UNIVERSIDADE FEDERAL DE SÃO CARLOS
DEPARTAMENTO DE COMPUTAÇÃO
ENGENHARIA DE COMPUTAÇÃO

Micael Valterlânio da Silva

Análise comparativa entre algoritmos de classificação de gênero musical baseados em diferentes representações visuais e transferência de aprendizado

São Carlos - SP

2023

Micael Valterlânio da Silva

Análise comparativa entre algoritmos de classificação de gênero musical baseados em diferentes representações visuais e transferência de aprendizado

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação da Universidade Federal de São Carlos, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientação Prof. Dr. Alan Demétrius Baria Valejo

São Carlos - SP

2023

Dedico este trabalho aos meus pais, Valter e Nécia, à minha irmã Micaelly, ao Milk, à minha namorada Deborah, à Neuza, à todas as pessoas que me acompanharam nessa jornada até aqui, que me apoiaram incondicionalmente, que riram e choraram comigo, que foram uma companhia, um abraço, um ombro amigo e o meu porto seguro quando eu mais precisei.

Agradecimentos

Agradeço primeiramente a Deus, que me deu saúde, forças, e conduziu a minha vida para que fosse possível ter todas as oportunidades que eu tive. Agradeço, imensamente, aos meus pais, Valter e Nécia, e à minha irmã Micaelly, pelo apoio incondicional, por todo o amor e paciência ao longo de todas as etapas da minha vida. Eles são os principais responsáveis pela minha formação acadêmica. Eu não tenho palavras para agradecer todo o esforço que vocês fizeram por mim ao longo de todos esses anos. Vocês são a razão pela qual eu luto diariamente, e sem vocês nada disso seria possível. Agradeço aos meus pais, Valter e Nécia, que se doaram de todas as formas possíveis para me criar e educar, sendo os melhores pais que alguém poderia ter a sorte de ter. Agradeço à minha irmã Micaelly, que foi a minha companheira em todos os momentos da vida, me corrigindo quando eu errei, me ouvindo, me apoiando, me ensinando, sendo um ponto de segurança quando eu mais precisei.

Agradeço à minha segunda mãe Neuza, que, juntamente com a minha mãe, orou por mim em todos os momentos da minha vida, que torceu por mim e sempre me apoiou nas minhas lutas.

Agradeço aos meus amigos que me acolheram durante a graduação, que me fizeram companhia, que dividiram os desafios de todos os anos de graduação comigo, que me ajudaram a refrescar a cabeça e a relembrar que tudo fazia parte de um processo.

Agradeço aos professores da UFSCar que foram essenciais para a minha formação, que me ensinaram, alguns de forma mais traumática, outros de uma forma mais calma e inspiradora, de forma a fomentar o desejo e a curiosidade pelo conhecimento. Agradeço aos professores da UFSCar, pois foi aqui onde aprendi a estudar, descobri quais métodos funcionam melhor para mim e aprendi a superar muitos limites que eu nem imaginava ter. Estou saindo da universidade como um pessoa mais completa em todos os âmbitos, com experiências ricas que me transformaram em uma pessoa melhor.

Também agradeço imensamente ao professor Dr. Alan Demétrius Baria Valejo pela orientação ao longo de todo trabalho e por todos os ensinamentos. Muito obrigado professor por toda a ajuda e orientação!

"Foi o Senhor que fez isto, e é coisa maravilhosa aos nossos olhos."
(Salmos 118:23)

Resumo

Organizar e recuperar informação musical automaticamente é uma atividade altamente requerida. Rotular músicas com informações que a descrevam de modo sucinto possui implicações nessas e em outras tarefas relacionadas. Uma das abordagens mais utilizadas para se rotular gravações musicais é por meio da informação de gênero. No entanto, esta tarefa é bastante difícil. Nos últimos anos, a literatura tem mostrado um avanço significativo nessa tarefa ao se aplicar algoritmos de aprendizado de máquina baseados em redes neurais profundas (RNPs). Nesse cenário, a prática comumente adotada é utilizar representações visuais de tempo-frequência do áudio como entrada para uma RNP. Diante disso, o objetivo deste trabalho é realizar uma análise comparativa entre o impacto da utilização de variadas representações visuais de música como, Espectrograma, Mel-espectrograma, Cromagrama, Tempograma e Tonnetz, obtidas a partir de seu áudio, e a transferência de aprendizado na classificação de gêneros por meio de RNPs. Serão apresentados os conhecimentos que embasam toda a pesquisa e os resultados obtidos, bem como uma análise dos resultados, para que seja possível compreender os processos que levaram a abordagem utilizando transferência de aprendizado ter resultados melhores e mais consistentes com relação a abordagem utilizando diferentes representações visuais.

Palavras-chave: Aprendizado da Máquina; Redes Neurais Profundas; Transferência de Aprendizado; Representações Visuais de Áudios.

Abstract

Organizing and retrieving musical information automatically is a highly demanded activity. Labeling music with information that succinctly describes it has implications for these and other related tasks. One of the most widely used approaches to label musical recordings is through genre information. However, this task is quite challenging. In recent years, the literature has shown significant progress in this task by applying machine learning algorithms based on deep neural networks (DNN). In this scenario, the commonly adopted practice is to use time-frequency visual representations of audio as input for a DNN. Therefore, the aim of this work is to perform a comparative analysis between the impact of using various visual representations of music, such as Spectrogram, Mel-spectrogram, Chromagram, Tempogram, and Tonnetz, obtained from its audio, and transfer learning in genre classification through DNNs. This research will present the foundational knowledge and acquired results, as well as an analysis of the outcomes. This analysis aims to understand the processes that contributed to the transfer learning approach outperforming the use of various visual representations in achieving better and more consistent results.

Keywords: Machine Learning; Deep Neural Networks; Transfer Learning; Visual Representations of Audio;

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Os grandes avanços tecnológicos das últimas décadas possibilitaram que a sociedade adquirisse novos hábitos com relação à utilização de serviços digitais, impactando diretamente a forma como as pessoas consomem música. Com o surgimento de grandes plataformas como YouTube[1], Spotify[2] e Deezer[3], armazenando e manipulando uma quantidade enorme de dados, se torna indispensável pensar no modo como é feita a organização e recuperação de toda essa quantidade de informação musical através de métodos computacionais. Uma das abordagens utilizadas ao lidar com a tarefa de organizar coleções de músicas é rotular cada uma das músicas com uma informação que a represente resumidamente, de modo que elas possam ser diretamente associadas a outras músicas que possuam essa mesma característica.

Uma das formas mais usuais de se rotular uma música é por meio de sua informação de gênero. A partir dessa informação é possível que as plataformas de reprodução online possam organizar seu acervo associando artistas e músicas com perfis parecidos. Essa forma de organizar os dados nos leva a possibilidades interessantes, como um sistema de recomendação mais eficiente, criação automática de playlists, meios de traçar um perfil de usuário de maneira mais assertiva, entre outras (KNEES; SCHEDL, 2015; WAN, 2016; LAMERE, 2008).

Entretanto, a classificação do gênero de uma música é uma tarefa difícil até para os humanos, devido a subjetividade à que a informação de gênero está associada. Além disso, é possível que músicas de gêneros diferentes soem similares em vários aspectos. Por exemplo, a formação das bandas de rock e de blues são muito parecidas, com instrumentos manipulados buscando timbres muito similares. Outro exemplo é a formação de trio, muito utilizada no Brasil usando Piano, Contrabaixo e Bateria, que pode ser facilmente encontrada tocando gêneros como MPB, Jazz e Bossa Nova.

Nos últimos anos, pesquisadores têm buscado mitigar esses problemas utilizando fontes de dados alternativas ao áudio da música que se deseja categorizar (SCHEDL et al., 2014). Um exemplo é a classificação multimodal, onde combinam-se diferentes tipos de dados, como áudio, imagem e texto, ou diferentes características para melhorar os resultados obtidos (ORAMAS et al., 2017; NANNI et al., 2017; NANNI et al., 2018). Ao se utilizar diversos tipos de dados de entrada, os resultados obtidos são comumente muito mais assertivos comparado aos obtidos a partir de apenas um tipo de entrada.

Apesar de ser uma abordagem bastante válida e ter se mostrado eficaz, nem

[1] <www.youtube.com>
[2] <www.spotify.com>
[3] <www.deezer.com>

sempre estão disponíveis diversos tipos de dados para a mesma música. Mesmo quando há disponibilidade desses dados, usar diferentes fontes pode tornar a tarefa mais custosa. Por esse e outros motivos, há diversos trabalhos em classificação de gêneros musicais que se limitam a utilizar apenas um tipo de dado. Indubitavelmente, o áudio é o tipo de dado mais utilizado nesta tarefa.

Nesse cenário, muitos dos trabalhos que definem o estado-da-arte em reconhecimento de gêneros musicais se baseiam na técnica de aprendizado por meio de redes neurais profundas (RNPs) (SIGTIA; DIXON, 2014; KERELIUK; STURM; LARSEN, 2015; JEONG; LEE, 2016; COSTA; OLIVEIRA; JR, 2017). Em sua maioria, esses trabalhos propõem métodos que se iniciam transformando o áudio de uma música em uma representação visual que destaca determinada característica. Essa representação visual é, então, utilizada como entrada em algoritmos de Aprendizado de Máquina baseados em RNPs.

Usualmente, utiliza-se apenas uma representação como entrada das redes. Mais especificamente, espectrogramas ou mel-espectrogramas. Essas representações destacam as frequências de pequenas janelas sequenciais de áudio em função do tempo. A principal diferença entre elas é o fato que a mel-espectrograma apresenta as frequências escaladas em um fator que simula a percepção humana das frequências de um sinal.

Ambas as representações citadas revelam características associadas ao timbre de uma música. Entretanto, há diversas representações visuais de áudio que evidenciam diferentes características musicais de uma gravação.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo realizar uma análise comparativa entre o impacto da utilização de variadas representações visuais de áudio e a utilização de transferência de aprendizado na tarefa de classificação de gêneros musicais. Para isso, serão estudadas diversas representações visuais de áudio, de modo a entender as características da música que elas buscam evidenciar. Em seguida, essas representações serão utilizadas como entrada para métodos de classificação de gêneros musicais encontrados na literatura. Para a classificação utilizando transferência de aprendizado foi selecionado como o modelo pré-treinado o YAMNet, um modelo treinado com um grande conjunto de dados, o AudioSet, uma base de dados com mais de 2 milhões de áudios de 10 segundos, rotulados em mais de 520 diferentes classes. Este modelo será utilizado para a construção de um novo modelo, analisado na tarefa de classificação de gêneros musicais.

1.1.2 Objetivos específicos

Os objetivos específicos podem ser elencados em:

- Analisar as diferentes representações visuais de áudio investigando como as diferentes informações musicais se refletem nas imagens;
- Utilizar as diferentes representações visuais de áudio para construir um modelo de classificação de gênero musical;
- A partir de um modelo pré-treinado em um grande conjunto de dados, utilizar da transferência de aprendizado para construir um novo modelo, treinado para classificação de gêneros musicais;
- Analisar os resultados obtidos, comparando o desempenho das duas abordagens na construção dos modelos para classificação de gêneros musicais a partir do áudio.

1.2 Organização do trabalho

Este trabalho está dividido em cinco capítulos. No Capítulo 1, são apresentados a proposta de pesquisa, os objetivos e a justificativa para a realização do mesmo. O Capítulo 2 abrange a fundamentação teórica necessária para cumprir com o objetivo proposto, passando pelos conceitos de organização e recuperação de informação musical, aprendizado de máquina, alguns dos diferentes tipos de aprendizado, redes neurais profundas, classificação de gênero musical, representação de sinais de áudio e transferência de aprendizado. No Capítulo 3, é exibida a metodologia utilizada para a execução dos experimentos, modelo e conjunto de dados utilizados nos experimentos. O Capítulo 4 destina-se a análise dos resultados obtidos, realizando a comparação entre o desempenho dos modelos construídos na utilização de cada uma das abordagens. Por fim, no Capítulo 5, são expostas as conclusões do trabalho e possíveis trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os principais conceitos que respaldam o presente trabalho. Sendo esses, organização e recuperação de informação musical, aprendizado de máquina, os diferentes tipos de aprendizado, redes neurais profundas, classificação de gênero musical, representação de sinais de áudio e transferência de aprendizado.

2.1 Organização e recuperação de informação musical

Levando em consideração como o consumo de música digital cresceu nas últimas décadas, se torna necessário pensar no modo como é feita a organização e recuperação de toda informação musical através de métodos computacionais. As informações que podem ser obtidas a partir de uma música precisam ser armazenadas de forma organizada, de modo que a recuperação dessas informações musicais ocorra de maneira eficiente e que a distância entre a informação que se pretende e a informação recuperada seja mínima.

A extração de informação musical tem impacto direto na resolução de uma ampla gama de problemas, como detecção de ritmo, transcrição automática de música, reconhecimento de artista, classificação de gênero e recomendação de músicas (LAMERE, 2008). A partir da extração dos dados, o próximo passo a ser pensado é o modo como essas informações são organizadas. Uma das formas de organizar coleções de músicas é através do uso de rótulos, nos quais cada música é associada a uma informação que a represente resumidamente, de modo que elas possam ser diretamente relacionadas a outras músicas que possuam essa mesma característica.

Muitos trabalhos voltam a sua atenção às diferentes formas de abordar a tarefa de organização e recuperação de informação (MIR), propondo algoritmos que utilizam diferentes critérios na categorização de uma música, como similaridade de gênero (TZANETAKIS; COOK, 2002), emoção ou sentimento (KIM et al., 2010), instrumento musical (BOYER et al., 2003) e utilização de tags mais gerais (LAMERE, 2008). Os diferentes critérios utilizados na categorização de uma música corroboram para que uma série de tarefas possam ser realizadas. Com as informações categorizadas de forma eficiente, é possível que plataformas de reprodução online possam organizar o seu acervo associando artistas e músicas com perfis parecidos, realizar uma criação automática de playlists, um sistema de recomendação mais eficiente, abre a possibilidade de traçar um perfil de usuário de modo mais assertivo, entre outras (KNEES; SCHEDL, 2015; WAN, 2016; LAMERE, 2008).

Considerando a complexidade atrelada à tarefa de categorização de uma música

com relação ao seu gênero, os trabalhos que definem o estado-da-arte nesta tarefa se baseiam na técnica de aprendizado por meio de redes neurais profundas (RNPs). Os dados relacionados à música são utilizados como entrada em algoritmos de Aprendizado de Máquina baseados em RNPs.

2.2 Aprendizado de máquina

O Aprendizado de Máquina (AM) é um subcampo da Inteligência Artificial (IA) que, através de métodos computacionais, possibilita que sistemas possam aprender padrões, generalizar conhecimento e realizar tarefas de forma automática (MITCHELL, 1997). O Aprendizado de Máquina possibilitou que muitas tarefas pudessem ser realizadas de forma eficiente, como: sistemas de recomendação, identificação de objetos em imagens, transcrição de áudio em texto, melhorando a forma como as pesquisas ocorrem na internet e os resultados, que, se baseando nos hábitos de consumo de um usuário, podem ser apresentados de forma cada vez mais individual, entre outros (LECUN; BENGIO; HINTON, 2015).

Uma tarefa de aprendizado pode ser entendida como uma tarefa em que buscamos melhorar algum indicador de performance para o resultado de uma atividade realizada, essa melhoria se dá através de algum tipo de treinamento (JORDAN; MITCHELL, 2015). Podemos utilizar como exemplo uma tarefa de detecção de fraude em cartões de créditos, a tarefa é indicar, para cada transação de cartão de crédito um rótulo como sendo “fraude” ou “legítima”. A métrica de performance a ser melhorada pode ser a acurácia, e os dados utilizados para treinamento podem ser um conjunto de transações de cartão de crédito. Cada rótulo atribuído de forma correta indica um caminho de melhoria no indicador de performance para o modelo, e cada rótulo atribuído de forma incorreta, indica um caminho a ser evitado.

2.2.1 Tipos de aprendizado

2.2.1.1 Aprendizado supervisionado

O aprendizado supervisionado envolve construir um modelo capaz de, a partir dos dados disponibilizados, identificar padrões e generalizar comportamentos em dados, de modo a aprender uma relação entre um conjunto de variáveis de entrada X e uma variável de saída Y (CUNNINGHAM; CORD; DELANY, 2008). Sendo assim, quando apresentada uma nova informação X, ainda não vista pelo modelo, será possível predizer a saída Y. Para que um modelo preditivo seja construído, o aprendizado supervisionado exige que os dados possuam uma classe. Existem dois tipos principais de algoritmo de aprendizado supervisionado: Regressão e Classificação. Os algoritmos de Regressão são

utilizados quando as classes possuem valores contínuos, já os algoritmos de Classificação são utilizados quando as classes possuem valores discretos (GUIDO S.; MULLER, 2016).

2.2.1.2 Aprendizado não-supervisionado

O aprendizado não supervisionado compreende em construir um modelo em que, ao receber os dados que serão utilizados para treinamento, as classes não são pré-definidas. De acordo com as características e padrões contidos nos dados, o modelo deve inferir o modo como eles podem ser agrupados, em quantas classes esses dados poderiam ser distribuídos, se baseando na similaridade entre os dados pertencentes a um mesmo grupo (DAUMÉ, 2017). Como os dados não são fornecidos para o algoritmo, com classes já definidas, é possível que eles sejam agrupados sem a necessidade de intervenção humana, ou que o algoritmo encontre um modelo capaz de agrupar os dados em novos grupos, encontrando padrões ocultos.

Os modelos de aprendizado não-supervisionado são frequentemente aplicados em tarefas de agrupamento, associação e redução de dimensionalidade. Na tarefa de agrupamento, o principal objetivo é dividir um conjunto de dados em grupos, de modo que os itens pertencentes a um mesmo grupo sejam mais similares entre si. A tarefa de associação, concentra-se em descobrir padrões, inferências, relações, entre diferentes itens em um conjunto de dados. A tarefa de redução de dimensionalidade consiste em, ao obter um conjunto com os dados em uma grande dimensão, realizar uma redução para uma dimensão menor, mantendo as características relevantes contidas nos dados (DAUMÉ, 2017; VLACHOS, 2010; KAUFMAN; ROUSSEEUW, 1990)

2.2.1.3 Aprendizado semi-supervisionado

No aprendizado supervisionado são utilizados conjuntos de dados rotulados, permitindo que o algoritmo construa um modelo a partir dos dados de exemplo, aos quais as classes já estão definidas. No aprendizado não-supervisionado a construção do modelo se dá pela ausência de rótulos nos dados utilizados para treinamento. Já o aprendizado semi-supervisionado combina as duas abordagens anteriores, o algoritmo tem como objetivo a construção de um modelo que é treinado com conjuntos de dados que incluem tanto exemplos rotulados quanto não rotulados. Dessa forma, o modelo pode aprender com os dados rotulados, mas também se beneficiando dos dados não rotulados, visando aprimorar a capacidade de generalização e o desempenho do modelo. Essa abordagem pode ser utilizada em casos que se torna muito custoso construir um conjunto de dados que esteja rotulado, mas que ainda é possível que uma parte do conjunto possa ser rotulada (KINGMA et al., 2014; XIE et al., 2020; BERTHELOT et al., 2019; YALNIZ et al., 2019).

Ainda existem outras abordagens, como aprendizado por reforço, aprendizado ativo, entre outros. Dentre as abordagens citadas, o foco deste trabalho é fazer uma análise

comparativa, se utilizando da abordagem de aprendizado supervisionado, para classificação de gêneros musicais, a partir da utilização de algoritmos de aprendizado de máquina baseados em RNPs.

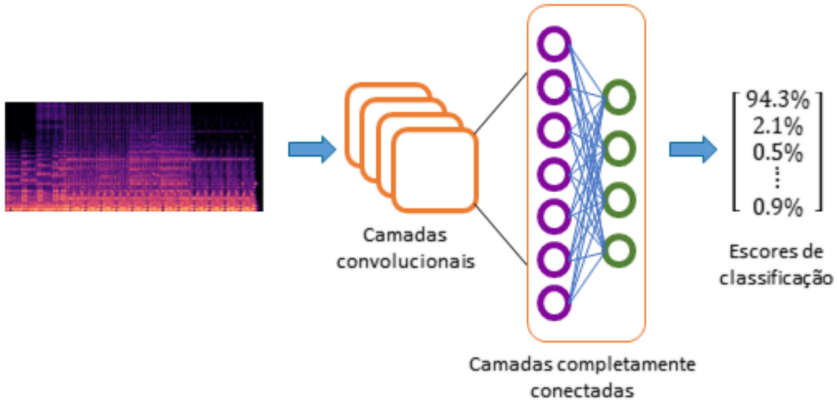
2.2.2 Redes neurais profundas

O conceito de Redes Neurais Profundas tem despertado o interesse de pesquisadores das mais diversas áreas. Isso se deve aos significativos avanços alcançados na classificação de uma variedade de tipos complexos de dados.

Os modelos de aprendizagem profundos são caracterizados por sua estrutura composta por diversas camadas, onde cada camada utiliza a saída da camada anterior como entrada. Essa abordagem tem o objetivo de adquirir distintos níveis de características ou representações dos dados de entrada, resultando na criação de uma hierarquia robusta de representações. Dentre as diversas arquiteturas de Redes Neurais Profundas (RNPs) propostas na literatura, a arquitetura mais utilizada tem sido a rede neural convolucional (CNN) (LECUN; BENGIO; HINTON, 2015). E um domínio dentre os quais essa classe de algoritmos tem se destacado é na tarefa de classificação de gêneros musicais (WAN, 2016).

Essencialmente, as Redes Neurais Convolucionais (CNN) são constituídas por camadas que aplicam, sequencialmente, diferentes filtros de convolução. Cada um desses filtros é responsável por transformar os dados de entrada de forma a evidenciar uma determinada característica. No geral, essas camadas são conectadas a uma rede neural completamente conectada, com o propósito de classificar os dados com base na transformação realizada na última camada convolucional. O esquema geral de classificação utilizando uma CNN é apresentado na Figura 1.

Figura 1 – Exemplo de classificação utilizando uma rede neural convolucional.



A maior limitação desse tipo de redes neurais reside no elevado custo computacional associado ao seu treinamento. Uma alternativa para agilizar o processo de treinamento é utilizar uma rede previamente treinada, exigindo apenas algumas atualizações nos parâmetros de seus filtros e nos valores de ativação dos neurônios. Mesmo que essa

rede possa ter sido treinada em outro contexto diferente, além de reduzir o custo de tempo do treinamento, essa abordagem pode resultar em melhorias significativas nos resultados. A essa técnica, é atribuído o nome de transferência de aprendizado (WEISS; KHOSHGOFTAAR; WANG, 2016).

No exemplo mostrado na Figura 1, a rede recebe como entrada um mel-espectrograma, uma representação visual de sinais de áudio intimamente relacionada ao timbre. Essa representação é comumente empregada na tarefa de classificação de gênero musical. O intuito deste trabalho é analisar os resultados envolvidos na classificação de gênero musical utilizando diferentes representações visuais de sinais de áudio e transferência de aprendizado.

2.3 Classificação de gênero musical

Os gêneros musicais são rótulos criados e muito frequentemente usados por humanos para categorizar e descrever o vasto universo musical em forma de grupos que compartilham características semelhantes. Devido a essa informação ser altamente subjetiva e estar ligada a diversos fatores comportamentais, culturais, históricos, e até ao marketing, os gêneros musicais não tem uma definição rigorosa e os limites não são objetivamente específicos. Porém, mesmo com a natureza característica subjetiva, é evidente que músicas pertencentes a um determinado gênero compartilham de características relacionadas à instrumentação, estrutura rítmica, tonalidade e outras características como escalas, acordes, melodias, atributos que criam uma identidade única para uma obra (TZANETAKIS; COOK, 2002).

Considere um conjunto de dados $X = \{x_1, x_2, x_3, ..., x_n\}$ onde $x_i = (A, r), i \in (1, 2, 3, ..., n)$, é um par ordenado contendo um conjunto de valores A e um rótulo r associado a ele. Cada elemento x_i é chamado exemplo, ou seja, o conjunto de dados X possui n exemplos nesse caso. No caso específico de classificação de gênero musical, um exemplo é uma música, definida pelo seu arquivo de áudio A e cujo gênero é r. Essa última informação é chamada de rótulo ou classe.

A tarefa de classificação compreende em atribuir rótulo para exemplos cuja classe é desconhecida. Formalmente, considere um conjunto de dados $Y = \{y_1, y_2, y_3, ..., y_m\}$ em que cada instância $y_i, i \in (1, 2, 3, ..., m)$, é definida apenas pelo conjunto de valores A, mas seu rótulo não é conhecido pelo modelo. Um algoritmo de classificação tem como objetivo induzir um modelo a partir de X, chamado conjunto de treinamento, capaz de induzir um rótulo para cada elemento em Y, chamado conjunto de teste.

É importante notar que a classe de cada exemplo deve pertencer a um conjunto pré-definido de rótulos. Por exemplo, o conjunto de rótulos possíveis para o problema de classificação de gêneros musicais pode ser representado por $r \in \{'rock', 'pop', 'mpb', 'jazz', ...\}$. Apesar de haver trabalhos em que as classes estão estruturadas hierarquicamente (SILLA;

FREITAS, 2011), em que cada exemplo pode possuir um ou mais rótulos (SILVA; WINCK, 2014) ou que considera-se a possibilidade de surgirem classes previamente desconhecidas (KRAWCZYK et al., 2017), tais variações da tarefa de classificação não serão consideradas neste trabalho.

Usualmente, para um dado exemplo com rótulo desconhecido, um modelo de classificação atribu escores para cada uma das classes definidas para o problema. Em outras palavras, a saída do modelo é um valor numérico relacionado ao grau de certeza que o modelo atribui para cada classe. Nesse cenário, a resposta final do processo de classificação é o rótulo com maior escore. Por exemplo, ao se classificar uma música com escore 0.75 para a classe 'hip-hop', 0.2 para 'pop' e valores menores para os demais gêneros, o classificador responde com a classe 'hip-hop'.

A literatura de aprendizado de máquina apresenta uma infinidade de algoritmos para induzir modelos de classificação (WITTEN et al., 2017). Como destacado anteriormente, nos últimos anos, a categoria de algoritmos baseados em RNPs têm se destacado no cenário de classificação de música em gêneros e outras categorias de rótulos.

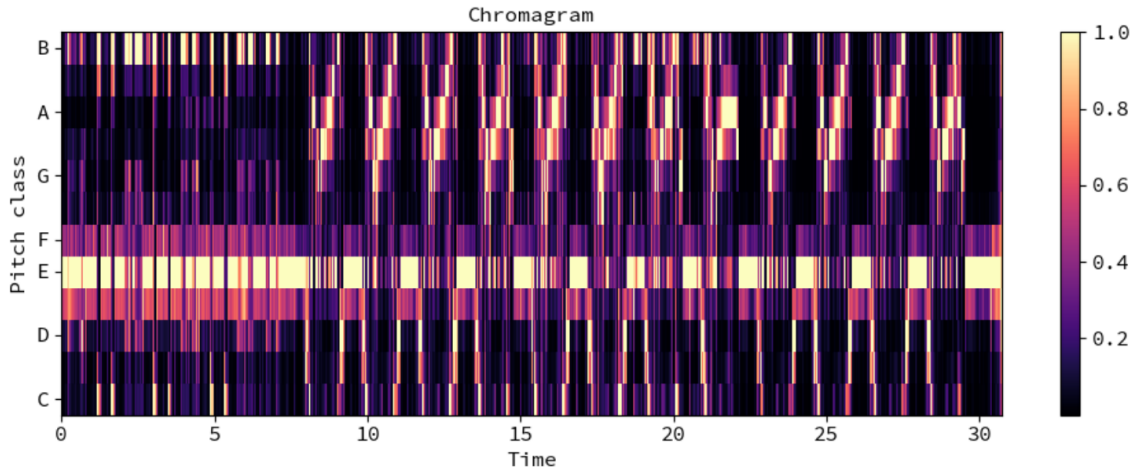
2.3.1 Representação de sinais de áudio

A eficácia dos algoritmos de classificação baseados em RNPs depende, entre outras características, da qualidade dos dados apresentados à rede. Esse fato se deve ao fenômeno comumente referido pelo jargão "garbage in, garbage out". Para que uma RNP consiga um desempenho satisfatório de classificação, é crucial que sejam fornecidos dados que contenham alguma relevância semântica, mesmo que oculta.

No contexto de classificação de gêneros musicais, o procedimento convencional para realizar essa tarefa envolve, primeiramente, a transformação do áudio em uma representação visual da música. Comumente, são utilizadas representações de tempo-frequência (como os espectrogramas) e algumas de suas variações (como os mel-espectrogramas). Essas representações apresentam uma forte relação com as características da música que definem seu timbre. A partir dessas representações, a RNP aplica filtros consecutivos para se estimar as características específicas dessas imagens que apresentam maior associação com cada gênero musical.

No entanto, diversas representações visuais da música já foram propostas na literatura e são utilizadas com o intuito de obter melhores resultados na tarefa de classificação de gêneros musicais (MÜLLER, 2015). Cada uma dessas representações busca evidenciar uma característica diferente. Tomemos como exemplo a representação cromagrama, ilustrada na Figura 2. Essa representação reflete a intensidade do sinal em frequências associadas a cada nota musical ao longo do tempo, ou seja, é uma representação que está intimamente ligada ao conceito de tonalidade.

Figura 2 – Exemplo de cromagrama.



Fonte: www.librosa.github.io/librosa

Observe que existem características nessa representação que podem ter relação direta com o gênero musical. Por exemplo, a complexidade da música (alta repetição ou grande variabilidade de notas), a duração das notas, o tom em que uma música foi executada, entre outras características que possui relação com esse tipo de representação.

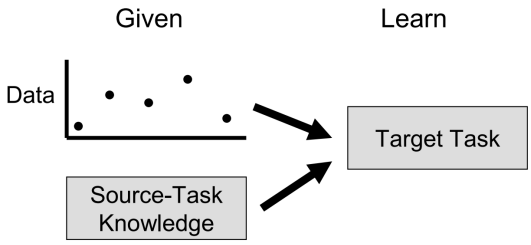
Note que, embora tenham sido abordadas representações associadas ao timbre e à tonalidade, existem representações diretamente vinculadas a outras características determinantes para definir o gênero de uma música, como tempo (GROSCHE; MÜLLER; KURTH, 2010), além de outras abordagens para representar as mesmas características, como o centróide tonal (HARTE; SANDLER; GASSER, 2006).

2.3.2 Transferência de aprendizado

A ideia central da transferência de aprendizado é melhorar o aprendizado em uma nova tarefa através da transferência de conhecimento já obtida em uma outra tarefa relacionada. Ao utilizar um modelo que já adquiriu conhecimento para resolver um problema (tarefa de origem), é possível aplicar esse conhecimento para resolver outros problemas relevantes (tarefa de destino). Por exemplo, se o modelo é capaz de realizar a identificação de instrumentos musicais, ou de emoções/sentimentos (tarefa de origem), o conhecimento adquirido seria útil para resolver a classificação de gêneros musicais (tarefa de destino). A ideia é que, embora as tarefas de origem e o destino não sejam idênticas, se o conjunto de dados para a tarefa de origem for muito maior do que o da tarefa de destino, a transferência pode levar a um desempenho maior (WON; SPIJKERVET; CHOI, 2021; CHOI et al., 2017b).

O objetivo ao se utilizar a transferência de aprendizado é obter uma melhor performance no aprendizado, se o método de transferência diminuiu a performance, então ocorreu uma transferência negativa. Um dos grandes desafios ao se treinar um novo modelo

Figura 3 – Ilustração da abordagem de transferência de aprendizado, recebendo o conhecimento prévio do modelo já treinado e novos dados para construção de um modelo para a tarefa de destino.



Fonte: (TORREY; SHAVLIK, 2010).

utilizando o método da transferência é produzir uma transferência positiva. É necessária a utilização de modelos prévios (tarefa de origem) que tenham relação com a tarefa de destino, evitando a transferência negativa.

3 Metodologia

Este capítulo descreve o conjunto de dados, os métodos, algoritmos e ferramentas que foram empregados para realizar os experimentos e obter os resultados. Os algoritmos foram elaborados na linguagem de programação Python[1], com o auxílio das bibliotecas: TensorFlow [2], Keras[3] e Librosa[4].

3.1 Conjunto de dados

Para a realização deste trabalho, foi utilizado o conjunto de dados GTZAN Genre Collection (TZANETAKIS; COOK, 2002). Este conjunto de dados é amplamente reconhecido e empregado na área de processamento de sinais de áudio, sendo especialmente adequado para tarefas de classificação de gênero musical e demais tarefas que demandem um conjunto de dados de música. A escolha do GTZAN Genre Collection para realizar a análise comparativa entre os algoritmos de classificação de gênero musical, empregando técnicas de aprendizado de máquina, foi motivada pela sua representatividade e qualidade.

O GTZAN Genre Collection é constituído por 1000 trechos de áudio, cada áudio com duração de 30 segundos, distribuídos em 10 diferentes gêneros musicais. Cada gênero é representado por 100 amostras, proporcionando um equilíbrio na quantidade de dados entre as classes. A presença de 100 amostras por classe é bastante importante, pois contribui para evitar desequilíbrios que poderiam impactar negativamente o desempenho do modelo. O conjunto de dados inclui músicas dos seguintes gêneros musicais: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, e rock.

É importante destacar que a escolha do conjunto de dados é um elemento crucial na condução de experimentos de aprendizado de máquina. O GTZAN Genre Collection, por sua notável presença na literatura científica e sua qualidade, proporciona uma base sólida para o desenvolvimento e avaliação de modelos de classificação de gênero musical utilizando técnicas como RNPs. O conjunto de dados GTZAN é bastante utilizado como uma base de dados de referência para medir performance dos diversos trabalhos que estudam o tema de classificação de gênero musical e demais tarefas que demandem um conjunto de dados músicais.

[1] <www.python.org>
[2] <www.tensorflow.org>
[3] <www.keras.io>
[4] <www.librosa.org>

3.2 Avaliação

A avaliação dos modelos será conduzida por meio de métricas específicas, como Precision, Recall, F-Measure e Acurácia, fornecendo uma visão abrangente do desempenho dos algoritmos implementados para construção de cada modelo. As métricas escolhidas para avaliar o desempenho dos modelos são fundamentais para compreender e analisar a eficácia dos modelos, e estas são métricas amplamente reconhecidas e utilizadas em trabalhos relacionados, além de oferecerem uma análise detalhada das capacidades dos modelos.

A métrica Precision (Precisão) representa a proporção de instâncias corretamente classificadas como pertencentes a uma determinada classe em relação ao total de instâncias pertencentes a esta classe. Uma Precision elevada indica poucos falsos positivos. Falsos positivos ocorrem quando o modelo, ao classificar um dado, assume uma classe como sendo positiva, porém, ao observar o rótulo legítimo é verificado que a classe é negativa.

A métrica Recall (Revocação) refere-se à proporção de instâncias corretamente classificadas como pertencentes a uma determinada classe em relação ao total de instâncias que realmente pertencem a essa classe. Indica a frequência em que o modelo é capaz de encontrar as músicas de uma classe. Um Recall elevado indica poucos falsos negativos. Falsos negativos ocorrem quando o modelo, ao classificar um dado, assume uma classe como sendo negativa, porém, ao observar o rótulo legítimo é verificado que a classe é positiva.

A métrica F-Measure (ou F_1-Score) é a média harmônica entre Precision e Recall e proporciona uma visão equilibrada do desempenho do modelo, sendo especialmente útil quando há um desequilíbrio entre as classes. Uma F-Measure alta indica uma alta qualidade geral do modelo.

A métrica Acurácia representa a proporção de instâncias corretamente classificadas em relação ao total de instancias. Embora seja uma métrica importante, em casos de desequilíbrio nas classes, pode não fornecer uma visão completa do desempenho.

As métricas mencionadas acima podem ser descritas matematicamente da seguinte forma:

$$Precisão = \frac{VerdadeirosPositivos(TP)}{VerdadeirosPositivos(TP) + FalsosPositivos(FP)}$$

$$Recall = \frac{VerdadeirosPositivos(TP)}{VerdadeirosPositivos(TP) + FalsosNegativos(FN)}$$

$$F_1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Acurácia = \frac{VerdadeirosPositivos(TP) + VerdadeirosNegativos(TN)}{Total}$$

Além das métricas de avaliação mencionadas anteriormente, a análise estatística desempenha um papel muito importante na validação dos resultados obtidos. O desvio padrão será utilizado para medir a dispersão dos dados, permitindo uma compreensão mais aprofundada da consistência e estabilidade do desempenho dos modelos analisados.

3.3 Método proposto

Inicialmente, com o propósito de entender as representações de áudio mais utilizadas pela comunidade científica, e quais características musicais elas evidenciavam, foi realizado um processamento em diversos áudios com músicas de diferentes gêneros. Analisando as representações geradas a partir desse procedimento, concluiu-se que era uma tarefa muito difícil gerar um entendimento pleno das representações a partir dos áudios extremamente ricos em informação, de forma a se ter uma evidência clara e intuitiva de seu impacto na classificação.

Nas Figuras 4, 5 e 6 é possível ver a complexidade encontrada ao tentar entender as informações representadas nos gráficos. Era possível entender o que estava acontecendo, é claro, porém, não estavam explícitas as consequencias visuais de nuances e características básicas da música, e de seus instrumentos, que poderiam caracterizá-las quanto ao seu gênero.

Figura 4 – Cromagrama

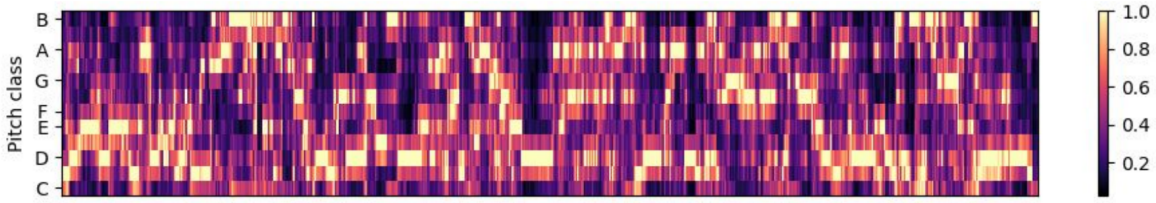
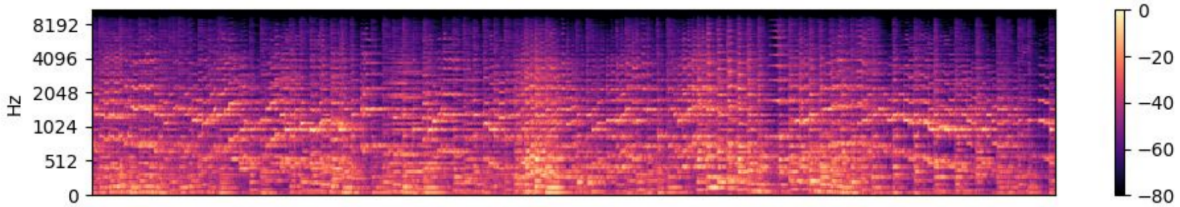
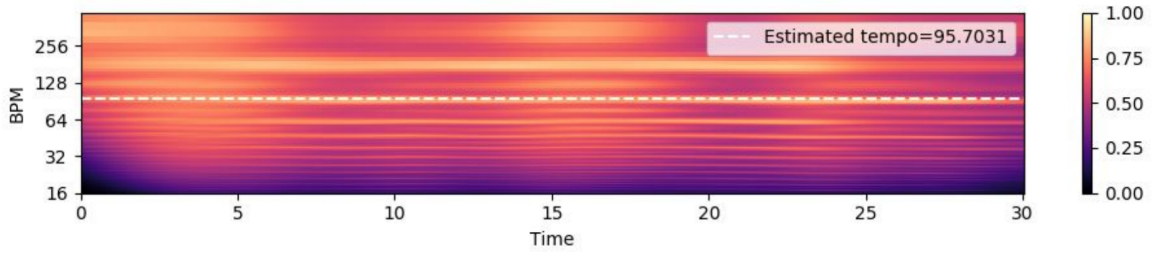


Figura 5 – Mel-espectrograma



A partir deste cenário, com o propósito de entender o que de fato eram as informações disponíveis nessas imagens, como cada característica musical se refletia nessas mesmas representações visuais e o quão evidenciada essas características se mostravam,

Figura 6 – Tempograma



voltou-se à base da teoria musical e, processando áudios simples, na maior parte das vezes áudios com poucos instrumentos tocando e eventos ocorrendo de forma isolada.

A princípio, para o entendimento da representação Tempograma, foram testados áudios com diversas características, como:

- músicas com uma fórmula de compasso composta
- músicas compostas com figuras rítmicas de tempo maior (músicas rigorosamente mais lentas)
- músicas com figuras rítmicas de tempo mais curto (músicas mais rápidas)
- músicas que possuíam, ao seu decorrer, uma alteração de andamento, como: *rallentando* (ou *ritardando*) e *accelerando*.

Na Figura 7 podemos ver a representação visual extraída de um trecho de 1 minuto da música *French Song* composta por *Pyotr Il'yich Tchaikovsky* sendo executada por um piano, para facilitar a visualização dos gráficos. Esta música foi escolhida por conta da variação de tempo que ela apresenta (*rallentandos* e *accelerandos*), para contribuir na observação de como o gráfico se comporta.

A linha tracejada encontrada na Figura 7 sinaliza o tempo estimado do áudio (a pulsação). Essa informação é calculada pela própria biblioteca *librosa*, através da função *librosa.beat.tempo*. Essa informação é calculada de acordo com a identificação dos tempos fortes encontrados na música. Porém, neste caso, em particular, não é uma informação totalmente correta, do ponto de vista musical. Com a utilização de um metrônomo pode-se observar que o tempo estimado seria em torno de 67-69 bpm (batidas por minuto).

Isso acontece por causa da identificação dos beats (tempo forte da música, o tempo em que ocorre uma batida mais forte). A extração do gráfico e dessas informações de tempo é feita pela identificação dos *beats* no tempo da música, e esssa identificação acaba considerando outras subdivisões das batidas, que não necessariamente caracterizam a pulsação da música. Por esse motivo, são consideradas outras batidas com uma maior frequência, ocasionando uma estimativa um pouco mais alta. No gráfico podem ser

observadas algumas outras linhas horizontais, que podem ser mais facilmente identificadas na Figura 8, que mostra, com uma menor evidência, as outras estimativas de tempo observadas no mesmo período. Essas outras estimativas são reflexos das subdivisões de tempo encontradas na música. É por este motivo que alguns gráficos ficam com a cor mais amarelada, porque existem muitas notas sendo executadas simultaneamente, como o exemplo do gênero clássico, e com ataques mais rápidos em determinado trecho, enquanto para outras músicas o gráfico apresenta uma cor mais azulada, mostrando possibilidades de tempo (pulsação) menos variadas.

Figura 7 – Tempograma obtido a partir da execução da música *French Song - Pyotr Il'yich Tchaikovsky*

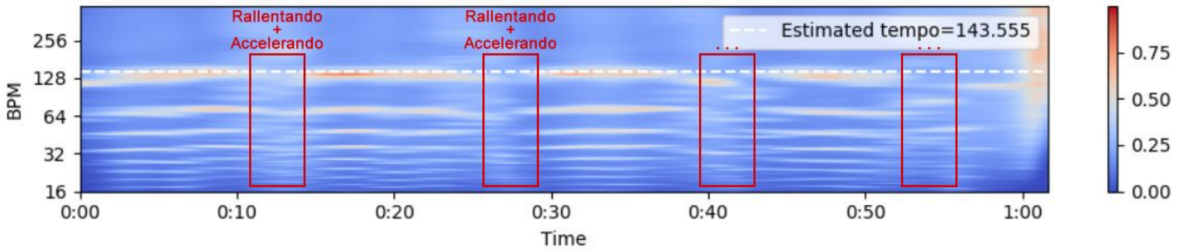
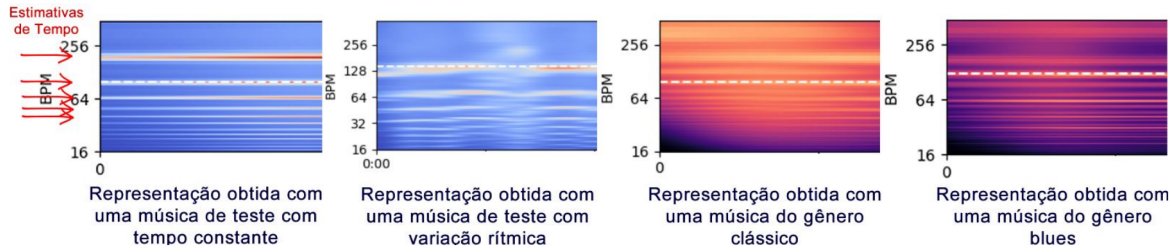


Figura 8 – Estimativas de tempo musical para diferentes áudios



Para entendimento das representações baseadas na frequência, como cromagrama, espectrograma, mel-espectrograma, foram testados áudios com diversas características, como:

- áudios com acordes, para testar o quão clara e precisa era a representação
- áudios com notas em diferentes oitavas (pra evidenciar a disparidade na frequência).
- áudios com *vibrato*
- áudios com uma escala definida, escala cromática, e algumas cadências musicais. A utilização deste teste foi para uma maior compreenão da representação Tonnetz.

Dessa forma, entendendo as informações que cada representação visual poderia evidenciar contribuindo para a classificação de um gênero musical, foram consideradas

diferentes representações visuais para a tarefa de classificação de gênero musical. As diferentes representações visuais utilizadas neste trabalho foram as seguintes: Espectrograma, Mel-espectrograma, Cromagrama, Espectrograma harmônico, Tempograma e Tonnetz. As Figuras 9, 10, 11, 12, 13 e 14 mostram um exemplo de cada representação para uma música do gênero classico.

A representação visual Espectrograma é uma representação visual que exibe a intensidade das diferentes frequências ao longo do tempo em um sinal de áudio. Em um espectrograma, o eixo horizontal representa o tempo, enquanto o eixo vertical representa a frequência, e a intensidade de cada frequência é ilustrada por meio de cores ou nível de brilho. Essa representação visual busca evidenciar diversas informações musicais, como: distribuição de frequências, timbre, e eventos específicos como vibratos, harmônicos, ruídos e outras variações sonoras relacionadas à frequência.

Já a representação visual Mel-espectrograma é uma variação do Espectrograma que utiliza uma escala de frequência mel para melhor representar as características que podem ser percebidas pelo ouvido humano, se alinhando melhor à forma como os seres humanos percebem as diferenças de tonalidade. O Mel-espectrograma busca evidenciar as mesmas características do Espectrograma, porém pode contribuir de forma mais específica para o âmbito musical, evidenciando melhor informações perceptíveis à audição humana.

A representação visual Cromagrama destaca a presença e intensidade das doze notas musicais ao longo do tempo em uma peça musical. Cada linha no cromagrama corresponde a uma nota específica, e a intensidade indica a presença da nota em um determinado momento. A principal contribuição do cromagrama é evidenciar as informações relativas às alturas tonais ou cromáticas em uma gravação musical, auxiliando na identificação de progressões de acordes, identificação de melodias, identificação dos modos presentes na música (análise modal) e identificação de notas e acordes.

A representação visual Espectrograma Harmônico destaca as componentes harmônicas em um sinal de áudio. Enquanto o espectrograma convencional exibe todas as componentes de frequência presentes em um sinal, o espectrograma harmônico destaca especificamente as frequências harmônicas relacionadas a uma nota fundamental, ajudando a evidenciar o relacionamento harmônico das notas sendo tocadas simultaneamente e a identificação das notas fundamentais e seus múltiplos harmônicos.

A representação visual Tempograma exibe a variação de andamento ao longo do tempo em uma peça musical, destacando as mudanças de velocidade, ou ritmo, revelando os momentos em que o andamento da música acelera, desacelera ou permanece constante. O Tempograma busca evidencias informações como andamento (tempo), acentuações rítmicas e mudanças de ritmo.

A representação visual Tonnetz é uma tecnica utilizada na análise musical para

visualizar relacionamentos tonais e harmônicos entre as notas musicais. O Tonnetz é um espaço geométrico que representa acordes, evidenciando as relações de consonância e dissonância entre acordes, destacando padrões tonais e conexões harmônicas. É bastante útil para entender a estrutura tonal de uma peça musical e como os acordes se movem e se conectam.

Figura 9 – Representação visual Espectrograma de uma música pertencente ao gênero clássico

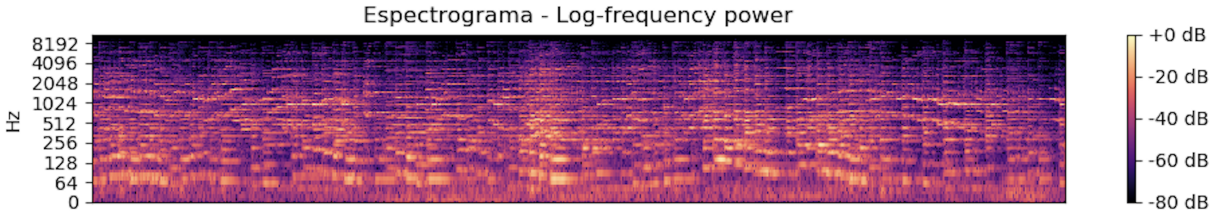


Figura 10 – Representação visual Mel-espectrograma de uma música pertencente ao gênero clássico

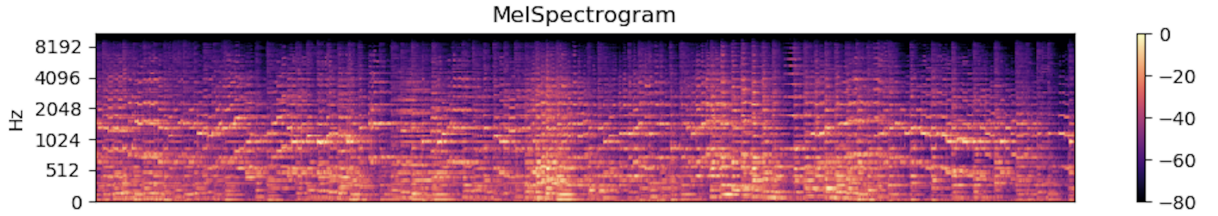


Figura 11 – Representação visual Cromagrama de uma música pertencente ao gênero clássico

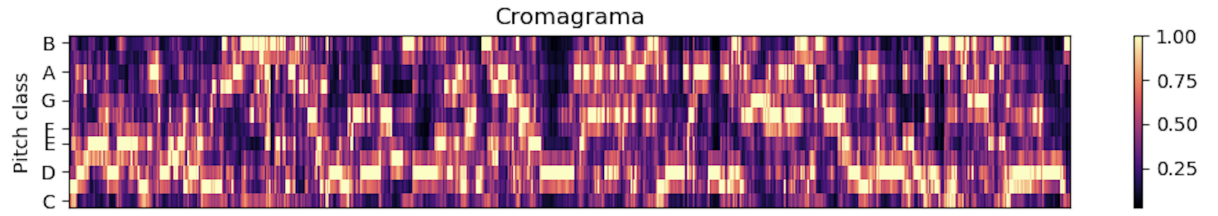
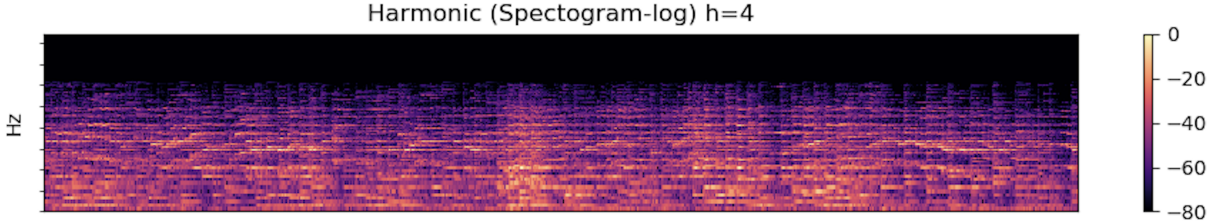


Figura 12 – Representação visual Espectrograma Harmônico de uma música pertencente ao gênero clássico



Para obtenção dos melhores resultados de cada representação também foram realizados testes em diversas arquiteturas de rede. É possível visualizar as principais arquiteturas testadas na Figura 15.

Figura 13 – Representação visual Tempograma de uma música pertencente ao gênero clássico

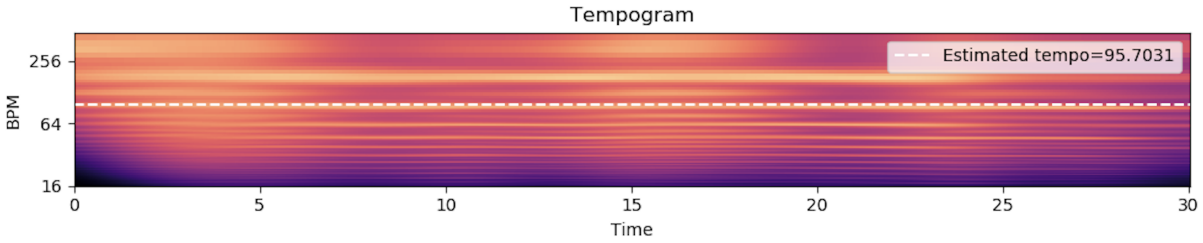


Figura 14 – Representação visual Tonnetz de uma música pertencente ao gênero clássico

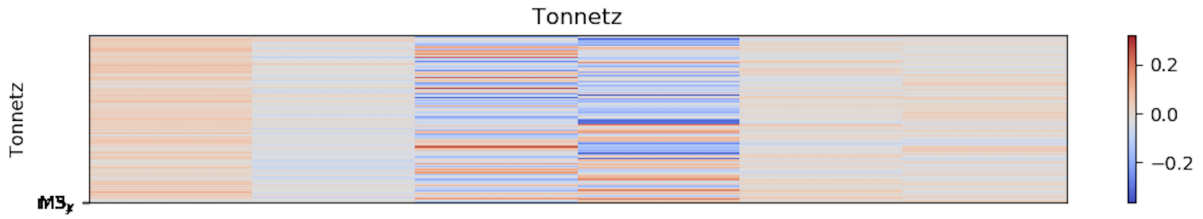
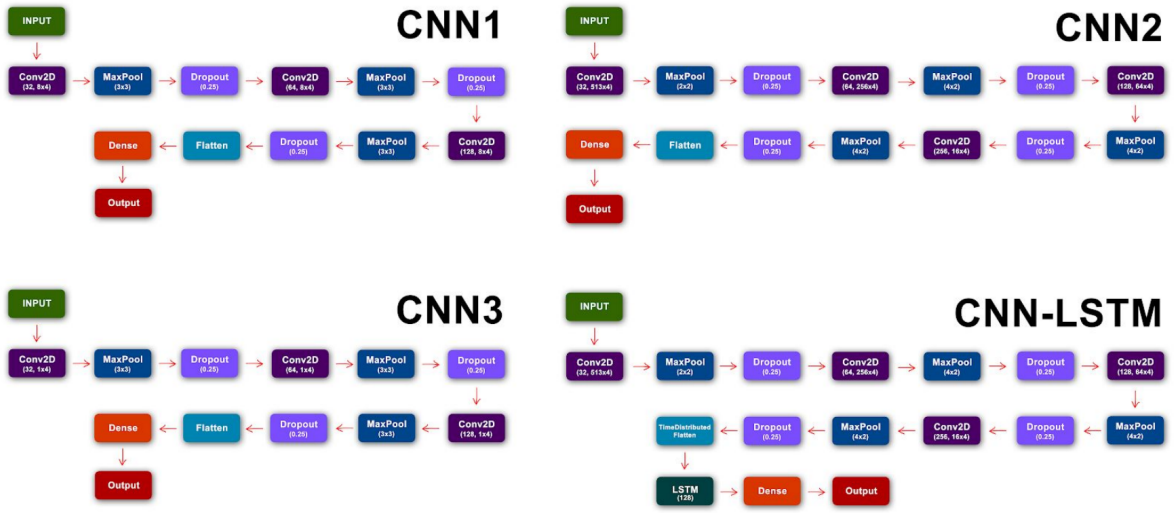


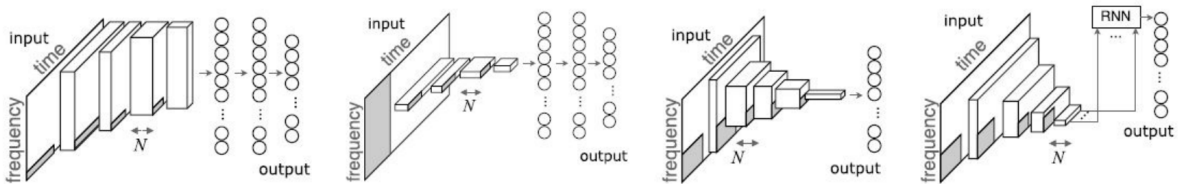
Figura 15 – Representação visual das arquiteturas testadas



Um dos principais aspectos levados em consideração para a escolha das arquiteturas foi o tamanho do *kernel* utilizado na rede. Foram escolhidos tamanhos de *kernels* de forma a obter um significado musical nas representações. Foi considerado um tamanho de *kernel* maior na horizontal, com o objetivo de que o modelo possa considerar a forma como a construção das frases foram realizadas na música, ou para considerar os intervalos musicais reproduzidos em uma determinada ordem. Também foram considerados filtros maiores verticalmente, com uma altura maior, para considerar notas sendo reproduzidas simultaneamente, ou para considerar instrumentos que emitem sons em frequências mais distantes sendo tocados simultaneamente. Dessa forma, com esses filtros seria possível obter um padrão de características de um gênero musical. Na Figura 16, é apresentada a ideia proposta em um trabalho de etiquetação automática de música, muito similar à

classificação de gênero, por meio da utilização conjunta de redes neurais convolucionais e recorrentes (CHOI et al., 2017a).

Figura 16 – Ilustração de diversos tamanhos de filtros considerados na arquitetura da rede



Uma outra estratégia que colaborou para a melhoria dos resultados foi cortar o áudio em pequenos trechos, fazendo cada um deles ser utilizado como um exemplo de treinamento. Os áudios da base de dados utilizada têm 30 segundos de duração, com exceção de pequenas flutuações. Cada áudio foi separado em pequenos trechos. Após alguns experimentos para a definição do comprimento de cada trecho, foi adotado um tamanho de janela de 1.5 segundo. Isso contribui em muito para a generalização dos dados, uma vez que aumenta-se a quantidade de áudios para o treino da rede neural. Pelo tamanho do áudio ser menor, a quantidade de informação presente em cada trecho é menor, facilitando a assimilação dos áudios processados e seus gêneros.

Para a construção do modelo utilizando transferência de aprendizado, foi utilizado como modelo pré-treinado o YAMNet (Yet Another Mobile Network). Este é um modelo que foi treinado com o conjunto de dados AudioSet, que é uma base de dados com mais de 2 milhões de áudios de 10 segundos rotulados em mais de 520 diferentes classes. As classes cobrem desde sons de animais, instrumentos musicais, gêneros musicais, e outros áudios presentes no ambiente (GEMMEKE et al., 2017). A Figura 17 mostra as diferentes classes de áudio presentes na base de dados.

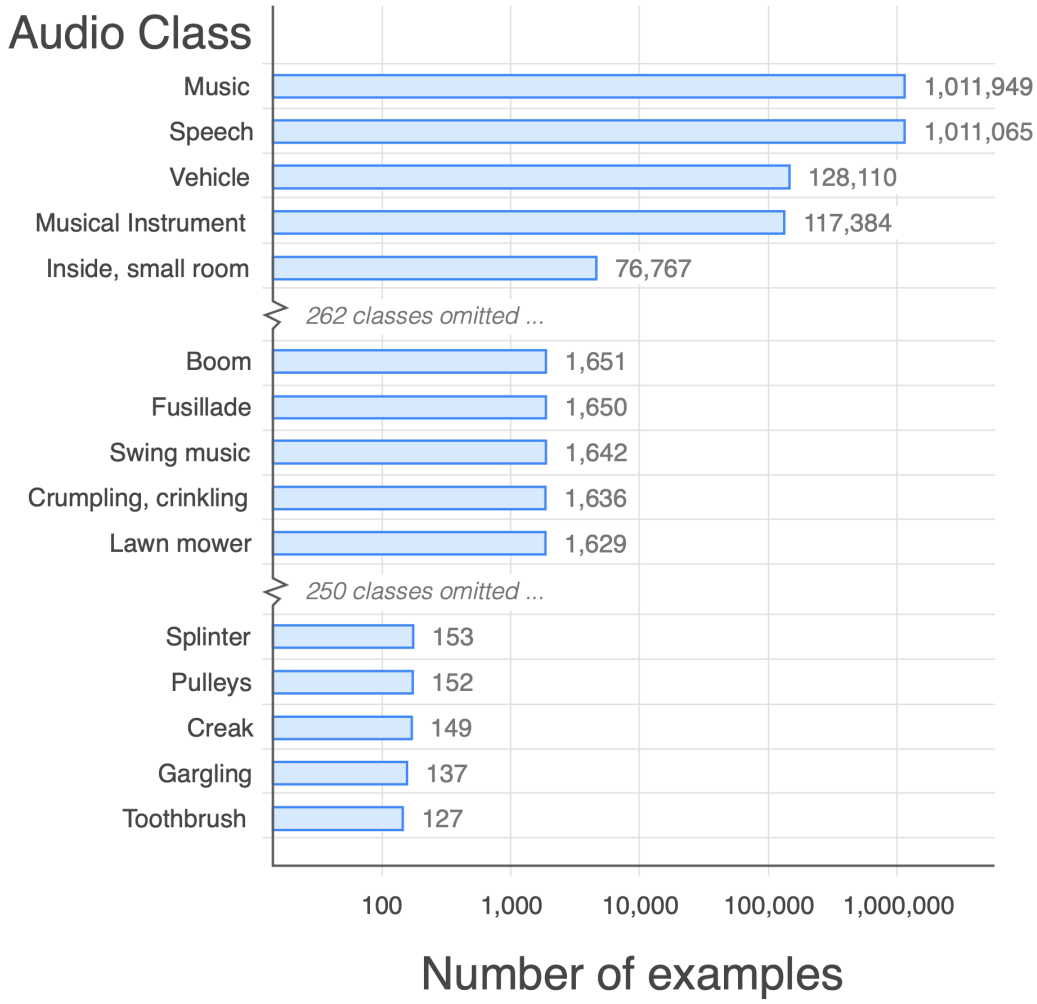
O modelo pré-treinado YAMNet emprega a arquitetura do modelo MobileNet_v1, com a última camada adaptada para a quantidade de classes presentes no conjunto de dados. O MobileNet é um modelo aplicado para o campo de computação visual, que é baseado em convoluções separáveis em profundidade (depthwise separable convolutions) (HOWARD et al., 2017). Na Tabela 1 é possível ver a arquitetura MobileNet, utilizada pelo modelo YAMNet.

A separação dos dados para o treinamento e teste das redes neurais foi feita da seguinte forma:

- Treinamento da Rede

  1. Foram separados 80% do conjunto de áudios disponíveis na base de dados (800 áudios).

Figura 17 – Rótulos e distribuição dos dados presentes no conjunto de dados AudioSet, utilizado para treinamento do modelo YAMNet



Fonte: (GEMMEKE et al., 2017)

2. Desses 800 áudios, 80% (640 áudios) foram utilizados para o treinamento da rede, e 20% (160 áudios) foram utilizados para validação.

3. Cada áudio foi separado em pequenos trechos, como descrito anteriormente.

4. Cada trecho de áudio foi utilizado como entrada para treinar a rede neural.

- Teste

1. Foram utilizados os 20% do conjunto de áudios disponíveis na base de dados (200 áudios) que não haviam sido utiliizados na fase de treinamento.

2. Cada áudio foi separado em pequenos trechos, do mesmo tamanho dos trechos que foram utilizados no treino.

3. Cada trecho de áudio foi utilizado como entrada para a validação.

4. As predições de cada trecho de áudio (de uma determinada música foram avaliadas como predição sendo um voto para uma categoria.

Tabela 1 – Arquitetura MobileNet_v1, utilizada pelo modelo YAMNet

Type / Stride	**Filter Shape**	**Input Size**
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
5× Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Fonte: (GEMMEKE et al., 2017)

5. Ao final, combinamos esses fotos de algumas formas diferentes: As principais foram:

I. uma votação simples, contando cada voto de forma fria, sem um peso estabelecido.

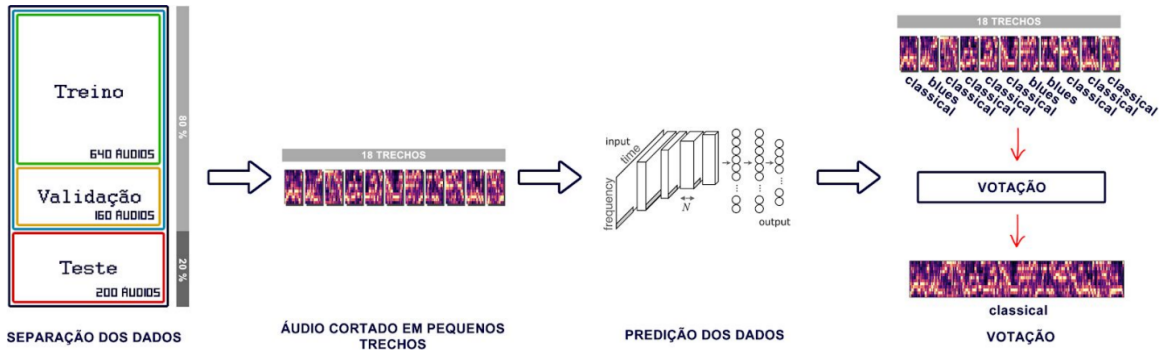
II. uma votação baseada no grau de certeza para cada possibilidade de gênero para cada trecho de áudio.

Todos os áudios foram escolhidos de forma aleatória, ou seja, foram randomizados, para a utilização no treinamento e validação da rede.

A Figura 18 demonstra de uma forma visual as etapas descritas acima.

Após as redes já treinadas com os dados da base de dados escolhida, então foram utilizados os dados de teste, definidos anteriormente, para obtenção das predições de cada rede, especialista em sua própria representação. A partir disso, as predições da mesma música de cada representação visual utilizada foram combinadas, de forma a obter ainda assim um melhor resultado.

Figura 18 – Processo de predição dos áudios



A Figura 19 ilustra a forma como foi realizado o processo de predição e combinação das representações visuais, e a Figura 20, a forma como ocorreu o processo utilizando transferência de aprendizado.

Figura 19 – Processo de predição e combinação das representações visuais

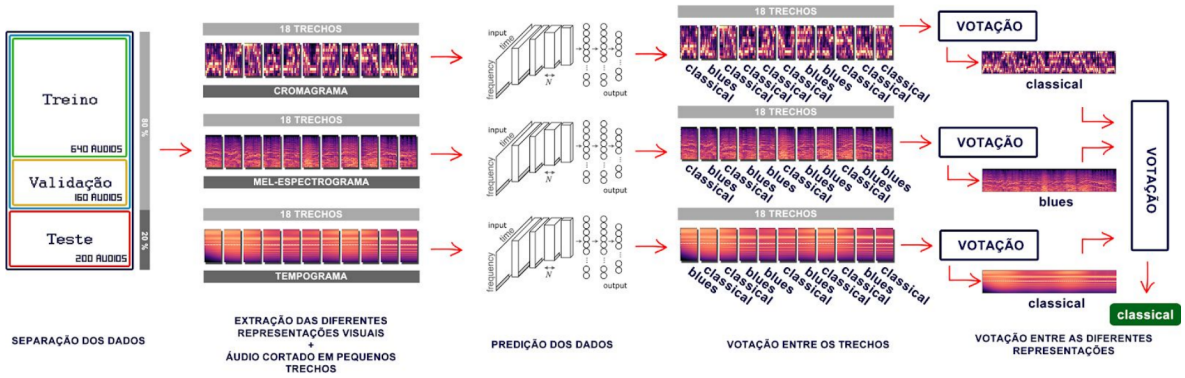
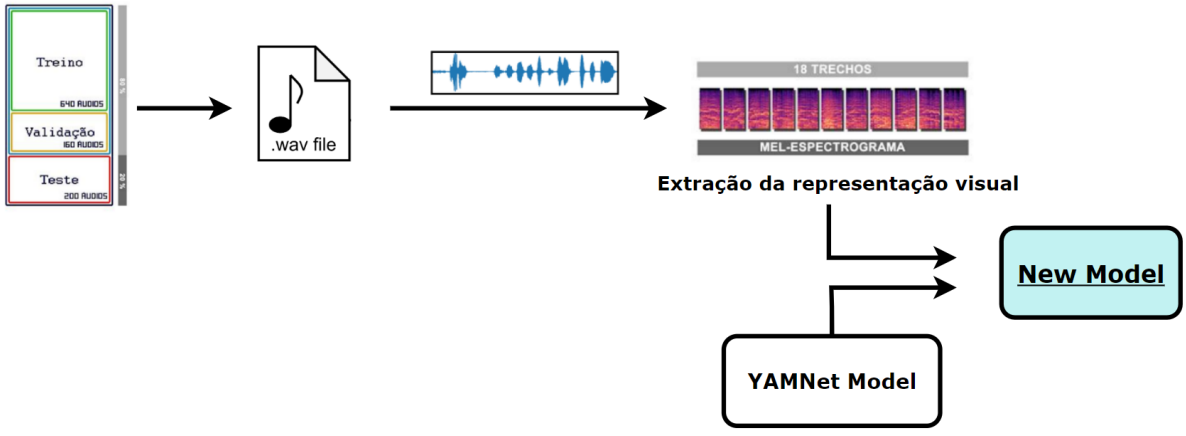


Figura 20 – Processo utilizando transferência de aprendizado



4 Análise e discussão dos resultados

Este capítulo tem o propósito de mostrar os resultados experimentais adquiridos a partir da utilização dos modelos construídos conforme discutido no Capítulo 3. Utilizando métricas como *Precision, Recall, F-Measure* e Acurácia, será analisado o ganho de desempenho obtido a partir das duas abordagens, utilização de diferentes representações visuais da música e transferência de aprendizado.

Para a classificação utilizando diferentes representações visuais da música, foi gerado um modelo para cada tipo de representação visual, após atingir os melhores resultados para cada rede neural responsável por cada representação, os resultados obtidos são apresentados na Tabela 3.

Tabela 2 – Resultados obtidos para cada representação visual

Representação	**Acurácia (%)**	
	p/ cada trecho (1.5s)	**p/ áudio inteiro (voting)**
espectrograma	75,35	83,50
log frequency spectrum	70,13	83,50
mel-espectrograma	73,55	79,00
mfcc	60,29	68,50
cromagrama	48,05	63,00
cromagrama24	52,41	60,00
espectrograma harmonico	38,23	52,00
tempograma	38,08	49,00
tonnetz	39,02	55,00

Todos os dados apresentados na Tabela 3 mostram resultados para cada tipo de representação visual utilizando a arquitetura CNN1. Dentre as arquiteturas apresentadas anteriormente no Capítulo 3, os melhores resultados foram obtidos utilizando a arquitetura CNN1. Podemos ver na Tabela 3, como exemplo, uma comparação dos resultados de cada arquitetura para a representação *log frequency spectrum*.

Tabela 3 – Resultados obtidos para cada arquitetura testada

Representação	**Arquitetura da Rede**	**Acurácia na validação (%)**	**Acurácia no teste (%)**	
			p/ cada trecho (1.5s)	**p/ áudio inteiro (voting)**
log frequency spectrum	CNN1	77,57	70,13	83,50
log frequency spectrum	CNN2	10,21	10,00	10,00
log frequency spectrum	CNN3	70,10	64,16	75,00
log frequency spectrum	CNN-LSTM	16,74	17,38	18,50

O fato dessa arquitetura ter contribuído melhor com os resultados contradiz um pouco o que era esperado, uma vez que a forma em que os filtros foram escolhidos nas

outras arquiteturas dão enfase ao significado musical contido nas representações. Os filtros foram escolhidos de forma a considerar a construção de frases musicais, a ordem em que as informações ocorriam na música e a altura (frequência) em que elas eram apresentadas. A hipótese admitida para que esse resultado tenha ocorrido é que, devido ao número baixo de músicas disponíveis para treino e validação, ou seja, o tamanho da base de dados utilizada, tornou-se muito difícil por parte da rede neural, a generalização das informações disponíveis em cada exemplo. Com uma base de dados maior, acredita-se ser possível um melhor resultado com as outras arquiteturas testadas, considerando as informações obtidas em cada representação visual, de forma a obter um significado musical sobre cada exemplo examinado.

A partir das predições fornecidas pelas melhores redes neurais em cada representação visual para todos os dados de validação, foi feita a combinação dessas predições. Assim como foi feito com as predições de cada trecho da música, combinando-as como se fossem votos para a classificação da música inteira, foram combinadas as predições de cada representação visual para que fosse possível obter os resultados referente a abordagem utilizando diferentes representações visuais para a tarefa de classificação de gênero musical. Na Figura 19 é possível conferir visualmente o processo descrito. O resultado da combinação, descrito anteriormente, é apresentado na Tabela 4. Onde os tipos de combinação foram realizados da seguinte forma:

- Voting: trata-se de uma votação simples, considerando o resultado de cada classificador referente a representação visual em específico.
- Scores Average: é uma média das predições de cada classificador para todas as classes. Cada classificador fornece a probabilidade de todas as classes, e então é realizada uma média para cada classe utilizando o resultado de todos os classificadores.
- Voting by Scores: cada modelo fornece uma probabilidade para uma classe, e então é realizada uma votação, considerando o grau de certeza que cada modelo teve.
- Scores Multiply: cada modelo recebe um multiplicador para determinadas classes, para que o seu palpite tenha um peso maior comparado aos demais classificadores, uma vez que a representação visual dele tem um melhor desempenho na classificação de determinados gêneros. Então é realizada uma votação considerando os palpites de todos os classificadores e os seus pesos.

A partir dos resultados mostrados na Tabela 4, podemos observar, comparando com os resultados obtidos utilizando apenas uma representação visual de áudio de forma isolada (Tabela 3), que houve uma melhora no desempenho do modelo de classificação quando são utilizadas diferentes representações visuais de áudio.

Tabela 4 – Resultados obtidos a partir da combinação das representações visuais

COMBINAÇÃO	
Tipo de Combinação	**Resultado**
Voting	86
Scores Average	87
Voting by Scores	85
Scores Multiply	85,5

Para uma análise mais completa dos resultados, o experimento de classificação de gênero musical utilizando diferentes representações visuais foi executado um total de 10 vezes. A partir dos resultados obtidos nas execuções, são apresentadas na Tabela 5 as métricas *Precision, Recall, F-Measure* e Acurácia obtidas para esta primeira abordagem.

Tabela 5 – Média e Desvio Padrão das métricas de desempenho para os resultados obtidos a partir da combinação das variadas representações visuais

Métrica	**Média**	**Desvio Padrão**
Precision	0.8718293348	0.0144122613
Recall	0.8540692692	0.0209433688
F-Measure	0.8628579235	0.0223844409
Acurácia	0.8523975363	0.0213371099

Para a classificação utilizando transferência de aprendizado, como mencionado anteriormente no Capítulo 3, foi utilizado como modelo pré-treinado o modelo YAMNet, que já foi treinado com o conjunto de dados AudioSet, para construção de um novo modelo, especialista na classificação de gêneros musicais. O conjunto de dados, como mencionado, foi o mesmo conjunto utilizado para a abordagem utilizando diferentes representações visuais de áudio, o *GTZAN Genre Collection*.

Assim como para a abordagem anterior, para uma análise mais completa dos resultados, o experimento de classificação de gênero musical utilizando transferência de aprendizado foi executado um total de 10 vezes. A partir dos resultados obtidos nas execuções, são apresentadas na Tabela 6 as métricas *Precision, Recall, F-Measure* e Acurácia obtidas para esta abordagem.

Tabela 6 – Média e Desvio Padrão das métricas de desempenho para os resultados obtidos a partir da transferência de aprendizado

Métrica	**Média**	**Desvio Padrão**
Precision	0.9030747773	0.0076267051
Recall	0.8978269078	0.0075093597
F-Measure	0.9004431963	0.0077806325
Acurácia	0.8972464829	0.0077298876

Na Tabela 7 são apresentados os resultados das duas abordagens lado a lado, para uma melhor visualização.

Tabela 7 – Tabela comparativa com os resultados das duas abordagens analisadas

Métrica	**Variadas Representações Visuais**		**Transferência de Aprendizado**	
	Média	**Desvio Padrão**	**Média**	**Desvio Padrão**
Precision	0.8718293348	0.0144122613	0.9030747773	0.0076267051
Recall	0.8540692692	0.0209433688	0.8978269078	0.0075093597
F-Measure	0.8628579235	0.0223844409	0.9004431963	0.0077806325
Acurácia	0.8523975363	0.0213371099	0.8972464829	0.0077298876

É possível notar que, para ambas as abordagens, combinação de diferentes representações visuais e transferência de aprendizado, as métricas Precision, Recall, F-Measure e Acurácia ficaram muito próximas, apenas na primeira abordagem, utilizando diferentes representações visuais, que a métrica Precision se destacou um pouco mais das demais métricas analisadas. Isso significa que o modelo está equilibrando bem a capacidade de classificação, indicando poucos falsos positivos (Precision) e poucos falsos negativos (Recall). Isso se reflete na métrica F_1-Score (Ou F-Measure), que é a média harmônica entre Precision e Recall. A Acurácia também se manteve próxima das demais, representando a quantidade de predições realizadas de forma correta em relação ao total de instâncias. É provável que o equilíbrio encontrado no dataset *GTZan Genre Collection*, que possui a mesma quantidade de dados para cada gênero, possa ter contribuído para um modelo equilibrado com essas métricas próximas.

Observa-se também que, o desvio padrão das métricas relacionadas a abordagem utilizando diferentes representações visuais (1), é, consideravelmente, maior que o desvio padrão das métricas da classificação utilizando transferência de aprendizado (2). Isso indica que houve uma maior dispersão nos resultados das métricas de desempenho para a primeira abordagem, aponta que os dados variam mais com relação a média. Esse fator mostra que o modelo (1) possui uma maior inconsistência no desempenho, uma maior sensibilidade aos dados de treinamento, ou uma menor estabilidade, com relação ao modelo (2).

Também é possível observar que as métricas de desempenho mostram resultados melhores para a classificação de gênero musical utilizando transferência de aprendizado. Acredita-se que o conjunto de dados utilizado pelo modelo pré-treinado, o conjunto AudioSet, possa ter contribuído de forma bastante significante. A Figura 17 mostra a quantidade de dados utilizados para algumas das classes presentes no conjunto de dados utilizado pelo modelo pré-treinado YAMNet. A classe *Music* possui uma quantidade de dados muito grande, 1.011.949 músicas, se comparada a quantidade de dados presente no *GTZAN Genre Collection*, um conjunto de apenas 1000 músicas. Porém, essa é a vantagem da transferência de aprendizado, poder utilizar um modelo que já foi treinado com uma quantidade muito maior de dados e que foi possível obter uma generalização maior dos padrões presentes nos dados.

5 Conclusão

Este trabalho compara a utilização de duas abordagens para a tarefa de classificação de gênero musical, a primeira abordagem é a construção do modelo utilizando diferentes representações visuais de áudio, e a segunda abordagem compreende na construção de um novo modelo utilizando transferência de aprendizado.

Tendo analisado o desempenho dos modelos obtidos no experimento, não só utilizando acurácia, mas também métricas como *Precision*, *Recall* e *F-Measure*, verificou-se que a transferência de aprendizado se demonstrou mais efetiva na tarefa de classificação de gênero musical, além de mostrar métricas de desempenho consideravelmente maiores, os números mostram uma maior consistência no desempenho dos modelos.

É válido ressaltar que o conjunto de dados utilizado na classificação através de diferentes representações visuais foi muito menor, com relação aos milhões de dados presentes na base de dados utilizada pelo modelo pré-treinado da transferência de aprendizado. Embora esta seja uma vantagem esperada do aprendizado de máquina, obter proveito de modelos já treinados e com uma grande quantidade de dados, é interessante fazer uma análise utilizando uma base de dados musical maior, como o conjunto *Free Music Archive* (FMA) (DEFFERRARD et al., 2016). Este conjunto de dados disponibiliza mais de 100 mil gravações livres de direitos autorais, cujo gênero foi anotado pelo próprio autor de cada música.

Além disso, é possível identificar uma curva de melhoria dos resultados com relação a utilização de diferentes representações visuais quando atribuímos pesos nos "palpites" de cada representação visual. Através de alguns testes realizados manualmente, obtivemos um resultado melhor comparado aos resultados da melhor rede neural para uma única representação visual. Mas não é possível garantir que este resultado seja o valor máximo da curva apresentada. Fica como possibilidade de trabalho futuro, a utilização de uma rede neural, para um aprendizado de como combinar esses pesos de cada representação visual, e maximizar a melhora obtida a partir da combinação das representações visuais para a classificação de gênero musical.

Por fim, considerando todos os dados do experimento, é possível concluir que a utilização de transferência de aprendizado na tarefa de classificação de gêneros musicais a partir do áudio, se demonstrou uma alternativa melhor, com modelos resultando em métricas de desempenho mais consistentes e consideravelmente maiores, com relação a utilização de diferentes representações visuais.

5.1 Trabalhos Futuros

Para continuação deste trabalho, algumas possibilidades de trabalhos futuros que poderiam contribuir a fim de melhorar o estudo e resultados obtidos são:

- Utilizar uma rede neural para aprender como combinar os pesos de cada representação visual, qual representação deve ter um palpite melhor para cada gênero musical, maximizando a melhora obtida a partir da combinação de diferentes representações;
- Realizar o experimento com um conjunto de dados maior, como o *Free Music Archive* (FMA);
- O gênero musical é uma informação que possui uma natureza hierárquica, um gênero pode ser dividido em diferentes sub-gêneros, como música popular, pode se dividir em gêneros como rock, que pode ser dividido em sub-gêneros como rock progressivo, metal, entre outros. Realizar um experimento para classificação hierárquica de gênero musical por meio de redes neurais profundas, considerando as abordagens utilizando diferentes representações visuais e transferência de aprendizado.

Referências

BERTHELOT, D. et al. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, v. 32, 2019. Citado na página 25.

BOYER, H. et al. Automatic classification of musical instrument sounds. *Journal of new music research. 2003; 32 (1): 3-21*, Taylor & Francis, 2003. Citado na página 23.

CHOI, K. et al. Convolutional recurrent neural networks for music classification. In: IEEE. *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. [S.l.], 2017. p. 2392–2396. Citado na página 39.

CHOI, K. et al. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017. Citado na página 29.

COSTA, Y. M.; OLIVEIRA, L. S.; JR, C. N. S. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied soft computing*, Elsevier, v. 52, p. 28–38, 2017. Citado na página 20.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: *Machine learning techniques for multimedia: case studies on organization and retrieval*. [S.l.]: Springer, 2008. p. 21–49. Citado na página 24.

DAUMÉ, H. *A course in machine learning*. [S.l.]: Hal Daumé III, 2017. Citado na página 25.

DEFFERRARD, M. et al. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016. Citado na página 47.

GEMMEKE, J. F. et al. Audio set: An ontology and human-labeled dataset for audio events. In: IEEE. *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. [S.l.], 2017. p. 776–780. Citado 3 vezes nas páginas 39, 40 e 41.

GROSCHE, P.; MÜLLER, M.; KURTH, F. Cyclic tempogram—a mid-level tempo representation for musicsignals. In: IEEE. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.], 2010. p. 5522–5525. Citado na página 29.

GUIDO S.; MULLER, A. C. *Introduction to Machine Learning with Python: a guide for data scientists*. [S.l.]: O'reilly, 2016. Citado na página 25.

HARTE, C.; SANDLER, M.; GASSER, M. Detecting harmonic change in musical audio. In: *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. [S.l.: s.n.], 2006. p. 21–26. Citado na página 29.

HOWARD, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. Citado na página 39.

JEONG, I.-Y.; LEE, K. Learning temporal features using a deep neural network and its application to music genre classification. In: *Ismir*. [S.l.: s.n.], 2016. p. 434–440. Citado na página 20.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. Citado na página 24.

KAUFMAN, L.; ROUSSEEUW, P. *Finding Groups in Data: An Introduction To Cluster Analysis*. [S.l.: s.n.], 1990. ISBN 0-471-87876-6. Citado na página 25.

KERELIUK, C.; STURM, B. L.; LARSEN, J. Deep learning, audio adversaries, and music content analysis. In: IEEE. *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. [S.l.], 2015. p. 1–5. Citado na página 20.

KIM, Y. E. et al. Music emotion recognition: A state of the art review. In: *Proc. ismir*. [S.l.: s.n.], 2010. v. 86, p. 937–952. Citado na página 23.

KINGMA, D. P. et al. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, v. 27, 2014. Citado na página 25.

KNEES, P.; SCHEDL, M. Music retrieval and recommendation: A tutorial overview. In: *Proceedings of the 38th International ACM SIGIR conference on research and development in information retrieval*. [S.l.: s.n.], 2015. p. 1133–1136. Citado 2 vezes nas páginas 19 e 23.

KRAWCZYK, B. et al. Ensemble learning for data stream analysis: A survey. *Information Fusion*, Elsevier, v. 37, p. 132–156, 2017. Citado na página 28.

LAMERE, P. Social tagging and music information retrieval. *Journal of new music research*, Taylor & Francis, v. 37, n. 2, p. 101–114, 2008. Citado 2 vezes nas páginas 19 e 23.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015. Citado 2 vezes nas páginas 24 e 26.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Education(ISE Editions), 1997. Citado na página 24.

MÜLLER, M. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. [S.l.]: Springer, 2015. v. 5. Citado na página 28.

NANNI, L. et al. Ensemble of deep learning, visual and acoustic features for music genre classification. *Journal of New Music Research*, Taylor & Francis, v. 47, n. 4, p. 383–397, 2018. Citado na página 19.

NANNI, L. et al. Combining visual and acoustic features for audio classification tasks. *Pattern Recognition Letters*, Elsevier, v. 88, p. 49–56, 2017. Citado na página 19.

ORAMAS, S. et al. Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint arXiv:1707.04916*, 2017. Citado na página 19.

SCHEDL, M. et al. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc., v. 8, n. 2-3, p. 127–261, 2014. Citado na página 19.

SIGTIA, S.; DIXON, S. Improved music feature learning with deep neural networks. In: IEEE. *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. [S.l.], 2014. p. 6959–6963. Citado na página 20.

SILLA, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, Springer, v. 22, p. 31–72, 2011. Citado na página 28.

SILVA, V. da; WINCK, A. T. Multi-label classification of music into genres. *Encontro Nacional de Inteligencia Artificial e Computacional (ENIAC 2014)*, 2014. Citado na página 28.

TORREY, L.; SHAVLIK, J. Transfer learning. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. [S.l.]: IGI global, 2010. p. 242–264. Citado na página 30.

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 3 vezes nas páginas 23, 27 e 31.

VLACHOS, M. Dimensionality reduction. In: _____. *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. p. 274–279. ISBN 978-0-387-30164-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_216>. Citado na página 25.

WAN, Y. *Deep learning for music classification*. Tese (Doutorado), 2016. Citado 3 vezes nas páginas 19, 23 e 26.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. *Journal of Big data*, SpringerOpen, v. 3, n. 1, p. 1–40, 2016. Citado na página 27.

WITTEN, I. H. et al. Practical machine learning tools and techniques. *Data Mining. Fourth Edition, Elsevier Publishers*, 2017. Citado na página 28.

WON, M.; SPIJKERVET, J.; CHOI, K. Music classification: beyond supervised learning, towards real-world applications. *arXiv preprint arXiv:2111.11636*, 2021. Citado na página 29.

XIE, Q. et al. Self-training with noisy student improves imagenet classification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 10687–10698. Citado na página 25.

YALNIZ, I. Z. et al. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. Citado na página 25.