

BEAT: A Feature-Enhanced and Balanced Prediction Model for Music Popularity

Jiangbei Wang^{1,a,#}, Yuxiang Fei^{1,b,#}, Yifu Zhao^{1,c,#}, Wenbo Wang^{1,d,#},
Wenyi Xiao^{1,e,#}

¹Beijing 21st Century School, Beijing, China

^aloganwang926@gmail.com, ^bfjx1003pidan@163.com, ^cYveszhao2011@163.com,

^d15600697505@163.com, ^e15810190630xwy@sina.com

Abstract: With the rapid development of streaming media, the forecast of the music popularity has practical value for platform optimization and recommendation as well as for creators to grasp market tendency. Existing studies generally train models with limited features and overlook the inherent class imbalance and feature scarcity of this task, which tends to produce undesirably low Precision or Recall. To solve the problems above, we proposed BEAT model. Considering the influences of artists and music genres on music popularity, we newly increased corresponding aggregated features. To tackle class imbalance, higher weights are assigned to minority class samples, and grid search is employed to determine optimal hyper parameter values. We conduct comparative experiments against eight machine learning models including SVM, Random Forest, and K-Nearest Neighbors. Experimental results demonstrate that our model achieves state of the art performance in terms of Accuracy, AUC and F1-score, reaching 94.99%, 97.39% and 79.15% respectively. It also obtains competitive Precision and Recall of 78.50% and 79.81%. These results verify that our model delivers favorable performance for music popularity prediction.

Keywords: Music Popularity Prediction, Machine Learning, Class Imbalance, Feature Engineering

1. Introduction

With the proliferation of music streaming platforms, the way of consuming music has been transformed from searching subjectively to being recommended by algorithms[1]. Such transformation makes the forecast of popularity become the key problem that platforms and creators pay attention to collectively. Platforms need it to optimize the strategy of recommendation, for enhancing the user retention while creators need to grasp the trend of aesthetic to enlarge the influence of their work. Hence, music popularity-classification research bears considerable practical value.

Existing studies adopt machine learning models to perform classification and prediction based on features such as audio information. However, existing methods ignore the impacts of artist information and song genres on music popularity, resulting in unsatisfactory classification capability. Furthermore, the popularity classification task suffers from severe class imbalance, i.e., the number of popular songs is much less than the number of unpopular ones. Consequently, models yield quite low Precision or Recall.

To solve those problems, this paper proposed a brand-new model called **Balanced and Enhanced Aggregated Trees (BEAT)**. This model, based on the LightGBM, achieved a combination improvement regarding feature engineering and imbalanced sample weight setting. From the feature engineering perspective, noticing that average popularity of artists and average popularity of song styles have significant impacts on music popularity, we added these two aggregated features to the model, thus enhancing the accuracy in model prediction. For the class-imbalance problem, we adjusted the weight of samples of high music popularity so that the model can focus on data with fewer samples, which improves the model's performance while predicting samples in minority and, to a great extent, alleviates the issue that during training, the model tends to prioritize unpopular music samples. The experiments demonstrate that BEAT substantially outperforms baseline models on core metrics including Accuracy, F1-score and AUC, and achieves competitive results in terms of Precision and Recall.

The contributions of this study are summarized as follows. First, BEAT proposed by this research mixes the characteristics from many dimensions, from audio feature, text feature to interaction feature,

23 dimensions in total, making an effective fusion of a variety of information, thus improving the accuracy of prediction obviously. Second, baseline models can only attain competitive performance in either Precision or Recall, while the proposed model effectively balances and yields high values for both metrics, with Precision of 78.50% and Recall of 79.81%, demonstrating its superiority. Third, since the proposed model exhibits prominent effectiveness in identifying whether a song will gain popularity, it can serve as a practical analytical tool for music streaming platforms and music creators.

The remainder of this paper is organized as follows. Chapter 2 reviews the related work. Then, dataset and the proposed model are described in Chapter 3. Chapter 4 elaborates on comparative experiments and Chapter 5 concludes this paper.

2. Related Work

In music popularity prediction, researchers adopt various algorithms and features in existing studies. Sangeetha et al. [2] employed Random Forest, K-Nearest Neighbors, and Decision Tree classifiers to predict music popularity on the Spotify platform. Experimental results demonstrate that Random Forest achieves the highest prediction accuracy and outperforms other competitors by a considerable margin. Kamal et al. [3] compared six classification algorithms including Random Forest and Support Vector Machines for song-popularity prediction. Their work is built upon song metadata and performed innovative analysis of swear words and emotional cues within lyrics, enriching the feature set for prediction tasks. Soundharya et al. [4] constructed a music-popularity prediction model based on XGBoost and LightGBM. They incorporated Optuna-driven hyperparameter optimization and cross-validation to enhance model generalization capability. Furthermore, they carried out a systematic feature-importance analysis to identify key musical attributes that determine a song's appeal. Dong et al. [5] utilized Ridge Regression, K-Nearest Neighbors and multi-layer neural networks. Their study mainly investigated the correlation between acoustic physical attributes of songs and their popularity. The experimental results suggest no significant correlation between the selected physical features and music popularity, implying that these attributes alone provide limited practical guidance for song composition. Vanita Jain et al. [6] adopted four machine-learning models including Support Vector Machine and Random Forest for music-popularity prediction with a dataset of over 40000 songs. This work applied Principal Component Analysis to reduce audio features from 15 dimensions to 13 dimensions. Support Vector Machine yielded the optimal performance on the full dataset, whereas Random Forest achieved the best results for songs released in the recent 30 years. Tama Shil et al. [7] constructed a dataset with diverse audio-feature characteristics across different genres via the Spotify Web API. They deployed seven models such as Logistic Regression and Random Forest for prediction. Experimental results indicate that XGBoost and Support Vector Machines obtain satisfactory performance.

Nevertheless, all the above six studies have certain limitations. They mainly rely on audio features while ignoring other factors such as artist-related attributes and musical mode, which restricts their capacity to capture the underlying complex patterns. Moreover, none of them systematically discuss the class-imbalance problem and its influence on model performance, leading to deviations from real-world application requirements.

3. Our Method

In this study, we used a publicly available Spotify dataset comprising 114,000 valid samples. The original dataset includes basic audio and metadata features for each song, such as artist, album name, track name, popularity, duration_ms, explicit, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time signature, and track genre. We formulate popularity prediction as a binary classification task by establishing a threshold of 60, classifying songs as "popular" if their predicted popularity exceeds 60, and "unpopular" otherwise. The distribution of sample labels is illustrated in Figure 1, which highlights a substantial class imbalance between the popular and unpopular categories.

At the feature engineering level, we designed a multi-dimensional feature enhancement scheme to overcome the limitation of relying solely on the original acoustic features. In particular, we introduced two extended features—artist average popularity (AAP) and genre average popularity (GAP). These features endow our model with a form of prior inference. For example, if the model computes that a singer's past albums have an average popularity of 85, then when confronted with a new song by that singer, the model can infer that the new song's popularity is likely in the high range, rather than making

a blind guess based purely on audio signals. In addition, these two features capture a typical Matthew effect in the music industry that the market performance of a song depends largely on creator identity, genre choice and other prerequisite conditions, rather than solely on the song's quality and acoustic characteristics. With our method, the effects of star power and genre selection can be quantified as numerical proxy variables.

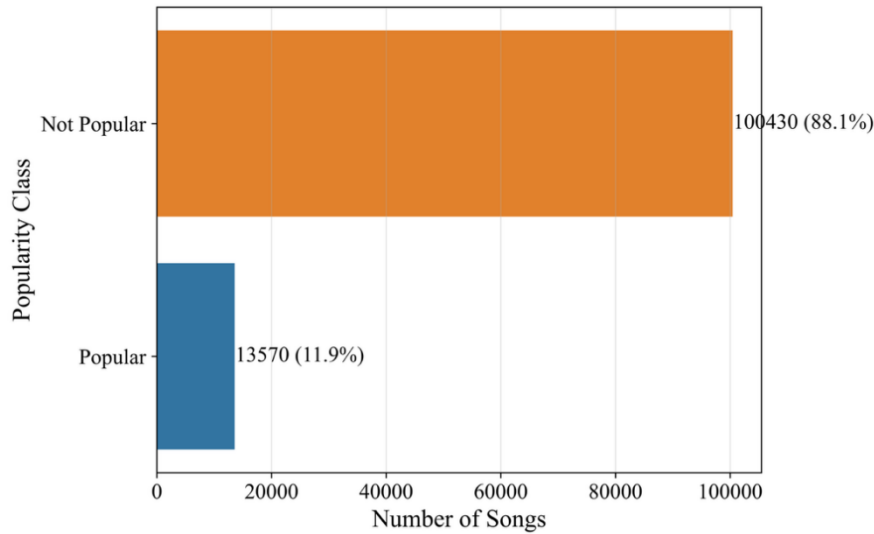


Figure 1. The distribution of sample labels.

For unstructured text data—song names, album names, and artist names—all exist in natural language form in the original data and cannot be directly fed into the model. Therefore, we adopted a technical pipeline combining Term Frequency-Inverse Document Frequency (TF-IDF) vectorization with Truncated Singular Value Decomposition (TruncatedSVD) to convert these texts into numerical vectors. TF-IDF vectorization splits each text into individual words, counts the term frequency of each word within a single document, and multiplies it by the inverse document frequency of that word across the entire corpus. The more frequently a word appears in a specific document but rarely in the overall corpus, the higher its discriminative value for that document. After TF-IDF processing, each song title is transformed into a high-dimensional sparse vector (typically exceeding 10,000 dimensions), where the number of dimensions equals the total number of distinct words in the corpus.

However, high-dimensional sparse vectors present two problems. First, excessive dimensions lead to low computational efficiency. Second, correlations among a large number of words cause information redundancy and interfere with model learning. To address this, we introduced TruncatedSVD for dimensionality reduction. SVD compresses the representation from one column per word to one column per latent topic. Taking album titles as an example, SVD may merge and condense highly co-occurring words such as "Remix", "Deluxe", and "Edition" into a positive loading on the third latent dimension. That is, a high value for the feature `album_svd_3` effectively indicates that the album title has a semantic tendency toward "deluxe/re-release".

At the algorithmic level, BEAT uses LightGBM as the base learner. After splitting the dataset into positive samples, i.e., popular songs, and negative samples, i.e., unpopular songs, we found that the proportion of popular songs is far lower than that of unpopular ones, at a ratio of approximately 1:3. If this imbalance is not addressed, prediction model tends to classify all samples as negative to achieve a higher overall accuracy, but at the cost of an extremely low Recall rate—missing a large number of popular songs. This is also why the accuracy of baseline is relatively low. To rectify this, BEAT pays more attention to samples that are incorrectly predicted as negative by increasing the penalty for misclassification. Specifically, the loss function for a single sample is defined as:

$$L_{weighted} = -\alpha y \log(p) - (1-y) \log(1-p), \quad (1)$$

and the corresponding weighted total loss is

$$L_{total}^{weighted} = \frac{1}{N} \sum_{i=1}^N [-\alpha y_i \log(p_i) - (1-y_i) \log(1-p_i)], \quad (2)$$

where α denotes the penalty weight coefficient for the positive class, calculated as $\frac{N_{neg}}{N_{pos}}$, with N_{neg} and N_{pos} representing the total counts of negative and positive samples, respectively. y represents the ground-truth label, and p denotes the predicted probability of the song being popular. This mechanism forces the model to focus more on positive samples during training.

Additionally, the original data was split into a training set and a test set at a 7:3 ratio. The test set is fully held out and used only for final evaluation. From the training set, an additional 20% is further separated as a validation set. During model training, we enabled the Early Stopping callback—if the validation AUC does not improve after 100 consecutive iterations, training is terminated early to prevent overfitting on the training set. The final number of boosting rounds is dynamically determined by the validation performance rather than being fixed in advance, allowing the model complexity to adaptively match the data scale.

4. Experiment

4.1 Dataset Preprocessing

Based on the original dataset, our study further constructs eight extended features to improve the model's interpretability for music-popularity trends. These features include artists' historical commercial appeal, baseline audience acceptance for different music genres, sample proportions of each genre, basic length and structural features of titles, special version identifiers for songs, attributes of fine-grained scene-specific word groups, the string length of artist names, and spelling rules or special-character patterns in artist names.

Music popularity is binarized according to the popularity metric. A song is labeled as 1 when its popularity value exceeds 60, and 0 vice versa. For model-performance evaluation, we split the complete dataset into a training set and a test set following a 7:3 ratio. Stratified random sampling is adopted to guarantee that the proportion of popular and unpopular samples in each subset remains consistent with the original dataset, mitigating the impact of sampling bias on model training and evaluation.

4.2 Evaluation Metrics and Baseline

This study employs five evaluation indicators, namely Accuracy, AUC, F1-score, Precision, and Recall, to comprehensively assess the overall performance of the model.

- **Accuracy** calculates the percentage of correctly predicted samples out of all testing data, i.e., $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.
- **AUC** quantifies the probability that the model assigns a higher predictive probability to a randomly selected positive sample than a randomly selected negative sample, corresponding to the area under the ROC curve.
- **Precision** measures the proportion of genuinely popular samples among all cases predicted as popular, i.e., $Precision = \frac{TP}{TP+FP}$, demonstrating how trustworthy the model's positive predictions are.
- **Recall** describes the proportion of real popular songs successfully retrieved by the model among all true popular samples in the dataset, i.e., $Recall = \frac{TP}{TP+FN}$, evaluating the model's capability to identify target popular items.
- **F1-score** is the harmonic mean of Precision and Recall, i.e., $F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$. A value closer to 1 signifies better overall performance across both precision and recall dimensions. Given the severe class imbalance between popular and unpopular song samples in this experiment, the F1-score is used to balance the two conflicting metrics and serves as the core optimization objective.

In our experiments, we adopt 8 classic machine-learning algorithms as baseline models: Decision Tree, Random Forest, XGBoost, LightGBM, Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes.

A controlled setting is employed to guarantee fair comparison. First, we keep the dataset and feature configuration consistent across all models. Specifically, all baselines share identical training, test, and

validation sets. Baseline models only utilize audio acoustic features, text dimensionality-reduced features and manually constructed interaction features, excluding the AAP and GAP features proposed in this paper. Second, for non-tree-based models including Logistic Regression, K-Nearest Neighbors and SVM, we use the default hyper-parameters rather than conducting grid-search-based hyper-parameter tuning. Third, for tree-based models such as Decision Tree, Random Forest and XGBoost, only the maximum tree depth is lightly optimized, while other hyper-parameters remain unchanged.

4.3 Classification Results and Analysis

Table 1: Classification results.

Model	Accuracy	AUC	F1-score	Precision	Recall
Decision Tree	0.8685	0.7155	0.5015	0.4944	0.5088
Random Forest	0.8997	0.9140	0.4164	0.8545	0.2753
XGBOOST	0.6780	0.5206	0.1744	0.3617	0.1149
LightGBM	0.6990	0.5363	0.0131	0.2222	0.0068
Logistic Regression	0.8745	0.8294	0.1910	0.5897	0.1140
K-Nearest Neighbors	0.8725	0.7755	0.3566	0.5189	0.2717
Support Vector Machine	0.8774	0.8036	0.2080	0.6498	0.1239
Naive Bayes	0.5314	0.7262	0.3496	0.2133	0.9686
BEAT(Ours)	0.9499	0.9739	0.7915	0.7850	0.7981

As shown in Table 1, our model demonstrates higher advantages across every indicator. In particular, the Accuracy, AUC and F1-score reach the highest levels among all models, at 94.99%, 97.39%, 79.15% respectively, showing significant improvement over other models. The extremely high Accuracy brings significant practical value, meaning that nearly 95 out of 100 songs can be correctly classified. The gap between our model and the second-best model is greater than 20%. The AUC far exceeds that of other models, indicating that our model has a strong ability to distinguish different samples and to produce highly reliable predictions.

The F1-score surpasses that of the second-best model, Decision Tree, by 29%, reflecting a comprehensive trade-off between Precision and Recall. This also suggests that our model achieves an optimal balance between the two metrics, avoiding the issue seen in Random Forest and Naive Bayes, where one metric is excessively high while the other is relatively low. Both the precision and recall of our model remain at high levels, at 78.50% and 79.80% respectively. This not only reduces misjudgments but also minimizes missed detections.

In conclusion, our model performs best in this task. Across all evaluation dimensions, our model achieves the highest values, and the balance between Recall and Precision is the most consistent, demonstrating a clear advantage over other models.

4.4 Feature Set Analysis

Based on the verification of model prediction accuracy, this section further analyzes the results by examining the feature importance scores generated by LightGBM, in order to understand why these features are significant and how they influence the predictions.

As shown in Figure 2, the top ten most important features output by the LightGBM model are as follows: AAP, GAP, genre_train_frequency (frequency of genres in the training set), album_svd_3 (third latent semantic dimension of album names), artist_char_length (length of artist names), album_svd_7 (seventh latent semantic dimension of album names), album_svd_0 (first latent semantic dimension of album names), instrumentality, artist_svd_4 (fourth latent semantic dimension of artist names), and liveness. Among them, the total importance of AAP and GAP exceeds 50%, which clearly indicates that the newly constructed aggregate statistical extended features in this study play a crucial role in the prediction framework. The strong correlation between the AAP and popularity suggests that the historical commercial appeal of top artists exhibits a significant inertial continuity effect. Even when new song snippets are not widely known, these artists still enjoy high levels of expected market interest. This perfectly aligns with the "superstar effect" commonly observed in the music industry, where a small number of top artists dominate most of the market attention and consumption traffic. The strong correlation between the GAP and popularity confirms the crucial role of genre selection in business decisions. There are systematic differences across musical genres in terms of listener base size, usage scenarios, and algorithm-based recommendation weights. Choosing the appropriate genre is itself an

important market strategy. Meanwhile, as representatives of original audio characteristics, instrumentality and liveness play a significant role in the "hard threshold filtering" effect—that is, when instrumentality is too high or liveness is too strong, the likelihood of a song becoming popular decreases—rather than contributing to any continuous improvement in popularity scores.

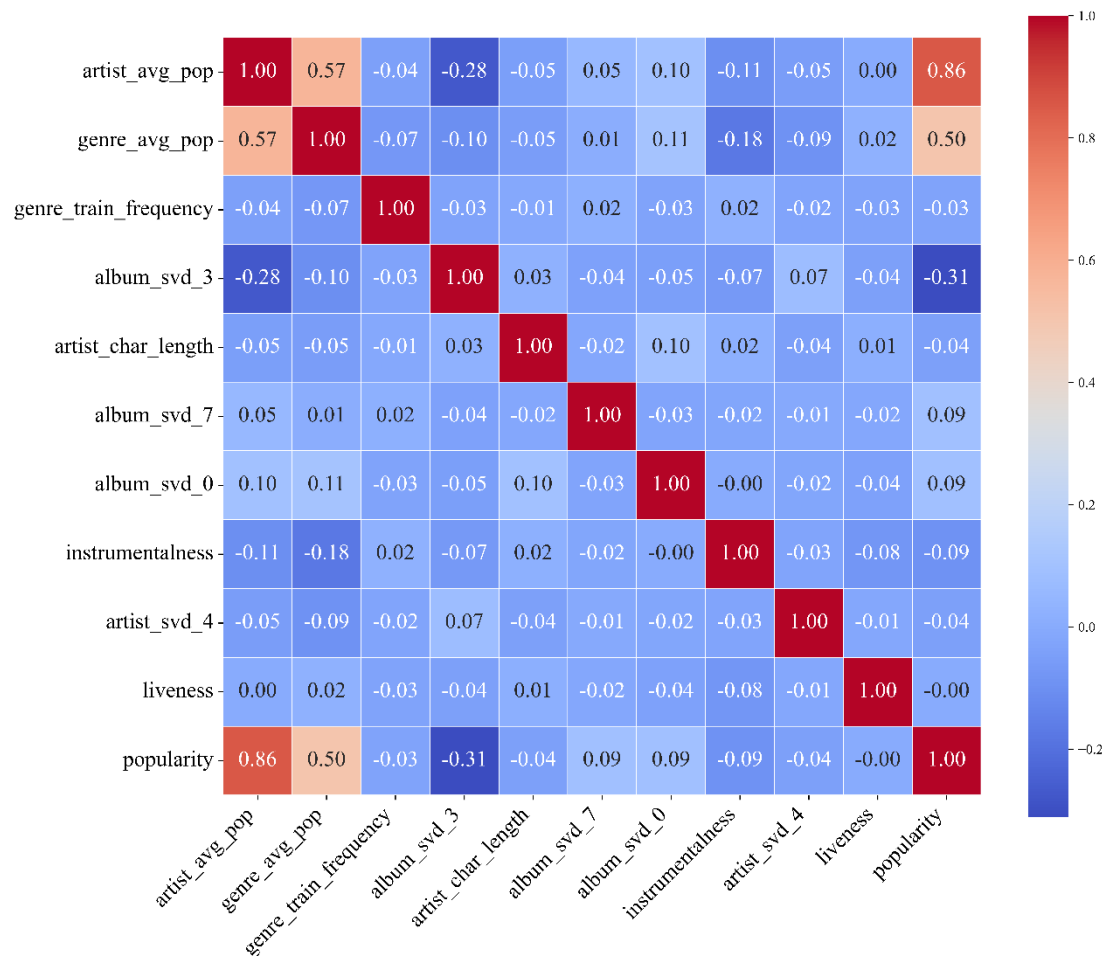


Figure 2. Top 10 features with popularity correlation

Additionally, among the remaining features, we identified three distinct mechanisms for generating predictive signals. First, `artist_char_length` (length of artist names) reflects the number of contributing authors, indicating that the scale of collaboration in music creation has significant predictive value for music popularity. Collaborative work among multiple artists enables the intersection and overlap of fan communities. As a result, a song gains exposure from multiple independent communities from its very initial release, creating a potential for non-linear amplification in visibility and dissemination. Moreover, the presence of multiple artist names serves as a signal to streaming platforms and potential listeners that the song involves substantial production costs and high-quality effort. Second, four text-extended latent semantic features—`album_svd_3`, `album_svd_7`, `album_svd_0`, and `artist_svd_4`—rank among the top ten. Their total importance even surpasses that of all the original acoustic features except for those mentioned above. This effectively demonstrates the effectiveness of the TF-IDF + TruncatedSVD text vectorization approach adopted in this study. Market-related labels such as "Remix" or "Deluxe" in album titles, as well as special character patterns in artist names (e.g., official channel identifiers), serve as "non-acoustic commercial signals" that cannot be captured by pure audio features alone.

5. Conclusion

This paper focuses on two crucial problems in music popularity prediction model, i.e., limited feature dimensions and unbalanced genres. To address these, this paper proposed a multi-dimensional prediction model, namely BEAT. At the feature level, two extended features (AAP and GAP) are introduced, effectively capturing the Matthew Effect in music industry, since the cumulative importance ratio of the

two extended features we proposed exceeds 50%. Meanwhile, the combination of TD-IDF and TruncatedSVD enables the model to conduct potential semantic extraction from unstructured text, further enriching the feature space. At algorithm level, by weighing the positive examples, the problem of low Recall caused by category imbalance has been effectively alleviated. The experimental results show AUC of this model reaches 97.39%, F1-score reaches 79.15%, and Accuracy reaches 94.99%, which is significantly higher than eight baseline models such as Decision Tree, Random Forest, and XGBoost. The Precision and Recall also achieve better effects, reaching 78.50% and 79.81% respectively. This study provides an effective solution for music popularity prediction, which can serve as a reference for the optimization of recommendation strategies by streaming media platforms and the market positioning of creators.

References

- [1] Moreau, F., Viotto, J., & Wikström, P. (2026). *Consumer satisfaction with new content discovery through algorithmic and human recommendations*. *Applied Economics Letters*, 1–5. <https://doi.org/10.1080/13504851.2026.2655917>
- [2] Sangeetha, A., Tejashree, M. , & Roshini, B.. (2025). *Spotify Dataset: Recommendation of Popularity Songs (Genre) Using Machine Learning Classification Techniques*. *International Conference on Hybrid Intelligent Systems*. Springer, Cham.
- [3] Kamal, J., Priya, P., Anala, M. R., & Smitha, G. R. (2021, September). *A classification based approach to the prediction of song popularity*. In *2021 international conference on innovative computing, intelligent communication and smart electrical systems (ICSES)* (pp. 1-5). IEEE.
- [4] Velmuruga, D. A. K.. (2023). *Machine learning approaches for predicting song popularity: a case study in music analytics*. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 07(11), 1-11.
- [5] Dong, A. , Qiu, R., & Ye, Z.. (2023). *Regression analysis of song popularity based on ridge, k-nearest neighbors and multiple-layers neural networks*. *Highlights in Science, Engineering and Technology*.
- [6] Jain, V. , Gupta, R. , Chauhan, A. , & Dubey, E. . (2025). *Cracking the Code of Spotify Hits: The Impact of Data Preprocessing and Machine Learning Models*. *2025 14th International Conference on System Modeling & Advancement in Research Trends (SMART)*. IEEE.
- [7] Shil, T. , Afrin, S., Hossain, N. , Saha, P., Maitra, P., & Anwar, M. . (2025). *Predicting Song Popularity by Analyzing Audio Features of Spotify Bengali Tracks across Diverse Genres*. *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE. pp1-6.