BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

USO DE VISUALIZAÇÃO DE INFORMAÇÃO PARA INVESTIGAÇÃO DA CLASSIFICAÇÃO SUPERVISIONADA DE GÊNEROS MUSICAIS

ARTHUR DE MELO REZENDE

Rio Verde, GO

2025

INSTITUTO FEDERAL GOIANO - CAMPUS RIO VERDE
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

USO DE VISUALIZAÇÃO DE INFORMAÇÃO PARA INVESTIGAÇÃO DA CLASSIFICAÇÃO SUPERVISIONADA DE GÊNEROS MUSICAIS

ARTHUR DE MELO REZENDE

Trabalho de Conclusão de Curso apresentado ao Instituto Federal Goiano - Campus Rio Verde, como requisito parcial para a obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Prof. Douglas Cedrim Oliveira

Rio Verde, GO

2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

de Melo Rezende, Arthur
D278u Uso de visualização de informação para investigação da classificação supervisionada de gêneros musicais / Arthur de Melo Rezende. Rio Verde 2025.

40f. il.

Orientador: Prof. Dr. Douglas Cedrim Oliveira.
Tcc (Bacharel) - Instituto Federal Goiano, curso de 0219201 - Bacharelado em Ciência da Computação - Integral - Rio Verde (Campus Rio Verde).
1. Redes Neurais Artificiais. 2. Perceptron Multicamadas. 3. Inteligência artificial explicável. 4. Projeção multidimensional. 5. Coeficientes Cepstrais de Frequência Mel. I. Título.

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAR PRODUÇÃO TÉCNICA NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Repositório Institucional do IF Goiano - RIIF Goiano Sistema Integrado de Bibliotecas

- Profissional de Educação do IF Goiano -

Com base no disposto na Lei Federal nº 9.610/98, e manual sobre a Produção Técnica, publicado pela DAV/CAPES/MEC*, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano, a disponibilizar gratuitamente o documento no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada eletronicamente abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

Identificação da Produção Técnica – DAV/CAPES

[ ] Editoria [ ] Material Didático

[ ] Curso de Formação Profissional [ ] Projetos de Extensão à Comunidade

[ ] Relatório Técnico Conclusivo [ ] Atividade Técnica/Tecnológica

[ ] Disseminação do Conhecimento Técnico/Tecnológico [ ] Produto Bibliográfico

[X] Outras Produções Técnicas - Tipo: TCC (Graduação)

Nome Completo do Autor/a: Arthur de Melo Rezende

Matrícula: 2020102201940024

Título do Trabalho: Uso de visualização de informação para investigação da classificação supervisionada de gêneros musicais

Restrições de Acesso ao Documento

Documento confidencial: [X] Não [ ] Sim

Justifique: _____

Informe a data que poderá ser disponibilizado no RIIF Goiano: 19 / 02 / 2025

O documento está sujeito a registro de patente? [ ] Sim [X] Não

O documento pode vir a ser publicado como livro e/ou artigo? [ ] Sim [X] Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a docente e/ou autor/a declara que:

1 - o documento é seu trabalho original, detém os direitos autorais da produção técnica e não infringe os direitos de qualquer outra pessoa ou entidade;

2 - obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

3 - cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Rio Verde, 19 de fevereiro de 2025.

(Assinado Eletronicamente)

Arthur de Melo Rezende (Autor)

(Assinado Eletronicamente)

Douglas Cedrim Oliveira (Orientador)

1058004

(Assinatura do Docente, Autor e/ou Detentor dos Direitos Autorais)

Documento assinado eletronicamente por:

- **Douglas Cedrim Oliveira, PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 19/02/2025 17:11:09.
- **Arthur de Melo Rezende, 2020102201940024 - Discente**, em 20/02/2025 09:25:13.

Este documento foi emitido pelo SUAP em 19/02/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 678201
Código de Autenticação: 7bac0717f0

INSTITUTO FEDERAL GOIANO
Campus Rio Verde
Rodovia Sul Goiana, Km 01, Zona Rural, 01, Zona Rural, RIO VERDE / GO, CEP 75901-970
(64) 3624-1000

Regulamento de Trabalho de Conclusão de Curso (TCC) – IF Goiano - Campus Rio Verde

ANEXO V - ATA DE DEFESA DE TRABALHO DE CURSO

Aos seis dias do mês de fevereiro de dois mil e vinte e cinco às dezesseis horas e trinta minutos, reuniu-se a Banca Examinadora composta por: Prof. Dr. Douglas Cedrim Oliveira (orientador), Prof. Dr. Adriano Soares de Oliveira Bailão (membro interno) e Prof. Dr. Danilo Pereira Barbosa (membro interno), para examinar o Trabalho de Conclusão de Curso (TCC) intitulado "Uso de visualização de informação para investigação da classificação supervisionada de gêneros musicais" de Arthur de Melo Rezende, estudante do curso de bacharelado em Ciência da Computação do IF Goiano – Campus Rio Verde, sob Matrícula nº 2020102201940024. A palavra foi concedida ao estudante para a apresentação oral do TC, em seguida houve arguição do candidato pelos membros da Banca Examinadora. Após tal etapa, a Banca Examinadora decidiu pela APROVAÇÃO do estudante. Ao final da sessão pública de defesa foi lavrada a presente ata, que segue assinada pelos membros da Banca Examinadora.

Rio Verde, 06 de fevereiro de 2025.

(Assinado eletronicamente)

Douglas Cedrim Oliveira

Orientador(a)

(Assinado eletronicamente)

Adriano Soares de Oliveira Bailão

Membro da Banca Examinadora

(Assinado eletronicamente)

Danilo Pereira Barbosa

Membro da Banca Examinadora

Observação:

Documento assinado eletronicamente por:

- **Douglas Cedrim Oliveira**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 06/02/2025 21:53:58.
- **Danilo Pereira Barbosa**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 06/02/2025 23:09:46.
- **Adriano Soares de Oliveira Bailao**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 08/02/2025 09:49:11.

Este documento foi emitido pelo SUAP em 06/02/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 672913
Código de Autenticação: f0dc13fb9a

AGRADECIMENTOS

Os meus sinceros agradecimentos pelos últimos 5 (cinco) anos e pelo longo período de escrita deste trabalho para conclusão do curso de Ciências da Computação.

Inicialmente agradeço a Deus, pelo cuidado e provisão. Agradeço pelas bênçãos e pelos livramentos durante as viagens entre Goiânia e Rio Verde.

Quero agradecer à minha família: meu pai, Adriano, que sempre esteve ao meu lado com seu suporte, seu sorriso no rosto e seus ensinamentos; minha mãe, Sheila, que me ensinou grande parte do que sei hoje, teve paciência para me auxiliar na escrita deste trabalho e sempre esteve disposta a me ajudar em qualquer situação; e aos meus irmãos, Vithor e Helena, que compartilharam bons momentos e risadas comigo, trazendo leveza e apoio ao fim do dia.

Gostaria também de expressar minha gratidão aos meus tios, Adelzuita e Adalberto, que abriram as portas de sua casa para que eu pudesse morar com eles durante minha estadia em Rio Verde. Ambos foram fundamentais ao longo de todo o processo de graduação, não apenas pela estadia e hospitalidade, mas também por causa dos ensinamentos valiosos, os quais levarei para o resto da vida. Durante os quatro anos que vivi com eles, foram como segundos pais para mim, e só tenho a agradecer.

Agradeço também a todos os meus familiares pelo amor, incentivo e apoio. Em especial, aos meus tios, Carlos Eduardo e Evilásio Júnior, que me deram a oportunidade de realizar o estágio obrigatório. Quero agradecer especialmente ao meu primo Daniel, que sempre esteve pronto para conversar quando eu precisava.

Agradeço ao IFGoiano, Campus Rio Verde, em especial, aos Coordenadores do curso, aos docentes e técnicos administrativos do CORE, aos meus colegas de classe, pois proporcionaram um rico ambiente de conhecimento para conclusão do curso.Agradeço, em especial, ao meu Orientador do TCC, prof Dr Douglas Cedrim, pela orientação, pelos encontros semanais por meio do Google Meet, pela amizade, paciência e parceria durante todo o processo da escrita. Também quero expressar minha gratidão aos meus amigos fora do IF: Adrian, João Vitor, Jordan, Juliana, Kauan e Stephanie. Obrigado a todos por fazerem companhia durante quase todas as semanas dos últimos quatro anos, compartilhando boas risadas e histórias memoráveis. Vocês são incríveis

Por fim, mas não menos importante, agradeço ao meu cachorro, Bob, e à minha gata, Emmy. Eles sabem muito bem por que estou agradecendo.

RESUMO

Arthur, Rezende. **Uso de visualização de informação para investigação da classificação supervisionada de gêneros musicais.** 2025. 40 f. Monografia – (Curso de Bacharel em Ciência da Computação), Instituto Federal Goiano - Campus Rio Verde. Rio Verde, GO.

Este trabalho de conclusão de curso propõe apresentar os resultados dos testes referente a aplicação de Redes Neurais Artificiais (RNAs) na classificação de músicas em diferentes gêneros utilizando o conjunto de dados GTZAN. O objetivo principal é utilizar a visualização de informação para investigação da classificação supervisionada de gêneros musicais. Para tanto utilizou-se os Coeficientes Cepstrais de Frequência Mel (MFCCs) para extrair as características dos áudios utilizados para o treinamento e teste de um modelo de Perceptron Multicamadas (MLP), aliado a técnicas de visualização de informações para análise do processo de treinamento.
O estudo demonstra que a utilização de 15 coeficientes MFCC resulta na maior acurácia de 99,25%, distinguindo efetivamente entre os gêneros musicais. Os resultados destacam a capacidade da RNA de generalizar entre diferentes entradas de áudio e seu potencial uso em sistemas de recomendação musical. Os achados reforçam a relevância do ajuste de parâmetros, das técnicas de extração de características e da aplicação de métodos de visualização para melhorar o desempenho e a interpretabilidade dos modelos de redes neurais na classificação de gêneros musicais.
Palavras-chave: Redes Neurais Artificiais (RNAs); Perceptron Multicamadas (MLP); Inteligência artificial explicável (xAI); Projeção multidimensional; Coeficientes Cepstrais de Frequência Mel (MFCC).

ABSTRACT

Arthur, Rezende. **Uso de visualização de informação para investigação da classificação supervisionada de gêneros musicais.**. 2025. 40 f. Trabalho de Conclusão de Curso – Bacharel em Ciência da Computação, Instituto Federal Goiano - Campus Rio Verde. Rio Verde, GO, 2025.

This work investigates the application of Artificial Neural Networks (ANNs) for classifying music into different genres using the GTZAN dataset. The primary focus is on leveraging Mel-Frequency Cepstral Coefficients (MFCCs) as features for training and testing a Multilayer Perceptron (MLP) model, combined with information visualization techniques to analyze the training process. The study demonstrates that using 15 MFCC coefficients results in the highest accuracy of 99.25%, effectively distinguishing between musical genres. Additionally, multidimensional visualization using the t-SNE technique provided a detailed understanding of how the model processes various audio features, aiding in pattern identification and training improvements. The results highlight the ANN's ability to generalize across different audio inputs and its potential application in music recommendation systems. The findings underscore the relevance of parameter tuning, feature extraction techniques, and visualization methods to improve the performance and interpretability of neural network models in music genre classification.

Keywords: Artificial Neural Networks; Multilayer Perceptron; eXplainable Artificial Intelligence; Multidimensional projection; Mel frequency cepstral coefficients.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
DCT	*Transformada Discreta do Cosseno*
GTZAN	Nome de um *dataset* específico para classificação de gêneros musicais
IA	Inteligência Artificial
K-NN	*k*-nearest neighbors
MFCC	*Mel-frequency Cepstral Coefficients* (Coeficientes Cepstrais de Frequência Mel)
MLP	*Multilayer Perceptron* (Perceptron Multicamadas)
MNIST	Modified National Institute of Standards and Technology
RNA	*Redes Neurais Artificiais*
ReLu	*Rectified Linear Unit*
STFT	*Short-time Fourier transform* (Transformada de Fourier de tempo curto)
SVM	Support vector Machine
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
XAI	*eXplainable Artificial Intelligence* (Inteligência Artificial Explicável)

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

As Redes Neurais Artificiais (RNAs) representam um método eficaz para a classificação de gênero musical, assim como, para a organização de bibliotecas de música, recomendação de música em plataformas de *streaming* e para a análise de tendências musicais. Estudos anteriores demonstram o bom desempenho das RNAs, na tarefa de classificação de gênero musical. Por exemplo, têm-se as pesquisas realizadas por Yang e Zhang (2019) que registrou um aumento significativo na capacidade de classificação de músicas devido a adição de camadas convolucionais repetidas em suas arquiteturas de Redes Neurais Convolucionais (CNN).

Observa-se que o uso de métodos de aprendizado profundo nas características de sinal de áudio mudou a extração e análise de características musicais. Zhang et al. (2016) implementaram CNNs no conjunto de dados GTZAN e apresentaram melhorias nos métodos de classificação automatizada de gêneros musicais. Esses métodos não apenas atualizam a capacidade dos classificadores de fornecer resultados, mas também ajudam na investigação minuciosa dos atributos de gêneros musicais variados em termos de timbre, ritmo e composição instrumental.

Diante disso, o problema que norteou esta pesquisa foi: Como utilizar técnicas de visualização da informação para melhor compreensão do processo de classificação realizado pelas redes neurais?

A pesquisa desenvolvida tem como objetivo geral utilizar a visualização de informação para investigação da classificação supervisionada de gêneros musicais.

Para alcançar o objetivo geral, foram elencados os seguintes objetivos específicos:

1. Identificar características de áudio que permitam a diferenciação entre diferentes gêneros musicais;
2. Implementar um modelo para a classificação supervisionada de gêneros musicais, garantindo que a arquitetura do modelo suporte a visualização das decisões de classificação;
3. Avaliar a eficácia do modelo utilizando métricas de desempenho para classificação supervisionada;
4. Utilizar técnicas de visualização de informação para análise qualitativa dos resultados, para auxiliar na explicação dos resultados das métricas.

A relevância dessa pesquisa está relacionada com o contínuo crescimento das plataformas de *streaming* de música, tendo em vista a capacidade de categorizar automaticamente grandes volumes de dados musicais. Portanto a combinação de técnicas avançadas de processamento de áudio com aprendizado de máquina e visualização da informação pode oferecer uma nova dimensão na análise da classificação de gêneros musicais.

Para tanto, foi necessário desenvolver uma RNA capaz de realizar na classificação de áudios de música em gêneros musicais utilizando o conjunto de dados GTZAN.

Ao incorporar as técnicas de visualização da informação, espera-se alcançar precisão a

classificação de gêneros musicais, assim como, promover uma melhor compreensão e confiança nos modelos de RNAs através do uso do XAI.

Gunning et al. (2019) definem XAI como um ramo emergente da inteligência artificial que visa aumentar a transparência e interpretabilidade dos modelos de aprendizado de máquina. No contexto da classificação de gêneros musicais, entender como as RNAs chegam às suas decisões é fundamental para garantir que o modelo seja confiável e livre e não tendencioso. Este termo será abordado posteriormente no capítulo referente a **Fundamentação Teórica**

Este trabalho está estruturado da seguinte forma:

Neste **Capítulo 1** são apresentados o tema, a justificativa e o problema de pesquisa, assim como o objetivo geral e os específicos.

No **Capítulo 2** são descritas as principais técnicas e modelos utilizados na classificação de gêneros musicais, com ênfase no uso de Redes Neurais Artificiais e técnicas de extração de características de áudio.

No **Capítulo 3** são apresentados os trabalhos relacionados, abordando as pesquisas e artigos utilizados para o entendimento dos temas e para a análise de outros trabalhos que investigam temas semelhantes a pesquisa proposta por esse trabalho.

No **Capítulo 4** são discutidos os materiais e métodos utilizados, incluindo a descrição do conjunto de dados GTZAN, as ferramentas de software empregadas e o processo de extração e padronização das características dos sinais de áudio. Além disso, um breve detalhamento da arquitetura, da função de ativação e do método de treinamento relacionados ao modelo de Rede Neural.

No **Capítulo 5** são apresentadas as discussões e os resultados obtidos com o modelo proposto, incluindo a análise das projeções t-SNE e a avaliação da métrica de silhueta.

Por fim, no **Capítulo 6** são apresentadas as considerações finais do trabalho, bem como sugestões para trabalhos futuros que possam explorar outras abordagens e técnicas na classificação de gêneros musicais.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo buscou-se apresentar alguns conceitos para fundamentar teoricamente os resultados dos testes referentes a aplicação de Redes Neurais Artificiais (RNAs) na classificação de músicas em diferentes gêneros utilizando o conjunto de dados GTZAN. Dentre eles: o conceito de música, de gêneros musicais, de espectrograma, de Inteligência Artificial, de aprendizado de máquina, dos modelos preditivos, das redes neurais artificiais, de visualização de informação, de projeção multidimensional, de eXplainable IA. Além dos conceitos, busca apresentar as características principais de uma música e as métricas de consistências.

2.1 Música

De acordo com Perlovsky (2015), a música é uma forma de arte universal que transcende barreiras culturais e temporais. A apreciação da beleza musical é uma experiência humana que envolve a apreciação de harmonias, melodias e ritmos complexos, tendo em vista que proporciona prazer estético. A música é considerada uma poderosa ferramenta para a expressão e regulação emocional, capaz de evocar e intensificar uma ampla gama de emoções, desde alegria e excitação até tristeza e nostalgia. A música tem o poder de unir pessoas, fortalecendo laços comunitários e culturais. É frequentemente utilizada em eventos sociais, celebrações e rituais para promover a coesão social e a identidade coletiva.

O referido autor ainda apresenta o conceito de musicoterapia. Trata-se de uma prática reconhecida que utiliza a música para promover a saúde mental e emocional, ajudando a reduzir o estresse, a ansiedade e a dor, além de melhorar o bem-estar geral (PERLOVSKY, 2015). Outros estudos desenvolvidos sobre a música apresentam os seus benefícios que estão relacionados aos aspectos cognitivos, melhorando a capacidade de memorização, de concentração e de aprendizagem. A prática musical também pode estar associada ao desenvolvimento de habilidades motoras finas e de coordenação (CABRAL; CORRÊA; NETO, 2023).

2.1.1 Características Principais de uma Música

Perlovsky (2015) afirma que a análise musical envolve a extração e o estudo de várias características que definem e diferenciam os gêneros musicais. Entre as principais características de uma música estão o timbre, a frequência e tonalidade, o ritmo, a dinâmica, a harmonia e a estrutura. De acordo com Blake (2012), timbre é a qualidade sonora que distingue diferentes fontes sonoras, mesmo quando estas produzem a mesma nota em termos de frequência fundamental. Ele é determinado por harmônicos, envelope e outras propriedades do som, sendo essencial para a identificação de instrumentos musicais e vozes. Os estudos de Moylan (2014) destacam que o timbre é uma combinação de características físicas do som, como amplitude e espectro, que são analisadas por meio de espectrogramas e técnicas de extração como os Coeficientes Cepstrais

de Frequência Mel (MFCC). Para Perlovsky (2015), o timbre é a qualidade sonora que permite distinguir diferentes fontes sonoras que produzem a mesma nota.

De acordo com Tzanetakis e Cook (2002), a melodia é definida cono uma sequência de notas musicais organizadas de forma linear e percebidas como uma entidade única. Também pode ser definida por elementos como altura, ritmo e andamento, sendo a base para a identificação de padrões musicais em diferentes gêneros. Ainda de acordo com os referidos autores, tem-se a percepção melódia, que está intimamente ligada ao timbre e ao contexto harmônico, reforçando sua complexidade e importância na análise musical.

Em estudo realizado por Cordero (2014), o ritmo trata-se de uma estrutura temporal da música, incluindo o tempo, o compasso e os padrões rítmicos. O ritmo incorpora a percepção da regularidade e do movimento na música.

De acordo com Constantinescu e Brad (2023), som é definido como uma onda mecânica que se propaga através de meios físicos, como ar, água ou sólidos, sendo percebido pelo ouvido humano como variações de pressão no tempo. A análise do som em contextos musicais frequentemente envolve sua representação em espectros de frequência, o que facilita a identificação de seus componentes básicos.

Para Perlovsky (2015), o som pode ser classificado em agudo ou grave por meio da análise de sua frequência. Enquanto isso, a tonalidade está diretamente relacionada à organização das notas musicais dentro de um sistema tonal, influenciando a harmonia e a melodia. Para analisar as propriedades físicas do som é necessário o uso de espectrogramas e coeficientes cepstrais.

Além disso, existe uma diferença entre som digital e sinal de áudio. Som digital refere-se à conversão de sinais de áudio analógicos em uma forma digital que pode ser manipulada e processada por computadores. Este processo inclui amostragem, que determina a taxa na qual os dados são capturados, e quantização, que define a precisão dos valores registrados (MCLOUGHLIN, 2016). A análise do sinal de áudio digital é amplamente utilizada em estudos sobre processamento de fala e música, utilizando técnicas como os MFCCs para capturar características tímbricas e melódicas.

2.1.2 Gêneros musicais

De acordo com Tzanetakis e Cook (2002), os gêneros musicais são categorias que agrupam músicas com características semelhantes, sendo importante na forma como a música é percebida, produzida e consumida. Cada gênero musical possui características que o diferenciam dos demais, como instrumentos utilizados, ritmo, harmonia, melodia e contexto cultural. Utilizando alguns dos gêneros musicais presentes no conjunto de dados usado durante a pesquisa, como exemplo têm-se "Hip Hop", "Raggae", "Clássica", "Rock", "Country e "Metal".

O Hip Hop, por exemplo, é conhecido pelo rap sobre batidas eletrônicas e samples, com ritmos sincopados e loops repetitivos. Já o Reggae tem um ritmo sincopado característico conhecido como "skank"e utiliza instrumentos como guitarra, baixo, bateria e teclado. O rock é

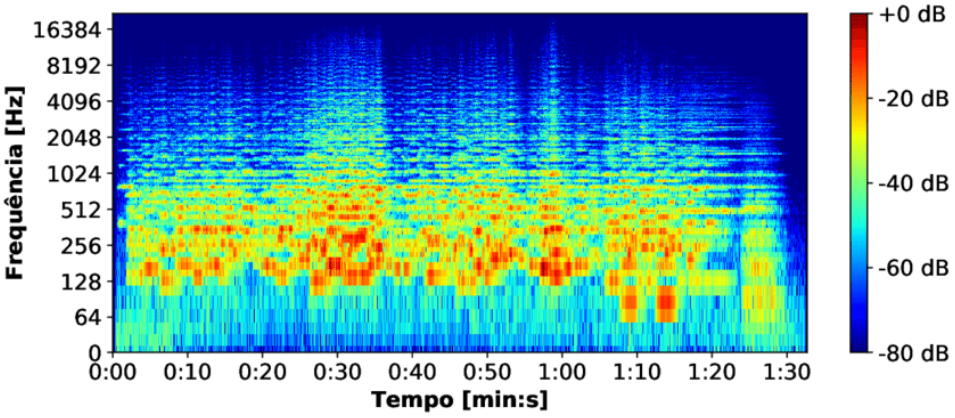
caracterizado por um forte ritmo e uso de guitarras elétricas, baixo elétrico e bateria, com uma batida marcada e ênfase nos tempos fortes. Em contraste, a música clássica utiliza uma variedade ampla de instrumentos como cordas, madeiras, metais e percussão, com ritmos e harmonias mais complexos. O Country destaca-se por suas melodias simples e letras narrativas, utilizando violão, banjo, violino e pedal steel guitar, com ritmos estáveis e harmonias simples. Em contraste, o Metal é intenso e distorcido, com guitarras elétricas altamente distorcidas, baixo, bateria e vocais gritados, ritmos rápidos e complexos e harmonias que utilizam modos menores e riffs complexos (TZANETAKIS; COOK, 2002).

Diante do exposto, observa-se que cada gênero musical carrega consigo um conjunto de elementos estilísticos que são identificáveis e distintos. A compreensão dessas diferenças é essencial para a análise musical, pois ajuda a contextualizar e analisar motivos pelo qual um modelo de Redes Neurais Artificias possa ter gerado dúvidas ou erros referentes a análise e classificação de músicas para gêneros musicais.

2.2 Espectrograma

Os espectrogramas são, por definição, representações gráficas da evolução do sinal na frequência em função do tempo e são bem aplicados no processamento de sinais de áudio. O trabalho de Ayvaz et al. (2022) descrevem que um espectrograma é feito a partir de uma transformação de Fourier, que desdobra um sinal de áudio em seus constituintes de frequência. Mapas 2D em que a dimensão horizontal representa o tempo, a dimensão vertical representa a frequência, e a intensidade/cor do tom em cada pixel correspondem a informações superpostas do sinal de áudio usando a Transformada de Fourier. Possibilitando uma descrição visual de como a energia é distribuída ao longo do tempo, e também a capacidade de descrever características relevantes no sinal, como, por exemplo, os formantes na fala ou os padrões rítmicos na música. A Figura 1 demonstra uma representação visual de um espectrograma sobre as dimensões da frequência em Hz pela dimensão do tempo em minutos.

Figura 1 – Exemplo de um espectrograma.



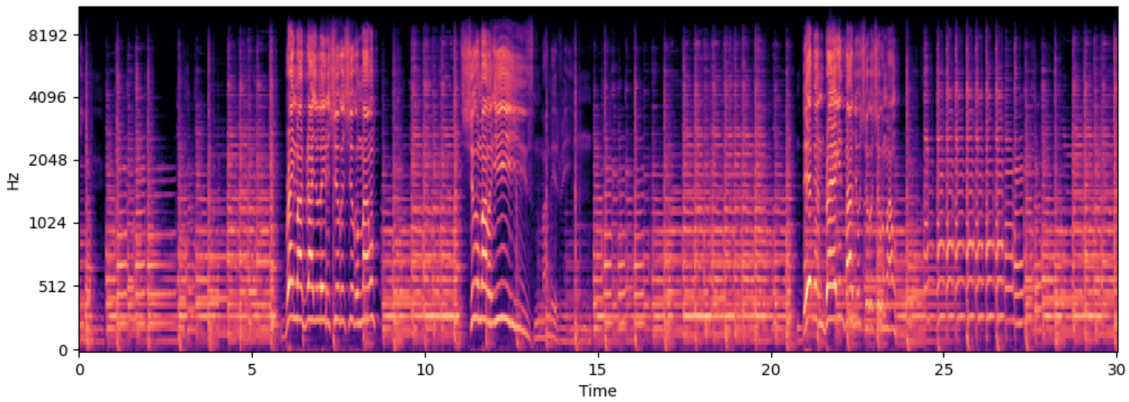
Fonte: Extraída de Spadini (2020).

A análise do sinal de áudio digital é amplamente utilizada em estudos sobre processamento de fala e música, utilizando técnicas como os Coeficientes Cepstrais de Frequência Mel (MFCCs) para capturar características tímbricas e melódicas.

Há muitos métodos para extrair espectrogramas, e a Transformada de Fourier de Curto Prazo (STFT) é a mais comumente usada. De acordo com o trabalho feito por Ayvaz et al. (2022), a STFT aplica a Transformada de Fourier em janelas curtas superpostas no sinal de áudio para que a informação do tempo e da frequência possa ser extraída ao mesmo tempo.

Contudo no presente trabalho têm-se um foco no espectrograma Mel, no qual é um tipo de espectrograma que utiliza a escala Mel, uma escala de frequência perceptual baseada na percepção do som pelo ouvinte humano. A escala Mel é usada com a intenção de simular a resposta não linear do ouvido humano às frequências para que as frequências mais altas sejam comprimidas e mais resolução seja usada para as frequências mais baixas (GHODASARA et al., 2015). Para obter um espectrograma Mel, primeiro um espectrograma é derivado do sinal de áudio, e então as frequências resultantes são convertidas em frequências Mel. Isso é feito passando as características resultantes do áudio por um banco de filtros triangulares espaçados de forma semelhante na escala Mel. Da representação da escala mel até os MFCCs há muitos mais cálculos. Primeiro, o logaritmo das amplitudes da escala mel é pego para que a escala da intensidade se torne uma escala logarítmica, que é um método relativamente próximo de como os humanos ouvem a intensidade dos sons. Em segundo lugar, a Transformada Discreta do Cosseno (DCT) é aplicada às amplitudes logarítmicas. MFCCs são coeficientes que representam a forma compacta e altamente informativa da envoltória espectral de um sinal sonoro, capturando as informações mais pertinentes para processamentos de alto nível, como reconhecimento de fala e classificação de som. A Figura 2 demonstra uma representação visual um arquivo de música do gênero blues em escala mel.

Figura 2 – Exemplo de áudio na escala mel.

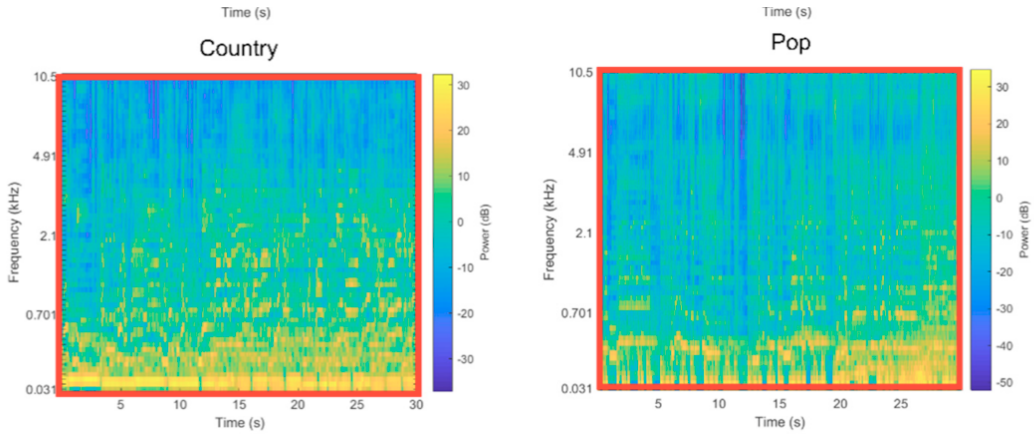


Fonte: Autoria própria.

Os MFCCs são amplamente usados no processamento de dados de áudio devido à capacidade deles de extrair boas características discriminativas dos sinais de áudio. O trabalho realizado por Cheng e Kuo (2022) faz de uso de técnicas de MFCC para extrair as características

do áudio, chegando em bons resultados, indicando essa técnica para as tarefas de classificação e reconhecimento. Ainda no mesmo artigo, os MFCCs ajudam a extrair o detalhe dos timbres e melodias sutis que distinguem um gênero musical de outro. A Figura 3 demonstras essas diferenças entre gêneros musicais extraidas com o MFCC, utilizando os gêneros Country e Pop como comparativo entre a intensidade

Figura 3 – Características MFCC retirada de áudios referentes aos gêneros musicais "Country"e "Pop".



Fonte: Extraída de Cheng e Kuo (2022).

2.3 Inteligência Artificial (IA)

De acordo com Russell e Norvig (2016), IA é um ramo da ciência da computação que possui a capacidade de simular a inteligência humana, com o objetivo de criar soluções para problemas do cotidiano. Os referidos autores sistematizaram a IA em 4 (quatro) dimensões: 1) Pensar como homem; 2) Atuar como um homem; 3) Pensar de forma racional; 4) Atuar de forma racional. A IA então promove a triangulação entre o processo de pensamento, raciocínio e comportamento humano.

De acordo com Xu et al. (2021), IA é uma tentativa de recriar em meio computacional a forma de pensamento humano. Provida de um conjunto de dados e parâmetros que possibilita o aprendizado e tomada de decisão em diversas áreas da iniciativa privada e do setor público. Entre as aplicações estão: análise preditiva; análise do comportamento dos usuários; os *chatbots* na área da medicina e saúde; no comércio online; na área de RH; no marketing; no ensino e aprendizagem corporativo; no campo do direito; na área do entretenimento; nas industrias etc (MORAIS; BRANCO, 2023).

2.3.1 Aprendizado de Máquina

Aprendizado de Máquina refere-se à prática de criar uma infraestrutura algorítmica que ensine um computador a aprender a realizar uma tarefa específica ou prever e decidir por

si mesmo, através da análise dos dados transmitidos a ele (BISHOP; NASRABADI, 2006). De acordo com Başţanlar e Özuysal (2014), o objetivo subsequente do Aprendizado de Máquina é criar sistemas que possam aprender independentemente a partir do fornecimento de dados, tornando-se cada vez melhores e mais inteligentes com o tempo. A IA provou ser muito útil em vários campos, tais como o reconhecimento de padrões, diagnóstico de doenças, previsão de mercados e processamento de dados de áudio (PURWINS et al., 2019), porque consegue lidar com grandes quantidades de dados e identificar correlações complexas. (DOMINGOS, 2012).

O aprendizado de máquina pode ser classificado em três categorias gerais de acordo com a categorização descrita em Sharma, Sharma e Jindal (2021): aprendizado supervisionado, não supervisionado e por reforço.

No **aprendizado supervisionado**, o modelo é treinado tanto com a entrada quanto com a saída correspondente. Isso significa que ele aprende apenas com as informações fornecidas, sabendo de antemão qual deveria ser o resultado esperado. Pretende-se que o modelo seja treinado de tal forma que possa mapear entradas para saídas viáveis e, ao mesmo tempo, generalizar esse conhecimento para uma nova observação de dados e situações invisíveis. Um método utilizado é a árvore de decisão, que divide o conjunto de dados em grupos menores, dependendo das características que possuem o maior ganho de informação.

No **aprendizado não supervisionado**, o modelo aprende com dados não rotulados, o que significa que a tarefa é encontrar agrupamentos ou tendências nos dados. Uma das técnicas mais utilizadas são a de agrupamento, onde o algoritmo coloca uma coleção de dados semelhantes em um cluster e com isso a análise de componente principal é aplicada para reduzir a dimensionalidade dos dados, assim mantendo a maior parte da variabilidade nos dados. O aprendizado não supervisionado é utilizado em problemas onde a etiquetagem dos dados é impossível de se ter acesso, assim, em problemas onde a estrutura dos dados precisa ser explorada em detalhes.

O **aprendizado por reforço** é uma abordagem onde um agente aprende a tomar decisões através de um processo de tentativa e erro, recebendo recompensas ou penalidades com base nas ações que toma. Este tipo de aprendizado é utilizado em robótica e jogos, onde o agente deve aprender a navegar em um ambiente complexo para alcançar um objetivo específico.

2.3.2 Métricas de avaliação

As métricas de avaliação desempenham um papel fundamental no contexto da Inteligência Artificial (IA), pois permitem medir a eficácia e o desempenho de modelos de aprendizado de máquina (LIU et al., 2014). Essas métricas oferecem uma abordagem quantitativa para avaliar a performance do modelo, identificando não apenas sua capacidade preditiva, mas também apontando áreas que necessitam de aprimoramento. Entre as principais métricas de avaliação, destacam-se a precisão, o recall e a acurácia.

A **precisão** é uma métrica que indica a proporção de previsões positivas corretas em relação ao total de previsões positivas realizadas pelo modelo. Em outras palavras, ela mede a

exatidão das previsões positivas, sendo particularmente relevante em cenários onde os custos de um falso positivo são elevados (POWERS, 2008).

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}} \tag{1}$$

A **acurácia** é a métrica mais utilizada para avaliar modelos de classificação. Ela representa a proporção total de previsões corretas, considerando tanto os verdadeiros positivos quanto os verdadeiros negativos em relação ao conjunto total de amostras (KULSKI, 2016). A acurácia é particularmente útil em conjuntos de dados equilibrados, onde o número de exemplos das classes positivas e negativas é semelhante. Sua fórmula é definida como:

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos} + \text{Verdadeiros Negativos}}{\text{Total de Amostras}} \tag{2}$$

O **recall**, também conhecido como sensibilidade ou taxa de detecção, mede a capacidade do modelo de identificar corretamente todas as ocorrências da classe positiva no conjunto de dados. Ele é particularmente importante em aplicações onde a identificação correta dos positivos é crucial, como na detecção de doenças ou fraudes (HOWARD et al., 2020). O recall é definido matematicamente como:

$$\text{Recall} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \tag{3}$$

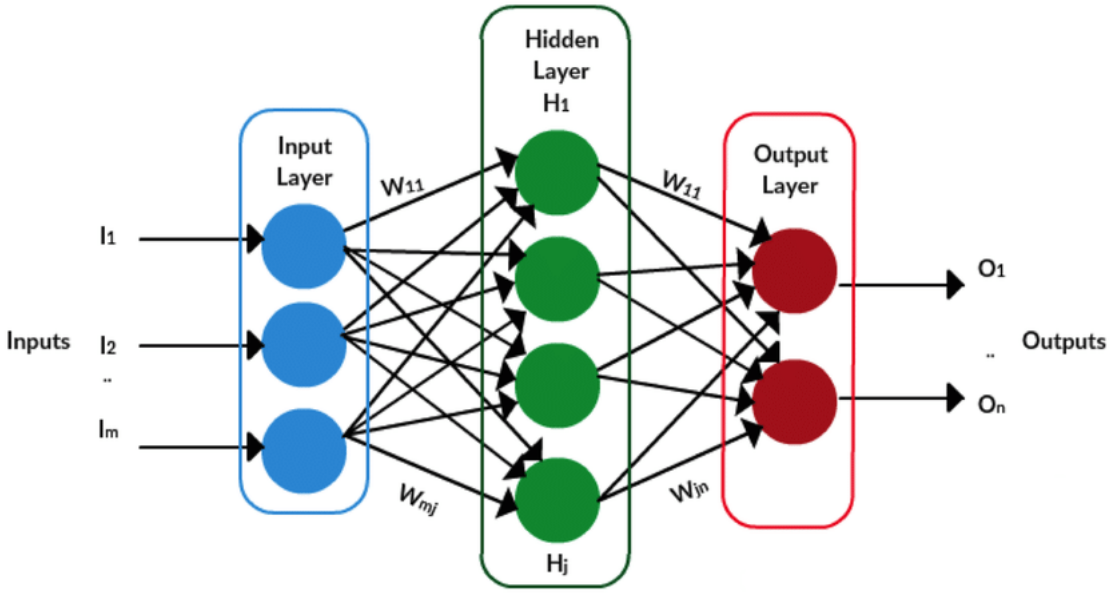
2.4 Modelos Preditivos

2.4.1 Redes Neurais Artificiais

De acordo com Tjoa e Guan (2020), as RNAs são modelos computacionais baseados nos princípios de processamento de informações do cérebro. Elas são compostas por unidades interconectadas com uma função de neurônios artificiais organizadas em camadas. As camadas são uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída (RAUBER et al., 2016). Uma característica única das RNAs é sua capacidade de aprender e generalizar a partir dos dados, incluindo o reconhecimento de padrões, classificação e previsão.

De acordo com Popescu et al. (2009), o processamento de uma rede neural artificial simples começa com a camada de entrada, através da qual os dados são introduzidos no modelo. Os dados são então transmitidos a todos os neurônios na camada de entrada e movidos para a primeira camada oculta, e durante cada transmissão, a entrada é multiplicada por uma ponderação. As ponderações são ajustáveis e servem para modular a contribuição relativa de cada entrada para o neurônio. O neurônio então aplica uma função de ativação à soma ponderada das entradas, de modo que o modelo se torne não linear e seja capaz de aprender representações complexas dos dados. A Figura 4 exemplifica como seria a estrutura desse modelo preditivo, com os vértices representando os neurônios das diversas camadas e as arestas representando as ligações entre cada neurônio e os pesos atribuídos a cada ligação.

Figura 4 – Arquitetura de um modelo RNA.



Fonte: Extraída de Alboaneen, Tianfield e Zhang (2017).

Um exemplo de uma função de ativação é o *ReLU (Rectified Linear Unit)*, que, de acordo com Banerjee, Mukherjee e Jr (2019), devolve zero para qualquer entrada negativa e a própria entrada para qualquer entrada positiva. Esta função de ativação tem sido utilizada na resolução de diversos problemas, tendo em vista a sua simplicidade e menor uso de recurso computacional. Em seguida, a camada intermediária envia essa informação para a camada de saída, e uma função de ativação adicional, como a função *Softmax*, pode ser usada para converter as saídas para probabilidades, no qual torna a ferramenta útil para classificação, por exemplo. Para cada iteração no treinamento, a saída da rede é medida em relação à saída desejada (rótulo), e o erro é propagado de volta à rede para atualizar os pesos. A rede tenta, com base nesse erro, imitar as falhas e alterar as ponderações de um modo tal que, pela utilização de técnicas de otimização como o gradiente descendente, o erro é reduzido.

Desta forma, processo por meio do qual esses pesos são atualizados, isto é, o treinamento, é um tipo de retropropagação, permitindo que o erro seja propagado de volta pela rede forçando os pesos a minimizar o erro nas próximas previsões (POPESCU et al., 2009). Isso fornece a capacidade das RNAs para aprender até mesmo padrões muito complexos que são não-lineares nos dados.

Além da utilização em aplicações de reconhecimento de imagens e processamento de linguagem natural, as RNAs também são utilizados em trabalho de interpretação de dados de áudio. Métodos, como representar espectrogramas e MFCCs como entradas para as RNAs, possibilitaram um grande número de aplicações, como reconhecimento de fala e classificação de gênero musical.

2.5 Visualização de Informação

A visualização de informação é uma área de estudo que se concentra em transformar dados complexos em representações visuais que facilitam a compreensão e a análise. Essa área é fundamental para diversas disciplinas, incluindo ciência de dados, estatística e aprendizado de máquina, pois permite que os usuários identifiquem padrões, tendências e anomalias nos dados de forma intuitiva e eficaz. A habilidade de visualizar dados complexos e multidimensionais é essencial para a tomada de decisões informadas e para a comunicação de descobertas de maneira clara e acessível (RAUBER et al., 2016).

A visualização de informação desempenha um importante papel na interpretação de grandes volumes de dados. Os estudos realizados por Rauber et al. (2016) mostram que representações visuais podem melhorar significativamente a compreensão e a tomada de decisão, tanto de RNAs quanto usuários. A visualização permite que esses usuários absorvam grandes quantidades de informação de forma rápida e intuitiva, facilitando a identificação de relacionamentos e a comunicação eficaz dos resultados.

A aplicação de técnicas de visualização de informação em aprendizado de máquina é particularmente benéfica para a análise de desempenho de modelos. Ao visualizar os erros de classificação é possível identificar áreas onde o modelo pode ser aprimorado. Além disso, a visualização das saídas de camadas ocultas em redes neurais pode oferecer um entendimento sobre como o modelo está processando os dados. Isso pode ajudar a diagnosticar problemas e melhorar a eficiência e a eficácia dos modelos de aprendizado de máquina (RAUBER et al., 2016). Essa necessidade de uma melhor interpretação de algoritmos de aprendizado de máquina é a base de uma sub-área conhecida como *eXplainable IA*.

2.5.1 *eXplainable IA*

Para Lipton (2016), a Inteligência Artificial Explicável (*eXplainable AI - XAI*) refere-se ao desenvolvimento de métodos e técnicas que tornam os processos internos de modelos de IA compreensíveis para humanos. Com o aumento da complexidade dos modelos de IA, particularmente os baseados em *deep learning*, a necessidade de explicações claras e interpretáveis das decisões do modelo se torna-se importante para o seu entendimento. A falta de transparência pode levar à desconfiança dos usuários e limitar a adoção de sistemas de IA em setores críticos, dentre eles, a área da saúde, finanças e segurança. De acordo com Lipton (2016), a XAI auxilia na identificação e mitigação de vieses nos modelos IA, aumentando a confiança dos usuários e cumprindo os requisitos regulatórios e éticos. Além disso, o referido autor argumenta que explicações interpretáveis são necessárias para identificar erros e melhorar os modelos de IA de maneira iterativa e segura. Diversas estratégias são utilizadas para esse propósito de interpretabilidade, sendo uma delas a projeção multidimensional.

2.5.2 Projeção multidimensional

A projeção multidimensional é uma técnica utilizada para visualizar dados de alta dimensionalidade em um espaço de duas ou três dimensões. Isso é particularmente útil em problemas de aprendizado de máquina onde os dados possuem muitas características. Uma das técnicas comuns de projeção multidimensional é a *t-Distributed Stochastic Neighbor Embedding* (t-SNE) (MAATEN; HINTON, 2008).

Os autores ainda explicam que a técnica de t-SNE é amplamente utilizada para visualização de dados complexos devido à sua capacidade de preservar a estrutura local dos dados, tornando-a ideal para identificar agrupamentos naturais. O t-SNE projeta dados de alta dimensão em duas ou três dimensões de forma que pontos semelhantes fiquem próximos uns dos outros no espaço projetado, enquanto pontos dissimilares ficam mais distantes. Isso é particularmente útil para visualizar dados em áreas como reconhecimento de padrões e aprendizado de máquina.

2.5.3 Métricas de consistência

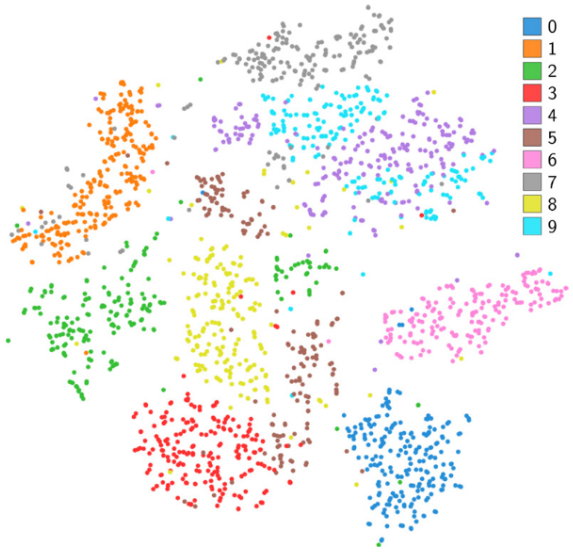
Para avaliar a qualidade das projeções multidimensionais são utilizadas métricas de consistência. Neste trabalho utilizou-se *silhouette*. que consiste em uma métrica que mede o quão semelhante um objeto é ao seu próprio grupo em comparação com outros grupos, sendo uma indicação da densidade e separação dos clusters formados. Essa métrica varia de -1 a 1, onde valores próximos a 1 indicam que os pontos estão bem agrupados, enquanto valores próximos a -1 indicam que os pontos estão mal agrupados (ŘEZANKOVÁ, 2018).

3 TRABALHOS RELACIONADOS

Rauber et al. (2016) exploram o uso da técnica de redução de dimensionalidade t-SNE para a visualização das atividades internas de RNA, com o objetivo de entender melhor como essas redes categorizam dados e identificam padrões. O estudo demonstra que a visualização de projeções t-SNE pode revelar *insights* valiosos sobre a estrutura interna de redes neurais, permitindo aos projetistas de redes identificar agrupamentos e relações que não são facilmente perceptíveis.

Ainda na pesquisa de Rauber et al. (2016) os autores apresentam uma projeção t-SNE das ativações da última camada oculta de uma rede neural treinada no conjunto de dados *Modified National Institute of Standards and Technology* (MNIST), conforme ilustrada na Figura 5. Essa visualização destaca a separação entre as classes de dígitos, mostrando a eficácia da técnica na representação das relações de alta dimensão em um espaço bidimensional. Esta técnica de visualização desenvolvida pelos referidos autores foi utilizado neste trabalho, adaptando-as para os dados de áudio.

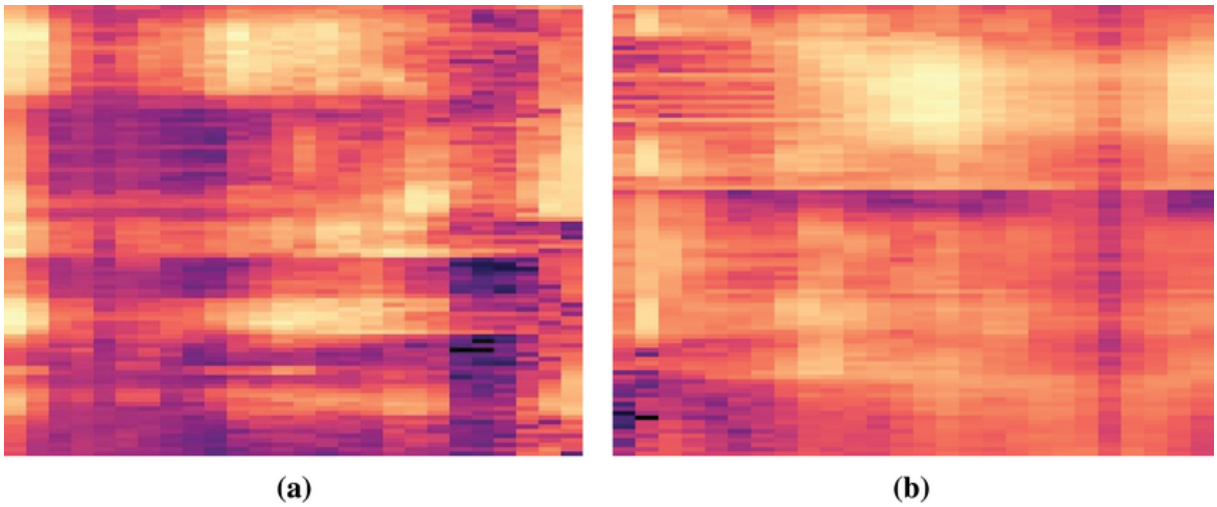
Figura 5 – Visualização t-SNE sobre a análise de um modelo de Rede Neural.



Fonte: Extraída de Rauber et al. (2016)

No trabalho deAyvaz et al. (2022), os autores exploram o uso de Mel espectrograma na identificação de padrões em dados de áudio, particularmente para o reconhecimento de fala e classificação de gêneros musicais. O estudo demonstra que Mel espectrograma são eficazes na captura das características temporais e de frequências do áudio, o que facilita a distinção entre diferentes tipos de sons. Na Figura 6 é apresentada a captação das características de duas vozes distintas utilizando espectrograma na escala Mel, assim podendo notar a intensidade das cores e a forma como padrão coletado é diferente entre uma e outra. O trabalho proposto por Ayvaz et

Figura 6 – Comparação entre áudios de vozes diferentes através de seus respectivos espectrogramas em escala Mel.



Fonte: Extraída de Ayvaz et al. (2022).

al. (2022) contribuiu para o entendimento sobre como é o funcionamento desse espectrograma, assim como mostra possíveis formas de compreender a captação das características de áudio.

O trabalho realizado por Logan et al. (2000) expandiu o uso de MFCCs para a análise musical, demonstrando que esses coeficientes capturam com eficácia características timbrais para a classificação de áudio. Utilizando um conjunto de dados composto por trechos de músicas em diferentes gêneros, os autores implementaram os MFCCs como entrada para modelos de aprendizado. Os experimentos revelaram que os MFCCs, particularmente em combinações com modelos baseados em *k-means clustering*, conseguem agrupar músicas com características similares, proporcionando uma base para tarefas de classificação. Os resultados destacaram a robustez dos MFCCs na captura de padrões timbrais, com acurácia de até 90% nos experimentos.

Um dos primeiros trabalhos desenvolvidos sobre a utilização de MFCC com o conjunto de dados GTZAN foi realizada por Tzanetakis e Cook (2002), para o desenvolvimento de um sistema automático de categorização de músicas. A abordagem utilizada pelos autores incluiu a extração de 20 (vinte) coeficientes MFCC e o uso de algoritmos como máquina de vetores de suporte (SVM) e K-NN. O sistema alcançou uma acurácia média de 61% ao classificar os 10 gêneros musicais presentes no conjunto de dados, destacando a eficácia dos MFCCs na captura de características sonoras diferenciadoras. Dessa forma, o estudo apresentou a possibilidade de combinar MFCCs com algoritmos de aprendizado de máquina para a classificação musical e ter resultados satisfatórios.

Utilizando o conjunto de dados GTZAN, Ghodasara et al. (2015) apresentaram o resultado de desempenho de uma abordagem onde a extração de características baseadas em MFCC é realizada em blocos. O mesmo estudo destaca o uso de uma SVM como classificador, implementando uma validação cruzada de três dobras para avaliação. Os resultados indicam que a abordagem baseada em blocos, especialmente com 40 (quarenta) bancos de filtros e 3

(três) blocos combinada com o uso da SVM oferece uma acurácia de até 98,43% na classificação de música/fala, superando as abordagens tradicionais. O referido trabalho é relevante para o presente estudo, pois ambos utilizam o conjunto de dados GTZAN e os coeficientes MFCC como principais características, embora o modelo proposto no presente estudo utilize MLP ao invés de SVM, possibilitando a implementação da utilização do conjunto de dados GTZAN com técnicas MFFC.

Patil e Nemade (2017) propuseram um sistema automatizado para a classificação de gêneros musicais, utilizando coeficientes cepstrais de frequência Mel (MFCC) juntamente com diversas outras arquiteturas de RNA. O estudo se destaca pelo uso do conjunto de dados GTZAN e pela comparação entre diferentes classificadores, como *K-Nearest Neighbors* (K-NN) e SVM com kernel linear e polinomial. Por mais que esse trabalho utiliza o conjunto de dados GTZAN, foi feito o treinamento de apenas 9 (nove) gêneros musicais, diferente dos 10 (dez) gêneros presentes no GTZAN. Dentre os classificadores avaliados, o SVM com kernel polinomial demonstrou ser o mais eficiente, atingindo uma acurácia de 78%, precisão de 79% e recall de 78%, superando tanto o K-NN quanto o SVM com kernel linear. Também é considerado um trabalho relevante para a pesquisa desse TCC pois demonstra a eficácia de outras técnicas de classificação de gêneros musicais, nesse caso em específico a técnica de SVM, especialmente com kernel polinomial, quando aplicada à classificação de gêneros musicais usando o conjunto de dados GTZAN, demonstrando assim uma importância da seleção adequada de kernels para maximizar a performance na classificação de áudio.

A Tabela 1 sintetiza essas abordagens analisadas.

Tabela 1 – Técnicas avaliadas para classificação de gêneros musicais usando a base de dados GTZAN.

Ano	**Característica do áudio**	**Classificador**	**Métricas**	**Referência**
2002	FFT, MFCC, outras feat.	k-NN	60% (Acurácia)	Tzanetakis e Cook (2002)
2002	FFT, MFCC, outras feat.	GMM	61% (Acurácia)	Tzanetakis e Cook (2002)
2015	Blocos de MFCC	SVM	98,4% (Acurácia)	Ghodasara et al. (2015)
2017	MFCC	SVM Linear	78% (Acurácia)	Patil e Nemade (2017)
2023	MFCC, e todas as features	SVM	60% (Acurácia)	Tzanetakis e Cook (2002)

Fonte: Autorial própria.

Na pesquisa realizada por Xu (2023) sobre a classificação de gêneros musicais realizou-se uma comparação de diferentes algoritmos de aprendizado de máquina para reconhecimento de gênero musical, incluindo SVM e CNN (Convolutional Neural Network), destacando que a precisão dos modelos de *deep learning*, como CNN, depende do processamento do áudio em espectrogramas. No entanto, a performance dessas técnicas não superou métodos tradicionais como SVM e *random forest*, como apresentado na Figura 7. Esses resultados descobertos mostram a qualidade das técnicas modernas utilizadas para classificação de gêneros musicais, contudo, também mostraram a complexidade envolvida na escolha do algoritmo adequado para esse tipo de tarefa.

Figura 7 – Resultados obtidos por Xu (2023) comparando diversas técnicas de classificação de gêneros musicais.

Table 2. Comparison of SVM with different kernels.

Algorithm	Accuracy with all features	Algorithm with top 20 features
SVM with polynomial kernel	69.0%	66.0%
SVM with RBF kernel	74.0%	65.5%

Table 3. Comparison of conventional machine learning algorithms.

Algorithm	Accuracy with 30-second input features	Algorithm with 3-second input features
Logistic Regression	66.5%	67.5%
Random Forests	74.5%	80.3%

Table 4. Comparison of neural network-based algorithms.

Algorithm	Accuracy with data processing	Algorithm without data processing
Convolution Neural Network	82.0%	25.0%
Feed-forward Neural Network	54.0%	33.0%

Fonte: Extraída de Xu (2023).

Observa-se que enquanto Xu (2023) compara o desempenho de modelos de *deep learning* com técnicas tradicionais, Patil e Nemade (2017) evidencia a eficiência de diferentes classificadores, como SVM com kernel polinomial. Sendo assim, os estudos realizados por Xu (2023) e Patil e Nemade (2017) são relevantes para o presente trabalho, tendo em vista a utilização do conjunto de dados GTZAN e da classificação de gêneros musicais por meio de técnicas de arquitetura como SVM e CNN, juntamente com a implementação de MFCCs.

No capitulo 5 propõe-se uma analise comparativa entre os resultados do modelo MLP obtidos neste trabalho e os estudos supracitados.

4 MATERIAIS E MÉTODOS

Neste capítulo, são discutidas os materiais e métodos utilizados, incluindo a descrição do conjunto de dados GTZAN, as ferramentas de software empregadas e o processo de extração e padronização das características dos sinais de áudio. Além disso, um breve detalhamento da arquitetura, da função de ativação e do método de treinamento relacionados ao modelo de Rede Neural.

4.1 Ferramentas

As ferramentas utilizadas neste trabalho foram descritas abaixo:

4.1.1 Python

A linguagem de programação Python foi desenvolvida por Guido van Rossum em 1989, sendo uma linguagem multiparadigma, orientada a objetos, funcional e interpretada. A utilização dessa tecnologia é utilizada devido a sua facilidade de uso, como ferramenta em casos de visualização e análise de dados, juntamente com *Machine Learning*, sendo assim, a principal linguagem utilizada na ferramenta online Google Colab. Neste presente trabalho, utilizou-se a versão 3.10.

4.1.2 Google Colab

O Google Colab é uma ferramenta desenvolvida pela Google, tendo a função de um ambiente de desenvolvimento interativo, capaz de gerar protótipos de modelos de aprendizado de máquina sem a necessidade do desenvolvedor dispor de um computador com os recursos necessários para que o modelo seja feito, utilizando assim os dos recursos na nuvem disponibilizada pela Google. Neste presente trabalho, utilzou-se a versão 1.0.0.

4.1.3 Biblioteca Sklearn

Scikit-Learn (Sklearn) é uma biblioteca popular de Python para aprendizado de máquina e ciência de dados. Representa o repositório de diferentes técnicas de agrupamento, classificação, regressão, pré-processamento de dados, seleção de modelo e validação cruzada (PEDREGOSA et al., 2011). Neste presente trabalho, utilizou 2.2.0., na prática dos métodos de aprendizagem supervisionada, usado no preparo dos dados e até mesmo na comparação dos desempenhos dos modelos com métricas de precisão, acurácia, recall e F1-score.

4.1.4 Biblioteca Numpy

O NumPy é uma biblioteca base do Python necessária para todas as operações científicas e de processamento de dados. Ele suporta arrays e matrizes de várias dimensões, pois trata-se de uma coleção de implementações de funções matemáticas que permitem ao usuário realizar operações rápidas e eficientes (HARRIS et al., 2020). Foi utilizada a versão 1.26.4 durante as implementações para o processamento e manipulação de dados

4.1.5 Biblioteca Librosa

A biblioteca Librosa é uma ferramenta utilizada para análise e processamento de áudio em Python. Neste trabalho, a Librosa foi empregada para carregar e extrair características de arquivos de áudio, facilitando a análise subsequente com técnicas de aprendizado de máquina (MCFEE et al., 2024). Neste presente trabalho, utilizou a versão 0.10.2.post1.

Para carregar os arquivos de áudio, utilizamos a função librosa.load, que lê o arquivo de áudio especificado e retorna a forma de onda (um array NumPy) e a taxa de amostragem.

Para a extração de características, como os coeficientes cepstrais de Mel-frequência (MFCCs), foi utilizando a função librosa.feature.mfcc. Os MFCCs foram calculados a partir da forma de onda e da taxa de amostragem e, achatados para formar um array adequado para entrada em modelos de aprendizado de máquina.

4.1.6 Biblioteca Tensorflow/Keras

O TensorFlow é uma biblioteca de código aberto desenvolvida pelo Google, utilizada para o desenvolvimento e treinamento de modelos de aprendizado de máquina e RNA (ABADI et al., 2015). Neste trabalho, foi empregada a API Keras (versão 3.5.0) ao TensorFlow (versão 2.17.1) para a criação e treinamento de uma RNA (CHOLLET et al., 2015).

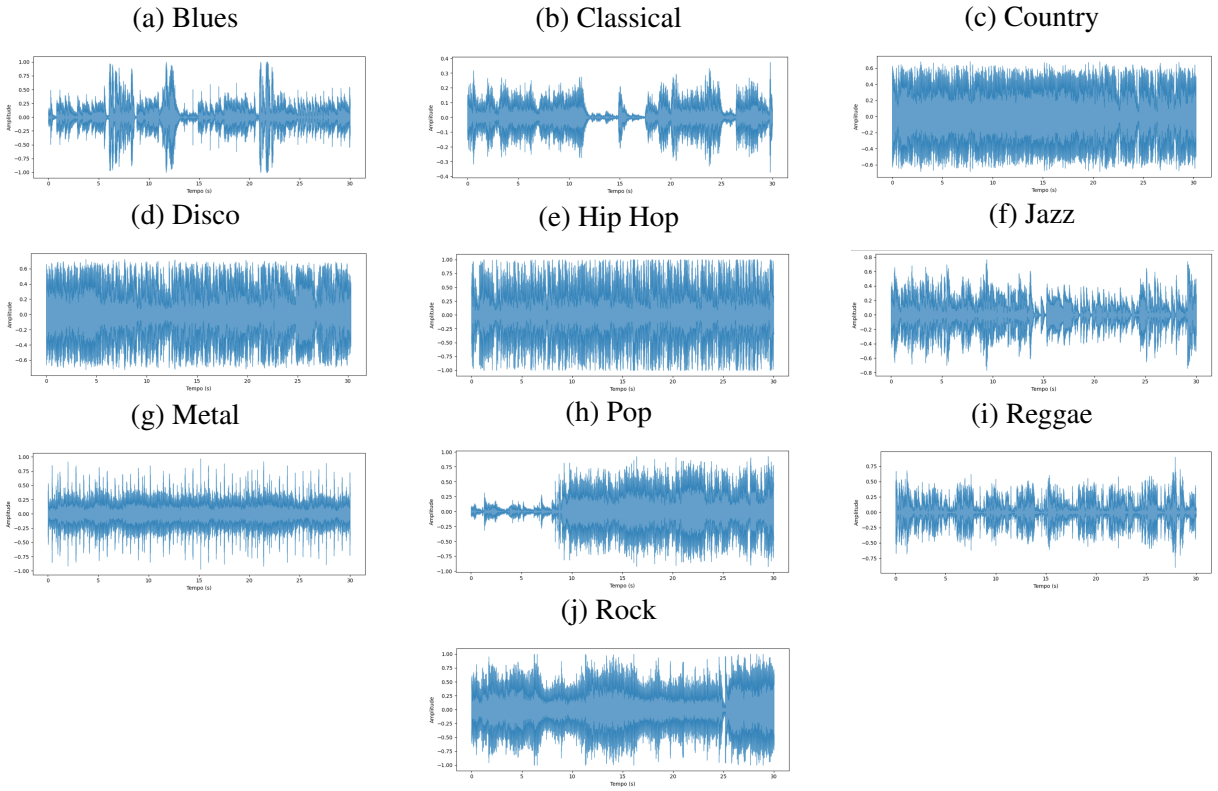
O modelo foi desenvolvido usando a classe `Sequential`, contendo múltiplas camadas densas (*fully connected*) configuradas com funções de ativação como ReLU e Softmax. A função de perda *sparse categorical crossentropy* foi utilizada para tarefas de classificação, enquanto o otimizador *Adam* foi selecionado para acelerar a convergência durante o treinamento. O TensorFlow/Keras foi escolhido devido à sua flexibilidade, alto desempenho e sua documentação, que facilitam a implementação de modelos complexos, além de sua integração com outras ferramentas utilizadas no trabalho.

4.2 Conjunto de dados

Para a realização deste trabalho, utilizou-se o conjunto de dados GTZAN, desenvolvido no ano de 2002, por Tzanetakis e Cook (2002) na Universidade de Columbia, para auxiliar no desenvolvimento de pesquisa de classificação automática de músicas. O conjunto de dados GTZAN é um conjunto de dados utilizado em pesquisas acadêmicas na área de processamento

de sinais de áudio e reconhecimento de padrões. Dispondo de 1000 (mil) trechos de áudio de 30 segundos cada, divididos igualmente em 10 gêneros musicais diferentes. Sendo coletado a partir de CDs comerciais, com cada trecho de áudio armazenado em formato WAV com uma taxa de amostragem de 44,1 kHz e uma profundidade de bits de 16. Além disso, cada trecho de áudio foi rotulado com o gênero musical correspondente, que inclui: *blues*, clássico, *country*, disco, hip hop, jazz, metal, pop, reggae e rock. Como exemplo, a Figura 8 ilustra todos os gêneros musicais presentes no conjunto de dados em um formato de ondas sonoras. O conjunto de dados é conhecido por sua diversidade, abrangendo uma ampla gama de estilos musicais que refletem diferentes características rítmicas, harmônicas e tímbricas. No entanto, estudos como Sturm (2013) destacaram certas limitações do GTZAN, como a presença de duplicatas e a possível falta de representatividade de alguns gêneros musicais mais modernos ou regionais. Apesar dessas limitações, o GTZAN continua sendo uma ferramenta valiosa para a pesquisa em classificação de gêneros musicais.

Figura 8 – Diversas frequências em um formato de Waveform de todos os gêneros musicais no conjunto de dados GTZAN.



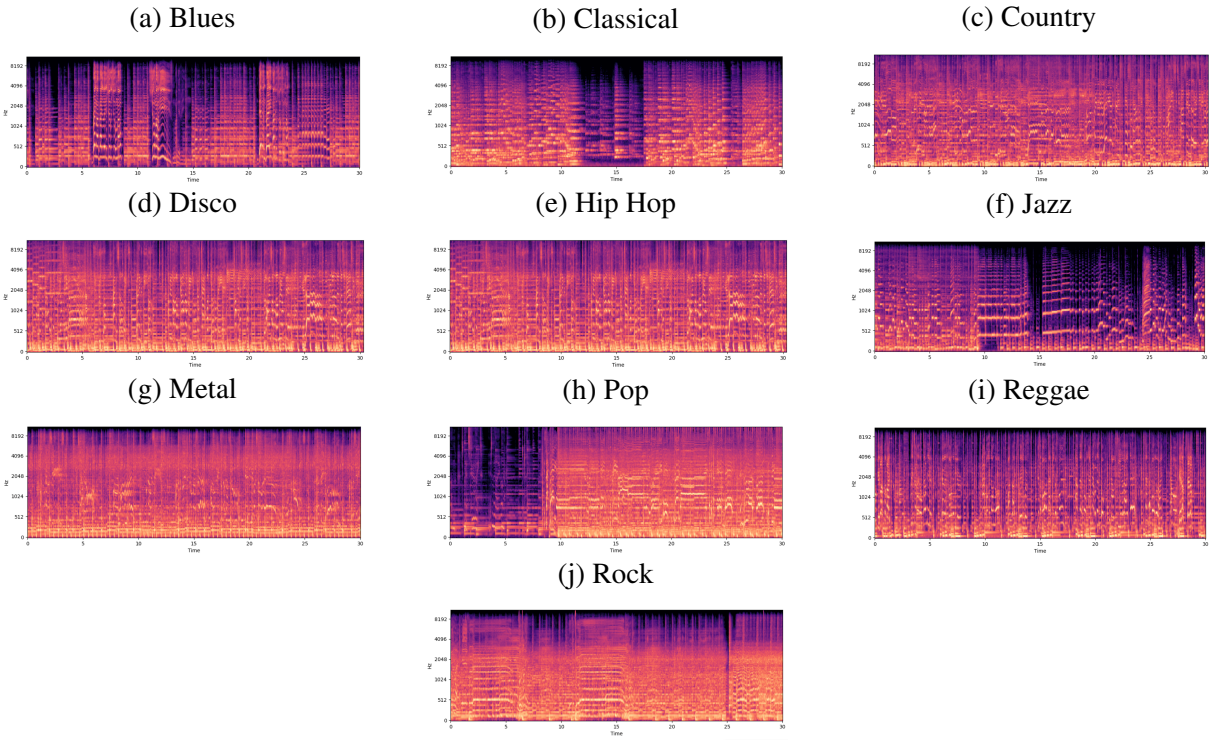
Fonte: Autoria própria.

O GTZAN tem sido amplamente utilizado para desenvolver e avaliar algoritmos de aprendizado de máquina para a classificação de gêneros musicais. CNN e MLP são comumente aplicados a este conjunto de dados devido à sua capacidade de aprender características complexas e não lineares dos dados de áudio.

4.3 Características de áudio

A extração e análise das características de áudio foram realizadas utilizando os coeficientes MFCC que são amplamente reconhecidos por sua eficácia em capturar informações relevantes sobre o timbre e a textura do som, essenciais para a classificação de gêneros musicais. As Figuras 9 e 10 apresentam faixas de áudio retiradas do conjunto de dados GTZAN no formato de Mel e MFCC.

Figura 9 – Exemplos do espectrograma em escala Mel de todos os gêneros muscais no GTZAN.



Fonte: Autoria própria.

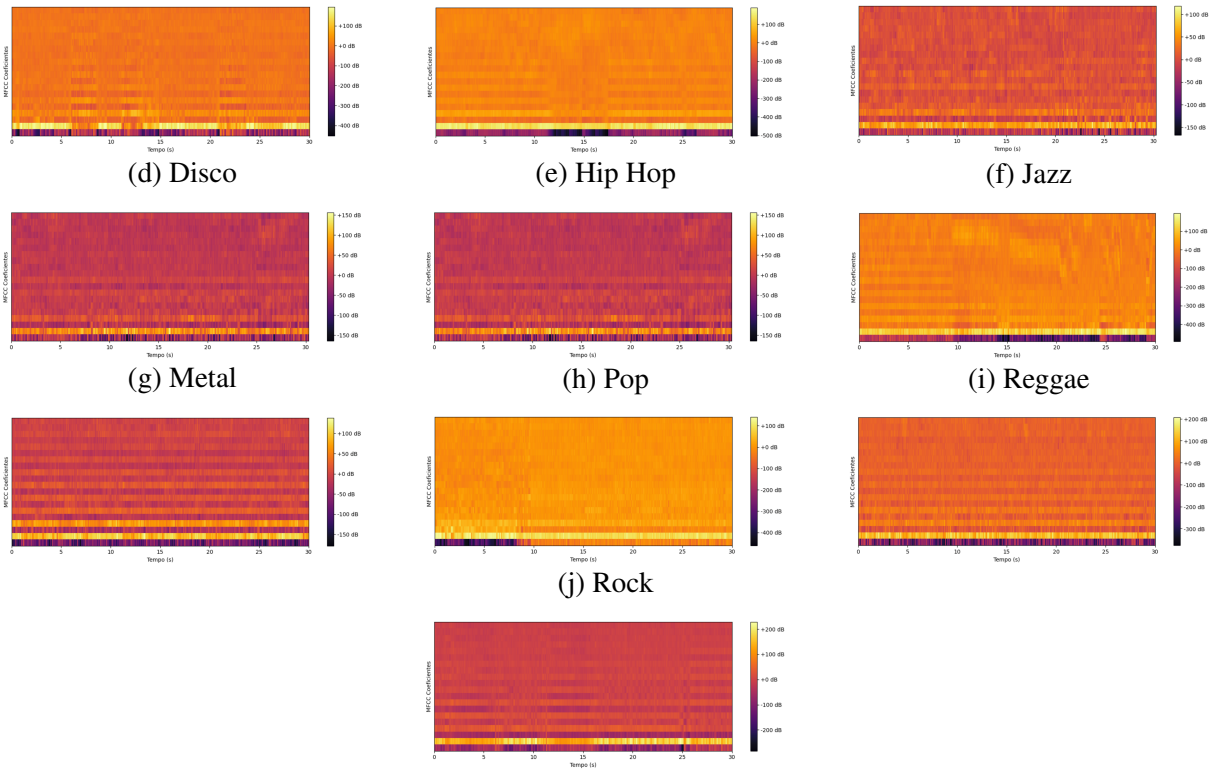
Os arquivos de áudio foram carregados utilizando a função librosa.load, que retorna a forma de onda do sinal e a taxa de amostragem. A partir desses sinais, os coeficientes MFCC foram extraídos usando a função librosa.feature.mfcc. Testou-se diferentes números de coeficientes MFCC (20, 15, 13, 10 e 5) para determinar a configuração que oferecia o melhor desempenho em termos de acurácia e qualidade dos clusters formados.

Dois métodos principais foram empregados para pré-processar os coeficientes MFCC. O primeiro método, denominado **achatamento (Flatten)**, no qual transformou o conjunto de coeficientes MFCC em um vetor unidimensional utilizando a função achatamento do Numpy. Essa técnica preservou a informação espectral em uma única dimensão, criando um vetor de características que foi utilizado como entrada para o modelo de RNA.

O segundo método calculou a média dos valores em cada linha dos coeficientes MFCC, gerando um **vetor de características médias**. Essa abordagem simplifica o vetor de entrada ao condensar as informações em uma única média representativa por coeficiente.

Figura 10 – Exemplos dos coeficientes MFCC de todos os gêneros muscais no GTZAN.

(a) Blues (b) Classical (c) Country



(d) Disco (e) Hip Hop (f) Jazz

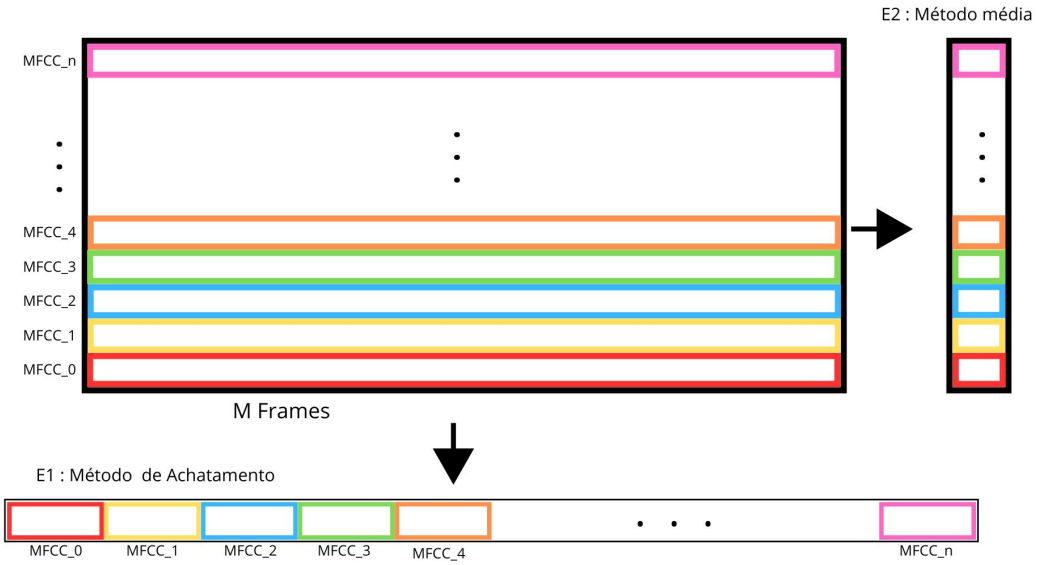
(g) Metal (h) Pop (i) Reggae

(j) Rock

Fonte: Autoria própria.

Na Figura 11 tem-se uma representação visual dessas duas técnicas e como foram extraídas as características de um espectrograma, com o método de média coletando apenas a

Figura 11 – Estratégias utilizadas para montagem do vetor de características de áudio.



Fonte: Autoria própria.

normalidade de uma linha MFCC, assim como o método de média transformou toda a linha de MFCC em um vetor de uma dimensão.

O código descrito no Algoritmo 1 dispõe como foi implementada a extração das características de áudio. A função **load_audio** utiliza a biblioteca Librosa para carregar os arquivos de áudio, retornando a forma de onda e a taxa de amostragem original do arquivo. Na sequência, a função **extract_features** implementa a extração das características de áudio por meio dos coeficientes MFCC, utilizando a função **librosa.feature.mfcc**. No código, o parâmetro **n_mfcc** representa a quantidade de coeficientes de MFCC que são extraídos, no qual o código está como **n_mfcc=3** mas durante os experimentos foram utilizados os coeficientes 5,10,13,15,20. Além disso, as duas abordagens foram aplicadas para o pré-processamento dos coeficientes, a abordagem **achatamento** e a de vetor de média.

Algoritmo 1: Extração de características do áudio.

```
def load_audio(file_path):
    y, sr = librosa.load(file_path, sr=None)
    return y, sr

def extract_features(y, sr, mode):
    mfccs = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=3)

    if (mode == 'media'):
        result = np.average(mfccs, axis=1)
    elif (mode == 'achatamento'):
        result = mfccs.flatten()
    else:
        result = []

    return result
```

Fonte: Autoria própria.

Para assegurar que todos os vetores de características tivessem o mesmo comprimento, foi utilizada a técnica de preenchimento com zeros (np.pad). Isso garantiu que todos os dados pudessem ser processados em *batch* pelo modelo de RNA. Após a padronização, os vetores de características foram armazenados em um *DataFrame* da biblioteca pandas, juntamente com suas respectivas etiquetas de gênero musical. A Tabela 2 mostra a variação dos dados de entradas gerado pelas duas técnicas propostas por neste trabalho, no qual mostra a variação de dimensões utilizando a técnica de achatamento, comparado ao de média que é constante com o coeficiente de MFCC. Essa variação de dimensão usando a estratégia de achatamento foi o motivo pelo preenchimento de zeros (np.pad), conforme descrito anteriormente. Os dados foram então divididos em conjuntos de treino e teste utilizando a função train_test_split da biblioteca

Sklearn, com 20% dos dados reservados para teste, permitindo uma avaliação confiável do desempenho do modelo em dados não vistos.

Tabela 2 – Variação na dimensão dos vetores de características.

Estratégia	**Dimensões**
Achatamento	$[16770, 33605]$
Médias	$\{5, 10, 13, 15, 20\}$

Fonte: Autoria própria.

4.4 Modelo de classificador

A classificação de gêneros musicais pode ser realizada utilizando técnicas de aprendizado de máquina, como RNA, portanto foi desenvolvido um modelo RNA para realizar essas classificações, assim como também um processo de extração e análise das características musicais do áudio.

4.4.1 Rede Neural Artificial

Para a construção do modelo de MLP, utilizou-se uma RNA *feedforward* composta por 3 (três) camadas densas e 1 (uma) camada de saída. A estrutura do modelo é descrita a seguir:

Inicialmente, definiu-se a camada de entrada, que recebeu o vetor de características extraídas do áudio. A dimensão da entrada foi ajustada para corresponder ao número de características após a padronização da entrada, como demonstrada na Tabela 2. Em seguida, implementou-se a primeira camada oculta com 256 (duzentas e cinquenta e seis) neurônios e uma função de ativação ReLU. Esta camada foi responsável por capturar interações complexas entre as características de entrada. Adicionou-se a segunda camada oculta, composta por 128 (centro e vinte e oito) neurônios também com função de ativação ReLU, que continuou a processar a informação das características aprendidas pela primeira camada oculta. Depois, incluiu-se a terceira camada oculta com 64 (sesenta e quatro) neurônios e função de ativação ReLU, que refinou as representações aprendidas para a classificação final. Finalmente, a camada de saída consistiu em 10 (dez) neurônios com função de ativação softmax, correspondendo aos 10 (dez) gêneros musicais. A função softmax converteu as saídas da rede em probabilidades de cada classe. A Figura 12 demonstra como ficou a arquitetura final da RNA.

O modelo foi compilado utilizando o otimizador Adam e a função de perda sparse categorical crossentropy adequada para problemas de classificação multiclasse. A métrica de acurácia foi utilizada para avaliar o desempenho do modelo durante o treinamento e a validação.

4.5 Projeção multidimensional

A construção e o treinamento do modelo MLP, utilizando o conjunto de dados GTZAN, possibilitaram a obtenção de dados referente a classificação de gêneros musicais. Nessa seção,

Figura 12 – Formato da arquitetura feita.

Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 256)	13,235,456
dense_5 (Dense)	(None, 128)	32,896
dense_6 (Dense)	(None, 64)	8,256
dense_7 (Dense)	(None, 10)	650

Total params: 13,277,258 (50.65 MB)
Trainable params: 13,277,258 (50.65 MB)
Non-trainable params: 0 (0.00 B)

Fonte: Autoria própria.

são explicadas os métodos utilizados para realizar a projeção multidimensional dos dados obtidos pelo modelo MLP.

4.5.1 Formação do vetor multidimensional

A formação do vetor multidimensional foi realizada com base na saída da penúltima camada oculta do modelo de rede neural desenvolvido. Esta abordagem foi escolhida devido à relevância das características aprendidas nas camadas ocultas, que representam abstrações de alto nível dos dados originais e são ideais para análise de agrupamentos e visualização, conforme Rauber et al. (2016). O código descrito no Algoritmo 2 foi utilizado para realizar essa formação do vetor.

Algoritmo 2: Formação do vetor da *Hidden Layer*.

```
1 # Obtendo a saida da penultima camada oculta do modelo
2 hidden_layer_model = tf.keras.Model(inputs=model.input, outputs=model.
      layers[-2].output)
3
4 # Extraindo as caracteristicas para o conjunto de treino e teste
5 train_features = hidden_layer_model.predict(X_train)
6 test_features = hidden_layer_model.predict(X_test)
```

Fonte: Autoria própria.

Inicialmente, foi criado uma instância de um modelo intermediário utilizando a API do TensorFlow/Keras. Esse modelo, denominado hidden_layer_model, foi configurado para ter como entrada os mesmos dados da rede neural original, chamada de model.input, e como saída a ativação da penúltima camada oculta, chamada model.layers[-2].output. Essa etapa permitiu acessar diretamente os valores das ativações produzidas na penúltima camada, preservando as informações processadas pelo modelo.

Logo em seguida, foram extraídas separadamente as características para os conjuntos de treino e teste. Por meio do método predict aplicado aos dados de entrada (X_train e X_test), os vetores de características multidimensionais foram gerados, correspondendo aos padrões identificados pela rede. Esses vetores servem como representação compacta e informativa de cada amostra de áudio no espaço de características de alta dimensão.

Por fim, os vetores extraídos foram utilizados como entrada para a técnica t-SNE, que os projetou para um espaço bidimensional. Essa projeção facilita a visualização das relações entre as amostras, como agrupamentos naturais e possíveis *outliers*, contribuindo para uma análise mais aprofundada do desempenho do modelo na tarefa de classificação de gêneros musicais. O código descrito no Algoritmo 3 apresenta como foi feito a formação do vetor para o t-SNE.

Algoritmo 3: Formação do vetor para a projeção da t-SNE.

```
# Ajustar e transformar separadamente com T-SNE
tsne_train = TSNE(n_components=2, random_state=42)
train_tsne_features = tsne_train.fit_transform(train_features)

tsne_test = TSNE(n_components=2, random_state=42)
test_tsne_features = tsne_test.fit_transform(test_features)
```

Fonte: Autoria própria.

4.5.2 Análise das Projeções Multidimensionais

Para analisar e visualizar as características extraídas da MLP, utilizou-se a técnica t-SNE. Esta técnica foi aplicada para projetar os dados de alta dimensionalidade em um espaço bidimensional, permitindo a visualização dos *clusters* formados pelos diferentes gêneros musicais.

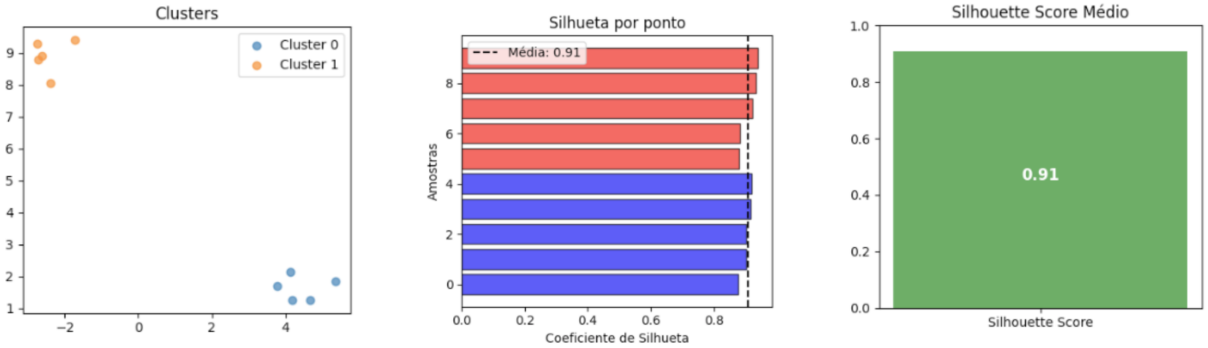
A técnica t-SNE foi escolhida por sua capacidade de preservar a estrutura local dos dados, o que foi útil para identificação dos agrupamentos naturais em dados de alta dimensionalidade. Os parâmetros utilizados na implementação do t-SNE foram os padrões fornecidos pela Sklearn, com exceção do número de iterações, que foi ajustado para garantir a convergência das projeções. Além disso, foi utilizada a métrica de *silhueta* para avaliar a qualidade das projeções t-SNE. A métrica de *silhueta* varia de -1 a 1, onde valores próximos a 1 indicam *clusters* bem definidos e valores próximos a -1 sugerem agrupamentos inadequados. Esta métrica foi calculada utilizando a função sklearn.metrics.silhouette_score, oferecendo uma análise complementar à acurácia, proporcionando uma avaliação mais abrangente da performance do modelo (PEDREGOSA et al., 2011).

Na Figura 13 é exemplificado o comportamento dessa métrica. A análise visual revela que os *clusters* encontrado no E1 apresentam uma separação clara, sem sobreposição significativa

entre os pontos. Isso se reflete nos valores de silhueta obtidos visto no E2, onde a maioria dos pontos exibe valores próximos de 1, indicando uma alta coesão interna e uma separação adequada entre os grupos. Além da análise visual, o gráfico de barras da silhueta visto no E3 confirma essa observação, pois a distribuição dos valores demonstra que os pontos estão corretamente atribuídos a seus respectivos grupos. A média da silhueta resultou em **0.91**, um indicativo de que o método de agrupamento aplicado foi eficiente para organizar os dados de maneira estruturada.

Figura 13 – Experimento de silhueta.

(a) E1 (b) E2 (c) E3



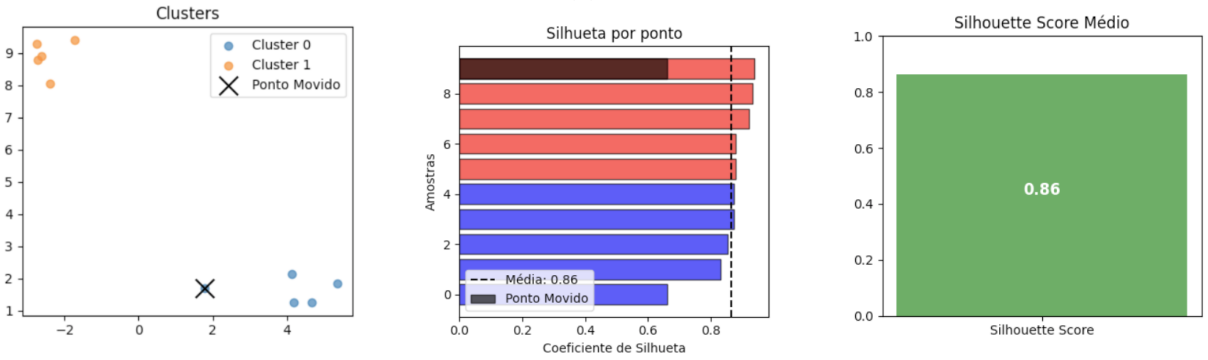
Fonte: Autoria própria.

Agora, a Figura 14 vemos o mesmo exemplo mas com uma alteração nos pontos, afastando um dos pontos do *cluster* 0 para mais próximo do *cluster* 1, como vemos em E4, assim, diminuindo o valor da silhueta individual do ponto movido, visto em E5, e também o valor geral médio, visto em E6.

Esses resultados reforçam a importância da métrica de silhueta como uma ferramenta complementar na avaliação da qualidade dos agrupamentos.

Figura 14 – Experimento de silhueta modificado.

(a) E4 (b) E5 (c) E6



Fonte: Autoria própria.

5 RESULTADOS E DISCUSSÃO

A partir dos modelos e técnicas discutidas nas seções anteriores, foi possível compilar resultados sobre a classificação de gênero musical utilizando RNA. Durante os testes realizados, um dos resultados obtidos foi utilizando o MFCC com 20 (vinte) coeficiente com a técnica de achatamento, com uma acurácia acima de 97% e resultados de silhueta maior que 0,65.

Realizou-se 10 (dez) testes com o modelo RNA proposto, com o intuito de analisar quais aspectos do modelo foram de maior importância para alcançar o melhor resultado, e também para a uma análise sobre gêneros musicais. A Tabela 3 mostra os resultados obtidos durante os testes realizados.

Tabela 3 – Resultados obtidos através da variação do número de coeficientes e do método de construção do vetor de características.

Número de coeficientes	Método	Acurácia	Silhueta
20	Achatamento	97.62%	**0,6567**
15	Achatamento	**99.25%**	**0,4792**
13	Achatamento	95.62%	**0,4036**
10	Achatamento	**98.37%**	0,3096
5	Achatamento	**98.87%**	0,0668
20	Média	95,92%	0,0373
15	Média	95,75%	0,0140
13	Média	95.62%	-0,0115
10	Média	94.34%	-0,0358
5	Média	91.29%	–0.1248

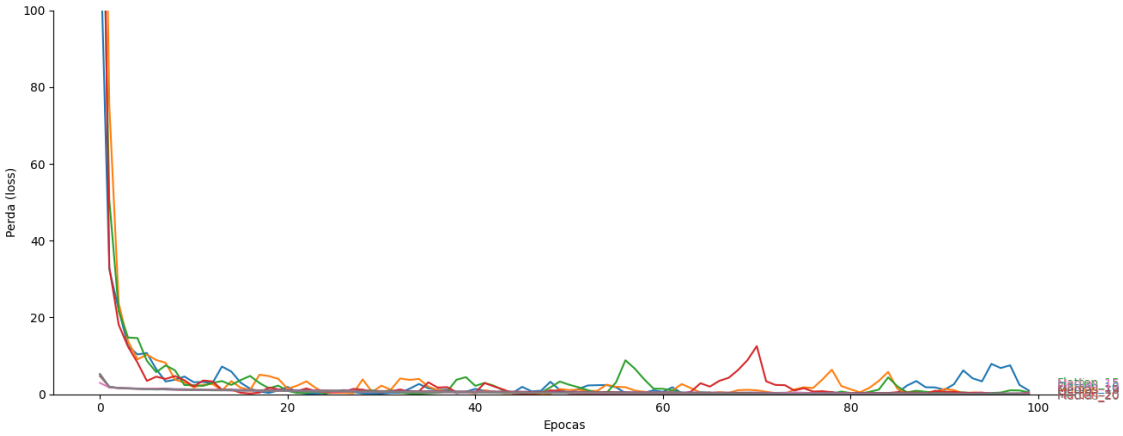
Fonte: Autoria Própria.

A acurácia foi utilizada como métrica de avaliação por causa que o conjunto de dados GTZAN é balanceado. O processo de separação de amostragem de treino/teste manteve o balanceamento das classes. Por mais que o valor de acurácia seja elevado em todas as instâncias e com ambas as técnicas, o valor de *silhueta* apresenta que a formação dos *cluster* não estão distintos uns dos outros, ou seja, a acurácia por mais que tenha importância no momento de avaliar se o modelo proposto está classificando corretamente os gêneros musicais. Sendo assim, a acurácia não é uma métrica válida para garantir a qualidade do modelo durante a visualização.

5.1 Função de Perda (*loss*)

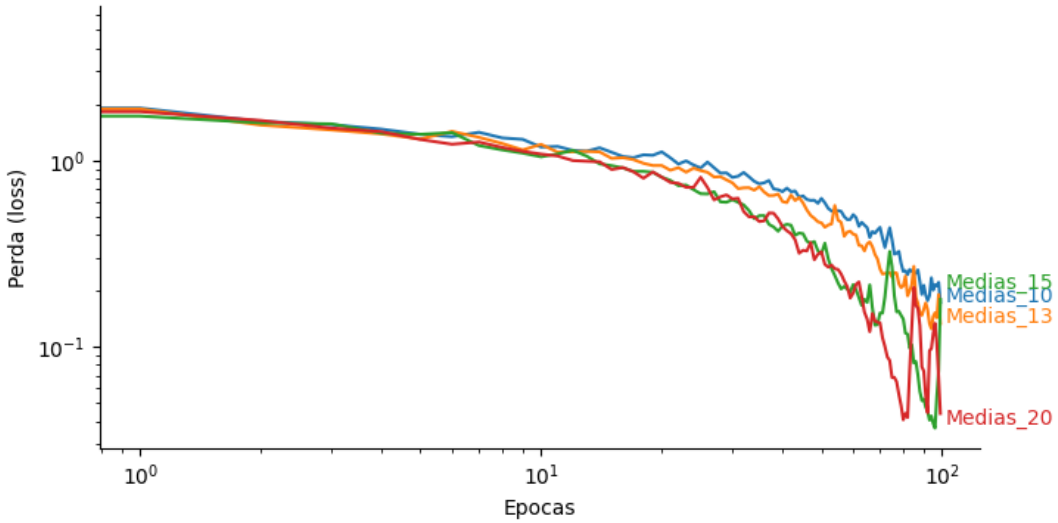
Para avaliar e visualizar a função de perda do modelo quando utilizado as configurações da RNA proposta, a Figura 15 demonstra a evolução da função de perda. Contudo, o formato linear dessa figura acabou não demonstrando as mudanças e o desenvolvimento com um maior detalhe. Enquanto que a Figura 16 permitiu uma visualização mais detalhada das mudanças de magnitude na perda durante o treinamento, devido ao gráfico no formato log-log. As linhas coloridas representam diferentes configurações de camadas da RNA, especificamente variações na camada de achatamento e combinações de médias.

Figura 15 – Análise dos valores de perda (loss) utilizando escala linear.



Fonte: Autoria própria.

Figura 16 – Análise dos Valores de Perda (Loss) ao Longo das Épocas.



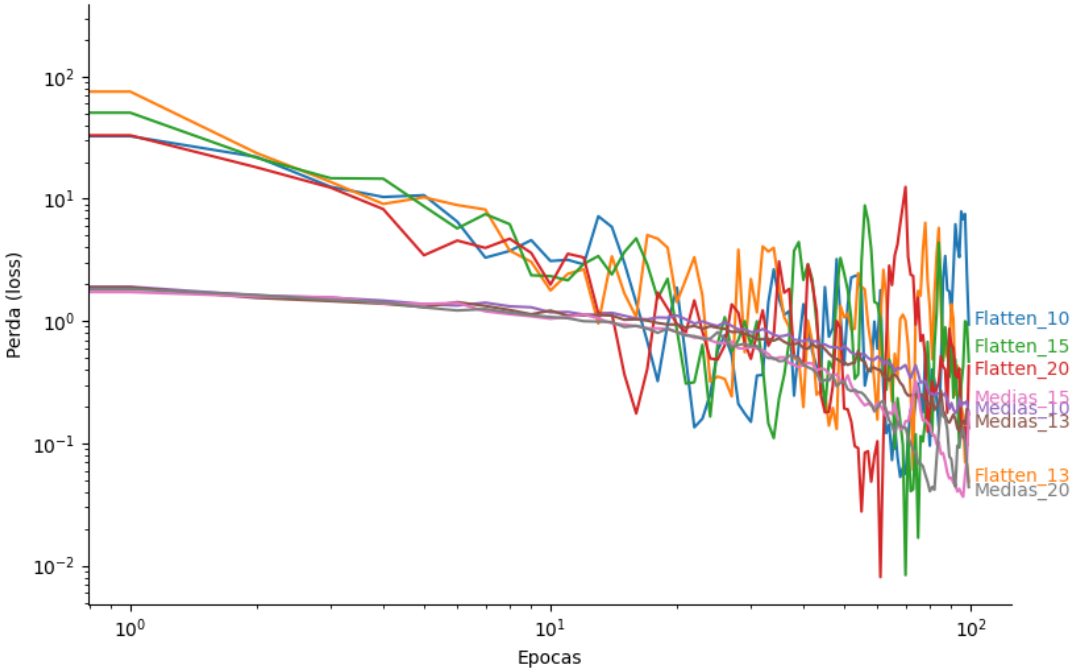
Fonte: Autoria própria.

A Figura 17 mostra a progressão da taxa de erro, a partir dos testes feitos com a técnica de achatamento. Ao utilizar essa técnica, os valores de perda (loss) diminuiram significativamente ao longo das épocas de treinamento, conforme os ajustes dos pesos realizados pelo modelo. Apesar de algumas flutuações observadas, especialmente em configurações como achatamento com 20 coeficientes MFCC, a tendência geral de queda indicou que o modelo está aprendendo e se adaptando bem, embora a convergência seja um pouco mais instável em comparação com outras configurações.

A Figura 18, apresenta os valores de perda ao longo das épocas utilizando a técnica de médias, exibindo uma curva de perda mais suave e consistente, sugerindo que essa técnica seja eficaz em suavizar as variações nos dados.

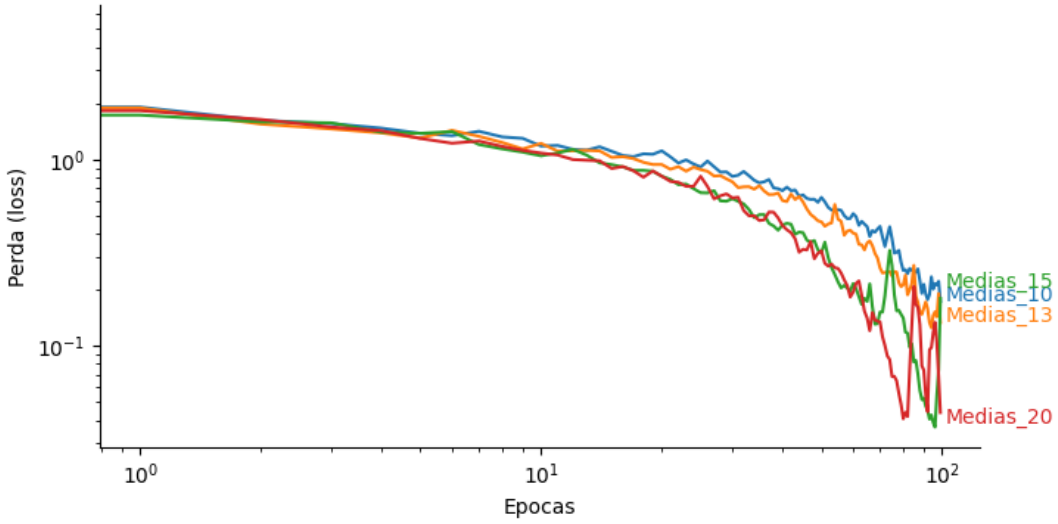
Diante do exposto, foi possível observar que, em todas as configurações, a perda inicia em um valor relativamente alto, progressivamente sofre uma redução conforme o aumento

Figura 17 – Análise dos valores de perda (loss) ao longo das épocas, somente da estratégia de achatamento, em escala logaritmica.



Fonte: Autoria própria.

Figura 18 – Análise dos valores de perda (loss) ao longo das épocas, somente da estratégia das médias, em escala logaritmica.



Fonte: Autoria própria.

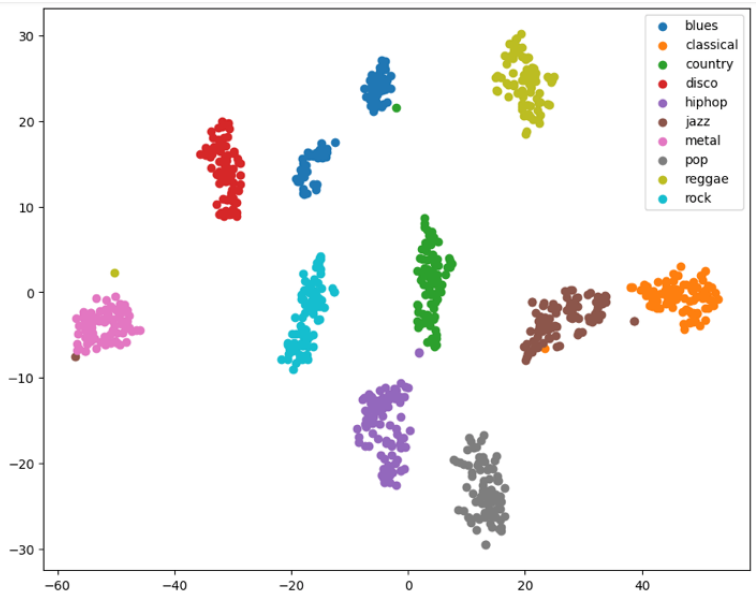
do número de épocas. Esse comportamento é esperado, pois a RNA ajustou seus pesos para minimizar o erro de predição com base nos dados de treinamento. As Figuras 17 e 18 indicam que algumas configurações (achatamento_10 e Médias_13) apresentam uma queda mais suave e consistente na perda, sugerindo uma convergência mais estável durante o treinamento. Vale ressaltar que configurações, como achatamento_20, apresentam flutuações maiores.

A análise da Figura 16 indica que diferentes configurações de achatamento e a aplicação de médias influenciaram diretamente a trajetória de aprendizado da RNA. Pressupõe que as configurações com menores valores de achatamento alcançaram uma convergência mais rápida e estável, enquanto as médias ajudaram a suavizar o ruído e identificar tendências subjacentes. No entanto, a redução dos valores de perda durante o treinamento da RNA não está necessariamente relacionada à qualidade da definição dos *clusters* quando os dados são projetados em 2D usando t-SNE.

5.1.1 Projeção das *Features*

Para entender melhor como o modelo separou os diferentes gêneros musicais, utilizou-se a técnica t-SNE para projetar as características extraídas em um espaço bidimensional. A Figura 19 apresenta o agrupamento dos gêneros musicais utilizando 20 coeficientes MFCC com a técnica de achatamento. Os *clusters* formados foram bem definidos, indicando uma separação eficaz dos gêneros musicais.

Figura 19 – Gráfico t-SNE com os testes realizados com coeficiente MFCC 20 e utilizando a técnica achatamento.

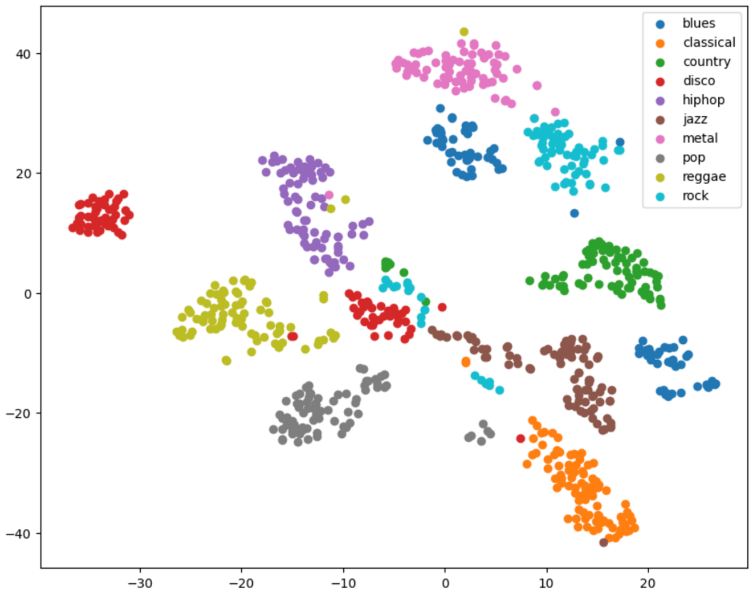


Fonte: Autoria própria.

Em contraste, a Figura 20 apresenta a projeção t-SNE utilizando 15 coeficientes MFCC com a técnica de achatamento. Embora a acurácia seja maior, os *clusters* são menos distintos, corroborando os resultados da métrica de *silhueta*. Isso reforça que a mudança no número de

coeficientes de MFCC favorece o treinamento, contudo diminui a separação entre os dados, deixando assim a projeção mais complexa de analisar.

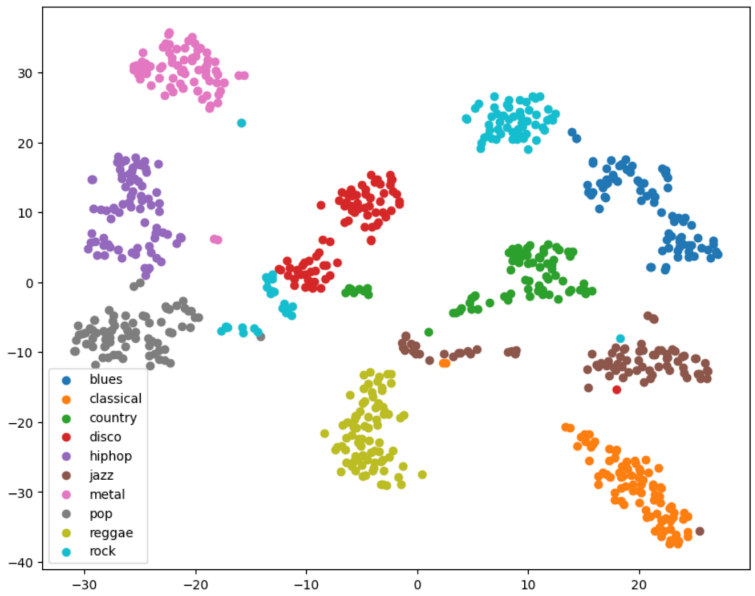
Figura 20 – Gráfico t-SNE com os testes realizados com coeficiente MFCC 15 e utilizando a técnica achatamento.



Fonte: Autoria própria.

As Figuras 21, 22 e 23 apresentam as projeções t-SNE para os coeficientes MFCC restantes (13, 10 e 5) da técnica de achatamento.

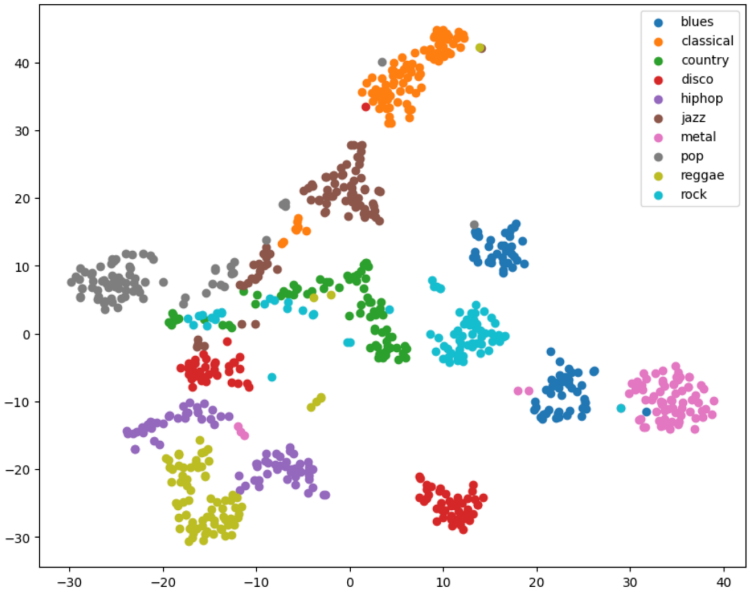
Figura 21 – Gráfico t-SNE com os testes realizados com coeficiente MFCC 13 e utilizando a técnica achatamento.



Fonte: Autoria própria.

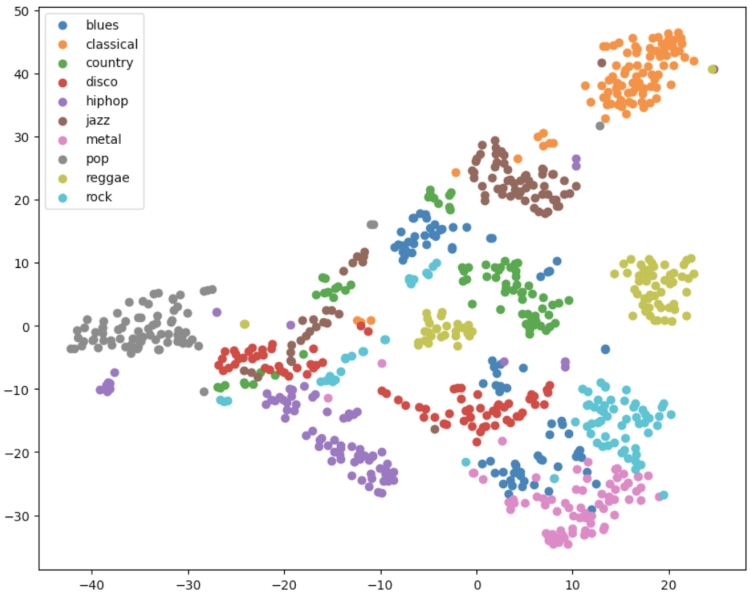
À medida que o número de coeficientes diminui, observa-se uma redução na qualidade da separação dos *clusters*, o que impacta negativamente a capacidade do modelo de distinguir entre os gêneros musicais.

Figura 22 – Gráfico t-SNE com os testes realizados com coeficiente MFCC 10 e utilizando a técnica achatamento.



Fonte: Autoria própria.

Figura 23 – Gráfico t-SNE com os testes realizados com coeficiente MFCC 5 e utilizando a técnica achatamento.

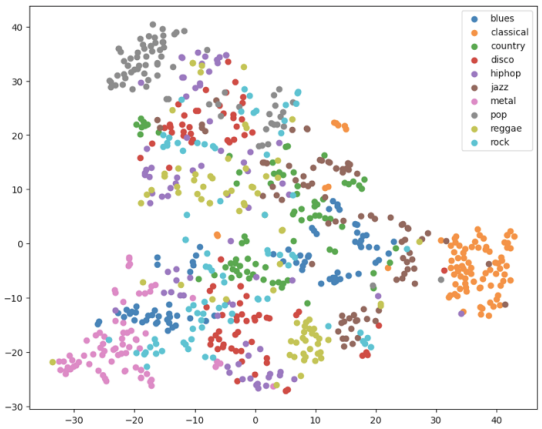


Fonte: Autoria própria.

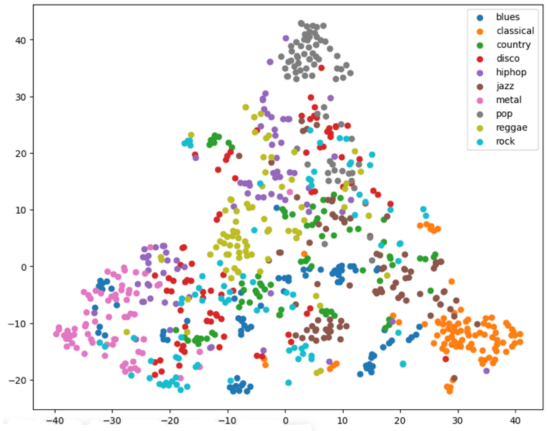
Foram geradas também projeções utilizando a técnica de média dos coeficientes MFCCs extraídos das músicas, cujos resultados podem ser vistos na Figura 24.

Figura 24 – Gráficos da t-SNE utilizando a técnica de média e utilizando 20, 15, 13, 10 e 5 coeficientes MFCC, respectivamente.

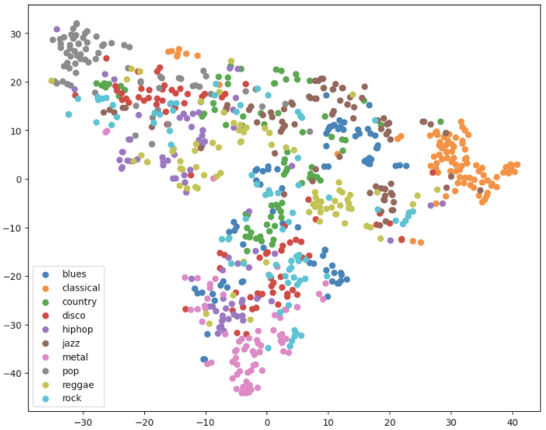
(a) 20



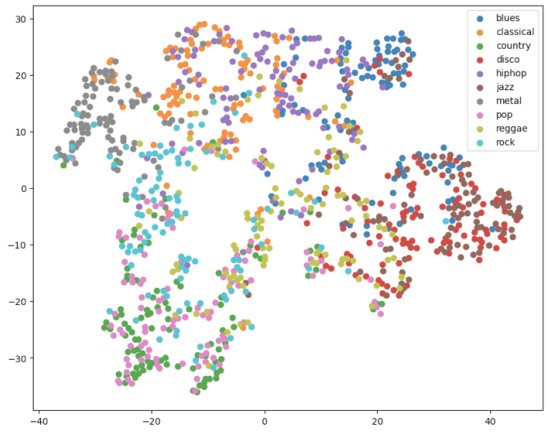
(b) 15



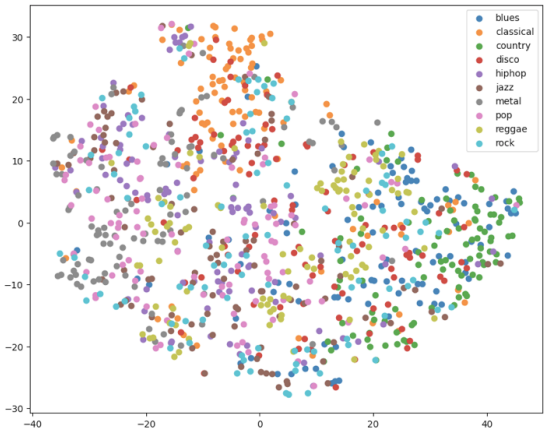
(c) 13



(d) 10



(e) 5



Fonte: Autoria própria.

5.1.2 Métricas de Consistência

A métrica de *silhueta* foi utilizada para avaliar a qualidade das projeções multidimensionais e a coerência dos clusters formados. Como demonstrado nas tabelas e figuras anteriores, o valor de *silhueta* para 20 coeficientes MFCC com a técnica de achatamento foi o mais elevado, indicando que essa configuração foi a que apresentou as melhores resultados dentro desta pesquisa para distinguir entre diferentes gêneros musicais.

Sendo assim, os resultados sugerem que a escolha do número de coeficientes MFCC é relevante para o sucesso do modelo, tanto em termos de acurácia quanto de separação dos *clusters*. A análise das projeções t-SNE, combinada com a métrica de *silhueta* proporcionou uma compreensão mais aprofundada do desempenho do modelo, permitindo identificar configurações que ofereceram um equilíbrio entre precisão e qualidade na separação dos gêneros musicais.

5.1.3 Analise de Caso

A partir das representações visuais criadas pelo t-SNE, analisou-se casos específicos para verificação das possíveis tomadas de decisões da RNA durante a classificação dos áudios de músicas presentes no conjunto de dados.

Observando a Figura 19, baseada no gráfico t-SNE utilizando os coeficientes MFCC 20 e a técnica de achatamento, observou-se uma proximidade entre músicas de *jazz* e *reggae* com o cluster principal de *metal*. Essa ocorrência pode ser atribuída a características compartilhadas, como timbres, padrões rítmicos ou elementos harmônicos. Por exemplo, músicas de *jazz* com forte presença de instrumentos de percussão e linhas de baixo marcadas podem compartilhar atributos acústicos com *metal*, especialmente se o método de extração MFCC não capturar adequadamente as nuances melódicas e harmônicas mais distintas. Além disso, é possível analisar outro caso: uma música do gênero *classical* situada próxima do cluster de *jazz*. Além disso, uma música de *jazz* aparece em uma posição intermediária entre os clusters de *jazz* e *classical*. Esses casos podem ser explicados por semelhanças na estrutura tonal ou na complexidade harmônica, características presentes em ambos os gêneros. É possível que a música *classical* próxima ao *jazz* contenha elementos improvisatórios ou uma instrumentação similar, enquanto a música *jazz* intermediária pode ter adotado formas composicionais mais rígidas, típicas da música *classical*.

Embora essas anomalias possam ser classificadas como "erros"em métricas de acurácia, a visualização com t-SNE destaca que tais classificações muitas vezes refletem características reais compartilhadas entre os gêneros.

Agora, analisando a Figura 24a, no qual apresenta uma visualização t-SNE baseada em 20 coeficientes MFCC do banco de dados GTZAN, utilizando a técnica de média, revelou-se que gêneros como *classical* e *pop* formam clusters bem definidos, devido às suas características acústicas como timbres e padrões harmônicos distintos. Em contraste, gêneros como *rock*, *metal* e *blues* apresentam uma sobreposição significativa, o que reflete elementos comuns entre eles, como o uso de guitarras elétricas, ritmos marcados e estruturas harmônicas similares. O gênero

disco se destaca dos demais pela formação de dois clusters distintos, sugerindo a existência de subgrupos dentro do estilo, possivelmente influenciados por variações em ritmo, instrumentação ou técnicas de gravação. As semelhanças entre gêneros musicais, como *rock*, *metal* e *blues*, podem ser explicadas por compartilharem características estruturais e instrumentais, tornando mais complexa a distinção baseada apenas nos coeficientes MFCC.

A técnica de média contribuiu para a suavização dos dados e a formação de clusters, mas também ressaltou essas diferenças internas em alguns gêneros. A partir disso, sugere-se trabalhos futuros para realizar uma análise aprofundada de cada situação apresentada acima.

5.2 Comparação entre o Desempenho do Modelo e Trabalhos Relacionados

A partir dos resultados obtidos pelos testes apresentados neste trabalho, propõe-se uma análise comparativa com os trabalhos de Patil e Nemade (2017) e Xu (2023). Ambos os trabalhos foram descritos no **Capítulo 3**.

Patil e Nemade (2017)

O trabalho de Patil e Nemade (2017) propôs a classificação de gêneros musicais utilizando um vetor de características com 28 (vinte e oito) dimensões, criado a partir das técnicas de extração de características disponíveis na biblioteca Librosa. Esse vetor inclui: a média de 13 (treze) coeficientes MFCCs, 1 (um) *Chroma Frequency*, 12 (doze) *Spectral Centroids*, 1 (um) *Spectral Roll-off* e 1 (um) *Zero Crossing Rate*. Com este vetor, os autores avaliaram classificadores como K-NN e SVM, alcançando uma acurácia de 78% com SVM e kernel polinomial no conjunto de dados GTZAN.

A partir desses estudos, realizou-se um novo experimento, adaptando o vetor de características com os ajustes no cálculo do Spectral Centroid. Para tanto, optou-se pela media simples, mantendo a essência do vetor original. Como resultado tem-se uma acurácia de treino de 49,00%, inferior a acurácia de 99,25% obtida no presente trabalho.

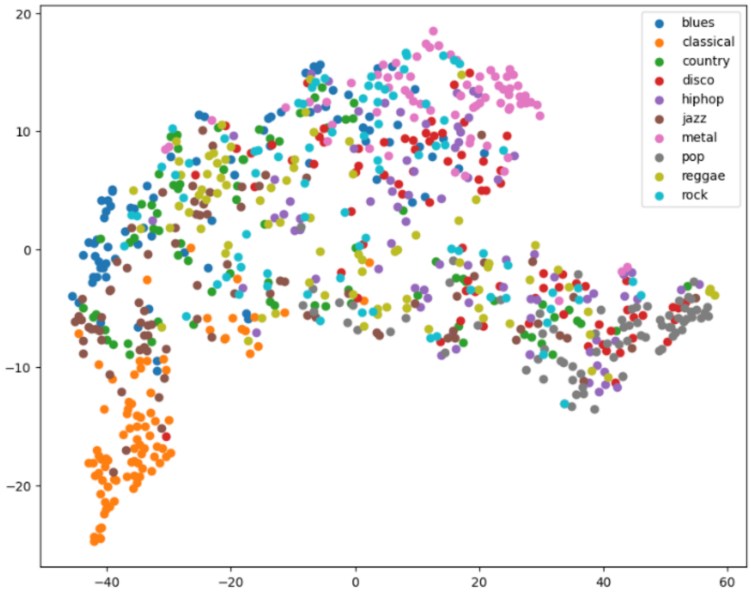
A projeção t-SNE dos clusters resultantes, ilustrada na Figura 25, apresenta uma dispersão elevada, indicando maior dificuldade do modelo em encontrar separabilidade entre os gêneros musicais com o vetor proposto por Patil e Nemade (2017). Utilizando a métrica de *silhoute*, foi chegado ao resultado de 0.18, também inferior aos melhores resultados obtidos no presente trabalho.

Os resultados dos testes realizados sugerem que o vetor de características proposto pelo trabalho do Patil e Nemade (2017) não é eficaz quando utiliza-se apenas o MFCC e uma maior quantidade de coeficientes, nas condições descritas no seu trabalho.

Xu (2023)

O trabalho de Xu (2023) comparou diferentes algoritmos de aprendizado de máquina, como SVM e CNN, na classificação de gêneros musicais. Embora as CNNs tenham apresentado

Figura 25 – Gráfico t-SNE com os testes realizados utilizando a metodologia proposta pelo trabalho de Patil e Nemade (2017).



Fonte: Autoria própria.

resultados promissores, os resultados de Xu (2023) foram limitados pelo uso de apenas 2 (dois) segundos de áudio de cada música do conjunto de dados GTZAN, reduzindo significativamente a quantidade de informações temporais que o modelo poderia captar.

Em contraste, o MLP utilizado neste trabalho, com base em MFCCs extraídos de trechos maiores (30 segundos), apresentou um desempenho superior. Isso sugere que, apesar da simplicidade do MLP em comparação com CNNs, sua combinação com MFCCs, que são uma representação mais compactas do áudio, revela-se eficaz para a tarefa de classificação de gêneros musicais. A limitação de Xu (2023) no uso de espectrogramas de curta duração pode ter contribuído para o desempenho inferior dos seus resultados. Em contrapartida, o MLP deste trabalho treinado com dados temporais mais extensos obteve características musicais eficazes, tendo como resultado, uma acurácia maior.

6 CONCLUSÃO

O objetivo deste trabalho foi utilizar a visualização de informação para investigação da classificação supervisionada de gêneros musicais, a partir da aplicação de RNA, com MLP. O estudo fez uso dos MFCCs como características principais, o que permitiu a extração de informações sonoras relevantes e viabilizou a análise e classificação por meio do modelo de RNA.

Os resultados dos experimentos indicaram que a escolha e o ajuste do número de coeficientes MFCCs influenciaram significativamente o desempenho do modelo. Notou-se que o uso de 15 (quinze) coeficientes MFCC, utilizando a técnica de achatamento, resultou na maior precisão de 99,25%. Enquanto a análise de *silhouette* revelou uma melhor separação de *clusters* com 20 (vinte) coeficientes, utilizando a técnica de achatamento. Demonstrando que, por mais que a acurácia tenha sua importância na classificação, a visualização no t-SNE apresenta de forma geral como os dados ficaram agrupados e distintos uns aos outros na visualização. Evidenciando assim, a necessidade de um equilíbrio entre acurácia e qualidade de agrupamento. O modelo de RNA desenvolvido apresentou um resultado satisfatório quanto a capacidade de generalização e precisão na classificação dos gêneros musicais. Dentre as configurações testadas, incluindo o número de neurônios e camadas, observou-se a necessidade de um ajuste fino dos parâmetros para maximizar o desempenho do modelo.

A técnica t-SNE mostrou-se eficiente para visualizar os *clusters* formados pelos diferentes gêneros musicais, reforçando a eficácia das RNAs na tarefa de classificação. Além disso, a análise multidimensional foi fundamental para interpretar a capacidade do modelo de generalizar padrões sonoros, e a utilização da XAI proporcionando uma melhor compreensão dos resultados, garantindo transparência no processo de tomada de decisão.

Conclui-se que o modelo de RNA desenvolvido possui potencial para aplicação em sistemas de classificação de áudio. E a metodologia aplicada pode ser expandida para outras áreas como recomendação de áudio. Tendo em vista o contínuo crescimento das plataformas de *streaming* de música, e a importância de entender melhor os processos de classificação automática, pois esse entendimento não apenas contribui para o desenvolvimento de novas técnicas, mas também melhora a eficiência na categorização de grandes volumes de dados musicais, algo fundamental para atender às necessidades dos usuários.

Para trabalhos futuros, recomenda-se o uso de conjunto de dados diversificados e o teste de arquiteturas complexas, como CNN, para melhorar ainda mais o desempenho e a robustez do modelo. Além disso, para explorar e buscar novas metodologias realizadas por outros autores, e também para expandir e realizar novas análises aprofundadas em pesquisas já referenciadas nesse trabalho, a exemplo do trabalho de Patil e Nemade (2017) e Xu (2023).

Referências

ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>. Citado na página 18.

ALBOANEEN, D. A.; TIANFIELD, H.; ZHANG, Y. Glowworm swarm optimisation for training multi-layer perceptrons. In: *Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. [S.l.: s.n.], 2017. p. 131–138. Citado na página 10.

AYVAZ, U. et al. Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning. *Computers, Materials & Continua*, 2022. Disponível em: <https://api.semanticscholar.org/CorpusID:246055008>. Citado 4 vezes nas páginas 5, 6, 13 e 14.

BANERJEE, C.; MUKHERJEE, T.; JR, E. P. An empirical study on generalizations of the relu activation function. In: *Proceedings of the 2019 ACM Southeast Conference*. [S.l.: s.n.], 2019. p. 164–167. Citado na página 10.

BAŞTANLAR, Y.; ÖZUYSAL, M. Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, Springer, p. 105–128, 2014. Citado na página 8.

BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. v. 4. Citado na página 8.

BLAKE, D. K. Timbre as differentiation in indie music. *Music Theory Online*, v. 18, n. 2, 2012. Citado na página 3.

CABRAL, D. dos S.; CORRÊA, L. J.; NETO, I. P. F. A importância da música como instrumento do desenvolvimento da aprendizagem da criança na educação infantil. *REVISTA FOCO*, v. 16, n. 10, p. e3251–e3251, 2023. Citado na página 3.

CHENG, Y.-H.; KUO, C.-N. Machine learning for music genre classification using visual mel spectrum. *Mathematics*, MDPI, v. 10, n. 23, p. 4427, 2022. Citado 2 vezes nas páginas 6 e 7.

CHOLLET, F. et al. *Keras*. 2015. <https://keras.io>. Citado na página 18.

CONSTANTINESCU, C.; BRAD, R. An overview on sound features in time and frequency domain. *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences*, v. 13, n. 1, p. 45–58, 2023. Citado na página 4.

CORDERO, O. H. G. A música, o ritmo e a educação física. *Revista Científica da Faculdade de Educação e Meio Ambiente*, v. 5, n. 2, p. 173–186, 2014. Citado na página 4.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM New York, NY, USA, v. 55, n. 10, p. 78–87, 2012. Citado na página 8.

GHODASARA, V. et al. Speech/music classification using block based mfcc features. *Music Information Retrieval Evaluation eXchange (MIREX)*, 2015. Citado 3 vezes nas páginas 6, 14 e 15.

GUNNING, D. et al. Xai—explainable artificial intelligence. *Science robotics*, American Association for the Advancement of Science, v. 4, n. 37, p. eaay7120, 2019. Citado na página 2.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 18.

HOWARD, B. E. et al. Swift-active screener: Accelerated document screening through active learning and integrated recall estimation. *Environment International*, Elsevier, v. 138, p. 105623, 2020. Citado na página 9.

KULSKI, J. K. Next-generation sequencing — an overview of the history, tools, and "omic" applications. In: KULSKI, J. K. (Ed.). *Next Generation Sequencing*. Rijeka: IntechOpen, 2016. cap. 1. Disponível em: <https://doi.org/10.5772/61964>. Citado na página 9.

LIPTON, Z. The mythos of model interpretability. *Communications of the ACM*, v. 61, 10 2016. Citado na página 11.

LIU, Y. et al. A strategy on selecting performance metrics for classifier evaluation. *International Journal of Mobile Computing and Multimedia Communications*, v. 6, p. 20–35, 10 2014. Citado na página 8.

LOGAN, B. et al. Mel frequency cepstral coefficients for music modeling. In: PLYMOUTH, MA. *Ismir*. [S.l.], 2000. v. 270, n. 1, p. 11. Citado na página 14.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. 11, 2008. Citado na página 12.

MCFEE, B. et al. *librosa/librosa: 0.10.2.post1*. Zenodo, 2024. Disponível em: <https://doi.org/10.5281/zenodo.11192913>. Citado na página 18.

MCLOUGHLIN, I. V. *Speech and Audio Processing: a MATLAB-based approach*. [S.l.]: Cambridge University Press, 2016. Citado na página 4.

MORAIS, F.; BRANCO, V. A inteligência artificial: conceitos, aplicações e controvérsias. *SIMPÓSIO INTERNACIONAL DE CIÊNCIAS INTEGRADAS DA UNAERP*, v. 20, 2023. Citado na página 7.

MOYLAN, W. *Understanding and crafting the mix: The art of recording*. [S.l.]: Routledge, 2014. Citado na página 3.

PATIL, N. M.; NEMADE, M. U. Music genre classification using mfcc, k-nn and svm classifier. *International Journal of Computer Engineering In Research Trends*, v. 4, n. 2, p. 43–47, 2017. Citado 6 vezes nas páginas , 15, 16, 35, 36 e 37.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 17 e 25.

PERLOVSKY, L. *Origin of music and embodied cognition*. [S.l.]: Frontiers Media SA, 2015. 538 p. Citado 2 vezes nas páginas 3 e 4.

POPESCU, M.-C. et al. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point . . . , v. 8, n. 7, p. 579–588, 2009. Citado 2 vezes nas páginas 9 e 10.

POWERS, D. Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. *Mach. Learn. Technol.*, v. 2, 01 2008. Citado na página 9.

PURWINS, H. et al. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, v. 13, n. 2, p. 206–219, 2019. Citado na página 8.

RAUBER, P. E. et al. Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics*, IEEE, v. 23, n. 1, p. 101–110, 2016. Citado 4 vezes nas páginas 9, 11, 13 e 24.

ŘEZANKOVÁ, H. Different approaches to the silhouette coefficient calculation in cluster evaluation. In: *21st international scientific conference AMSE applications of mathematics and statistics in economics*. [S.l.: s.n.], 2018. p. 1–10. Citado na página 12.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. [S.l.]: Pearson, 2016. Citado na página 7.

SHARMA, N.; SHARMA, R.; JINDAL, N. Machine learning and deep learning applications-a vision. *Global Transitions Proceedings*, v. 2, n. 1, p. 24–28, 2021. ISSN 2666-285X. 1st International Conference on Advances in Information, Computing and Trends in Data Engineering (AICDE - 2020). Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666285X21000042>. Citado na página 8.

SPADINI, T. *Generative Adversarial Networks para Aprimoramento de Áudio e Voz*. Tese (Doutorado) — Checar, 02 2020. Citado na página 5.

STURM, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013. Citado na página 19.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, IEEE, v. 32, n. 11, p. 4793–4813, 2020. Citado na página 9.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, IEEE, v. 10, n. 5, p. 293–302, 2002. Citado 5 vezes nas páginas 4, 5, 14, 15 e 18.

XU, Y. A comparison of machine learning algorithms for music genre recognition. *Applied and Computational Engineering*, v. 8, p. 568–573, 08 2023. Citado 6 vezes nas páginas , 15, 16, 35, 36 e 37.

XU, Y. et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, Elsevier, v. 2, n. 4, 2021. Citado na página 7.

YANG, H.; ZHANG, W.-Q. Music Genre Classification Using Duplicated Convolutional Layers in Neural Networks. In: *Proc. Interspeech 2019*. [S.l.: s.n.], 2019. p. 3382–3386. ISSN 2958-1796. Citado na página 1.

ZHANG, W. et al. Improved music genre classification with convolutional neural networks. In: *Interspeech*. [s.n.], 2016. Disponível em: <https://api.semanticscholar.org/CorpusID:27853460>. Citado na página 1.