УДК 004.032.26

Д. В. Плаксин, магистрант

ФГБОУ ВО «МГУ им. Н.П. Огарёва»

ГИБРИДНАЯ АРХИТЕКТУРА НЕЙРОННОЙ СЕТИ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ МУЗЫКАЛЬНОГО ЖАНРА

HYBRID ARCHITECTURE OF A NEURAL NETWORK FOR THE PROBLEM OF MUSIC CLASSIFICATION

Аннотация. Рассматривается вопрос классификации музыкальных жанров при помощи различных видов гибридной нейронной сети, основанной на комбинации свёрточной и рекуррентной нейронных сетей. Статья направлена на анализ возможностей нескольких моделей для классификации музыки и определение того, какая модель лучше подходит для этой задачи. Эти результаты проливают свет на дальнейшие исследования музыки.

Ключевые слова: нейронная сеть, свёрточная нейронная сеть, рекуррентная нейронная сеть, классификация музыки, GTZAN мел-спектрограмма

Abstract. The issue of classifying musical genres with the help of various types of hybrid neural network based on a combination of convolutional and recurrent neural networks is considered. The paper aims to analyze the two models' capability for music classification and determine which model is better suited for the task. These results shed light on guiding further exploration of computer music.

Keywords: neural network, convolutional neural network, recurrent neural network, music classification, GTZAN Mel-spectrogram

Введение

Распознавание и классификация музыкальных жанров по аудиоданным представляет собой важную задачу, известную как «музыкальная классификация». В связи с быстрым увеличением музыкальных архивов цель классификации музыки очевидна. Наблюдается значительное увеличение

количества музыкальных сэмплов, что затрудняет сохранение музыкального порядка вручную. Потенциальные применения классификации музыкальных жанров в системах музыкальных рекомендаций [1] и сервисах потоковой передачи музыки [2] привели к обширным исследованиям в этой области. Однако музыкальная классификация представляет собой сложную задачу, вызванную наличием нечеткого характера в различных музыкальных образцах. В результате стоит изучить классификацию музыки с неизменной точностью.

Нейросетевые технологии могут быть использованы для определения музыкального жанра путем анализа акустических характеристик аудиозаписей. В этом случае, нейронная сеть обучается на наборе данных, состоящем из аудиозаписей различных жанров, и на основе этого определяет жанр новой аудиозаписи.

Для этого в первую очередь необходимо произвести извлечение характеристик аудиозаписей. К ним относятся громкость, ритм [3,4], скорость пересечения нуля [4] и кепстральные коэффициенты мел-частоты (MFCC) [5,6,7]. Кроме того, исследователи также исследовали спектрограммы, основанные на преобразовании Фурье [8,9], вейвлет-преобразовании [10] или преобразовании с постоянной добротностью [11], которые содержат обширную частотно-временную информацию (например, временную информацию, периодическое биение, ритм и т. д.) и может обеспечить более удовлетворительную производительность.

Другим важным компонентом классификации музыкальных жанров является разработка алгоритмов классификации для обработки акустических характеристик. Классические алгоритмы машинного обучения включают статистические методы, такие как наивные байесовские классификаторы [12], случайные леса [13] и метод опорных векторов (support vector machine, SVM) [5,6]. Помимо традиционных подходов, с развитием глубокого обучения [14], для классификации музыки используют такие методы, как свёрточные нейронные сети (Convolutional Neural Networks, CNN) и рекуррентные

нейронные сети (Recurrent Neural Networks, RNN) , эти подходы являются двумя наиболее эффективными подходами к классификации музыкальных данных [15,16]. При этом CNN лучше записывает пространственные зависимости в доменах признаков [17,18], а RNN удовлетворительно обрабатывает временные зависимости последовательных данных [19,20].

Тем не менее, существуют некоторые ограничения существующих методов. Основываясь на анализе музыкальных сигналов, мы наблюдаем, что жанр музыки — это очень широкое понятие [21]. Музыкальные треки, принадлежащие к одному и тому же жанру, могут иметь различные акустические характеристики, такие как ритм и такты. Рисунок 1 иллюстрирует значительный разброс спектрограмм музыкальных треков, относящихся к жанру «рок».

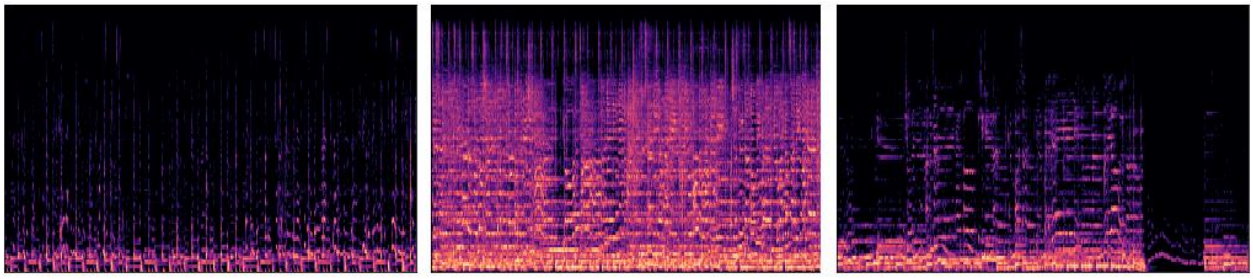

Рисунок 1 – Спектрограммы аудиозаписей в жанре «рок»

Недостатком существующих методов является то, что они плохо справляются с такими разнообразными распределениями данных с большими внутриклассовыми различиями. Чтобы правильно отнести различные аудиозаписи к одному и тому же жанру, модель должна дополнительно фиксировать глубокую скрытую информацию. Однако когда объем данных недостаточен для «копания глубже», это может привести к ложному индуктивному смещению и отрицательно сказаться на точности классификации.

Для построения архитектуры нейронных сетей в данной работе будут задействованы свёрточные нейронные сети и рекуррентные нейронные сети, рассмотрим их подробнее.

Свёрточная нейронная сеть (CNN) — специальная архитектура искусственных нейронных сетей, предложенная Яном Лекуном в 1988 годуи нацеленная на эффективное распознавание образов.

Название архитектура сети получила из-за наличия операции свёртки, суть которой в том, что каждый фрагмент изображения умножается на матрицу (ядро) свёртки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения

Рекуррентные нейронные сети (RNN) — вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. В данной статье рассмотрены четыре подхода к построению рекуррентных моделей: сеть с долговременной и кратковременной памятью (LSTM), управляемый рекуррентный блок (GRU), а также их двунаправленные версии (Bi-LSTM и Bi-GRU).

Актуальность данной работы состоит в том, что в условиях постоянного увеличения количества музыкальных композиций, необходима классификация музыки по жанрам для упрощения последующей обработки музыкальных композиций, в частности, для поиска музыки или рекомендаций.

Исследования по теме

Предложение полностью автоматической системы классификации музыкальных жанров было впервые выдвинуто Цанетакисом и Куком [3]. Они представили три набора функций для представления тембровой текстуры, ритмического содержимого и основного тона, а также обучили классификаторы распознавания статистических образов с использованием реальных аудиоколлекций. Выпущенный ими набор данных GTZAN стал эталонным набором данных для большинства последующих работ в этой области. С тех пор были предложены различные алгоритмы, основанные на

традиционном машинном обучении. Сюй и др. [4] использовали многослойный классификатор, основанный на машинах опорных векторов, для замены традиционных евклидовых методов, основанных на расстоянии, и других методов статистического обучения. Патил и Немаде [6] использовали акустические характеристики, такие как вектор MFCC, частоты цветности, спектральный спад, спектральный центроид и скорость пересечения нуля, и объединили алгоритмы SVM и k-NN (метод k ближайших соседей) для выполнения задачи классификации. Чаудхури и др. [22] исследовали наивный байесовский метод, деревья решений, логистическую регрессию и случайный лес для классификации музыкальных жанров.

С появлением алгоритмов глубокого обучения и вычислительных ресурсов в этой области все большее распространение получают алгоритмы распознавания, основанные на глубоких нейронных сетях. Лю и др. [23] использовали метод взаимодействия признаков обучения среднего уровня, основанный на сверточной нейронной сети; Абессер и Мюллер [24] описали музыкальную теорию джаза с помощью недавно предложенной архитектуры глубокой нейронной сети U-net; Чжуан и др. [25] разработал классификатор Transformer для анализа взаимосвязи между различными звуковыми кадрами; и Хасгивала и Тейлор [7] оценили производительность систем классификации, основанных на новом Vision Transformer, RNN-LSTM (сеть долгой краткосрочной памяти) и архитектуре на основе CNN с использованием MFCC в качестве акустической функции. Как сообщается в литературе, методы, основанные на глубоком обучении, обычно могут обеспечить более удовлетворительную производительность, чем методы, основанные на классическом машинном обучении.

Кроме того, как метод, управляемый данными, успех методов глубокого обучения не может быть отделен от создания крупномасштабных наборов данных. Кроме того, эффективность системы классификации зависит от качества и масштаба набора данных. Существует множество бесплатных наборов данных с открытым исходным кодом для классификации

музыкальных жанров, в том числе GTZAN Genre Collection (GTZAN), Free Music Archive (FMA) и Million Song Dataset (MSD). Коллекция жанров GTZAN — это широко используемый набор данных для классификации музыкальных жанров, который включает в себя предварительно извлеченные функции, такие как кепстральные коэффициенты Mel-частоты (MFCC) и характеристики цветности. Архив бесплатной музыки — это большая коллекция бесплатной легальной музыки со связанными метаданными, такими как имя исполнителя, альбом и жанр. Набор данных Million Song представляет собой набор аудиофункций и метаданных для миллиона современных популярных музыкальных треков. MSD включает в себя звуковые функции, такие как высота звука, тембр и ритм, а также метаданные, такие как исполнитель, год выпуска и популярность. Эти наборы данных вносят вклад в открытое сообщество и предоставляют достаточно обучающих данных для моделей глубокого обучения на основе данных.

Выбор методов и средств решения

Эта работа направлена на классификацию музыкальных жанров с помощью свёрточной и реккурентной нейронных сетей, а также на объединении этих двух подходов. Данные нейросети реализованы на основе данных, полученных с использованием мел-спектрограмм и MFCC.

Эксперименты проводились с помощью набора данных GTZAN, находящегося в свободном доступе [3]. Набор данных содержит 1000 музыкальных клипов, разделенных на 10 жанров, в каждом из которых по 100 песен, таких как «блюз, классика, кантри, диско, хип-хоп, джаз, металл, поп, регги и рок» продолжительностью 30 секунд каждая, а также таблицы с данными, полученными на основе отрезков длиной 3 секунд и 30 секунд. При частоте дискретизации 22050 Гц каждый клип имеет размер 16 бит для моноканала и кодируется в формате mp3.

Для задачи классификации в этой статье мы будем использовать несколько подходов: CNN, RNN (LSTM) и объединение подходов CNN и RNN: CNN и LSTM, CNN и Bi-LSTM, CNN и GRU, CNN и Bi-GRU.

Модели будут обучаться на двух наборах данных, полученных на основе трёх- и тридцатисекундных отрезков песен соответственно, каждый набор разделен на три части: тренировочные данные (70%), валидационные данные (20%) и тестовые данные (10%).

В качестве среды разработки была выбрана платформа Google Colab,в которую интегрированы все необходимые для библиотеки, такие как NumPy, Pandas, Keras, Tensorflow и другие.

Для каждой нейросети были выбраны оптимальные параметры, представленные в таблице 1.

Модель	Кол-во эпох	Кол-во слоёв (CNN)	Кол-во слоёв (RNN)	Размер ядра в свёрточных слоях	Размер батча
CNN	50	32, 64, 64	-	4, 8, 8	32
RNN (LSTM)	50	-	64, 64	-	32
CNN+LSTM	50	32, 64, 64	32, 64, 128	4, 8, 8	32
CNN+Bi-LSTM	50	32, 64, 64	32, 64, 128	4, 8, 8	32
CNN+GRU	50	32, 64, 64	32, 64, 128	4, 8, 8	32
CNN+Bi-GRU	50	32, 64, 64	32, 64, 128	4, 8, 8	32

Таблица 1 – параметры нейросетей.

Анализ результатов работы

В ходе работы был проанализирован набор данных GTZAN, а также была составлена его корреляционная тепловая карта, представленная на рисунке 2.

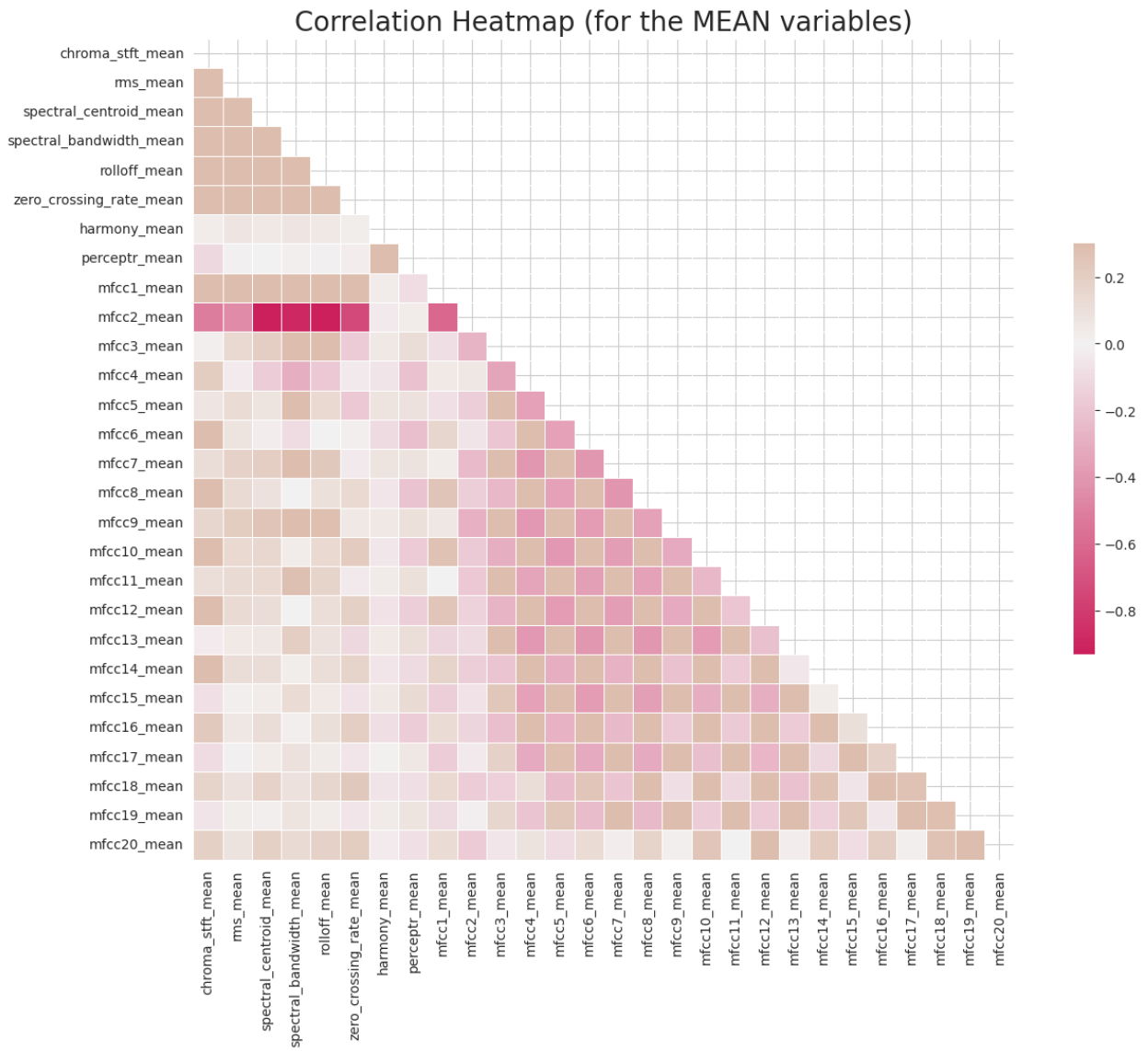

Рисунок 2 – Корреляционная тепловая карта

В ходе эксперимента было разработано шесть различных моделей нейронных сетей. На рисунке 3 представлена диаграмма для сравнения точности моделей для отрезков длины 3 секунды, на рисунке 4 – для отрезков длины 30 секунд.

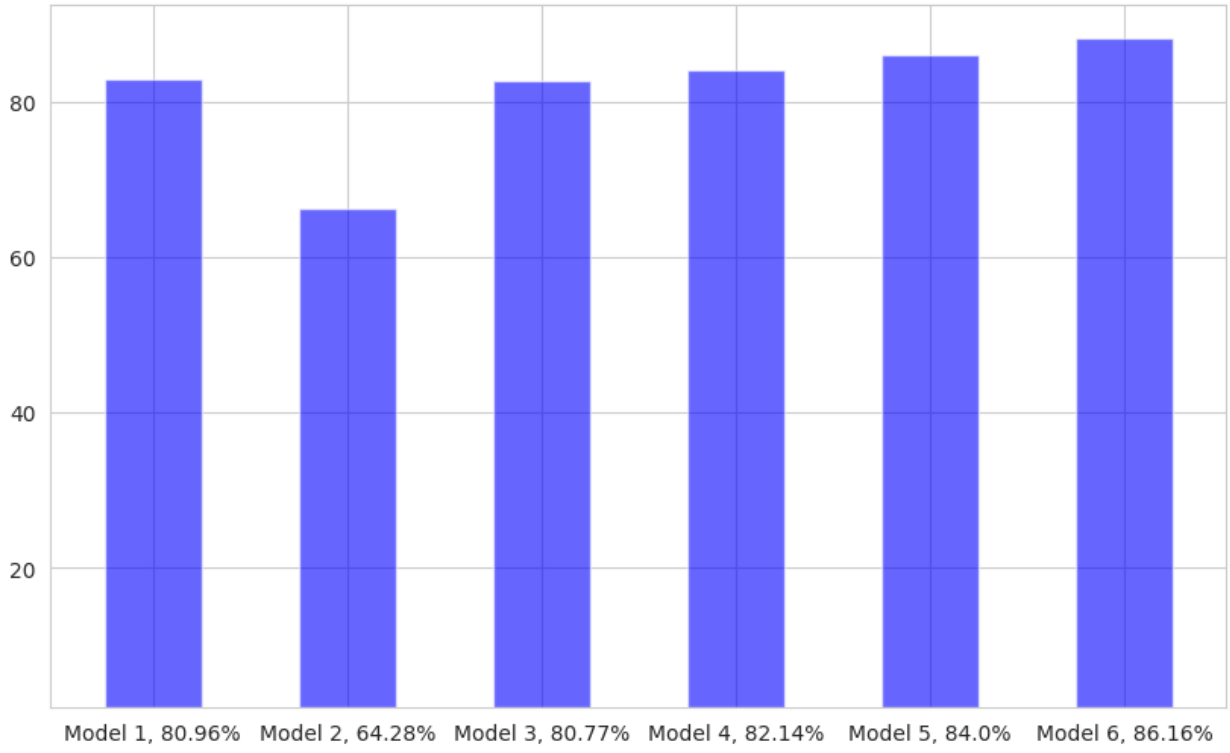

Рисунок 3 – Гистограмма для отрезков длины 3 секунды

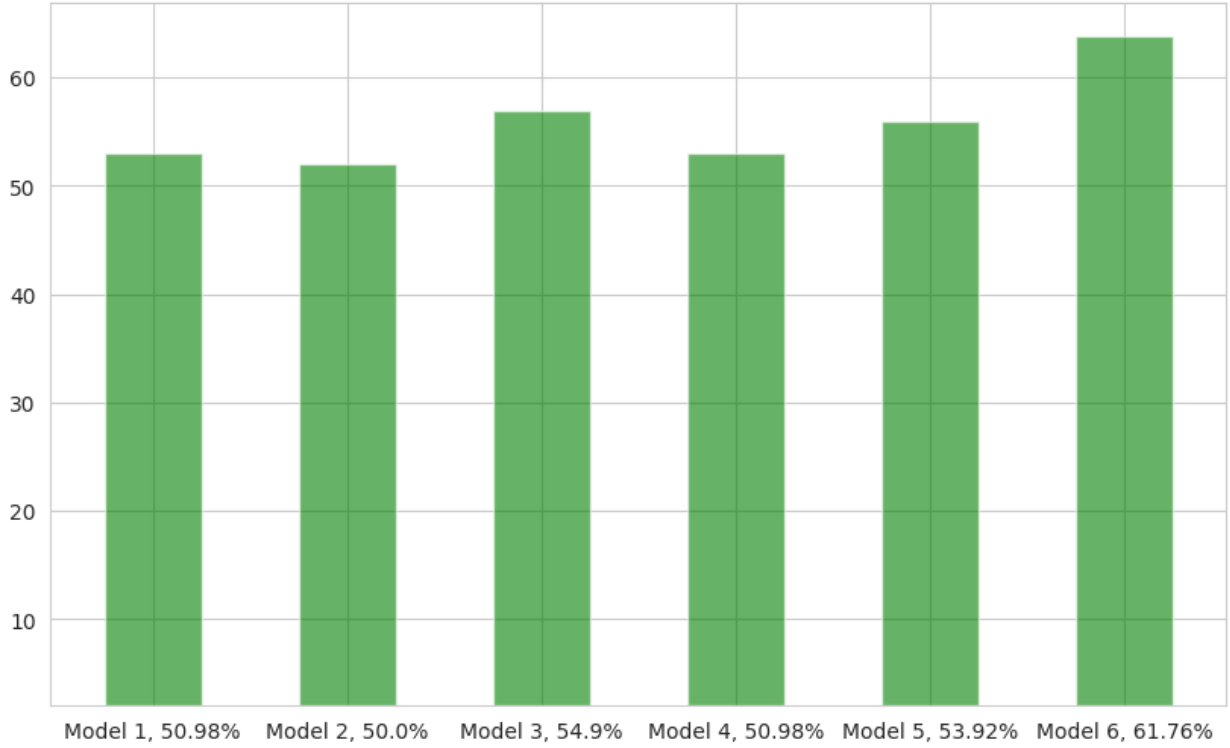

Рисунок 4 – Гистограмма для отрезков длины 30 секунд

Таким образом, получили, что на выбранном наборе данных наибольшей точности достигает модель CNN+Bi-GRU с точностью 86,16% в случае данных, полученных из отрезков длины 3 секунд и 61,76% на отрезках длины 30 секунд.

Из результатов можно сделать вывод, что гибридизация может улучшить точность работы модели нейронной сети, а значит, что для задачи классификации музыки гибридизация архитектур нейронных сетей является полезной.

Заключение

В этой статье мы выполнили задачу классификации музыки в общедоступном наборе данных под названием GTZAN, представив гибридизацию CNN и различных вариантов RNN. Для извлечения признаков мы использовали данные, предоставленные в наборе данных GTZAN вместе с шестью комбинациями нейронных сетей, а именно: CNN, RNN, CNN и LSTM, CNN и Bi-LSTM, CNN и GRU и, наконец, CNN и Bi-GRU. Для данных, полученных из отрезков длины 3 секунды, используя комбинацию CNN и Bi-GRU, мы получили самую высокую точность 86,16%, тогда как для отрезков в 30 секунд комбинация CNN и Bi-GRU показала самую высокую точность 61,76%, что так же было лучшим среди всех комбинаций. Также было проведено сравнение всех полученных моделей и было обнаружено, что гибридизация увеличивает точность работы нейронной сети, что положительно сказывается на классификации музыки.

Список использованных источников

1. Elbir, A.; Aydin, N. Music genre classification and music recommendation by using deep learning. Electron. Lett. 2020, 56, 627–629.

2. Rajanna, A.R.; Aryafar, K.; Shokoufandeh, A.; Ptucha, R. Deep neural networks: A case study for music genre classification. In Proceedings of the 2015

IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 9–11 December 2015; pp. 655–660.

3. Tzanetakis, G.; Cook, P. Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. 2002, 10, 293–302.

4. Xu, C.; Maddage, N.C.; Shao, X.; Cao, F.; Tian, Q. Musical genre classification using support vector machines. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 6–10 April 2003; Volume 5, pp. 429–432.

5. Kour, G.; Mehan, N. Music genre classification using MFCC, SVM and BPNN. Int. J. Comput. Appl. 2015, 112, 12–14.

6. Patil, N.M.; Nemade, M.U. Music genre classification using MFCC, K-NN and SVM classifier. Int. J. Comput. Eng. Res. Trends 2017, 4, 43–47.

7. Khasgiwala, Y.; Tailor, J. Vision transformer for music genre classification using mel-frequency cepstrum coefficient. In Proceedings of the 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 23–25 September 2021; pp. 1–5.

8. Pelchat, N.; Gelowitz, C.M. Neural network music genre classification. Can. J. Electr. Comput. Eng. 2020, 43, 170–173.

9. Cheng, Y.H.; Kuo, C.N. Machine Learning for Music Genre Classification Using Visual Mel Spectrum. Mathematics 2022, 10, 4427.

10. Jena, K.K.; Bhoi, S.K.; Mohapatra, S.; Bakshi, S. A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis. Neural Comput. Appl. 2023, 1–26.

11. Zhao, H.; Zhang, C.; Zhu, B.; Ma, Z.; Zhang, K. S3t: Self-supervised pre-training with swin transformer for music classification. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 606–610.

12. Silla, C.N.; Koerich, A.L.; Kaestner, C.A. A machine learning approach to automatic music genre classification. J. Braz. Comput. Soc. 2008, 14, 7–18.

13. Bahuleyan, H. Music genre classification using machine learning techniques. arXiv 2018, arXiv:1804.01149.

14. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444.

15. Bisharad, D.; Laskar, R.H. Music genre recognition using convolutional recurrent neural network architecture. Expert Syst. 2019, 36, 1–13.

16. Huang, A.; Wu, R. Deep Learning for Music. arXiv 2016, arXiv:1606.04930.

17. Abdoli, S.; Cardinal, P.; Koerich, A.L. End-to-end environmental sound classification using a 1D convolutional neural network. Expert Syst. Appl. 2019, 136, 252–263.

18. Murad, A.; Pyun, J.-Y. Deep Recurrent Neural Networks for Human Activity Recognition. Sensors 2017, 17, 2556.

19. Wu, W.; Han, F.; Song, G.; Wang, Z. Music Genre Classification Using Independent Recurrent Neural Network. In Proceedings of the 2018 Chinese Automation Congress (CAC), Xi'an, China, 30 November–2 December 2018; pp. 192–195.

20. Ashraf, M.; Ahmad, F.; Rauqir, R.; Abid, F.; Naseer, M.; Haq, E. Emotion Recognition Based on Musical Instrument using Deep Neural Network. In Proceedings of the 2021 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2021; pp. 323–328.

21. Rimmer, M. Beyond omnivores and univores: The promise of a concept of musical habitus. Cult. Sociol. 2012, 6, 299–318.

22. Chaudhury, M.; Karami, A.; Ghazanfar, M.A. Large-Scale Music Genre Analysis and Classification Using Machine Learning with Apache Spark. Electronics 2022, 11, 2567.

23. Liu, J.; Wang, C.; Zha, L. A middle-level learning feature interaction method with deep learning for multi-feature music genre classification. Electronics 2021, 10, 2206

24. Abeßer, J.; Müller, M. Jazz bass transcription using a U-net architecture. Electronics 2021, 10, 670.

25. Zhuang, Y.; Chen, Y.; Zheng, J. Music genre classification with transformer classifier. In Proceedings of the 2020 4th International Conference on Digital Signal Processing, Chengdu, China, 19–21 June 2020; pp. 155–159.