. Collecting Music Data

We compiled a list of songs from each region. Choosing songs is one of the most important parts of our research. To ensure a balanced representation of songs, our selection process focused on the most popular tracks within specific regions on Spotify, with a particular emphasis on rock music. We set the popularity of music threshold to 0, meaning we collected the songs that had more than zero popularity value. Which ensured that we included songs that had gained even the slightest attention. Initially, we handpicked 150 rock songs from Bangladesh, followed by an equivalent number from West Bengal, India.

To curate the final list, we enlisted the opinions of ten passionate music enthusiasts. These individuals thoughtfully ranked the songs based on their personal preferences. From these rankings, we meticulously identified the top 73 songs from each region [17].

We retrieved the popularity of the songs by using the Spotify Search API along with corresponding song track IDs [18].

Sample API Request for Retrieving Track ID:

```
curl --request GET\
--url https://api.spotify.com/v1/tracks/id\
--header 'Authorization: '\
--header 'Content-Type: application/json'
```

Sample API Response for Retrieving Track ID: [Only the section regarding the songs' popularity response is presented; the full response is not included here.]


```
{
   "name": "string",
   "popularity": 0,
   "preview_url": "string",
   "track_number": 0,
   "type": "track",
   "uri": "string",
   "is_local": true
}
```


B. Feature Extraction

The process of assessing and extracting valuable information from unprocessed data is known as feature extraction. Feature extraction in the context of music is examining the audio signal of a piece of music and identifying particular traits that can be used to categorize or evaluate the music.

First, we made a spreadsheet that has all the selected song names and their corresponding track IDs.

Bangladeshi songs and corresponding Spotify Track ID data set link : Bangladeshi songs and track ID file

West Bengal songs and corresponding Spotify Track ID data set link: West Bengal songs and track ID file

Subsequently, we embarked on the task of extracting a multitude of song features through the Spotify Web API. To initiate this process, we created a Node.js API capable of retrieving the Spotify track IDs for each song listed within the spreadsheet. By leveraging these IDs and facilitating API calls to the Spotify web interface, we were able to get all the features of the songs [19], [20].

Music data set with features: Music Data Set

Replication Package: Github Code

Sample API Request for Feature Extraction:

```
curl --request GET\
--url api.spotify.com/v1/audio-features/id\
--header 'Authorization: '\
--header 'Content-Type: application/json'
```

Sample API Response for Feature Extraction:


```
{
  "acousticness": 0.00242,
  "analysis_url": "https://api.spotify.com/v1/au
   dio-analysis/2takcwOaAZWiXQijPHIx7B",
  "danceability": 0.585,
  "duration_ms": 237040,
  "energy": 0.842,
  "id": "2takcwOaAZWiXQijPHIx7B",
  "instrumentalness": 0.00686,
```

```
  "key": 9,
  "liveness": 0.0866,
  "loudness": -5.883,
  "mode": 0,
  "speechiness": 0.0556,
  "tempo": 118.211,
  "time_signature": 4,
  "track_href": "https://api.spotify.com/v1/tracks/
   2takcwOaAZWiXQijPHIx7B",
  "type": "audio_features",
  "uri": "spotify:track:2takcwOaAZWiXQijPHIx7B",
  "valence": 0.428
}
```


After making API calls, we obtained all the aforementioned features for each song in JSON format. Using this JSON data, we created CSV file.

C. Dataset Description

The dataset utilized in this study comprises a comprehensive collection of popular rock songs from both Bangladesh and West Bengal(India), with musical features collected from Spotify [21]. The dataset is structured across multiple CSV files, each containing specific subsets of the data. Here, in table I we discussed our dataset fields that are related to musical features.

1. bangladesh_songs_features_csv.csv: This CSV file contains the musical features extracted from Spotify for popular rock songs originating from Bangladesh. Each entry includes information about the song's musical attributes, such as tempo, key, danceability, energy, acousticness, instrumentalness, valence, loudness, and more. In table I we discussed them in detail. There are also two additional fields named 'song name' and 'popularity,' indicating how popular a specific song is. Additionally, a categorical field labeled as "country" assigns a value of 1 to indicate that the song is from Bangladesh.

2. west_bengal_songs_features_csv.csv: Similarly, this CSV file comprises musical features obtained from Spotify for popular rock songs originating from West Bengal, India. Like the Bangladesh dataset, each entry includes a comprehensive set of musical attributes, alongside 'song name', 'popularity' and a categorical "country" field with a value of 0 to denote that the song is from West Bengal.

3. all_song_features_csv.csv: This CSV file is a combined dataset merging the musical features of rock songs from both Bangladesh and West Bengal. By consolidating the data from the individual CSV files, the combined dataset offers a holistic view of rock music across the two regions, facilitating comparative analysis and exploration of cross-cultural influences. The "country" field distinguishes between songs from Bangladesh (labeled as 1) and West Bengal (labeled as 0), enabling researchers to identify regional trends and differences in musical characteristics.

To ensure uniformity and comparability across different features, scaling methods have been applied to bring the features into specific ranges. This standardization process enables consistent analysis and interpretation of the data, mitigating the impact of differing feature scales on analytical outcomes. Additionally, the scaled features are consolidated into a single CSV file, facilitating streamlined access and analysis.

1. scaled_all_song_features_csv.csv: This CSV file contains the scaled musical features extracted from Spotify for popular rock songs from both Bangladesh and West Bengal. Each entry includes detailed information about the song's scaled musical attributes, ensuring that features are within a specific range for uniform analysis. The scaled features encompass a variety of dimensions, including tempo, key, danceability, energy, acousticness, instrumentalness, valence, loudness and more. In table I we discussed them in detail. There are two additional fields named 'song name' and 'popularity,' indicating how popular a specific song is. Additionally, the "country" field distinguishes between songs from Bangladesh (labeled as 1) and West Bengal (labeled as 0), enabling researchers to explore regional trends and differences in scaled musical characteristics.

D. Cleaning The Data

An essential preprocessing step in machine learning is data scaling, commonly referred to as feature scaling. It entails altering the values of the data to make them fall within a given range or distribution [22], [23].

Not all song features are in the equal value range. The variables that are used to measure a song's acousticity, danceability, energy, instrumentalness, liveness, speechiness, and valence are considered to be continuous ratio variables, all of which are rated on a scale of 0 to 1.

Another set of ratio variables that are quantified in their own units are duration (Ms), loudness (Db), and tempo (BPM).

Secondly, using the conventional pitch class nomenclature, our predictor key is measured as a (categorical) nominal variable with the keys ranging from 0 to 11. A third separate variable is called mode, which is dichotomous and has the values 0 for the minor mode and 1 for the major mode.

When qualities are measured using such disparate scales, comparisons might be challenging. That's why we used a feature scaling method to take all the features into a similar range[24].

1) Feature Scaling (Min-Max Scaling)

The min-max scalar method of normalization places all the data into a range between a specific min and max value using the mean and standard deviation.

We used the MinMaxScaler class from scikit-learn library in order to scale our data into a specific range. The fit

TABLE I. Music Dataset Features

Features	Range	Description
Acousticness	0.0-1.0	Acousticness is a feature that serves as a measure of confidence indicating whether a track is acoustic or not. A score of 1.0 signifies a high level of confidence that the track is acoustic.
Danceability	0.0-1.0	Spotify calculates a metric called ”Danceability” which gauges a track’s suitability for dancing by analyzing various musical elements.
Duration ms		This refers to the length of the track measured in milliseconds.
Energy	0.0-1.0	Spotify uses a scale called ”energy” to indicate the level of intensity and activity in a song. Songs that are more energetic and active tend to capture people’s attention, so popular songs likely have higher energy scores.
Instrumentalness	0.0-1.0	Spotify has a metric called ”instrumentalness” which determines whether a song has vocals or not. In this context, even sounds like ”ooh” and ”aah” are considered instrumental. Rap or spoken word tracks are considered ”vocal” songs.
Key	>=-1 and <=11	The key is represented by integers using the Pitch Class notation. If the system fails to detect the key, the value assigned is -1.
Liveness	0.0-1.0	”liveness” indicates the likelihood of the presence of a live audience in a recording. A score above 0.8 strongly suggests that the track is a live performance.
Loudness	-60 to 0 db	The metric ”loudness,” as calculated by Spotify, assesses the overall volume of a track in decibels (dB). Loudness is the characteristic of a sound that most directly relates to its physical strength (amplitude).
Mode	0-1	The metric ”mode,” indicates whether a track is in a major (1) or minor (0) key, which corresponds to the type of scale used for its melodic content.
Time signature	>=3 and <=7	The ”time signature” offers an approximation of the notational convention that specifies the number of beats present in each bar, also known as a measure.
Speechiness	0.0-1.0	”Speechiness” metric identifies whether a track contains spoken words. If a recording predominantly features spoken words, such as those in talk shows, audiobooks, or poetry, the metric value tends to be closer to 1.0. A value above 0.66 typically indicates that the track consists entirely of spoken words, while values between 0.33 and 0.66 suggest the presence of both speech and music, including genres such as rap. Conversely, values below 0.33 are more likely to represent tracks without spoken words, such as instrumental music.
Tempo	n BPM	The ”tempo” metric assesses the estimated pace of a track in beats per minute (BPM). Tempo is a musical term that refers to the speed or rate of a particular piece and is determined by the average duration of beats within it.
Valence	0.0-1.0	The ”valence” metric provided by Spotify reflects the degree of positivity conveyed by a particular track. Tracks with higher valence scores are generally perceived as more positive, happy, and euphoric, while those with lower scores are more likely to sound negative, sad, depressed, or angry.

method was used to compute the minimum and maximum values of the features in the dataset. The transform method was used to scale the features according to specific range. After Scaling the data, the value of our features are ranged between 0 and 1 [25].

E. Analyzing Popularity Between Two Regions

We measured the mean and median of Bangladeshi song popularity and West Bengal song popularity.

1) Mean Of The Popularity

Mean: The mean is the average of the values in the given set. It indicates that values in a particular data set are distributed equally.

$$\text{Fromula of Mean : } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We have generated a bar chart illustrating the average popularity comparison between Bangladeshi and West Ben-

gal songs.

Average popularity of Bangladeshi songs: 0.611457

Average popularity of West-Bengal songs: 0.457347

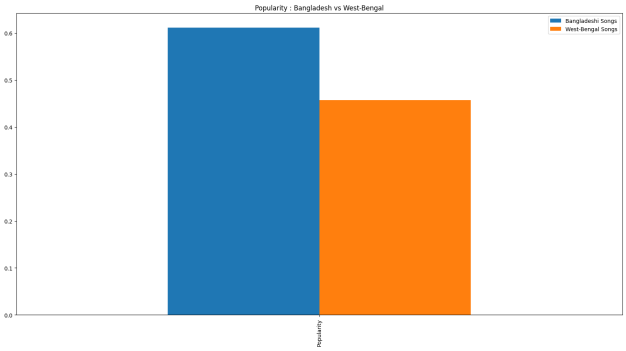

Figure 4. Average popularity comparison between Bangladesh and West-Bengal songs.

The calculated average distinctly favors Bangladeshi songs, indicating that, on the whole, Bangladeshi songs enjoy greater popularity compared to those from West Bengal.

2) Median Of The Popularity

Median: The middle number in an ordered sequence of numbers is called the median, and it can be more indicative of data collection than the mean.

if n is odd,

$$median = \left(\frac{n+1}{2}\right)^{th}$$

if n is even,

$$median = \frac{(\frac{n}{2})^{th} + (\frac{n}{2}+1)^{th}}{2}$$

n = number of terms
th = n(th) number

We have generated a box plot figure 5 to depict the difference in popularity median between Bangladeshi songs and West Bengal songs.

Popularity median of Bangladeshi songs: 0.5681

Popularity median of West Bengal songs: 0.4545

From this mathematical analysis, we can come to a conclusion that Bangladeshi rock songs are more popular than West Bengal rock songs.

F. Extracting Information From Scatter Plot

Now we know Bangladeshi Songs are more popular than West Bengal Songs. We now plot our features into a scatter plot in order to find any significant information that can lead us to find the reason why the popularity of songs differs between these two region. Here in figure 6 have some scatter plots with respect to the features.

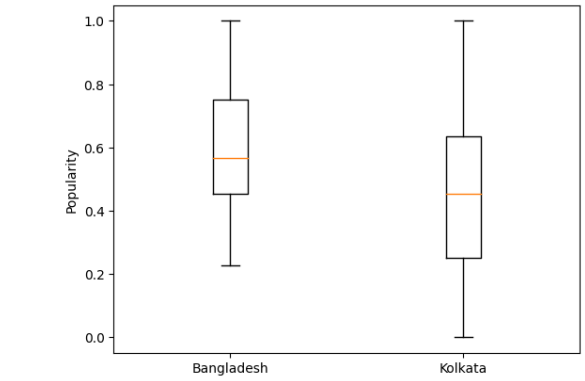

Figure 5. Median of popularity comparison between Bangladesh and West Bengal Songs.

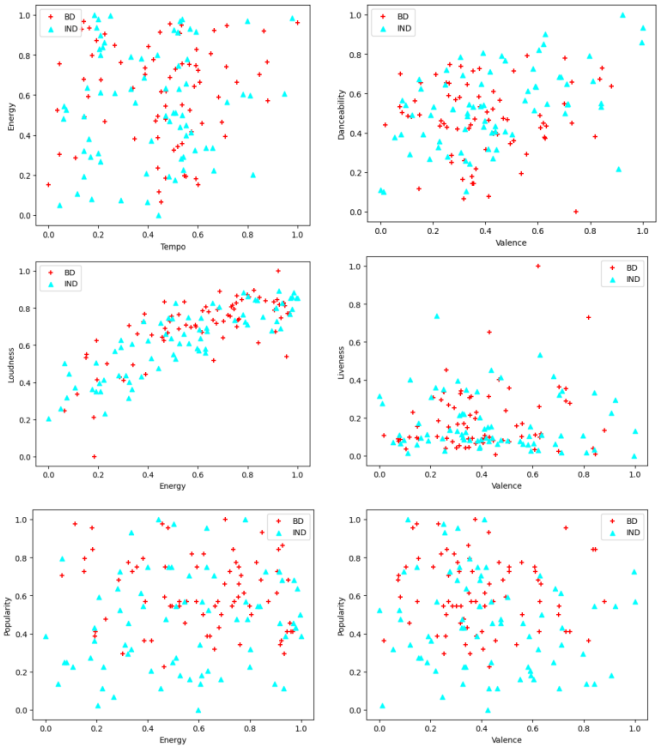

Figure 6. Scatter plots.

We have generated all the possible combinations of features through two-dimensional scatter plotting. From the graphical evidence, we didn't find any useful information that can lead us to determine why the popularity differs from Bangladesh to West Bengal.

G. Random Forest and Feature Importance using Mean Decrease Impurity (MDI)

Random Forest: Random Forest is a powerful ensemble learning technique used in both classification and regression

tasks. It is composed of multiple decision trees, each trained on a random subset of the data and using a random subset of features. This randomness helps to reduce overfitting and improve generalization performance. In classification tasks, the final prediction is determined by a majority vote from individual trees, while in regression tasks, it's typically the mean prediction of individual trees [26].

Feature Importance using MDI: Feature importance in Random Forest is often measured using Mean Decrease Impurity (MDI). MDI quantifies the importance of each feature by computing the average decrease in impurity (typically Gini impurity or entropy) caused by a feature when it is used for splitting across all trees in the forest. Features that result in large impurity decreases are considered more important, as they contribute significantly to the predictive power of the model.

Rationale for Choosing Random Forest Feature Importance Using MDI: Feature Importance with Mean Decrease Impurity (MDI) emerges as an apt choice for feature selection within our dataset, where the target variable denotes the country of origin (1 for Bangladesh, 0 for West Bengal). Its robustness in handling categorical target variables, coupled with its ability to analyze high-dimensional feature spaces, renders it particularly suitable. By constructing an ensemble of decision trees, each trained on a random subset of the data, Random Forest mitigates overfitting concerns and captures intricate relationships among features and the target variable. Moreover, the MDI metric provides a robust measure of feature importance, aiding in the identification of discriminative features that distinguish between the two countries.

Importance of Feature Selection: In the context of our research, the selection of relevant features is paramount for delineating the distinguishing characteristics between rock songs from Bangladesh and West Bengal. Effective feature selection not only reduces dimensionality and computational complexity but also enhances model interpretability by focusing on features that contribute significantly to the classification of songs based on country of origin. By excluding irrelevant or redundant features, feature selection facilitates the construction of parsimonious models while improving predictive performance and generalization capabilities.

Implementation of Random Forest: The implementation of Random Forest with MDI for feature importance and subsequent feature selection involved the following steps:

(a) Data Preprocessing: We used the dataset that we prepared for this study. We split the data into 75% and 25% for training and testing purposes.

(b) Random Forest Training: A Random Forest classifier was trained on the preprocessed data to compute feature importances using MDI [27].

(c) Feature Importance Ranking: Features were ranked in ascending order based on their MDI scores, and the top six features were selected for subsequent analysis.

Significance of Results from Random Forest Feature Importance using MDI: Features with higher MDI scores are indicative of their importance in distinguishing between the two countries. By prioritizing features based on their MDI scores, we gain a nuanced understanding of the salient musical attributes that contribute to the regional distinctions in rock music. These insights not only enrich our understanding of cultural preferences and musical landscapes but also have practical implications for music classification.

H. Support Vector Machine

After identifying important features through Random Forest feature importance using Mean Decrease Impurity (MDI), we proceeded to assess the predictive power of these features in distinguishing between rock songs from Bangladesh and West Bengal using a Support Vector Machine (SVM) model. We are using SVM classification to distinguish songs from Bangladesh and West Bengal using the features based on MDI score. If we achieve high accuracy, it means that these features are important for telling the differences between songs in Bangladesh and West Bengal. Then we can say, these features might be the reason why songs from Bangladesh are popular.

Rationale for Choosing Support Vector Machine : Support Vector Machine (SVM) is an optimal choice for our classification task due to its ability to handle non-linear relationships in high-dimensional data efficiently [28]. Its robustness to outliers ensures reliable predictions, while its flexibility in kernel selection allows adaptation to diverse data structures. Moreover, SVM's inherent design for binary classification aligns perfectly with our target variable representing two distinct classes (Bangladesh and West Bengal). Overall, SVM offers a powerful framework for analyzing the relationship between musical features and the popularity of rock songs from the two regions, enabling accurate predictions and insightful interpretations within a concise and efficient model.

Feature Selection: We selected the top six most important features identified through Random Forest feature importance analysis using MDI. These features were chosen based on their respective MDI scores, which quantified their importance in discriminating between the two countries.

SVM Model Training: We trained an SVM classifier using the selected important features as input variables and the country of origin (1 for Bangladesh, 0 for West Bengal) as the target variable. The SVM model was chosen for its ability to handle binary classification tasks and its flexibility in handling high-dimensional feature spaces.

Implementation of SVM : The implementation of Support Vector Machine involved the following steps:

(a) Feature Selection: We identified the top six influential features using Random Forest with Mean Decrease Impurity (MDI) analysis. These features were chosen based on their importance in distinguishing between rock songs from Bangladesh and West Bengal.

(b) Feature Extraction: Top six features were selected as the independent variables (X) for our classification task. These features represent key musical attributes such as tempo, energy, danceability, valence, instrumentalness, and acousticness.

(c) Target Variable: We defined the target variable (Y) as the country of origin, where 1 represents Bangladesh and 0 represents West Bengal. The target variable serves as the label for classification, indicating the country to which each rock song belongs.

(d) Dataset Formation: Constructed the dataset by pairing the selected six features (X) with their corresponding country of origin labels (Y). Each data instance comprises a feature vector representing the musical attributes of a rock song and its associated country of origin label.

(e) Train-Test Split: We Split our dataset into training and testing sets(70% for training and 30% for testing purposes) using a function like 'train_test_split' from scikit-learn. This helps evaluate the performance of the SVM model on unseen data.

(f) Selection of Kernel Method : We used Radial Basis Function (RBF) Kernel in our SVM. It is ideal for non-linear data with no prior knowledge of data distribution. It's versatile and widely used due to its ability to capture complex relationships.

(g) Model Training: Fit the SVM classifier to the training data using the fit method. This step involves learning the optimal decision boundary that separates the different classes in the feature space.

(I) Model Evaluation: Once trained, we evaluated the performance of the SVM classifier on the test data. We used evaluation metrics such as accuracy to assess the model's predictive performance.

I. Hypothesis Questions

Hypothesis: When an assumption is backed by evidence, it becomes a hypothesis. Hypotheses are used by researchers to determine relationships between variables and make predictions based on theoretical principles and empirical data. By using statistical tests, researchers can assess the evidence in support of the alternative and null hypotheses, which are two opposing claims.

Null Hypothesis (H0): States that there is no relationship between the two variables

Alternate Hypothesis (H1): States that there is a relationship between the two variables.

We formulated some hypotheses to determine any relation between the musical features and popularity.

1) Hypothesis 1

Null Hypothesis: There is no relationship between Acousticness and Popularity.
Alternate Hypothesis: There is a relationship between Acousticness and Popularity.

2) Hypothesis 2

Null Hypothesis: There is no relationship between Danceability and Popularity.
Alternate Hypothesis: There is a relationship between Danceability and Popularity.

3) Hypothesis 3

Null Hypothesis: There is no relationship between Duration and Popularity.
Alternate Hypothesis: There is a relationship between Duration and Popularity.

4) Hypothesis 4

Null Hypothesis: There is no relationship between Energy and Popularity.
Alternate Hypothesis: There is a relationship between Energy and Popularity.

5) Hypothesis 5

Null Hypothesis: There is no relationship between Instrumentalness and Popularity.
Alternate Hypothesis: There is a relationship between Instrumentalness and Popularity.

6) Hypothesis 6

Null Hypothesis: There is no relationship between Liveness and Popularity.
Alternate Hypothesis: There is a relationship between Liveness and Popularity.

7) Hypothesis 7

Null Hypothesis: There is no relationship between Loudness and Popularity.
Alternate Hypothesis: There is a relationship between Loudness and Popularity.

8) Hypothesis 8

Null Hypothesis: There is no relationship between Speechiness and Popularity.
Alternate Hypothesis: There is a relationship between Speechiness and Popularity.

9) Hypothesis 9

Null Hypothesis: There is no relationship between Tempo and Popularity.
Alternate Hypothesis: There is a relationship between Tempo and Popularity.

10) Hypothesis 10

Null Hypothesis: There is no relationship between Valence and Popularity.
Alternate Hypothesis: There is a relationship between Valence and Popularity.

J. Pearson Correlation Test

The Pearson correlation is a statistical measure that evaluates the linear relationship between two continuous variables. It is often used to measure the strength and direction of the relationship between two variables [29].

The Pearson correlation coefficient denoted as "r", ranges from -1 to 1, where -1 indicates a perfect negative correlation (as one variable increases, the other decreases), 0 indicates no correlation, and 1 indicates a perfect positive correlation (as one variable increases, the other one will also increase).

By assessing the r values for both the musical features and popularity of corresponding song, we can effectively evaluate the relationship between these elements. This exploration will provide valuable insights into how changes in certain musical attributes might influence the overall popularity of a music.

The Pearson correlation is widely used in data analysis and machine learning because it is easy to interpret and provides a quick summary of the relationship between two variables. It is especially useful when trying to determine if there is a linear relationship between two variables.

The Pearson correlation can be used to answer questions such as and many more:

- Is there a relationship between Energy and Popularity?
- Is there a relationship between Loudness and Popularity?
- Is there a relationship between Danceability and Popularity?

Pearson Correlation Formula :

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Where,
r = Pearson Correlation Coefficient
X_i = Y variable samples
Y_i = Y variable samples
$\bar{X}$ = mean of values in X variable
$\bar{Y}$ = mean of values in Y variable

Significance of Pearson Correlation Coefficient (r) value :

TABLE II. Significance of r value

Pearson correlation coefficient (r) value	Significance	Direction
Greater than 0.5	Strong Significance	Positive
Between 0.3 and 0.5	Moderate Significance	Positive
Between 0 and 0.3	Weak Significance	Positive
0	No Significance	None
Between 0 and –0.3	Weak Significance	Negative
Between –0.3 and –0.5	Moderate Significance	Negative
Less than –0.5	Strong Significance	Negative

We have used the Pearson Correlation method in order to find the significance and direction of two variables.

To investigate the relationship between various musical features collected from Spotify and the popularity of songs, we employed the pearsonr function from the scipy.stats module. This function enabled us to compute the Pearson correlation coefficient, a measure of linear correlation, between pairs of continuous variables [30].

K. Chi Square Test

The chi-square test (X2) is a statistical technique used to compare anticipated and actual results. Its objective is to determine if the difference between the observed and predicted data is due to chance or if it is connected to the variables being examined. This makes the chi-square test an effective tool for exploring and interpreting the relationship between the two variables[31].We used the Chi-Square Test to find evidence of our proposed hypotheses.

The formula of Chi-Square :

$$X^2 = \sum \frac{(O - E)^2}{E}$$

X^2 is the chi-square test statistic
$\sum$ is the summation operator
O is the observed frequency
E is the expected frequency

Calculating The P Value: To investigate the relationship between music features and song popularity, we conducted a chi-square test using data collected from Spotify. Our aim was to determine whether certain music features are associated with the popularity of songs. We used the chi2_contingency function from the scipy.stats module to calculate the p-value [32]. The significance of p-value is shown in table III. When the p-value is less than 0.05, the null hypothesis is rejected, indicating that there is a relationship between the two variables. If the p-value is more than 0.05, the data are not statistically significant and

our null hypothesis is not rejected.

TABLE III. Significance of P value

Significance Level	Specification
p > 0.05	Not Significant
p ≤ 0.05 (5%)	Significant
p ≤ 0.01 (1%)	Very Significant
p ≤ 0.001 (0.1%)	Highly Significant

4. RESULT

A. *Results From Random Forest Feature Importance Using MDI*

The Mean Decrease in Impurity(MDI) score from the Random Forest Feature Importance is shown in figure 7.

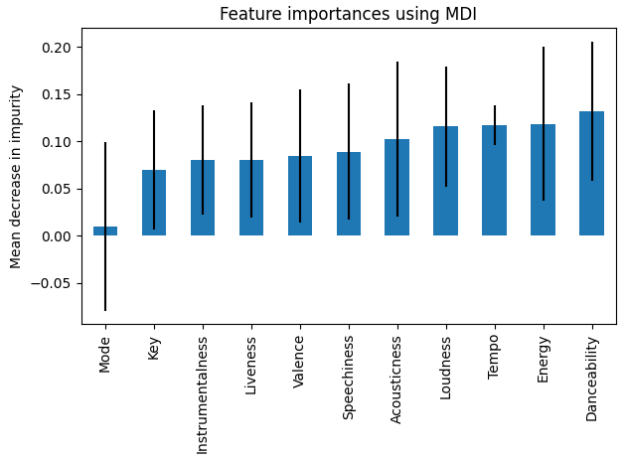

Figure 7. Result from random forest

The result is shown in ascending order based on the MDI score which reflects the relative importance of each feature in the classification process. Features with higher MDI scores are deemed more influential and contribute more substantially to the decision-making process within the Random Forest model. We used this data in the Support Vector Machine to find out the classification.

B. *Results from Support Vector Machine*

We used the top six features based on the MDI score from Random Forest Feature Importance analysis. The selected features are 'Tempo' ,'Acousticness', 'Speechiness', 'Loudness', 'Energy', 'Danceability'.

TABLE IV. SVM classification accuracy

Kernel	Accuracy
rbf	45.45%

Our SVM classification accuracy is 45.45%. If the accuracy is high, it suggests SVM can effectively classify rock songs from Bangladesh and West Bengal based on these features. So, these features could be the reason for the higher popularity of Bangladeshi songs. But here we got lower accuracy which is insufficient to assert that the selected features significantly distinguish between rock songs from Bangladesh and West Bengal. That is why we conducted further analysis to highlight additional factors that may better discern the differences in popularity between rock songs from Bangladesh and West Bengal.

C. *Hypotheses Testing*

To find out the reasons behind the higher popularity of Bangladeshi Rock Songs over West-Bengal Rock Songs, we first need to know whether our proposed hypotheses are accepted or rejected.

We used two methods to validate our proposed hypotheses.
1. Pearson Correlation Test.
2. Chi-Square Test.

1) *Results From Pearson Correlation Test*

To calculate the Pearson Correlation of each feature with respect to popularity, we generated a correlation heatmap figure 8, using Python Seaborn and matplotlib library.

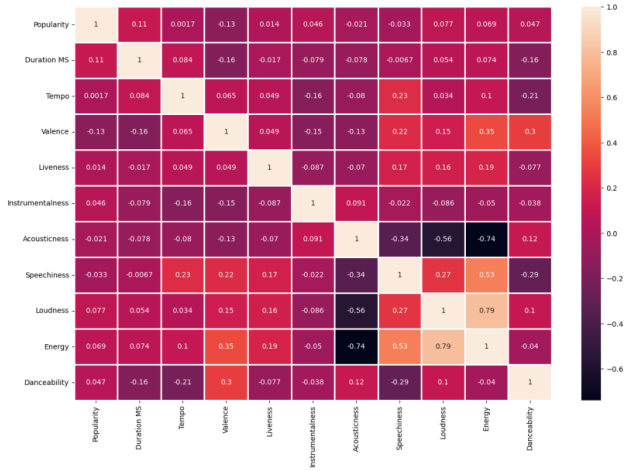

Figure 8. Pearson Correlation Heatmap

We have given a detailed verdict of our proposed hypotheses in the following. We have added a table V, that tells the correlation coefficient value of each feature with respect to popularity and depicts the significance and direction of the two features.

Here, we come to know that all the features are weakly correlated with the popularity feature.

2) *Results From Chi-Square Test*

Pearson Correlation only describes the linear relationship of a given dataset. The chi-square Test calculates the association relation of a given dataset. That's why we take the approach of the Chi-Square Test to find whether our

TABLE V. Pearson Correlation coefficient value of each feature with respect to popularity

Hypothesis No	Features With Respect To Popularity	Correlation Coefficient Value (r)	Significance	Direction
1	Acousticness	-0.021	Weak Significance	Negative
2	Danceability	0.047	Weak Significance	Positive
3	Duration	0.11	Weak Significance	Positive
4	Energy	0.069	Weak Significance	Positive
5	Instrumentalness	0.046	Weak Significance	Positive
6	Liveness	0.014	Weak Significance	Positive
7	Loudness	0.077	Weak Significance	Positive
8	Speechiness	-0.033	Weak Significance	Negative
9	Tempo	0.0017	Weak Significance	Positive
10	Valence	-0.13	Weak Significance	Negative

proposed hypotheses are acceptable or not. Here the table VI depicts the result.

TABLE VI. Chi Square Testing P value

Hypothesis No	Features With Respect To Popularity	Chi Square Testing P-Value (p)	Null Hypothesis
1	Acousticness	0.268	Retain
2	Danceability	0.1287	Retain
3	Duration	0.131	Retain
4	Energy	0.037	Reject
5	Instrumentalness	0.1312	Retain
6	Liveness	0.455	Retain
7	Loudness	0.133	Retain
8	Speechiness	0.149	Retain
9	Tempo	0.131	Retain
10	Valence	0.107	Retain

Hypothesis 1 :
Null Hypothesis: There is no relationship between Acousticness and Popularity.
Alternate Hypothesis: There is a relationship between Acousticness and Popularity.
We found the P-value, $p = 0.268$, So the evidence says we retain our null hypothesis.

Hypothesis 2 :
Null Hypothesis: There is no relationship between Danceability and Popularity.
Alternate Hypothesis: There is a relationship between Danceability and Popularity.
We found the P-value, $p = 0.1287$, So the evidence says we retain our null hypothesis.

Hypothesis 3 :
Null Hypothesis: There is no relationship between Duration and Popularity.
Alternate Hypothesis: There is a relationship between Duration and Popularity.
We found the P-value, $p = 0.131$, So the evidence says we retain our null hypothesis.

Hypothesis 4 :
Null Hypothesis: There is no relationship between Energy and Popularity.
Alternate Hypothesis: There is a relationship between Energy and Popularity.
We found the P-value, $p = 0.037$, So we reject our null hypothesis.

Hypothesis 5 :
Null Hypothesis: There is no relationship between Instrumentalness and Popularity.
Alternate Hypothesis: There is a relationship between Instrumentalness and Popularity.
We found the P-value, $p = 0.1312$, So the evidence says we retain our null hypothesis.

Hypothesis 6 :
Null Hypothesis: There is no relationship between Liveness and Popularity.
Alternate Hypothesis: There is a relationship between Liveness and Popularity.
We found the P-value, $p = 0.455$, So the evidence says we retain our null hypothesis.

Hypothesis 7 :
Null Hypothesis: There is no relationship between Loudness and Popularity.
Alternate Hypothesis: There is a relationship between Loudness and Popularity.
We found the P-value, $p = 0.133$, So the evidence says we retain our null hypothesis.

Hypothesis 8 :
Null Hypothesis: There is no relationship between Speechiness and Popularity.

Alternate Hypothesis: There is a relationship between Speechiness and Popularity.
We found the P-value, $p = 0.149$, So the evidence says we retain our null hypothesis.

Hypothesis 9 :
Null Hypothesis: There is no relationship between Tempo and Popularity.
Alternate Hypothesis: There is a relationship between Tempo and Popularity.
We found the P-value, $p = 0.131$, So the evidence says we retain our null hypothesis.

Hypothesis 10 :
Null Hypothesis: There is no relationship between Valence and Popularity.
Alternate Hypothesis: There is a relationship between Valence and Popularity.
We found the P-value, $p = 0.107$, So the evidence says we retain our null hypothesis.

From the Chi-Square Test, we only found a relationship between "Popularity" and "Energy". Since the Chi-Square Testing P-value is not limited to only linear relationships but also calculates important data relations like "association", we only retain/reject our proposed hypotheses based on the Chi-Square Test P-value. So, our null hypothesis, "There is no relationship between Energy and Popularity" is rejected, meaning there is a relationship between "Energy and Popularity".

We discovered that music energy could be the reason for the popularity of Bangladeshi rock music, indicating a relationship with popularity. Song energy emerges as a key feature and might be playing a crucial role in the higher popularity of Bangladeshi rock music.

5. CONCLUSION AND FUTURE WORK

We aim to provide deeper insights into the implications and significance of our findings regarding the differences in song popularity between the two regions. By identifying the features that play a major role in these differences, our study offers valuable insights for the music industry. Music industry professionals can utilize our research to tailor their strategies and can make songs more popular among specific regions.

In this study, two questions motivated us to pursue the research. The first question was, "Is there any difference in popularity between Bangladeshi rock songs and West Bengal rock songs?" If yes, then which factors are responsible for making a song popular? And the last question was, "Is there any way to show the differences in popularity using musical features?". We had undertaken a statistical approach to solve these questions. We measured the mean and mode of the popularity of Bangladeshi Rock Songs and West Bengal Rock Songs to find out the answer to the first question.

TABLE VII. Bangladesh and West-Bengal song's popularity

	Bangladesh	WestBengal
Mean Value	0.611457	0.457347
Median Value	0.5681	0.4545

From the table VII, we can come to the conclusion that Bangladeshi songs are more popular than West Bengal songs.

As for our second question, we first extracted the musical features provided by Spotify. Then we scaled our features into a similar range which is [0,1].

We collected the top six features from Random Forest Feature Importance analysis and used them in the Support Vector Machine(SVM). The purpose of using SVM is to classify the songs of Bangladesh and West Bengal. The higher accuracy of SVM should indicate that the collected features based on MDI score are sufficient enough to conclude the differences between Bangladesh and West Bengal Songs, and these features might be the reason for the higher popularity of Bangladesh rock songs. The accuracy result we found from our SVM classification (45.45%) is too low to come to a conclusion. That is why we further continued our study with the Pearson Correlation Test and Chi-Square Test by proposing hypotheses.

We proposed 10 hypotheses. In order to find significant evidence for our proposed hypotheses, we pursued two types of hypothesis testing.

1. Pearson Correlation
2. Chi-Square Test

For our first, second, and third hypotheses: we couldn't find any evidence of a relationship from the Chi Square hypothesis testing. Hence we retain our null hypothesis.

For our fourth hypothesis, we found a satisfying p-value that was enough to show evidence of the relationship between Energy and Popularity features. So, we reject our null hypothesis. From the Pearson Correlation test, we came to know that Popularity and Energy have a positive relationship, meaning higher Energy will most likely lead to higher Popularity.

For our fifth, sixth, seventh, eighth, ninth, and tenth hypotheses, we couldn't find any evidence of a relationship from the Chi-Square hypothesis testing. Hence we retain our null hypothesis.

So, we found only one relationship which is "Energy and Popularity is positively related".

Our main objective in this study was to analyze the reasons behind popularity differences in two different regions using the audio features. After conducting this study we

have come to conclusion :

- We have not found a sufficient amount of features that have a significant relationship with song popularity. However, we found a relationship between energy and popularity, and energy likely leads to higher popularity of Bangladeshi rock music. However, it is almost impossible to precisely differentiate two different regional songs with just one feature which is energy.
- Bangladesh has a significantly larger number of rock genre bands than West Bengal and Bangladesh releases rock songs more frequently than West Bengal. This might be one of the reasons why Bangladeshi rock songs are more popular.
- Another important thing is that, unlike Bangladesh, the national language of India is not Bangla. So naturally Bangla language is practiced more in Bangladesh. This might be a reason for the musical popularity differences.
- India has her own musical heritage which is one of the richest in the world. Indian musical culture mostly revolves around sub-continental classical songs.
- The popularity of local music can also be influenced by how easily accessible it is.

In our study, we acknowledge several limitations that may affect the depth and scope of our findings. Firstly, we relied solely on the features provided by Spotify, which may not encompass all relevant musical characteristics influencing song popularity differences between regions. Additionally, our dataset may not be sufficiently large to capture the full spectrum of variations in song popularity. Moreover, we did not consider lyrical content in our analysis, which could provide valuable insights into regional preferences and cultural influences. These limitations highlight opportunities for future research to explore additional musical features and expand the dataset size to gain a more comprehensive understanding of regional differences in song popularity.

As we couldn't find enough relationship between Spotify song features and song popularity, which is evident enough to state that the higher popularity of Bangladeshi rock songs is a psychological and cultural matter. Nevertheless, our future work could include the following tasks:

- In future work, we plan to use a larger dataset.
- This research does not encompass lyrical content. Future iterations of this research will incorporate lyrical information.
- We will continue to identify additional features related to music.

References

[1] J. Lee, J. Park, J. Nam, and J. Park, "Cross-cultural transfer learning using sample-level deep convolutional neural netowrks," 2017.

[2] B. Aarden and D. Huron, "Mapping european folksong: Geographical localization of musical features," 08 2001.

[3] S. Brown, P. E. Savage, A. M.-S. Ko, M. Stoneking, Y.-C. Ko, J.-H. Loo, and J. A. Trejaut, "Correlations in the population structure of music, genes and language," *Proc Biol Sci*, vol. 281, no. 1774, p. 20132072, Nov. 2013.

[4] P. E. Savage, "Cultural evolution of music," *Palgrave Communications*, vol. 5, pp. 1–12, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:60442180

[5] S. Brown and J. Jordania, "Universals in the world's musics," *Psychology of Music*, vol. 41, no. 2, pp. 229–248, 2013. [Online]. Available: https://doi.org/10.1177/0305735611425896

[6] P. E. Savage and S. Brown, "Mapping music: Cluster analysis of song-type frequencies within and between cultures," *Ethnomusicology*, vol. 58, no. 1, pp. 133–155, 2014. [Online]. Available: https://www.jstor.org/stable/10.5406/ethnomusicology.58.1.0133

[7] L. Rego, "A cross-cultural comparison of song lyrics using nlp techniques," 2020.

[8] K. Liew, Y. Uchida, H. Domae, and A. H. Q. Koh, "Energetic music is used for anger downregulation: A cross-cultural differentiation of intensity from rhythmic arousal," 2022.

[9] L. Xu, M. Xu, Z. Jiang, X. Wen, Y. Liu, Z. Sun, H. Li, and X. Qian, "How have music emotions been described in google books? historical trends and corpus differences," *Humanities and Social Sciences Communications*, vol. 10, pp. 1–11, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259213790

[10] M. Mia, P. Das, and A. Habib, "Verse-based emotion analysis of bengali music from lyrics using machine learning and neural network classifiers," *Int. J. Comput. Digit. Syst.*, vol. 15, no. 1, pp. 359–370, Jan. 2024.

[11] A. Hu, X. Lee, J. Choi, and K. Downie, "Title a cross-cultural study of mood in k-pop songs," 2014.

[12] S. C. Izen, R. Y. Cassano-Coleman, and E. A. Piazza, "Music as a window into real-world communication," *Frontiers in Psychology*, vol. 14, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259883943

[13] P. E. Savage, E. Merritt, T. Rzeszutek, and S. Brown, "Cantocore: A new cross-cultural song classification scheme," 2012.

[14] P. E. Savage, "Alan lomax's cantometrics project: A comprehensive review," *Music & Science*, vol. 1, p. 2059204318786084, 2018. [Online]. Available: https://doi.org/10.1177/2059204318786084

[15] X. Cai and H. Zhang, "Music genre classification based on auditory image, spectral and acoustic features," *Multimedia Systems*, vol. 28, pp. 779 – 791, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248724856

[16] W. Seo, S.-H. Cho, P. Teisseyre, and J. Lee, "A short survey and comparison of cnn-based music genre classification using

multiple spectral features," *IEEE Access*, vol. 12, pp. 245–257, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266566348

[17] *Mobile Web Player.* [Online]. Available: https://open.spotify.com/

[18] *Node JS*. [Online]. Available: https://nodejs.org/en/docs

[19] "Spotify for developers." [Online]. Available: https://developer.spotify.com/documentation/web-api/#spotify-uris-and-ids

[20] "Spotify search." [Online]. Available: https://developer.spotify.com/documentation/web-api/reference/get-track

[21] "Spotify song features." [Online]. Available: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

[22] "Why scaling is important in machine learning?" [Online]. Available: https://medium.com/analytics-vidhya/why-scaling-is-important-in-machine-learning-aee5781d161a

[23] "Feature scaling." [Online]. Available: https://en.wikipedia.org/wiki/Feature_scaling

[24] "Guide to data cleaning: Definition, benefits, components, and how to clean your data." [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning

[25] "Minmaxscaler." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[26] "Feature importances with a forest of trees." [Online]. Available: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html?fbclid=IwAR2qiUcBTJOQDiltUgty1gHvrZly9oCMT8Bg8aer8HWKvBf0ihPLot

[27] "Randomforestclassifier." [Online]. Available: https://scikit-learn.org/0.15/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[28] "svm." [Online]. Available: https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,than%20the%20number%20of%20samples

[29] "Pearson correlation coefficient." [Online]. Available: https://www.scribbr.com/statistics/pearson-correlation-coefficient/

[30] "pearsonr." [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

[31] *Chi-squared test*. [Online]. Available: https://en.wikipedia.org/wiki/Chi-squared-test

[32] "scipy.stats." [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

Moshiur Rahman Autul graduated from Shahjalal University of Science and Technology in 2023 with B.Sc.(Eng.) in Software Engineering major. He is currently working as a Lecturer in the Department of Software Engineering at Metropolitan University, Sylhet. His research interest comprises the study of Machine Learning (ML), Software Engineering, and Human-Computer Interaction.

Durjoy Dey is currently working as an Associate Software Engineer at Cefalo Bangladesh Ltd. He graduated from Shahjalal University of Science and Technology in 2023 with B.Sc.(Eng.) in Software Engineering major. His research interests include Human-Computer Interaction, Machine Learning (ML), and Software Engineering.

Partha Protim Paul is actively working as a Lecturer at the Institute of Information and Communication Technology of Shahjalal University of Science and Technology (SUST). Before joining SUST he worked as a software engineer at Orbitax. His research interests include Software Engineering, Automated Program Repair, Human-Computer Interaction, and Software Testing.

Mohammed Raihan Ullah is working as a lecturer at the Institute of Information and Communication Technology of Shahjalal University of Science and Technology (SUST). He also worked in Retune as a Software Engineer. His research interest includes Software Testing, Machine learning, and Artificial Intelligence.