CESAR SCHOOL

LUCAS ANDRADE ABRÃO

AVALIAÇÃO DE MODELOS DE GERAÇÃO MUSICAL NO ESTILO FREVO POR REDES NEURAIS

RECIFE

2024

LUCAS ANDRADE ABRÃO

AVALIAÇÃO DE MODELOS DE GERAÇÃO MUSICAL NO ESTILO FREVO POR REDES NEURAIS

Dissertação apresentada ao programa de Mestrado em Engenharia de Software do Centro de Estudos e Sistemas Avançados do Recife – CESAR School, como requisito para a obtenção do título de Mestre em Engenharia de Software.

Orientação: Prof. André Câmara Alves do Nascimento

Coorientação: Profa. Ana Paula Cavalcanti Furtado

LUCAS ANDRADE ABRÃO

AVALIAÇÃO DE MODELOS DE GERAÇÃO MUSICAL NO ESTILO FREVO POR REDES NEURAIS

Trabalho Aprovado. Recife: **09/04/2024**

Professor: André Câmara Alves do Nascimento
(CESAR School/UFRPE)
Orientador

Professor: Rafael Ferreira Leite de Mello
(CESAR School/UFRPE)
Avaliador Interno

Professor: Péricles Barbosa Cunha de Miranda
(UFRPE)
Avaliador Externo

RECIFE
2024

Dedicatória

Este trabalho é totalmente dedicado a Deus pela sua bondade, favor e pela obra maravilhosa do Senhor Jesus Cristo na cruz.

Eu te amo, Deus!

Dedico também aos meus pais por todo amor,

carinho e dedicação a todos os seus filhos.

Obrigado por tudo!

Agradecimentos

Agradeço primeiramente a Deus pelo dom da vida e por ter me abençoado em todos os instantes da minha vida, sempre com Sua graça, amor e proteção. Obrigado, Deus!

Aos meus pais, Edir Abrão e Elizabeth Abrão, por todo o carinho, dedicação, bom exemplo e honestidade, que mesmo nas maiores dificuldades da vida sempre demonstraram amor a Deus, pelos filhos e sendo um verdadeiro referencial de integridade e honestidade. Palavras não conseguem expressar a gratidão por tudo o que fizeram por todos os que lhes cercam!

Aos meus irmãos e irmãs por fazerem da nossa família ser única e especial, mesmo com todos os defeitos, as qualidades e amor que sentimos uns pelos outros sempre são maiores do que qualquer diferença.

Aos meus amigos que sempre me incentivaram e aos colegas que me ajudaram nesse percurso.

Ao meu Orientador, Professor Dr André Câmara, pelo incentivo, apoio, direção e profissionalismo na condução desta pesquisa.

Ao CESAR School, pelo excelente quadro de professores, orientadores e profissionais que fazem da Instituição um verdadeiro centro de pesquisa e inovação.

Resumo

Explorando a convergência entre avanços em Inteligência Artificial e expressão musical, este estudo tem por objetivo investigar a aplicação de técnicas avançadas de Deep Learning no contexto de geração automática de músicas. Com ênfase em modelos como LSTM, WaveGAN, CNN, Autoencoders e GRU, a pesquisa visa à criação de séries temporais de áudio, com o objetivo de compor músicas no estilo Frevo - um gênero musical autêntico e influente no Nordeste do Brasil. A base de dados, composta por 30 músicas de Frevo, servirá como peças-chave para o treinamento das redes neurais. A avaliação do desempenho das redes foi realizada por meio de métricas como MAE, MSE, RMSE, SNR e PSNR, complementadas por métricas espectrais, avaliação visual, SPL e SNDR, proporcionando uma análise abrangente da qualidade das séries temporais de áudio geradas. Ao final, foram realizadas análises e validações do desempenho de cada rede neural utilizando os mesmos hiperparâmetros ou dados próximos baseados em sua arquitetura, respectivamente. Essas análises mostraram que algumas redes têm uma geração de áudio melhor e mais próxima da realidade (como a rede GRU) do que outras (como os Autoencoders), com base nos valores atribuídos aos hiperparâmetros neste estudo. O estudo destaca não apenas a capacidade das redes neurais na criação personalizada de conteúdo musical, mas também busca enriquecer a expressividade artística ao conectar inovações tecnológicas à rica e única tradição musical do Frevo. Assim, ao ajudar as pessoas a compor músicas no estilo Frevo utilizando a assistência da inteligência artificial, o estudo aponta para a importância de identificar a melhor rede neural para esse fim entre as redes neurais analisadas pelas métricas utilizadas neste estudo.

Palavras-chave

Áudio; deep learning; inteligência artificial; métricas; redes neurais.

Abstract

Exploring the convergence between advances in Artificial Intelligence and musical expression, this study aims to investigate the application of advanced Deep Learning techniques in the context of automatic music generation. With an emphasis on models such as LSTM, WaveGAN, CNN, Autoencoders, and GRU, the research seeks to create temporal series of audio with the goal of composing music in the Frevo style - an authentic and influential musical genre in Northeast Brazil. The dataset, consisting of 30 Frevo songs, serves as key pieces for training the neural networks. The performance evaluation of the networks was conducted using metrics such as MAE, MSE, RMSE, SNR, and PSNR, complemented by spectral metrics, visual evaluation, SPL, and SNDR, providing a comprehensive analysis of the quality of the generated audio temporal series. In the end, analyses and validations of the performance of each neural network were made using the same hyperparameters or close data based on their architecture, respectively. These analyses showed that some networks have better audio generation and are closer to reality (such as the GRU network) than others (such as Autoencoders), based on the values assigned to the hyperparameters in this study. The study highlights not only the ability of neural networks in the personalized creation of musical content but also seeks to enrich artistic expressiveness by connecting technological innovations with the rich and unique musical tradition of Frevo. Thus, by helping people compose music in the Frevo style using the assistance of artificial intelligence, the study points to the importance of identifying the best neural network for this purpose among the neural networks analyzed by the metrics used in this study.

Key-words

Artificial intelligence; audio; deep learning; metrics; neural networks.

SUMÁRIO

Lista de ilustrações

Lista de tabelas

1 INTRODUÇÃO

A música é uma linguagem universal que transcende as barreiras do idioma e da cultura. Ela é uma parte essencial da experiência humana, e desempenha um papel importante em muitas esferas da vida, como a religião, a cultura e a sociedade. A música pode nos emocionar, nos inspirar e nos unir. Ela pode nos ajudar a expressar nossas emoções, a compartilhar nossas experiências e a construir laços sociais. Para além disso, especialistas em antropologia, musicologia e linguística observam a importância essencial que a linguagem e a música têm na nossa identidade coletiva, desde a formação da nossa espécie até a estruturação em grupos sociais como clãs, tribos, nações e culturas, é o que defende Diana Santiago no livro "*Prática musical, memória e linguagem*" (2018).

1.1 PROBLEMA

Avanços significativos têm sido feitos na geração de música por Inteligência Artificial (IA), como evidenciado por estudos recentes. Destaca-se o trabalho de Su e outros (2018), que apresentaram o MuseGAN, uma rede adversarial generativa (GAN) capaz de produzir música de alta qualidade em diversos gêneros, como também o trabalho de Yang e outros (2017), que produziram o MidiNet, que ajuda na geração musical no domínio simbólico e dentre vários outros, como o EmotionBox por Kaitong Zheng (2021), que produz música emocional automática, por exemplo.

O foco deste trabalho é gerar música no estilo Frevo pela falta de material que explore esse rítmo, que é singular, e que pela pouca produção sobre a geração musical neste estilo por IA, como também a sua importância por ser um estilo popular e muito utilizado na região Nordeste do Brasil, é interessante ter experimentos e estudos voltados para a produção musical neste contexto, tanto a níveis de pesquisa computacionais como a nível social.

A geração musical baseada em IA possui potencial para criar mais músicas em menos tempo, proporcionando uma cena musical mais diversa e vibrante. Ademais, a IA pode personalizar a música para atender aos gostos individuais de

cada ouvinte, tornando-a mais acessível, permitindo que qualquer pessoa crie música, independentemente de habilidades ou treinamento musical.

Diante da relevância da música, especialmente no contexto do Frevo, um dinâmico gênero musical brasileiro distinguido como Patrimônio Cultural Imaterial da Humanidade pela UNESCO (2012), a aplicação de inteligência artificial (IA) na geração de áudio se apresenta como um desafio intrigante. A fidedigna reprodução de elementos distintivos, como o ritmo acelerado, instrumentos de percussão e melodias vibrantes, emerge como uma exigência complexa. A compreensão dessas complexidades no contexto específico do Frevo se torna crucial para explorar plenamente o potencial da IA na criação musical.

Neste cenário desafiador, o problema de pesquisa se concentra na busca por soluções que superem as dificuldades intrínsecas à geração de áudio por IA no contexto do Frevo. Como lidar de maneira eficaz com a reprodução fiel de características tão distintivas desse gênero musical? Esta questão orienta a pesquisa em direção a estratégias inovadoras e abordagens que levem em consideração a singularidade do Frevo, contribuindo para avanços significativos na interseção entre música e a inteligência artificial analisando através de métricas em diferentes áreas de construção musical: Precisão, qualidade de áudio, como também análise espectral e visual-espectral.

Diante do desafio de empregar inteligência artificial (IA) na geração de áudio para o Frevo, várias questões orientam a pesquisa vindas de outras ideias abordadas por outros trabalhos sobre o tema de geração musical e áudio, levando em consideração o contexto do Frevo:

a) Reprodução do Ritmo Acelerado: Arquiteturas de geração musical são capazes de reproduzir com precisão o característico ritmo acelerado do Frevo? Esta pergunta destaca a necessidade de estratégias inovadoras para garantir a autenticidade na reprodução desse elemento crucial do gênero musical. Este tipo de abordagem já foi analisado por outros estudos, como o “Geração de Músicas Polifônicas utilizando Redes Neurais Artificiais” de Mendes (2019). Neste estudo, foi feita a geração de melodias através de Redes Neurais Artificiais levando em

consideração o contexto de músicas polifônicas. Não desenvolve o Frevo em si, mas é um contexto em que a aplicação para outros estilos possui forte aplicação, incluindo o Frevo.

b) Avaliação da Qualidade do Áudio: Como podemos determinar a eficácia da IA na geração de áudio no contexto do Frevo, considerando suas características únicas? Essa questão ressalta a importância de uma abordagem abrangente de avaliação, que englobe aspectos quantitativos, subjetivos e espectrais. A rápida cadência rítmica, variações melódicas e elementos culturais distintivos do Frevo demandam uma avaliação que vá além de métricas puramente técnicas. Portanto, além de métricas de precisão como MAE, MSE, RMSE e qualidade de sinal, como SNR e PSNR, por exemplo, é fundamental incorporar avaliações visuais-espectrais, que capturem a percepção humana da autenticidade e expressividade do áudio gerado. Além disso, métricas espectrais podem oferecer insights sobre a fidelidade na reprodução das características acústicas distintivas do Frevo. Essa abordagem multifacetada não apenas enriquece a compreensão do desempenho dos modelos de IA, mas também reflete a complexidade e riqueza do contexto musical do Frevo, contribuindo para o desenvolvimento de modelos mais contextualmente relevantes e eficazes

c) Análise Visual-Espectral: Como a análise visual-espectral pode complementar a análise da geração de áudio no contexto do Frevo? A incorporação de análises visuais-espectrais oferece uma perspectiva adicional na avaliação da fidelidade e autenticidade das produções musicais geradas. Ao examinar características como a distribuição de energia ao longo do espectro de frequência e a presença de padrões visuais que refletem aspectos intrínsecos do Frevo, como o brilho e a qualidade do ritmo, podemos obter ideias valiosas sobre a qualidade perceptual do áudio gerado, essa é a ideia defendida por Wyse (2017).

1.1.1 Objetivo geral

O objetivo central desta pesquisa é avaliar a aplicação de técnicas de geração de áudio baseadas em inteligência artificial (IA). Ao abordar especificamente o

contexto do Frevo, um gênero musical rico e acelerado, busca-se desenvolver estratégias inovadoras e aprimorar métodos analíticos, incorporando métricas específicas e considerando a percepção visual-espectral geradas. Essa abordagem visa não apenas atingir resultados técnicos avançados na geração de áudio por IA, mas também contribuir para o enriquecimento da interseção entre música e tecnologias inteligentes, promovendo avanços mais amplos no campo da inteligência artificial aplicada à produção musical.

1.1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

a) Levantamento Bibliográfico: Realizar uma revisão abrangente da literatura relacionada à aplicação de inteligência artificial na geração de áudio, com foco nas técnicas utilizadas para reproduzir características específicas do Frevo. Isso inclui a investigação de arquiteturas de redes neurais, algoritmos de processamento de áudio e métodos de avaliação de qualidade;

b) Mapeamento de Métricas Adequadas: Identificar e selecionar métricas adequadas para avaliar a fidelidade e autenticidade do áudio gerado pela IA no contexto do Frevo. Isso envolve não apenas métricas de precisão, como MAE, MSE, RMSE e de qualidade de sinal como SNR e PSNR, mas também métricas espectrais e avaliação visual-espectral para capturar aspectos perceptivos e de sonoridades relevantes. Esse é o entendimento do estudo *“Quality Enhancement for Highly Degraded Music Using Deep Learning-Based Prediction for Lost Frequencies”* por Serra e outros (2021);

c) Avaliação de Métodos do Estado da Arte: Analisar e comparar os métodos do estado da arte em geração de áudio por IA, com foco na sua capacidade de reproduzir o ritmo acelerado, instrumentos de percussão e melodias características do Frevo. Isso inclui experimentos utilizando diferentes arquiteturas de redes neurais, como LSTM, WaveGAN, CNN, Autoencoders e GRU, para determinar quais apresentam melhor desempenho na geração de áudio no estilo Frevo;

d) Avaliação Visual-Espectral: Realizar avaliações visuais-espectrais para entender a percepção visual dos diferentes tipos de espectrogramas gerados a partir das diferentes neurais e fazer uma comparação visual do que foi gerado em uma rede e que foi diferente em outra, como tamanho de onda, espaço e tempo gerado etc. Este tipo de avaliação foi utilizado no trabalho *"Audio spectrogram representations for processing with convolutional neural networks"* por Wyse (2017).

Esses objetivos específicos são essenciais para alcançar o objetivo central da pesquisa, que é avaliar as aplicações de técnicas de geração de áudio baseadas em inteligência artificial no contexto do Frevo. Ao realizar cada etapa de forma metódica e abrangente, espera-se contribuir significativamente para o avanço do conhecimento na interseção entre música e inteligência artificial, promovendo o desenvolvimento de modelos mais contextualmente relevantes e eficazes para a geração de áudio musical.

1.2 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está estruturado da seguinte forma: No Capítulo 1, é apresentada a contextualização do trabalho, do problema a ser estudado, as perguntas que sustentam a razão para essa pesquisa, como os objetivos gerais e específicos para este estudo. No Capítulo 2, é fornecido um embasamento teórico, onde é descrito em melhores detalhes sobre Aprendizagem de Máquina voltada para o contexto de geração musical e sobre os principais tipos de Aprendizagem de Máquina, bem como as redes neurais e as métricas utilizadas para a validação desta pesquisa. No Capítulo 3, os trabalhos relacionados ao tema de geração musical com Inteligência Artificial são explorados, mostrando outros casos em que os mais diferentes tipos de redes neurais são utilizados em diferentes contextos, mas com o principal foco na geração musical. No Capítulo 4, é apresentado o Método com a Proposta, a montagem e configuração de todo o ambiente de execução dos testes para essa pesquisa.

No Capítulo 5 é mostrado os experimentos e os resultados em si, destacando as Avaliações de Precisão, Qualitativas, Espectrais e Visuais-Espectrais na

produção musical no estilo Frevo em cada um dos diferentes tipos de redes neurais propostos e discutidos no Capítulo 2. As conclusões mais assertivas de forma mais abrangente também são apresentadas neste capítulo. No Capítulo 6, é apresentada a Conclusão Final, com as limitações encontradas nesta pesquisa e as ponderações sobre os trabalhos futuros do que pode ser feito. Algumas sugestões são colocadas como possíveis temas de pesquisa para que os pesquisadores possam ter uma melhor referência do que e como estruturar melhor suas respectivas pesquisas.

2 EMBASAMENTO TEÓRICO

Segundo o entendimento de Nguyen (2019), Aprendizagem de Máquina é a aplicação de estatística em modelos computacionais que fazem predições baseadas em uma base de dados fornecida previamente.

Essa abordagem sobre a capacidade dos computadores de aprender com a experiência e melhorar a sua performance automaticamente, também é defendida por Michie, Spiegelhalter e Taylor (1994), e tem sido aplicada em diversas áreas, incluindo a geração de música. Faul (2019) expande essa definição ao descrever o aprendizado como um fenômeno complexo que envolve a descoberta de novos conhecimentos através da observação e experimentação. Neste contexto, o aprendizado de máquina na geração musical automática pode ser compreendido como a pesquisa e digitalização dos processos de aprendizagem musical em diferentes níveis de representações.

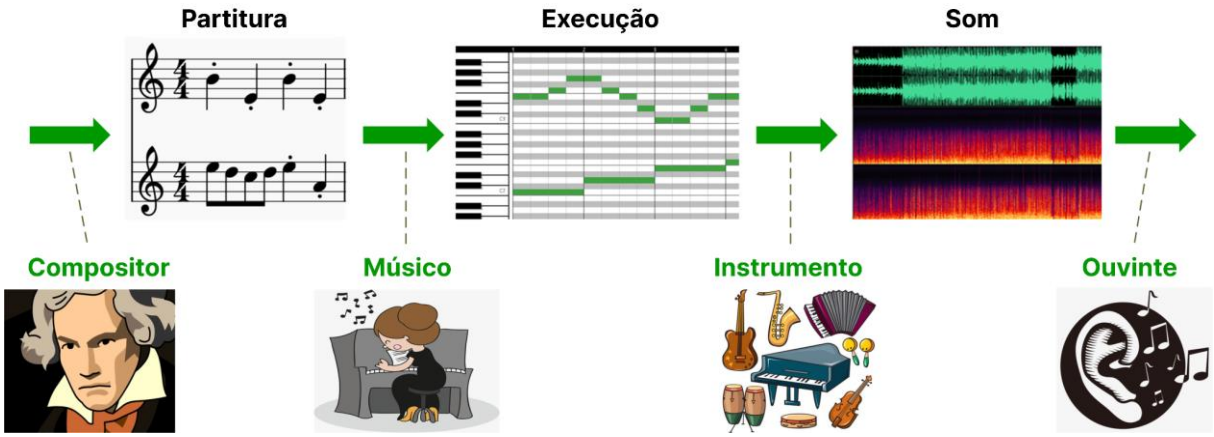
Agora, ao explorar a geração musical automática, podemos compreender como os princípios do aprendizado de máquina são aplicados para criar composições musicais de forma autônoma e inovadora. Vamos examinar mais detalhadamente como esses princípios se entrelaçam para moldar a paisagem da música atualmente.

2.1 GERAÇÃO MUSICAL AUTOMÁTICA

Cerca de duzentos anos atrás, o renomado poeta americano Henry Wadsworth Longfellow supostamente afirmou: "A Música é a linguagem universal da humanidade", uma ideia que recebeu validação de pesquisadores da Universidade de Harvard em 2019, liderados pelo cientista Samuel Mehr - membro da Harvard Data Science Initiative. Segundo o entendimento do trabalho “*A Comprehensive Survey on Deep Music Generation: Multi-Level Representations, Algorithms, Evaluations, and Future Directions*” de Shulei Ji e outros (2020), a correlação entre a notação musical e sua manifestação auditiva espelha a relação entre texto escrito e linguagem falada. A notação musical funciona como uma representação visual

profundamente simbólica e abstrata capaz de preservar e comunicar efetivamente ideias musicais, enquanto a experiência auditiva abrange um sinal contínuo e tangível que codifica todos os detalhes discerníveis. Essas duas formas podem ser delineadas em níveis distintos, com a notação ocupando o patamar superior e o som residindo abaixo. A interpretação e expressão da música predominantemente dependem de nuances de performance; por exemplo, alterar o tempo de uma composição animada pode conferir-lhe um tom melancólico. Consequentemente, um extrato intermediário pode ser introduzido entre as camadas superior e inferior para delinear complexidades de performance. Assim, o processo de geração de música geralmente se desdobra em três estágios, conforme ilustrado na Figura 1: durante o estágio inicial, os compositores elaboram partituras musicais; na fase subsequente, os intérpretes executam essas partituras, dando vida à música; por fim, no estágio final, a performance passa por sonificação, com timbres variados (instrumentos) sobrepostos, sendo percebida pelo público humano.

Figura 1 - Processo de geração musical



Fonte: Ji, Luo e Yang, 2020, p. 2.

No cenário musical contemporâneo, a inteligência artificial (IA) desempenha um papel significativo, especialmente na geração de áudio. Um exemplo notável dessa aplicação é o contexto do Frevo, um enérgico gênero musical brasileiro reconhecido pela UNESCO (2012) como Patrimônio Cultural Imaterial da Humanidade. No entanto, a geração de áudio para o Frevo por meio de IA apresenta desafios complexos, exigindo a reprodução precisa de elementos característicos, como o ritmo acelerado, os instrumentos de percussão e as melodias vibrantes.

Nesse cenário desafiador, diversos estudos têm contribuído para avanços significativos. O trabalho *"MuseGAN: A Musical Generative Adversarial Network"* (Su et al., 2018) destaca-se ao apresentar o MuseGAN, um modelo baseado em Redes Adversariais Generativas (GANs) capaz de gerar música de alta qualidade em diferentes gêneros. Métricas como Diferença Média Quadrática (MSE), Razão de Sinal para Ruído (SNR) e Razão de Sinal para Ruído Normalizado (PSNR) foram utilizadas para avaliar a qualidade das produções. Além disso, em *"A Comparative Study on Variational Autoencoders and Generative Adversarial Networks"* (Sami, 2019), a comparação entre duas técnicas de aprendizado de máquina para geração de dados, incluindo música – Variational Autoencoders (VAEs) e Generative Adversarial Networks (GANs) – revela insights valiosos. Os resultados sugerem que GANs são mais adequadas para aplicações que demandam originalidade e criatividade, enquanto VAEs destacam-se em cenários que exigem consistência com um conjunto de dados de treinamento.

Em paralelo, o artigo *"MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation"* (Yang et al., 2017) introduz uma Rede Neural Convolucional Generativa Adversarial (GAN) específica para a geração de música no domínio simbólico. Treinado em um conjunto de dados de melodias MIDI, o MidiNet demonstra a capacidade de criar melodias novas e originais. Esses estudos não apenas refletem a diversidade de abordagens na interseção entre música e IA, mas também indicam uma evolução contínua nesse campo promissor.

2.2 FREVO

Como nosso foco é gerar série temporal de áudio no ritmo de Frevo, vamos a um breve histórico para contextualização. Segundo Diniz (2012), a origem do frevo ainda é um tema controverso. Alguns estudiosos acreditam que o frevo surgiu no final do século XIX, como uma mistura de ritmos africanos e europeus. Outros acreditam que o frevo tem origens mais antigas, remontando ao período colonial. O frevo é um ritmo acelerado e contagiante. A música do frevo é geralmente tocada por bandas de frevo, que utilizam instrumentos como trombones, saxofones,

trompetes e bateria. A dança do frevo é caracterizada por passos rápidos e acrobáticos, realizados pelos foliões.

A dança do frevo é uma forma de expressão corporal que combina movimentos rápidos e acrobáticos com passos tradicionais da capoeira. Os foliões usam adornos coloridos, como chapéus, lenços e guarda-chuvas, que são utilizados para criar efeitos visuais. O frevo é um dos símbolos da cultura popular brasileira. É uma manifestação cultural viva, que é transmitida de geração em geração. O frevo é praticado em diversas ocasiões, como Carnaval, festas populares e eventos culturais. Segundo a UNESCO, o frevo foi inscrito na Lista Representativa do Patrimônio Cultural Imaterial da Humanidade em 2012. A inscrição é um reconhecimento da importância do frevo como manifestação cultural brasileira.

2.3 TIPOS DE APRENDIZAGEM DE MÁQUINA

O campo de Aprendizagem de Máquina, uma vertente da Inteligência Artificial, tem se destacado devido ao aumento do poder computacional e à disponibilidade de vastos conjuntos de dados. Conforme observa Latah e outros (2016), utiliza-se de dados disponíveis em seu ambiente para aprender com a experiência, e é usado para melhorar o desempenho. Há três tipos de abordagem de Aprendizagem de Máquina: Aprendizagem Supervisionada, Aprendizagem Não-Supervisionada e a Aprendizagem Por Reforço. As principais características de cada uma delas é exposta abaixo.

A Aprendizagem Não-Supervisionada, uma faceta relevante da Aprendizagem de Máquina, destaca-se por descobrir padrões ocultos sem intervenção humana, sendo amplamente aplicada em setores como marketing, segmentação de clientes e sistemas de recomendação, além de modelagem de clusterização e redução dimensional (Ghahramani, 2003). A Aprendizagem Supervisionada, por sua vez, segundo Carmo (2013), é uma classe de algoritmos dentro do aprendizado de máquina que possui a característica de inferir uma função dado um conjunto de dados que é chamado de conjunto de treinamento. Por meio desse algoritmo, os valores irão se corrigindo e adequando sua função com o valor esperado na saída.

Por outro lado, a Aprendizagem por Reforço, outra área crucial da Aprendizagem de Máquina, concentra-se na capacidade de um agente aprender a tomar decisões em ambientes dinâmicos. Xu (2020) amplia essa perspectiva, indicando que o aprendizado por reforço abrange tarefas de previsão e controle, sendo crucial para estimar valores de políticas e encontrar políticas ideais que maximizem as recompensas esperadas.

2.3.1 *Deep Learning* - técnica para implementar a Inteligência Artificial

Uma abordagem em forma de algoritmo do início de movimento de Aprendizagem de Máquina, Redes Neurais Artificiais, surgiu e desapareceu ao longo do tempo. Redes neurais são uma forma de representar o entendimento do cérebro humano – interconexões entre neurônios (perceptrons, no caso de rede neural). Todavia, diferente de um cérebro biológico onde qualquer neurônio pode se conectar com qualquer outro neurônio dentro de uma certa distância física, essas redes neurais artificiais têm camadas discretas, conexões e direções de propagação de dados e formas como esses dados serão tratados.

> Não há ainda um consenso real para a definição de Deep Learning. É um repertório de técnicas de Machine Learning baseados em redes neurais artificiais. O aspecto principal que todos concordam é no termo deep, ou seja, são muitas camadas de processando muitos níveis de abstração hierárquica, nos quais, são automaticamente extraídos do dado (Briot; Jean-Pierre, 2020).

Aplicações que usam *Deep Learning* têm feito tarefas das mais variadas dentro de *Machine Learning*, tipo: Classificação, predição e, recentemente, tradução. Para cada tipo de tarefa, existe uma melhor abordagem e tipo de rede neural que se adequa melhor à realidade desta rede neural. Os principais tipos de redes neurais existentes são: ANN (*Artificial Neural Network*), CNN (*Convolutional Neural Networks*) e a RNN (*Recurrent Neural Network*), sendo que elas também possuem subtipos.

2.4 GERAÇÃO MUSICAL BASEADA EM APRENDIZAGEM DE MÁQUINA

A Inteligência Artificial (IA) tem o potencial de revolucionar a indústria da Música. McFee (2019), em um estudo recente, identificou duas abordagens

principais para a geração musical baseada em IA. A primeira é a abordagem baseada em regras, que utiliza um conjunto de regras, podendo ser derivadas da teoria musical, padrões de uso comum ou preferências pessoais do usuário. A segunda, é a abordagem baseada em aprendizado de máquina, na qual um algoritmo aprende a gerar música a partir de dados de exemplos, podendo ser treinado em músicas existentes ou criado do zero.

McFee (2019) também identificou desafios a serem superados para que a geração musical baseada em IA alcance seu pleno potencial. Esses desafios incluem a necessidade de desenvolver algoritmos que possam criar músicas verdadeiramente originais, abordar questões de controle dando aos usuários a capacidade de especificar estilo e outras características, e considerar aspectos éticos, como o potencial de plágio e a exploração de músicos humanos. Apesar desses obstáculos, a geração musical baseada em IA tem o potencial de ser uma ferramenta poderosa para criar músicas novas e emocionantes. À medida que a tecnologia de IA continua a se desenvolver, podemos esperar ver ainda mais avanços na música impulsionada por Inteligência Artificial no futuro.

2.5 ABORDAGENS SUPERVISIONADAS

> Aprendizado supervisionado, é uma classe de algoritmos dentro do aprendizado de máquina que possui a característica de inferir uma função dado um conjunto de dados que é chamado de conjunto de treinamento. Esse conjunto é dado pelo par entrada e saída onde a entrada possui as características do exemplo em si e a saída o valor desejado. Através desse conjunto de dados o algoritmo irá se corrigindo e adequando a sua função com o valor esperado de saída (Carmo, 2013, p. 25).

Esse tipo de abordagem é caracterizado pela utilização de conjuntos de dados previamente rotulados para treinar algoritmos, os quais desempenham tarefas como a classificação de dados ou a previsão de resultados com alta precisão. Durante o processo de treinamento, à medida que os dados de entrada são introduzidos no modelo, suas ponderações são ajustadas iterativamente até que o modelo alcance um ajuste adequado, processo que frequentemente faz parte da validação cruzada. O aprendizado supervisionado desempenha um papel crucial na resolução de uma variedade de problemas do mundo real em grande escala,

exemplificado, por exemplo, na classificação de e-mails como *spam*, direcionando-os para uma pasta separada na caixa de entrada.

2.5.1 Redes LSTM (*Long Short-Term Memory*)

Redes LSTM são um tipo de RNN (*Recurrent Neural Network*) com modificações na estrutura em suas camadas. Em uma rede neural LSTM, busca-se propagar a entrada no tempo t, assim como o estado da rede no passo anterior através do uso de estruturas denominadas portões.

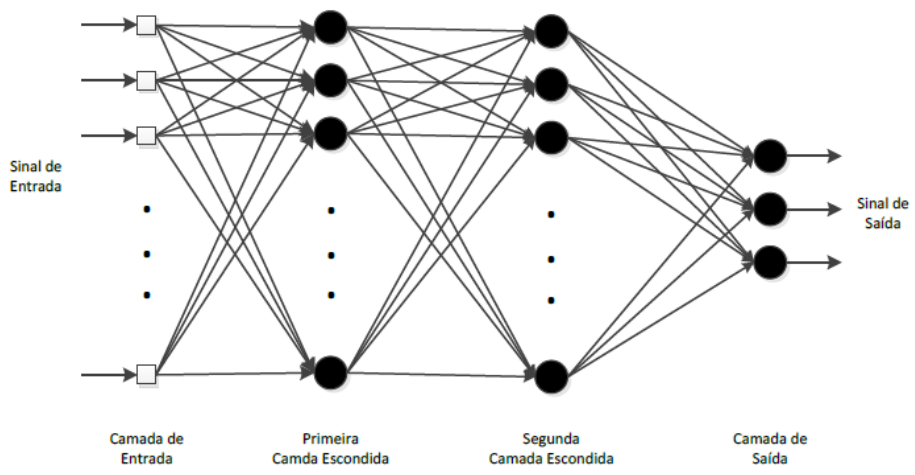
Especificamente, uma célula LSTM contém três tipos de portões de controle:

a) portão para decidir qual informação não é relevante, o portão de esquecimento (do inglês, *Forget Gate*);

b) portão para definir quais entradas serão utilizadas para atualizar as células de memória, o portão de entrada (do inglês, *Input Gate*);

c) portão que decide quais serão as saídas da célula, levando em conta qual a entrada e qual o estado da célula de memória, o portão de saída (do inglês, *Output Gate*) (Brownlee, 2017).

Figura 2 - Rede Neural Artificial Multicamadas

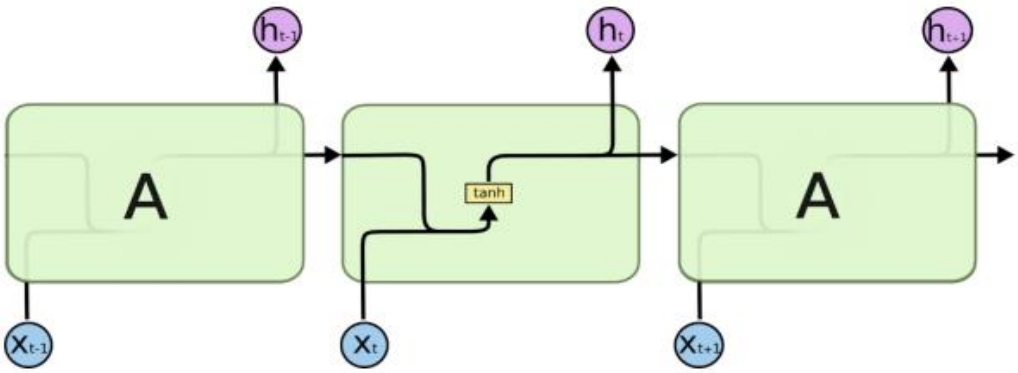


Fonte: Do Carmo, 2013, p. 37.

Cada uma das conexões numa rede neural tem um peso associado a ela. Um perceptron pode pesar altamente um input se ele tem um impacto significante. Como uma rede neural é treinada, ela já está ciente do que será o seu *output* depois de treinar o modelo. O *output* atual e as funções de perda (*loss functions*) pontuam quão bem a rede performou. As funções de perda podem ser a diferença absoluta esperada e a real ou o erro quadrático médio de muitas outras funções. A rede neural irá tentar minimizar as perdas ajustando os pesos das conexões através de um processo de chamada de retorno. As interações ANN envolvem os *inputs* serem calculados e passados pelos perceptrons até os *outputs* finais, na qual é usado em uma função de perda com o resultado esperado, e finalmente os pesos são ajustados através da propagação de retorno. Um realce além do ANN é a rede *Long Short-Term Memory* (LSTM). Esse tipo de rede mostrado na Figura 2 possibilita um modelo considerar o que ocorreu previamente bem como um conjunto normal de entradas. A rede LSTM terá um papel fundamental nesse projeto até mesmo porque as notas musicais têm uma relação lógica com a sequência de notas de notas que a precederam.

Segundo o entendimento de Nguyen (2019) e mostrado na Figura 3, este tipo de rede possibilita que um modelo considere o que ocorreu anteriormente, assim como o conjunto normal de entradas. Isso porque na rede LSTM, as notas musicais estão logicamente ligadas às notas que as precedem.

Figura 3 - Diagrama de uma *Long Short-Term Memory Network*



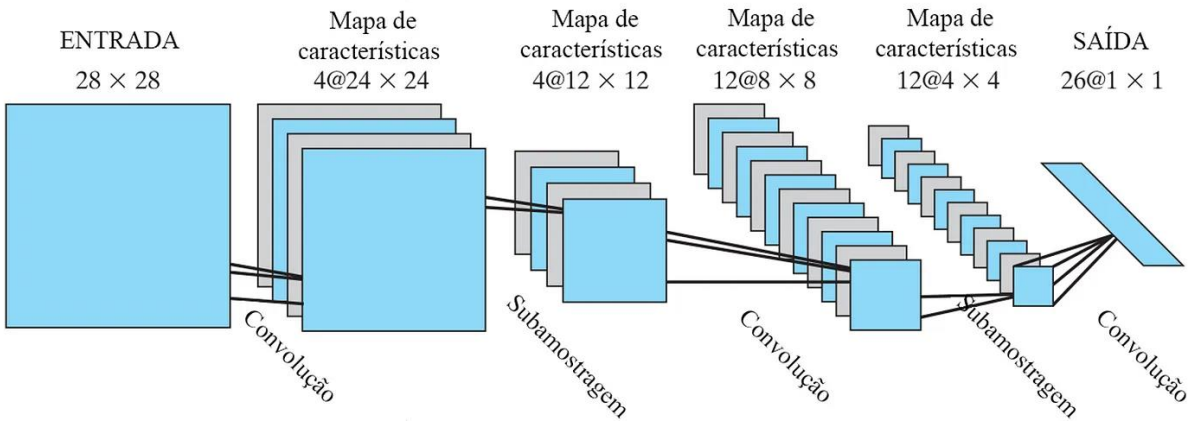
Fonte: Nguyen, 2019.

Neste caso, temos uma entrada **xt**, por exemplo, que passa por todo um circuito para, então, gerar um estado **ht**. Depois há um estado anterior, mas o que gera o próximo estado não é simplesmente uma mesma camada que está se retroalimentando.

2.5.2 Redes CNN (*Convolutional Neural Network*)

Redes neurais convolucionais (CNNs) são um tipo de rede neural artificial que é amplamente usada em tarefas de visão computacional. CNNs são inspiradas na estrutura do cérebro humano e são capazes de aprender representações de dados visuais que são relevantes para a tarefa em questão, conforme representado na Figura 4.

Figura 4 - Rede Neural Convolucional (LeNet)



Fonte: Imagem extraída, editada e traduzida de Lecun (1998).

Um trabalho importante sobre CNNs é o *"CNN Music Emotion Classification"*, de Xin Liu e outros (2017). Esse trabalho introduziu um método que combina espectrogramas de música com uma CNN para prever *tags* emocionais. Apesar do ganho de desempenho, há espaços para ajustes na arquitetura CNN e na seleção de segmentos de música. Além disso, concluiu-se que a falta de compreensão das saídas da CNN dificulta as interpretações das emoções inspiradas pela música.

Redes Neurais Convolucionais (CNNs) destacam-se em tarefas de visão computacional, como classificação de imagens, detecção de objetos e segmentação de imagens. Além de sua eficácia nessas áreas, sua relativa facilidade de treinamento através de técnicas de aprendizado supervisionado, permitindo a aprendizagem de mapeamentos de entradas para saídas desejadas, contribui para sua popularidade. A escalabilidade das CNNs é notável, possibilitando o treinamento em conjuntos de dados extensos. No contexto de aplicações, as CNNs desempenham papeis fundamentais em diversas áreas, como reconhecimento facial, identificação de objetos, segmentação de imagens, geração de imagens, processamento de linguagem natural e reconhecimento de voz, consolidando-se como uma tecnologia versátil e poderosa para resolver uma ampla gama de problemas.

2.5.3 Redes GRU (*Gated Recurrent Unit*)

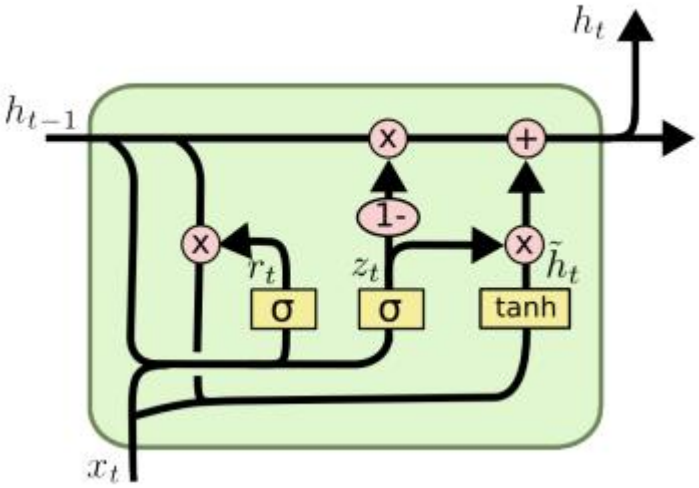
As *Gated Recurrent Units* (GRUs) representam uma classe de unidades recorrentes em redes neurais, notáveis por sua eficiência em comparação com unidades tradicionais, como *Long Short-Term Memory* (LSTM). Compostas por duas portas, a de atualização e a de reset, as GRUs oferecem controle sobre a retenção e descarte de informações da camada anterior na saída atual, representado na Figura 5. Sua versatilidade as torna aplicáveis em diversas áreas, desde processamento de linguagem natural até reconhecimento de padrões e tarefas de robótica.

No artigo *“Empirical Evaluation of Gated Recurrent Neural Networks of Sequence Modeling”*, dos autores Chung e outros (2014), foi direcionada a atenção para unidades mais avançadas que utilizam o mecanismo de portas, como a Unidade de Memória de Longo Prazo (LSTM) e a *Gated Recurrent Unit* (GRU). Foi investigado o desempenho dessas unidades em tarefas de modelagem de música polifônica e modelagem de sinais de fala. Os resultados demonstraram que essas unidades recorrentes mais avançadas superaram de fato as unidades recorrentes mais tradicionais, como as unidades *tanh*.

Outro exemplo destacado provém de uma pesquisa realizada por Rajib Rana e outros (2016) onde é comparada a eficácia da GRU e da LSTM na classificação de emoções a partir de fala ruidosa. Embora a LSTM tenha um desempenho ligeiramente superior em certos cenários, a GRU se destaca pelo tempo de execução mais curto. Essa relação entre precisão e complexidade temporal da GRU é especialmente vantajosa para aplicações embarcadas.

Em resumo, as GRUs emergem como uma ferramenta poderosa em aplicações diversas, sendo mais eficientes e mais facilmente treináveis do que unidades recorrentes convencionais.

Figura 5 – Arquitetura GRU (*Gated Recurrent Unit*)



Fonte: Abdulwahab e outros, 2017.

As unidades GRU (*Gated Recurrent Units*) são uma arquitetura de rede neural recorrente (RNN) utilizada para modelar sequências de dados. Elas consistem em dois componentes principais: o portão de *reset* e o portão de atualização. O portão de *reset* controla a quantidade de informação do estado anterior que é mantida no estado atual, enquanto o portão de atualização determina a quantidade de informação da entrada atual que é incorporada no estado atual.

Os símbolos utilizados incluem r_t para o portão de reset, z_t para o portão de atualização, h_t para o estado de saída candidato e h_t para o estado de saída real no instante t. A multiplicação elemento a elemento é representada por $\cdot \bar{\cdot} \cdot$ e a função sigmoide é representada por σ. Além disso, denota a concatenação de vetores. Os parâmetros incluem matrizes de pesos (W_r, W_z, W_h) e vetores de bias (b_r, b_z, b_h) para os portões e a saída. Os estados finais das unidades GRU direta e inversa são denotados por h_fn e h_bn, respectivamente.

Na explicação, é detalhado como o estado de saída do candidato é calculado usando a função sigmoide e os pesos da matriz W_h. O estado de saída real é obtido multiplicando o estado de saída candidato pelo portão de atualização e adicionando o bias b_h. Por fim, os estados finais das unidades GRU direta e inversa são concatenados para formar a entrada do *tweet*.

2.6 ABORDAGENS NÃO-SUPERVISIONADAS

> O aprendizado não-supervisionado é uma técnica de aprendizado de máquinas em que os usuários não precisam supervisionar o modelo. Ao invés disso, permite que o modelo trabalhe por conta própria para descobrir padrões e informações que não foram detectadas anteriormente, pois lida com dados não rotulados. Os dados rotulados poderiam ser descritos como um conjunto de dados que possui uma identificação, uma “etiqueta” para as observações (Araújo, 2021, p. 15).

Esse tipo de abordagem é caracterizado pela utilização de conjuntos de dados não rotulados para treinar algoritmos, os quais desempenham tarefas como a descoberta de padrões, a redução de dimensionalidade e a geração de dados. Durante o processo de treinamento, os algoritmos não supervisionados buscam identificar padrões nos dados sem a necessidade de orientações específicas.

2.6.1 Redes *Autoencoders*

Autoencoders são um tipo de rede neural artificial usada para aprender codificações eficientes de dados não rotulados. Um *autoencoder* aprende duas funções: uma função de codificação que transforma os dados de entrada e uma

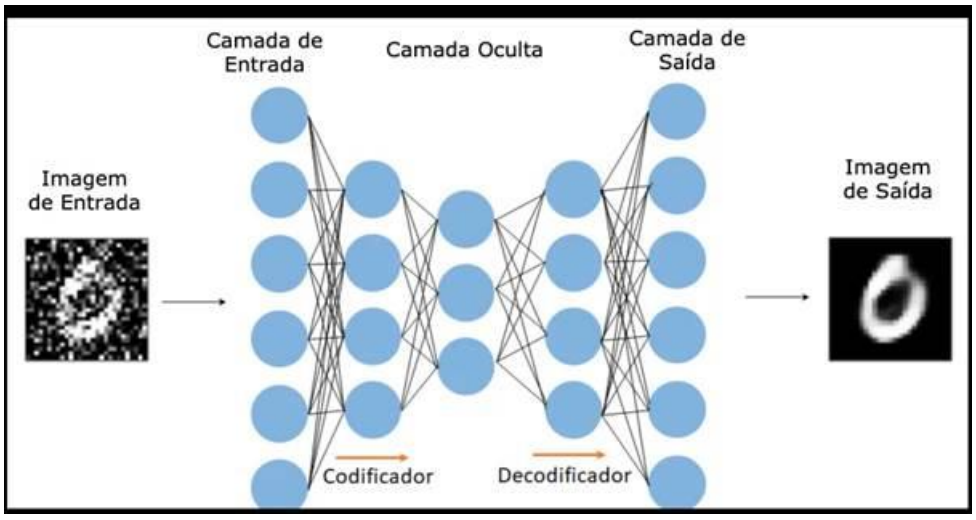
função de decodificação que recria os dados de entrada da representação codificada.

O conceito de *autoencoders* foi introduzido por Bernard Widrow e seus colegas em 1988. O artigo original, *“Adaptive Boltzmann Machines for Speech Recognition” (IEEE Transactions on Acoustics, Speech and Signal Processing, 1988)*, é considerado um dos trabalhos pioneiros em aprendizado de máquina.

A arquitetura básica de um *autoencoder* é composta de duas camadas principais: uma camada de entrada, uma camada de codificação e uma camada de decodificação. A camada de entrada recebe os dados de entrada, que podem ser imagens, áudio ou texto. A camada de codificação transforma os dados de entrada em uma representação codificada de menor dimensão. A camada de decodificação recria os dados de entrada da representação codificada.

O treinamento de um *autoencoder* é um processo iterativo. No início do treinamento, o *autoencoder* é inicializado aleatoriamente. Os dados de entrada são então alimentados no *autoencoder* e a saída da camada de decodificação é comparada com os dados de entrada originais. O *autoencoder* é então atualizado para minimizar a diferença entre a saída da camada de decodificação e os dados de entrada originais, conforme representado na Figura 6.

Figura 6 – Representação gráfica de um *autoencoder*



Fonte: The Keras Blog, 2024.

2.7 ABORDAGENS POR REFORÇO

Segundo o entendimento de Sutton (2018), o aprendizado por reforço é um tipo de aprendizado de máquina no qual o agente aprende a tomar decisões em um ambiente dinâmico. O agente recebe recompensas ou punições por suas ações, e aprende a tomar decisões que maximizam as recompensas.

Esse tipo de abordagem é caracterizado pela utilização de um agente que interage com um ambiente para aprender a tomar decisões. O agente recebe recompensas ou punições por suas ações, e aprende a tomar decisões que maximizem as recompensas. Em tarefas de controle, o agente não é dado uma política. O objetivo do agente é encontrar a política ideal, que é a política que maximiza a recompensa esperada.

Xu (2020) afirma que algoritmos de aprendizado por reforço podem ser vistos como uma forma de transformar métodos de programação dinâmica inviáveis em algoritmos práticos. A programação dinâmica é um método para resolver problemas de otimização em um ambiente dinâmico. No entanto, métodos de programação dinâmica podem ser inviáveis para problemas em larga escala.

A Figura 7 mostra um resumo das principais abordagens de Aprendizagem de Máquina vistas nesta pesquisa.

Figura 7 - Diagrama dos tipos de aprendizado em Aprendizagem de Máquina

Fonte: Dataat, GitHub, 2024.

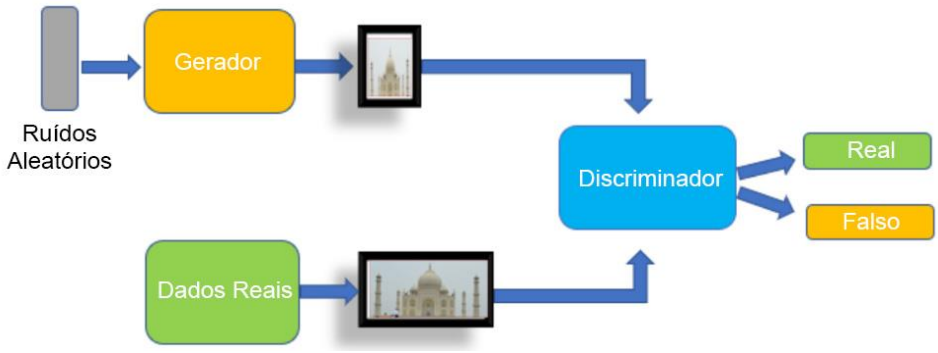
2.7.1 Redes GAN (*Generative Adversarial Network*)

GANs, ou redes neurais adversariais generativas, são um tipo de modelo de aprendizado de máquina que consiste em dois modelos, um gerador e um discriminador. O gerador é responsável por gerar dados novos e o discriminador é responsável por distinguir entre dados reais e gerados.

O conceito de GANs foi introduzido por Ian Goodfellow e seus colegas em 2014. O artigo original, *"Generative Adversarial Networks"*, é considerado um dos trabalhos mais influentes em aprendizado de máquina. O treinamento de um GAN é um processo iterativo. No início do treinamento, o gerador gera dados aleatórios. O discriminador então tenta distinguir entre os dados gerados e os dados reais. O gerador é então atualizado para gerar dados que sejam mais difíceis de distinguir do discriminador. O discriminador também é atualizado para distinguir melhor entre os dados gerados e os dados reais.

Ao longo do treinamento, o gerador aprende a gerar dados que são cada vez mais indistinguíveis dos dados reais. O discriminador também aprende a distinguir cada vez melhor entre os dados reais e gerados. A Figura 8 representa uma arquitetura GAN:

Figura 8 – Diagrama de uma GAN (*Generative Adversarial Network*)



Fonte: Lashgari e outros, 2020, p. 108885.

Nesta pesquisa iremos utilizar a arquitetura WaveGAN. WaveGANs são uma variação de GANs que são especificamente projetados para gerar dados de áudio.

WaveGANs usam um gerador baseado em WaveNet para gerar dados de áudio de alta qualidade.

3 TRABALHOS RELACIONADOS

Inicialmente, fizemos uma busca no *Google Scholar* utilizando as palavras-sobre em Português e Inglês, como “geração musical com inteligência”, “*music generation with artificial intelligence*” etc., e alguns trabalhos relevantes foram feitos e ainda estão sendo feitos, como o Soundraw (2021) desenvolvido pela empresa Soundraw, que é uma plataforma online que possibilita a criação de música com inteligência artificial, oferecendo ferramentas para gerar melodias, harmonias, ritmos e instrumentação.

Outro app mais recente é o Kompoz (2022) da empresa Kompoz, que é um app que utiliza inteligência artificial para ajudar os usuários a compor músicas, oferecendo sugestões de melodias, harmonias e ritmos. Como não foram encontrados dados com mais detalhes em artigos científicos desses trabalhos e alguns outros, não iremos reproduzir detalhes mais aprofundados sobre eles neste estudo.

O trabalho *“ML-Mandolin”* (Nguyen, 2019) foi desenvolvido para classificar músicas do mesmo estilo e a partir daí, gerar novas músicas no estilo vietnamita utilizando o instrumento mandolin através de uma rede LSTM. Este trabalho foi o texto e experimento-base para este trabalho atual onde estamos desenvolvendo o ML-Frevo, com significativas customizações e diferenças, como saída em apenas 1 canal no ML-Mandolin e número de músicas reduzido comparado ao ML-Frevo.

O estudo *"MuseGAN: A Musical Generative Adversarial Network"* (Su et al., 2018) se destaca por apresentar o MuseGAN, um modelo que utiliza Redes Adversariais Generativas (GANs) para criar música de alta qualidade em diversos gêneros. Para avaliar a qualidade das produções, foram utilizadas métricas como Diferença Média Quadrática (MSE), Razão de Sinal para Ruído (SNR) e Razão de Sinal para Ruído Normalizado (PSNR).

Ademais, em *" Comparative Study on Variational Autoencoders and Generative Adversarial Networks"* (Sami, 2019), foi realizada uma comparação entre

duas técnicas de aprendizado de máquina para geração de dados musicais: Autoencoders Variacionais (VAEs) e Redes Adversariais Generativas (GANs). Os resultados indicam que as GANs são mais adequadas para aplicações que exigem originalidade e criatividade, enquanto os VAEs se destacam em cenários que requerem consistência com um conjunto de dados de treinamento.

Paralelamente, o artigo *"MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation"* (Yang et al., 2017) apresenta o MidiNet, uma Rede Neural Convolucional Generativa Adversarial (GAN) desenvolvida especificamente para gerar música no domínio simbólico. Treinado com um conjunto de dados de melodias MIDI, o MidiNet demonstra sua capacidade de criar melodias novas e originais. Esses estudos não apenas demonstram a diversidade de abordagens na interseção entre música e inteligência artificial, mas também indicam um progresso contínuo nesse campo promissor.

No trabalho *“EmotionBox: a music-element-driven emotional music generation system using Recurrent Neural Network”* de Kaitong Zheng e outros (2021), é apresentado o *EmotionBox*, um gerador de música emocional automático impulsionado por elementos musicais. Diferentemente de métodos anteriores já propostos, esse modelo dispensa conjuntos de dados musicais rotulados com emoções. Em vez disso, utiliza a densidade de notas e o histograma de sua altura para representar o grau de excitação e a valência da música, respectivamente. A partir disso, diversas combinações de excitações e valência são mapeadas para diferentes emoções, conforme o “modelo de emoção de Russell”.

Os resultados experimentais revelaram que o método utilizado no trabalho *EmotionBox* tem um desempenho mais competitivo e equilibrado na geração de música emocional do que os métodos baseados em rótulos de emoção. Especialmente, foi testado um melhor desempenho na geração de músicas com baixa excitação. Foi concluído que, com base na densidade de notas e histograma de altura, o método foi capaz de gerar música envolvente e expressiva, podendo ser direcionado para uma emoção específica.

No artigo *“A Comprehensive Survey on Deep Music Generation”*, Ji e outros (2020) propõem fornecer uma visão geral abrangente das técnicas de aprendizado profundo usadas na geração de música. Classificam as tarefas de geração de música em três níveis: partituras, performance e áudio direto. Para cada nível, eles discutem diferentes algoritmos de aprendizado profundo usados, como redes neurais recorrentes (RNNs), redes neurais convolucionais (CNNs), autoencoders e redes neurais adversariais generativas (GANs). Eles também discutem diferentes representações musicais adequadas para cada tarefa. Para avaliar a qualidade da música gerada, eles discutem métodos subjetivos e objetivos.

Os autores concluem que a geração de música por *Deep Learning* é uma área promissora com o potencial de democratizar a criação musical e levar a novas formas de expressão musical.

As métricas utilizadas foram:

a) **Testes de audição:** ouvintes humanos avaliam a qualidade da música gerada;

b) **Análise de *feedback*:** ouvintes humanos fornecem feedback sobre a música gerada;

c) **Métricas de similaridade:** medem a semelhança entre a música gerada e a música humana;

d) **Métricas de estrutura:** medem a estrutura da música gerada;

e) **Métricas de conteúdo:** medem o conteúdo da música gerada.

Em resumo, o artigo fornece uma visão geral abrangente das técnicas de aprendizado profundo usadas na geração de música. O trabalho discute diferentes algoritmos, representações e métodos de avaliação, e aponta para direções futuras promissoras.

No trabalho *“MusPy: A Toolkit for Symbolic Music Generation”* (Hao-Wen e outros, 2020), os autores apresentam o MusPy, que é uma biblioteca em Python

para a geração de música que oferece ferramentas para o gerenciamento de dados, pré-processamento e avaliação de modelos. O artigo apresenta análises estatísticas de onze conjuntos suportados pelo MusPy, incluindo experimentos de generalização entre conjuntos.

Os resultados oferecem insights sobre a sobreposição de domínios entre os conjuntos de dados e indicam que alguns contêm amostras mais representativas de gêneros cruzados do que outros, fornecendo orientações para a escolha de conjuntos de dados em futuras pesquisas. Em suma, o MusPy é apresentado como uma ferramenta essencial para o desenvolvimento de sistemas para a geração de música, com análises estatísticas e experimentos que ajudam na escolha de conjuntos de dados e na melhoria da generalização de modelos de aprendizado de máquina ao combinar conjuntos heterogêneos.

Esses trabalhos diferem em relação ao foco do estudo proposto neste trabalho, pois por mais que eles tenham utilizado redes neurais para geração de música, o que também é utilizado neste estudo, eles não focaram na geração musical do estilo Frevo, pois isso é um estilo totalmente brasileiro, conforme já foi abordado anteriormente.

Também vale ressaltar, que neste estudo, a abordagem está na produção de novas músicas através das referidas redes neurais que foram utilizadas em vários outros estudos, inclusive nos mencionados neste trabalho, mas a customização e hiperparâmetros que estão sendo utilizados vêm de uma base já pronta de redes LSTM que é de 2019 e que desde então, nenhuma nova técnica foi implementada para Frevo a nível de pesquisas para saber qual a melhor rede neural poderia fornecer melhores resultados para este fim, especificamente.

A tabela abaixo apresenta um resumo do que foi abordado nos trabalhos relacionados e como eles se comparam com este trabalho em relação a produção musical no estilo Frevo, especificamente:

Tabela 1- Resumos dos trabalhos mencionados

Aspecto	EmotionBox (2021)	Ji et al (2020)	MusPy (Hao-Wen et al, 2020)	MidiNET (2017)	MuseGAN (2018)	ML-Mandolin (2019)	ML-Frevo (2024)
Aplicação	Geração de Música Emocional	Geração de Música	Geração de Música	Geração de música em diferentes formatos	Geração de música de vários tipos de instrumentos do zero.	Geração de Música Vietnamita no Mandolin	Geração de Música Frevo
Aplicação de Redes Neurais	GRU	RNN, CNN, Autoencoders, GAN	Transformer, CNN	CNN e GAN	GAN (gerador e discriminador)	LSTM	LSTM, WaveGAN, CNN, Autoencoder, GRU
Método de Avaliação	Julgamento de Qualidade	Testes de audição, análise de feedback, métricas de similaridade, estrutura e conteúdo	Avaliação subjetiva automática em composição melódica	Avaliação quantitativa e qualitativa, como também avaliações adicionais, como experimentação interativa	Avaliação humana, perda de discriminador e perda de fidelidade	Classificação de estilo musical	Análise de métricas de precisão, qualitativas, visuais-espectrais e espectrais
Resultados	Mais competitivo e equilibrado na geração de música emocional	Área promissora com potencial de democratizar a área de criação musical	Comparáveis aos de humanos em composição melódica e acompanhamento harmônico	Desempenho promissor na criação de música simbólica, com capacidade de criar músicas complexas e altamente fiel ao estilo musical desejado	Capaz de gerar música de alta qualidade e de também fazer acompanhamento de músicas geradas	Geração musical de série temporal de áudio no estilo de música vietnamita	Análise final de melhores opções de redes neurais para composição musical no estilo Frevo baseado nos dados coletados
Diferenças	Utiliza densidade de notas e histograma de altura para representar excitação e valência musical	Visão geral das técnicas de aprendizado profundo para geração musical	Ferramentas de pré-processamento e avaliação de modelos, análise estatística de conjunto de dados	Arquitetura CNN-GAN se tornando um mecanismo de condicionamento inovador e capacidade de gerar músicas em diferentes formatos	Arquitetura multi-trilha sequencial para geração e acompanhamento de música simbólica	Foco na geração musical de série temporal no estilo de música vietnamita e classificação do áudio no estilo musical antes de alimentar o dataset com dados para treinamento	Foco na geração de músicas no estilo Frevo, avaliação específica para este estilo e verificação das melhores métricas para este fim

Fonte: elaborado pelo autor, 2024.

Explicando melhor as diferenças destacadas na tabela são em relação aos aspectos principais de cada trabalho mencionado.

Aplicação: Refere-se ao propósito principal da aplicação das técnicas de geração de música. Enquanto *EmotionBox* visa a geração de música emocional, Ji et al foca na geração de música de forma geral, MusPy fornece uma ferramenta para geração de música com várias funcionalidades, *ML-Mandolin* se concentra na geração de música de vários tipos de instrumentos do zero, MuseGAN destaca-se na geração de música em diferentes formatos, e MidiNet concentra-se na geração de músicas no domínio simbólico.

Arquitetura de Redes Neurais: Indica os tipos de redes neurais utilizadas em cada trabalho. Enquanto *EmotionBox* usa GRU *(Gated Recurrent Unit)*, Ji et al emprega uma variedade de arquiteturas, como RNNs, CNNs, *Autoencoders* e GANs. MusPy utiliza *Transformer* e CNN. ML-Mandolin emprega CNN e GAN, MuseGAN utiliza GAN (gerador e discriminador), e MidiNet utiliza LSTM, WaveGAN, CNN, Autoencoder e GRU.

Método de Avaliação: Refere-se aos métodos utilizados para avaliar a qualidade das músicas geradas. Enquanto *EmotionBox* utiliza julgamento de qualidade baseado em densidade de notas e histograma de altura, Ji et al (2020) empregam testes de audição, análise de feedback, métricas de similaridade, estrutura e conteúdo. *MusPy* utiliza avaliação subjetiva automática em composição melódica. *ML-Mandolin* utiliza avaliação humana, perda de discriminador e perda de fidelidade, MuseGAN utiliza Diferença Média Quadrática (MSE), Razão de Sinal para Ruído (SNR) e Razão de Sinal para Ruído Normalizado (PSNR), e *MidiNet* utiliza avaliação quantitativa e qualitativa, como também avaliações adicionais, como experimentação interativa.

Resultados: Indica os resultados alcançados por cada trabalho. Enquanto *EmotionBox* alcança resultados competitivos e equilibrados na geração de música emocional, Ji et al aponta para a promissora democratização da criação musical, *MusPy* alcança resultados comparáveis aos de humanos em composição melódica e acompanhamento harmônico, *ML-Mandolin* é capaz de gerar música de alta

qualidade e de também fazer acompanhamento de músicas geradas, *MuseGAN* destaca-se na capacidade de gerar música em diferentes formatos, e MidiNet alcança resultados promissores na criação de música simbólica, com capacidade de criar músicas complexas e altamente fiel ao estilo musical desejado.

Essas diferenças são importantes para destacar as abordagens específicas de cada trabalho e como eles se diferenciam em seus objetivos, métodos e resultados.

4 MÉTODO

Neste capítulo será apresentada a Proposta, Material, *Setup* e Método adotados para a realização desta pesquisa.

4.1 PROPOSTA

Nesta pesquisa, com o foco em gerar e avaliar séries temporais de áudio no estilo Frevo, num primeiro passo para a geração musical, iremos carregar os dados e limitá-los a um intervalo específico. Em seguida, criaremos conjuntos de treinamento e teste dividindo os dados temporais em sequências. Posteriormente, treinaremos os modelos utilizando o conjunto de treinamento e geraremos previsões para os conjuntos separados com esse propósito. Após isso, salvaremos as previsões e a parte da música original. É importante salientar que os dados rodarão sempre em 2 canais para um som *stereo* ao invés de mono.

Esta pesquisa vem da base do trabalho feito por Nguyen (2019) em sua obra *“Machine Learning Mandolin”*, no qual ele cria músicas vietnamitas usando redes LSTM, mas ele não explora os dados de métricas, apenas a forma de classificação e arquitetura para a criação, é usada apenas do LSTM. Nesse estudo, são exploradas as métricas a partir do código-fonte do trabalho que ele utilizou para *Mandolin*, usando uma customização para Frevo adicionando mais músicas e explorando as métricas onde ele não explorou. As outras redes neurais foram a partir de análises de outros trabalhos de geração de dados com áudio no geral, não apenas de música e as métricas eram métricas próprias para áudio. Vale lembrar também, que as arquiteturas utilizadas, usam o LSTM como modelo de extração musical, etc já que esse processo é o mesmo para todas as redes neurais, mas o treinamento em si em diferentes redes neurais foi uma customização própria para testes e usando os mesmos hiperparâmetros ou parecidos, para que o estudo pudesse ter dados consistentes de comparação.

As métricas utilizadas para a avaliação tanto de qualidade, previsão e da série temporal de áudio são: *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE),

Root Mean Squared Error (RMSE), *Signal-to-Noise Ratio* (SNR) e *Peak Signal-to-Noise Ratio* (PSNR) para cada canal e de forma global. Também calcula métricas espectrais, como centroide espectral, largura de banda espectral e contraste espectral para cada canal.

Apresenta gráficos de espectrograma para avaliação visual da qualidade da música gerada em comparação com a original. Calcula métricas de Nível de Pressão Sonora (SPL) e Relação Sinal-Ruído e Distorção (SNDR) para cada canal e de forma global.

4.2 MÉTRICAS DE AVALIAÇÃO

A avaliação mais frequente da precisão de um modelo de rede neural, destinado a prever os futuros valores de ações, baseia-se na computação do erro do modelo. Quanto menor o erro de um modelo, mais bem adaptado ele está ao conjunto de dados e mais próxima sua previsão se aproxima dos valores reais. Essa afirmação é apoiada por trabalhos de pesquisa recentes em seus resultados, como o trabalho de Steurer e outros (2020).

Nesta seção, as métricas que serão usadas neste trabalho serão apresentadas para avaliação ampla da série temporal de áudio gerada.

4.2.1 MAE – *Mean Absolute Error* (Erro Absoluto Médio)

É uma métrica do tipo diferença absoluta. Essa métrica limita o espaço de pontos discrepantes no desempenho de modelos e é, entretanto, altamente eficiente onde existem problemas de qualidade no conjunto de dados (Steurer e outros, 2020). É uma medida de precisão simples e intuitiva, que é menos sensível a outliers do que outras métricas, como o erro quadrático médio (MSE). Quanto menor o valor do erro, mais próximo dos resultados esperados. A equação abaixo, expressa a fórmula do MAE:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y\_n - \hat{y}\_i|$$

Onde:

- n: número total de observações no conjunto de dados
- ∑: este símbolo denota a soma. A Fórmula MAE calcula a soma das diferenças absolutas entre os valores reais e as previsões para todas as observações no conjunto de dados.
- $\hat{y}\_i$: é o valor previsto ou esperado da i-ésima observação
- y_n : valor real da i-ésima observação
- | * | : é o valor absoluto

4.2.2 MSE – *Mean Squared Error* (Erro Quadrático Médio)

É o erro quadrático médio, que é uma medida mais pesada do erro do que o MAE. Quanto menor o MSE, melhor a precisão da previsão. O erro quadrático médio (MSE) é uma métrica de avaliação que mede a diferença quadrática média entre os valores reais e os valores previstos. Segundo o entendimento de Steurer e outros (2020), essa métrica é extremamente sensível a pontos extremos e é útil em situações em que é crucial minimizar erros significativos de previsão.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y\_n - \hat{y}\_i)^2$$

Onde:

- n: é o número total de amostras da base de dados
- y_n: é o valor real da i-ésima observação
- $\hat{y}\_i$: é o valor previsto ou estimado da i-ésima observação
- Σ: sigma, notação matemática que representa a soma
- $(y\_n - \hat{y}\_i)^2$: é a diferença entre o valor real e o valor previsto para a i-ésima amostra, elevada ao quadrado, garantindo que o resultado seja sempre positivo
- $\frac{1}{n}$: é o inverso do número total de amostras, o que transforma a soma do quadrado das diferenças em uma média, tornando a métrica independente do tamanho da base de dados

4.2.3 RMSE – *Root Mean Squared Error* (Raiz Quadrada do Erro Quadrático Médio)

RMSE - *Root Mean Squared Error* (Raiz Quadrada do Erro Quadrático Médio) é a raiz quadrada do MSE, que é uma medida mais intuitiva do erro do que o MSE. Quanto menor o RMSE, melhor a precisão da previsão, conforme Steurer e outros (2023).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y\_n - \hat{y}\_i)^2}$$

Onde:

- n: número total de observações ou amostras na base de dados
- y_n: é o valor real (ou observado) da i-ésima amostra
- $\hat{y}\_i$: representa o valor previsto (ou estimado) da i-ésima amostra pelo modelo
- Σ: sigma, símbolo de somatório
- $(y\_n - \hat{y}\_i)^2$: é a diferença entre o valor real e o valor previsto para a i-ésima amostra, elevada ao quadrado, garantindo que o resultado seja sempre positivo
- $\frac{1}{n}$: é o inverso do número total de amostras, o que transforma a soma do quadrado das diferenças em uma média, tornando a métrica independente do tamanho da base de dados
- √: raiz quadrada da expressão dentro do parêntesis

4.2.4 SNR – *Signal-to-Noise Ratio* (Relação Sinal-Ruído)

A relação sinal-ruído (SNR) é uma métrica de avaliação que mede a relação entre a energia do sinal e a energia do ruído. É uma medida de qualidade do sinal, que é expressa em decibéis (dB). É uma métrica fundamental para avaliar a qualidade de um sinal, sendo que quanto maior, melhor. É o que deixa a entender o trabalho *“Using audio quality to predict word error rate in an Automatic Speech Recognition System”* de Fish e outros (2006).

$$SNR = 10\ log_{10}(\frac{P_{signal}}{P_{noise}})$$

Onde:

- P_{signal} : Potência do sinal é potência do sinal de interesse
- P_{noise} : Potência do ruído é a potência do ruído presente no sinal
- log_{10} : é o logarítmo na base 10
- 10 : é um fator de conversão para a escala decibel (dB)

4.2.5 PSNR – *Peak Signal-to-Noise Ratio* (Relação Sinal-Ruído de Pico)

É a relação sinal-ruído de pico, que mede a diferença entre os valores máximos do sinal e do ruído. Quanto maior o PSNR, melhor a qualidade do sinal. É o que demonstra Ichigaya e outros (2006) em seu artigo *"A method of estimating code PSNR code using quantizing DCT coefficients"*.

$$PSNR = 10\ log_{10}(\frac{valor\ de\ pico\ do\ sinal^2}{MSE})$$

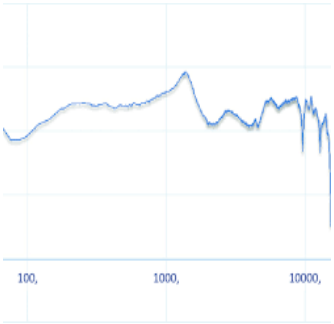
Onde:

- $valor\ de\ pico\ do\ sinal^2$: Valor de pico do sinal é o valor máximo que pode ser representado pelo sinal elevado ao quadrado.
- MSE : É o Erro Quadrático Médio (*Mean Squared Error*), que é a média dos quadrados das diferenças das amplitudes das amostras dos sinais de áudio original e processado/comprimido
- log_{10} : é o logarítmo na base 10
- 10: é um fator de conversão para a escala decibel (dB)

4.2.6 SPL – *Sound Pressure Level* (Nível de Pressão Sonora)

O nível de pressão sonora (SPL) é uma métrica de avaliação que mede a intensidade do som. É expresso em decibéis (dB), onde 120 dB é o limite da audição humana sem dor ou incômodo, ou seja, o SPL pode ser considerado bom.

Figura 9 - Gráfico do nível de pressão sonora em dB
em função da frequência em Hz



Fonte: Cunha, Julia e outros (2016), p. 954.

O SPL é calculado usando a seguinte fórmula:

$$SPL = 20\ log_{10}(\frac{\rho}{\rho_{ref}})$$

Onde:

- ρ : é a pressão sonora do som
- ρ_{ref} : é a pressão de referência , geralmente definida como 20 µPa
- log_{10} : é o logarítmo na base 10
- 20 : é um fator de conversão para transformar a relação de pressão em decibéis (dB)

4.2.7 SNDR – *Signal-to-Noise-Distortion Ratio* (Relação Sinal-Ruído-Distorção)

A relação sinal-ruído-distorção (SNDR) é uma métrica de avaliação que mede a relação entre a energia do sinal, a energia do ruído e a energia da distorção. É uma medida de qualidade do sinal, que é expressa em decibéis (dB).

$$SNDR = 10\ log_{10}(\frac{P_{signal}}{P_{noise} + P_{distortion}})$$

Onde:

- P_{signal} : é a potência do sinal de interesse
- P_{noise} : é a potência do ruído.
- $P_{distortion}$: é a potência da distorção.
- log_{10} : é o logarítmo na base 10
- 10: é um fator de conversão para a escala decibel (dB)

4.2.8 Avaliação Visual-Espectral

Realiza uma avaliação visual-espectral comparativa entre o áudio original e o áudio gerado pelos modelos testados. Em vez de depender apenas de métricas objetivas, como erro médio absoluto (MAE) ou erro quadrático médio (MSE) etc., a avaliação visual-espectral envolve a percepção humana. Nesse caso, é uma abordagem para avaliar a qualidade perceptual do áudio gerado em comparação com o áudio original. Em vez de depender apenas de métricas objetivas, como erro médio absoluto (MAE) ou erro quadrático médio (MSE) etc., a avaliação visual-espectral envolve a percepção humana.

No caso específico do projeto, ele usa gráficos de espectrograma para visualizar as características espectrais do áudio original e do áudio gerado. Esses gráficos ajudam na identificação visual de semelhanças e diferenças entre os dois sinais de áudio. O espectrograma exibe como a intensidade do som varia em diferentes frequências ao longo do tempo. Ao comparar os espectrogramas do áudio

original e do áudio gerado, é possível observar padrões e características importantes, como a presença de frequências específicas, mudanças na intensidade ao longo do tempo e outras propriedades espectrais. Essa análise visual permite uma avaliação mais intuitiva da qualidade do áudio gerado em comparação com o áudio original.

Em resumo, a avaliação visual-espectral busca capturar a percepção humana da qualidade do áudio, fornecendo uma perspectiva mais holística e intuitiva, complementando as métricas objetivas tradicionais. Essa abordagem é valiosa, especialmente quando o objetivo é criar modelos que gerem áudio com qualidade percebida semelhante ao áudio original.

4.2.9 Avaliação Visual-Espectral

As métricas espectrais objetivas são calculadas para uma sequência de áudio específica. Aqui algumas métricas espectrais mais específicas são usadas para calcular esses valores para uma sequência de áudio específica.

As métricas calculadas são:

a) ***Spectral Centroid:*** O centro de massa do espectro;

b) ***Spectral Bandwidth:*** A largura da faixa efetiva do espectro;

c) ***Spectral Contrast:*** Uma medida da diferença entre picos e vales no espectro.

As métricas MAE, MSE, RMSE, PSNR e SNDR são todas adequadas para avaliar a precisão de uma previsão de série temporal de áudio. As métricas SNR e SPL são menos usadas para avaliar a precisão de previsões de séries temporais de áudio. SNR mede a relação entre o sinal e o ruído, enquanto SPL mede o nível de pressão sonora do sinal. Essas métricas são mais frequentemente usadas para avaliar a qualidade do áudio, em vez da precisão da previsão. Ambas as funções são utilizadas para avaliar subjetivamente e objetivamente a qualidade do áudio gerado em comparação com o áudio original, fornecendo tanto uma visualização quanto métricas numéricas.

Então, podemos separar tudo o que foi discutido da seguinte forma, conforme é mostrado na Tabela 2.

Tabela 2 - Métricas de Avaliação de Precisão e Qualidade do Áudio

Categoria	Métrica	Descrição
Precisão	MAE	Erro Médio Absoluto
Precisão	MSE	Erro Quadrático Médio
Precisão	RMSE	Raiz Quadrada do Erro Quadrático Médio
Qualidade do Sinal	SNR	Relação Sinal-Ruído
Qualidade do Sinal	PSNR	Relação Sinal-Ruído de Pico
Qualidade do Áudio	SPL	Nível de Pressão Sonora (Média)
Qualidade do Áudio	Spectral Centroid	Centro Espectral Médio
Qualidade do Áudio	Spectral Bandwidth	Largura de Banda Espectral Média
Qualidade do Áudio	Spectral Contrast	Contraste Espectral Médio

Fonte: elaborado pelo autor, 2024.

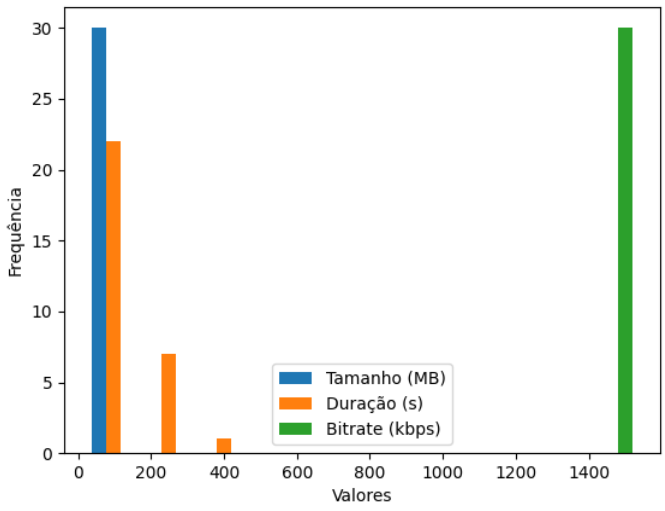
A métrica SPL foi adicionada à categoria de qualidade do áudio, pois é uma medida da intensidade do som. Essa medida é importante para a qualidade do áudio, pois sons mais altos são geralmente mais agradáveis de ouvir dentro de uma bandagem de decibéis apta ao ouvido humano.

4.3 MATERIAIS

Inicialmente, 30 canções instrumentais no estilo Frevo (de rua) de diferentes artistas foram baixadas de uma playlist do *YouTube* por meio do programa *aTube Catcher* que é disponibilizado gratuitamente na internet. As 30 canções selecionadas foram baseadas no foco do trabalho, que era gerar música no estilo Frevo e principalmente, do estilo apresentado nas ruas durante os Carnavais de Recife/PE. A quantidade foi definida por colocar, aproximadamente, um terço a mais de músicas do que o trabalho original é o *ML-Mandolin*, onde eles 19 músicas, inicialmente. Os arquivos foram salvos no formato .mp3 com alta qualidade sonora, sendo que os arquivos .mp3 ficam salvos comprimidos para reduzir o tamanho dele.

Para transformar o arquivo .mp3 em .wav um conversor online foi utilizado. Esse formato de dado é o que será utilizado pelo modelo na aprendizagem de máquina para os testes deste estudo. Para a preparação dos dados, foram coletadas 30 canções no ritmo Frevo em uma *playlist* do *YouTube* com duração entre 2min20s e 3min30s cada música. Elas foram colocadas em uma pasta chamada Frevo. As músicas foram lidas e utilizadas a partir desta pasta em todos os scripts para todos os testes das redes neurais (Fig. 10).

Figura 10 – Histograma das Médias do Tamanho,
Duração e Bitrate das Músicas



Fonte: elaborado pelo autor, 2024.

Analisando os dados, os dados abaixo foram obtidos rodando um *script* que também está à disposição para verificação no repositório:

Tamanho médio dos arquivos (MB): 30.23

Duração média das músicas: 2 minutos e 45 segundos

Bitrate médio das músicas (kbps): 1531.84

4.4 AMBIENTE DE EXECUÇÃO E EXPERIMENTOS

Um aplicativo foi escrito em Python 3.11 rodando sobre uma máquina rodando *Windows* 11 PRO 64 *bits*, processador 2.60 GHz Core i7-13650 HX 13th Gen Intel® Core(TM) com 32GB RAM e utilizando o IntelliJ IDEA 2023.3.2 *(Ultimate Edition)* da JetBrains. Vale lembrar que qualquer IDE pode ser usado nesse caso.

As redes neurais que são usadas nesta pesquisa são: LSTM *(Long Short-Term Memory)*, GAN *(Generative Adversarial Network)*, *Autoencoders*, CNN *(Convolutional Neural Networks)* e GRU *(Gated Recurrent Unit)*. As redes neurais escolhidas foram o LSTM - pois como mencionado, este estudo é uma continuidade do estudo anterior *"ML-Mandolin"* feito por Nguyen (2019), WaveGAN, CNN, *Autoencoders* e GRU - pois a maioria dos estudos feitos onde a geração musical por redes neurais é o foco, utilizavam estas redes como base de suas pesquisas.

Os hiperparâmetros utilizados para testar a geração de série temporal de áudio serão os mesmos ou próximos dos valores testados na maioria deles, tendo como base o LSTM como o parâmetro inicial de referência. Essa abordagem proporcionará uma análise mais aprofundada do comportamento das diferentes redes neurais em um determinado contexto específico, considerando os mesmos hiperparâmetros com os mesmos valores ou valores aproximados. Esse método possibilitará uma comparação mais precisa e significativa entre os modelos, fornecendo *insights* valiosos sobre suas performances relativas e potenciais vantagens ou limitações em relação às configurações de hiperparâmetros.

Os hiperparâmetros utilizados em cada uma das redes neurais, foram os mesmos ou com alguma mudança específica por causa da arquitetura da rede neural, mas de forma específica, conforme apresentado na Tabela 3.

Tabela 3 - Tabela com os dados dos hiperparâmetros em comum

Hiperparâmetro	**LSTM**	**CNN**	**GRU**
Tamanho do lote	50	50	50
Épocas	50	50	50
Formato de entrada	(1, 100)	-	(None, 1, 100)
Ativação das camadas densas	Linear	ReLU	Leaky ReLU
Função de perda	MSE	MSE	MSE
Otimizador	Adam	Adam	Adam

Fonte: elaborado pelo autor, 2024.

Tabela 4 - Tabela com os dados dos hiperparâmetros específicos

Hiperparâmetro	**WaveGAN**	**Autoencoders**
Tamanho da sequência	100	-
Número de interações	10000	500

Tamanho da janela	25	-
Passo do filtro	4	-
Dimensão do recurso	100	-
Taxa de aprendizado (discriminador)	Adam	-
Taxa de aprendizado (gerador)	Adam	-
Momentum do BatchNormalization	0.8	-
Tamanho do filtro	-	3
Número de filtros	-	64
Taxa de aprendizado	-	0.0001
Número de unidades por camada	-	256, 128, 64, 32, 64, 128, 256

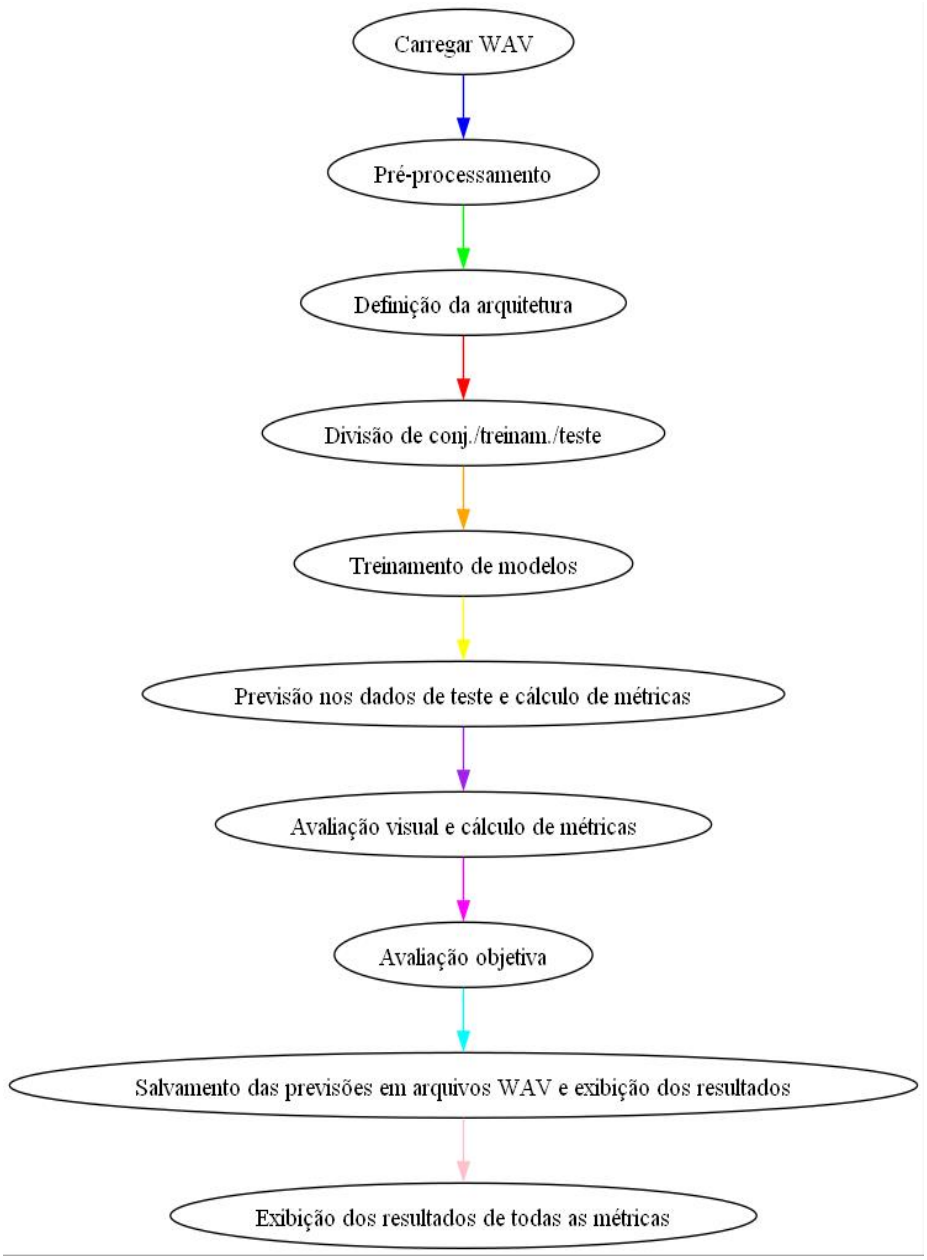
Fonte: elaborado pelo autor, 2024.

Esses dados representam os hiperparâmetros utilizados em cada modelo para testar a geração de série temporal de áudio, proporcionando uma base comparativa para análise do comportamento das diferentes redes neurais. Vale lembrar que o WaveGAN não tem um parâmetro de “*shape*” como os outros modelos, pois a sua arquitetura é baseada em redes convolucionais que operam diretamente sobre a sequência real de entrada.

5 EXPERIMENTOS E RESULTADOS

Ao rodar o programa, o fluxo seguido pelas cinco redes neurais testadas (LSTM, WaveGAN, CNN, *Autoencoder* e GRU) seguem o seguinte fluxo demonstrado na Figura 11.

Figura 11 - Fluxo seguido para a produção do áudio por meio das redes neurais



Fonte: elaborado pelo autor, 2024.

O programa carrega os 30 arquivos de áudio WAV e realiza algumas operações de pré-processamento nos dados, como dividir os conjuntos de treinamento e teste, criar conjuntos de dados com janelas deslizantes para alimentar a rede neural específica e remodelar os dados para o formato adequado.

Depois disso, ele define a arquitetura da rede neural, compila e treina dois modelos separados, um para cada canal de áudio – estamos sempre trabalhando com 2 canais de áudio já que o som é no formato estéreo.

Após o treinamento da rede neural, o programa faz previsões nos dados de teste usando os modelos treinados e calcula várias métricas de erro do áudio gerado em relação ao áudio original, ou seja, do áudio de teste em relação ao áudio de treinamento. Nesses cálculos, várias métricas são usadas, como: MAE (Erro Médio Absoluto), MSE (Erro Quadrático Médio), RMSE (Raiz do Erro Quadrático Médio), SNR (Relação Sinal-Ruído) e PSNR (Relação Sinal-Ruído de Pico). Ele também calcula métricas espectrais, como centroide espectral, largura de banda espectral e contraste espectral para cada canal de áudio e globalmente.

Ademais, é realizada uma avaliação visual-espectral, imprimindo os espectrogramas do áudio original e gerado para avaliação humana. Ele também calcula SPL (Nível de Pressão Sonora) e SNDR (Relação Sinal-Ruído e Distorção) para avaliação objetiva do áudio gerado.

Finalmente, as previsões geradas são salvas em arquivos WAV e exibe os resultados de todas as métricas calculadas, fornecendo uma análise abrangente do desempenho do modelo na geração de música e sua comparação com o áudio original.

5.1 AVALIAÇÃO DE PRECISÃO E QUALITATIVA

MÉTRICAS DE PRECISÃO					
Métrica	**LSTM**	**WaveGAN**	**CNN**	**Autoencoders**	**GRU**
MAE	526.11	**0.05**	436.84	663.49	511.75
MSE	847879.37	847879.37	**360776.93**	1466633.62	794021.75
RMSE	920.67	920.67	**600.55**	1211.04	890.70

MÉTRICAS DE QUALIDADE DE SINAL					
Métrica	**LSTM**	**WaveGAN**	**CNN**	**Autoencoders**	**GRU**
SNR (dB)	-59.28	-59.28	**-55.56**	-61.66	-58.99
PSNR (dB)	-11.15	-11.15	**-7.43**	-13.53	-10.86

MÉTRICAS DE QUALIDADE DO ÁUDIO					
Métrica	**LSTM**	**WaveGAN**	**CNN**	**Autoencoders**	**GRU**
SPL (dB)	-3.28	-3.28	-7.61	**150.77**	-3.15
SNDR (dB)	27.83	27.83	26.97	-9.54	**28.12**

Segundo o entendimento de Bishop (2006), em relação a métricas de Precisão de Áudio, quanto menor o valor do MAE, MSE e RMSE, melhor é a precisão do modelo. O menor valor do erro Erro Médio Absoluto (MAE) foi do WaveGAN, então ele foi mais preciso nesse contexto, obtendo um valor bem baixo de 0.05 apenas. Observa-se que o modelo CNN se destaca por sua melhor precisão, como evidenciado pelos menores valores de Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE), denotando uma notável capacidade de previsão. Contudo, é importante ressaltar que tal precisão é atenuada pela qualidade de sinal inferior, expressa pelos índices mais baixos de Relação Sinal-Ruído (SNR) e Pico de Relação Sinal-Ruído (PSRN), o que pode impactar a fidelidade dos resultados musicais.

Por outro lado, o modelo *Gated Recurrent Unit* (GRU) exibe um equilíbrio entre precisão e qualidade do sinal que se revela particularmente relevante na produção de música no estilo Frevo. Embora não lidere em termos de precisão absoluta, sua habilidade em preservar a fidelidade do sinal de áudio, pode ser mais valorizada em um contexto musical, onde a qualidade perceptual desempenha um papel crucial. Considerando a natureza enérgica e a complexidade rítmica do Frevo, é imprescindível a reprodução fiel dos timbres e nuances musicais, tornando o modelo GRU uma escolha promissora para a composição e produção musical nesse estilo.

Enquanto o modelo CNN pode ser preferível em situações que demandam precisão absoluta, como análise de dados críticos, o modelo GRU oferece uma abordagem mais equilibrada, priorizando a qualidade perceptual do áudio. Deste modo, ao empregar modelos de aprendizagem de máquina para produção e composição musical no estilo Frevo, torna-se premente ponderar não apenas a precisão, mas também a qualidade do sinal e a qualidade do sinal e a experiência auditiva global proporcionada por cada modelo, considerando, assim, as nuances inerentes a este contexto específico.

5.2 AVALIAÇÃO ESPECTRAL (MÉTRICAS ESPECTRAIS GLOBAIS)

CENTROIDE ESPECTRAL GLOBAL

Modelo	LSTM	WaveGAN	CNN	Autoencoders	GRU
Valor	2419.27 Hz	**273.75 Hz**	1413.36 Hz	3799.85 Hz	2419.27 Hz

LARGURA DE BANDA ESPECTRAL GLOBAL

Modelo	LSTM	WaveGAN	CNN	Autoencoders	GRU
Valor	1850.18 Hz	**96.28 Hz**	1619.85 Hz	3175.25 Hz	1850.18 Hz

CONTRASTE ESPECTRAL GLOBAL

Modelo	LSTM	WaveGAN	CNN	Autoencoders	GRU
Valor	18.69 Hz	**96.28 Hz**	19.03 Hz	17.95 Hz	18.69 Hz

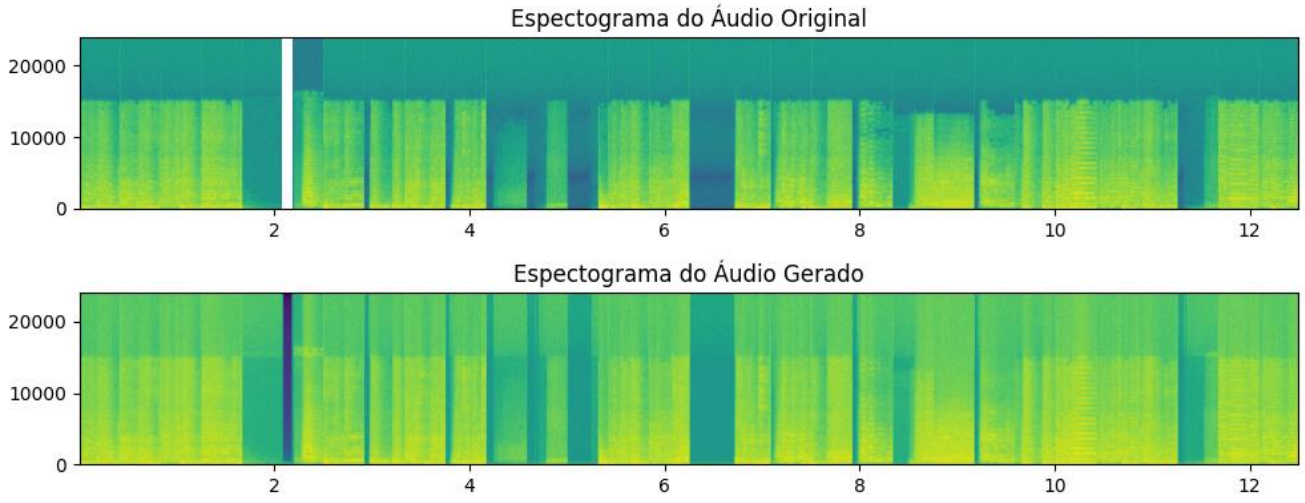
Com base nos resultados das métricas espectrais globais, a WaveGAN se destacou em relação aos outros modelos de geração de áudio, como LSTM, CNN, *Autoencoders* e GRU. Especificamente, a WaveGAN apresentou valores consideravelmente mais altos tanto para o centroide espectral quanto para o contraste espectral, indicando uma capacidade superior de capturar e reproduzir nuances detalhadas das características espectrais dos sinais de áudio. Por exemplo, o centroide espectral da WaveGAN foi de 273.75 Hz, enquanto os demais modelos variaram entre 1413.36 Hz (CNN) e 3799.85 Hz (Autoencoders), mostrando uma distância significativa em relação ao desempenho da WaveGAN.

Segundo alguns estudos feitos, mas não conclusivos, como *“A generative model for raw audio”* de OORD e outros (2016), e o *“GANSynth: Adversarial Neural Network Synthesis”*, de Engel e outros (2020), apontam que minimizar a informação espectral pode levar a uma generalização aprimorada, permitindo que o modelo capture e reproduza uma ampla variedade de características espectrais presentes nos dados de treinamento. Isso resulta em um modelo mais capaz de lidar com diferentes tipos de sinais de áudio. Ademais, a minimização da informação espectral pode ajudar a reduzir o *overfitting* ao evitar o foco excessivo em detalhes específicos dos dados de treinamento.

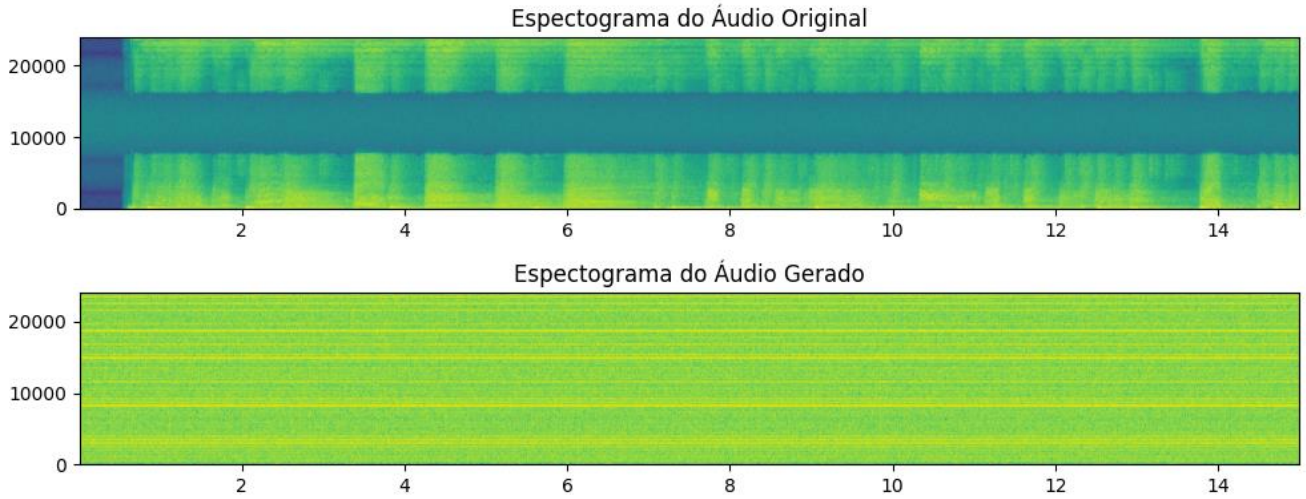
Outra vantagem é o controle aprimorado da qualidade do áudio gerado. Ao minimizar a informação espectral, o modelo se torna menos sensível a variações indesejadas no espectro de frequência, o que resulta em amostras de áudio mais realistas e agradáveis aos ouvidos. Portanto, a WaveGAN obteve o melhor resultado devido à sua capacidade de minimizar a informação espectral, o que permitiu uma melhor captura e reprodução das características gerais e relevantes do áudio.

5.3 AVALIAÇÃO VISUAL-ESPECTRAL

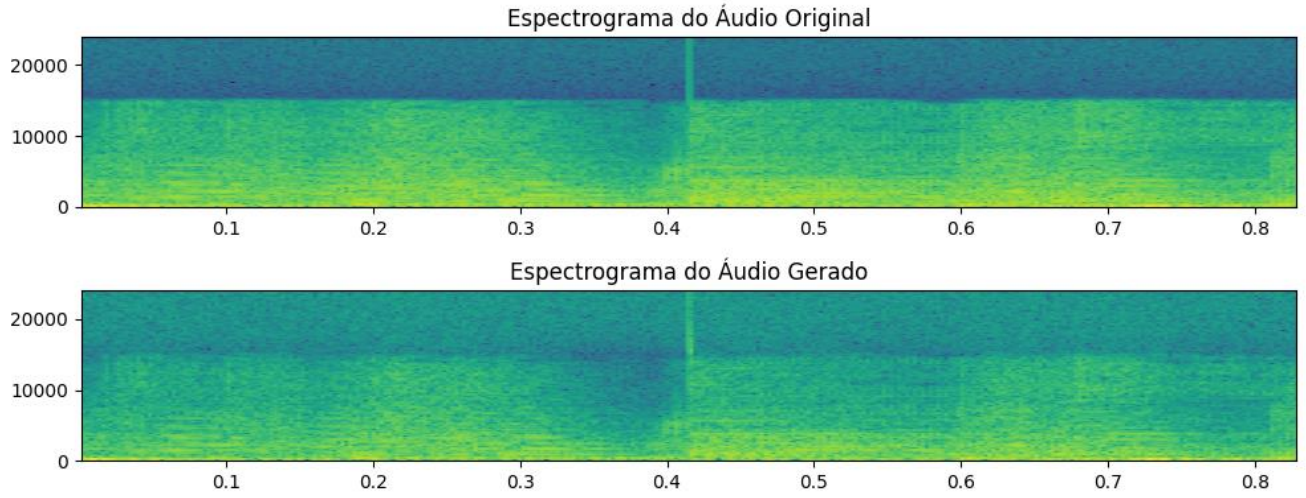
LSTM



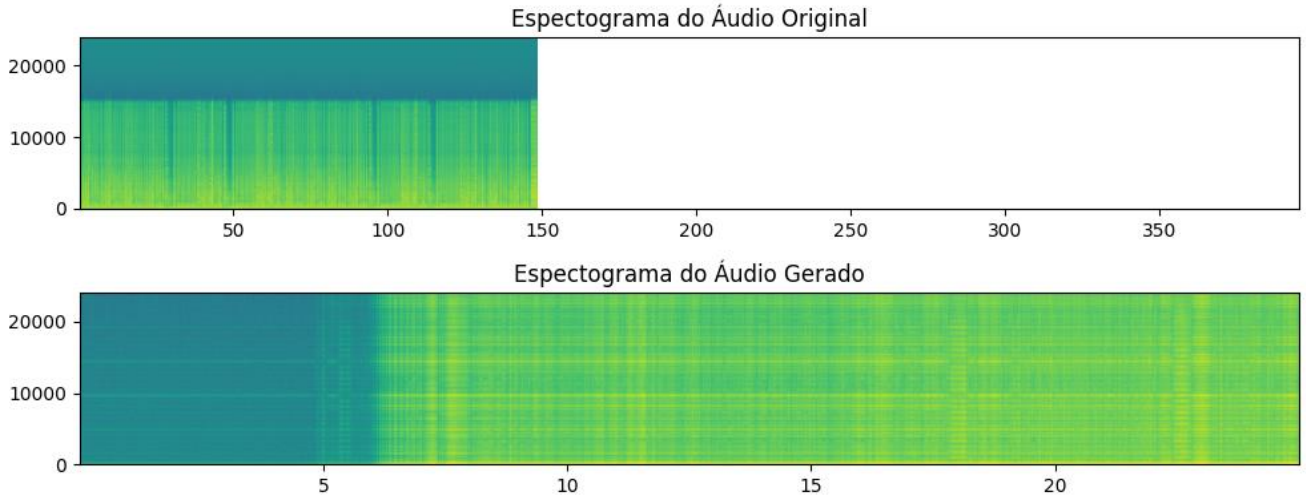
WaveGAN



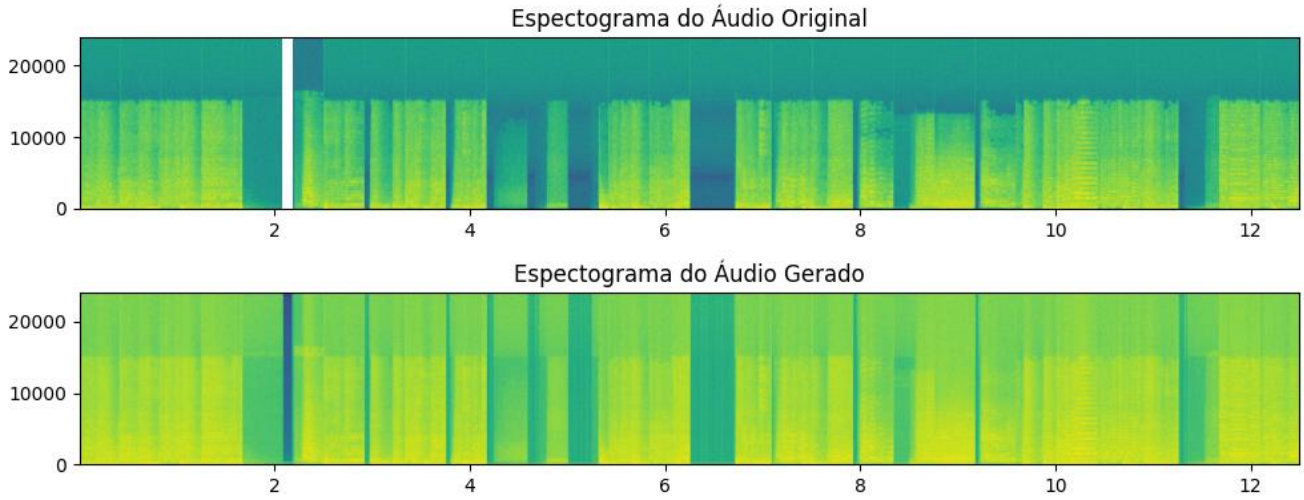
CNN



Autoencoders



GRU



Fonte: elaboração própria, 2024.

Aqui, é feita uma exibição dos espectrogramas desses sinais de áudio lado a lado para permitir uma comparação visual entre eles. Os espectrogramas são representações gráficas que mostram a intensidade das diferentes frequências de um sinal de áudio ao longo do tempo. Essa visualização pode ajudar na avaliação visual-espectral da qualidade do áudio gerado em comparação com o áudio original. É uma abordagem qualitativa para avaliar aspectos como fidelidade, clareza e semelhança entre o áudio original e o gerado.

Nota-se que nos *Autoencoders* há uma diferença enorme entre o áudio original e o áudio gerado. Essa diferença se dá pelo fato que de todas as redes neurais, foi onde o áudio produzido ficou longe de ser equivalente ao áudio original, que na verdade, não foi um áudio satisfatório com muitos ruídos e chiados no início e um ruído quase imperceptível na segunda após 2:30m de execução do ruído e chiado gerando dados nulos no espectrograma de áudio original.

5.4 CONSIDERAÇÕES FINAIS

Neste estudo, foram analisadas algumas redes neurais usadas na produção de áudio. No contexto analisado, foi utilizada para produção de música no estilo Frevo. As redes que foram objeto de estudo foram: LSTM, WaveGAN, CNN,

Autoencoders e GRU. As redes usam os mesmos hiperparâmetros ou bem parecidos dependendo de sua arquitetura para que os resultados possam ser comparados posteriormente.

Os resultados da tabela revelam métricas de precisão e qualidade do áudio para diferentes modelos de geração: LSTM, WaveGAN, CNN, *Autoencoders* e GRU. Observa-se que o modelo WaveGAN demonstra um desempenho superior em termos de MAE, indicando uma precisão mais alta em suas previsões, enquanto a CNN se destaca em MSE e RMSE, sugerindo uma menor dispersão e maior precisão. No entanto, todos os modelos exibem valores negativos de SNR e PSNR, indicando que o sinal de ruído é mais proeminente que o sinal de interesse. Notavelmente, o modelo *Autoencoders* apresenta o maior SPL, implicando uma intensidade sonora mais alta em suas previsões, enquanto o GRU exibe o maior SNDR, sugerindo uma melhor relação entre o sinal de interesse e o ruído/distorção. Essa análise fornece uma visão abrangente do desempenho relativo dos modelos em várias métricas de precisão e qualidade do áudio.

As tabelas “Métricas de Precisão” e “Métricas de Qualidade” fornecem uma análise abrangente das métricas de precisão e qualidade do áudio para diversos modelos de geração, destacando diferenças significativas entre elas. O modelo WaveGAN se destaca pelo menor Erro Médio Absoluto (MAE), sugerindo uma precisão superior em suas previsões. Além disso, a CNN apresenta o menor Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE), indicando uma precisão relativamente alta e uma dispersão menor em relação aos valores reais. No entanto, todos os modelos exibem Relação Sinal-Ruído (SNR) negativa, denotando que o sinal de ruído é mais proeminente que o sinal de interesse, com a CNN mostrando o menor valor de SNR. Quanto ao Nível de Pressão Sonora (SPL), o modelo *Autoencoders* se destaca com o maior valor, sugerindo um nível mais alto de pressão sonora em suas previsões, enquanto o GRU apresenta a melhor Relação Sinal-Ruído e Distorção (SNDR), indicando uma melhor relação entre o sinal de interesse e o ruído/distorção.

Acerca da Análise Visual-Espectral, vemos que os gráficos gerados possuem uma similaridade entre os gráficos do LSTM, CNN e GRU, indicando algum tipo de similaridade do áudio gerado com o áudio original em vários pontos de suas métricas, quer por exatidão ou por similaridade. WaveGAN e *Autoencoders* é nitidamente visto que houve uma disfunção do áudio gerado para o áudio original e esses dados são comprovados pelas outras análises qualitativas, de precisão e espectrais.

Considerando todos esses pontos discutidos e a importância da qualidade perceptual na produção musical, a rede mais adequada para produzir música no estilo Frevo, segundo os resultados obtidos nos experimentos deste estudo, seria o modelo Gated Recurrent Unit (GRU). Embora o *Autoencoder* possa ter uma precisão ligeiramente superior em termos de previsão, o GRU demonstra um equilíbrio entre a precisão e a qualidade do sinal, o que é essencial para a produção musical, especialmente em um contexto como o Frevo, onde a qualidade perceptual desempenha um papel crucial. Além disso, a análise dos espectrogramas revela que o modelo *Autoencoder* produziu resultados significativamente diferentes do áudio original, sugerindo uma menor fidelidade na reprodução das características musicais desejadas. Portanto, considerando não apenas a precisão, mas também a qualidade do sinal e a fidelidade perceptual, o modelo GRU emerge como a escolha mais promissora para a produção musical utilizando redes neurais para o estilo Frevo.

Em suma, enquanto cada modelo possui seus pontos fortes e fracos, a escolha do modelo mais adequado dependerá dos requisitos específicos do projeto, como a priorização da precisão, qualidade do sinal ou tolerância ao ruído. Essa análise comparativa fornece insights valiosos para orientar a seleção do modelo mais apropriado para diferentes aplicações de geração de áudio, considerando a complexa interação entre as métricas de desempenho avaliadas.

6 CONCLUSÃO

Os resultados obtidos fornecem uma compreensão detalhada das métricas de precisão e qualidade do áudio para uma variedade de modelos de geração, incluindo LSTM, WaveGAN, CNN, *Autoencoders* e GRU. Cada modelo exibe seu desempenho distintivo, refletido nas métricas de MAE, MSE, RMSE, SNR, PSNR, SPL e SNDR. O destaque fica para o modelo WaveGAN, que demonstra uma precisão superior em termos de MAE, enquanto a CNN se sobressai em MSE e RMSE, sugerindo uma menor dispersão e maior precisão em suas previsões. Entretanto, a análise revela que todos os modelos apresentam valores negativos de SNR e PSNR, indicando que o sinal de ruído é mais proeminente em relação ao sinal de interesse, sugerindo a necessidade de aprimoramentos na filtragem ou no processo de geração para reduzir o impacto do ruído nos resultados.

Além das métricas mencionadas, é crucial considerar o contexto específico de cada aplicação ao avaliar o desempenho dos modelos. Por exemplo, enquanto o modelo *Autoencoders* se destaca com o maior SPL, indicando uma maior intensidade sonora, o GRU exibe o maior SNDR, sugerindo uma melhor relação entre o sinal de interesse e o ruído/distorção. Essas nuances ressaltam a importância de uma abordagem personalizada na escolha do modelo mais adequado para cada cenário de uso, considerando as demandas específicas do projeto, como precisão, qualidade do sinal e tolerância ao ruído.

Além disso, é necessário abordar as limitações encontradas durante o estudo, como as restrições de hardware ao executar alguns modelos, principalmente *Autoencoders*, que resultaram em baixa velocidade e indisponibilidade temporária do sistema. Outra limitação importante foi a escassez de materiais específicos no contexto de geração de série temporal de áudio, o que tornou a pesquisa mais desafiadora e exigiu adaptações de hiperparâmetros com base em experimentos de outras áreas. Essas limitações destacam áreas potenciais para futuras melhorias e refinamentos nos métodos de geração de áudio.

Considerando esses resultados e limitações, há várias oportunidades para pesquisas futuras visando aprimorar os sistemas de geração de áudio. Estratégias como otimização de parâmetros, combinação de modelos e diversificação de conjuntos de dados podem potencializar ainda mais o desempenho dos modelos. Além disso, explorar modelos de aprendizado profundo mais avançados, como Redes Generativas Adversariais (GANs), e investigar técnicas de interpretabilidade dos modelos podem proporcionar insights valiosos e impulsionar o desenvolvimento de sistemas mais robustos e eficazes de geração de áudio para uma ampla gama de aplicações.

Respondendo de forma direta as perguntas que guiaram e motivaram esta pesquisa:

Arquiteturas de geração musical são capazes de reproduzir com precisão o característico ritmo acelerado do Frevo? Sim, mas com certo tipo de limitações de hardware, ajustes de hiperparâmetros, quantidade de dados para treinamento e teste etc. Ou seja, ajustando esses dados é esperada uma melhora significativa no resultado, uma vez que isso foi demonstrado através do trabalho inicial do ML-Mandolin que possuía 19 músicas e aumentando em pouco mais de um terço de música na base de dados, mesmo sendo outro estilo, o ML-Frevo destacou um melhor resultado tanto no LSTM que foi onde o ML-Mandolin atuou, como no GRU que foi onde o ML-Frevo teve melhor performance.

Como podemos determinar a eficácia da IA na geração de áudio no contexto do Frevo, considerando suas características únicas? Foi demonstrado através de avaliações de métricas de precisão, qualitativas, espectrais e visuais-espectrais, fazendo uma análise comparativa dentre as mais diferentes redes neurais onde houve ganho, perda e similaridade por dados dessas métricas como também análise visual também. Foram disponibilizados os áudios gerados para uma análise humana no repositório de código do projeto também.

Como a análise visual-espectral pode complementar a análise da geração de áudio no contexto do Frevo? Através da incorporação dos espectros gerados, foram feitas várias análises visuais-espectrais onde foram avaliadas a fidelidade e

autenticidade entre o áudio gerado e o áudio original. Ao examinar a distribuição de energia ao longo do espectro de frequência, foi visto que alguns padrões remetem similaridade ao áudio de frevo original, mas algumas nuances dependem de mais testes e mais dados serem coletados para treinamento.

6.1 LIMITAÇÕES

Os resultados apresentados revelam oportunidades significativas para aprimorar os sistemas de geração de áudio por meio da exploração de estratégias avançadas de otimização de parâmetros e combinação de modelos. Além disso, investigar técnicas de pré-processamento de dados mais sofisticadas e aquisição de conjuntos de dados diversificados pode enriquecer a análise e fornecer resultados mais abrangentes. Outra linha promissora de pesquisa é a adoção de modelos de aprendizado profundo mais avançados, como Redes Generativas Adversariais (GANs), para explorar seu potencial na produção de áudio de alta qualidade.

Por outro lado, é importante reconhecer as limitações enfrentadas durante o estudo, como restrições de hardware que causaram lentidão e “travamentos” durante a execução de certos modelos, especialmente os baseados em *Autoencoders*. Além disso, a escassez de materiais específicos sobre geração de áudio no estilo Frevo via inteligência artificial adiciona complexidade à pesquisa, exigindo a adaptação de técnicas de outras áreas e a experimentação com diferentes hiperparâmetros. Essas limitações ressaltam a necessidade de soluções inovadoras e estratégias adaptativas para lidar com desafios computacionais e de disponibilidade de dados.

Diante desses resultados e desafios, futuras investigações podem se concentrar não apenas na melhoria do desempenho dos modelos, mas também na interpretabilidade deles. Compreender os padrões aprendidos pelos modelos e integrar técnicas de explicabilidade pode torná-los mais transparentes e confiáveis, impulsionando seu uso em aplicações práticas que exigem interpretabilidade, como assistência médica e análise forense de áudio, por exemplo. Essa abordagem integrada promete avanços significativos na geração de áudio e sua aplicação em diversos domínios.

6.2 TRABALHOS FUTUROS

Considerando os trabalhos futuros, há diversas oportunidades para aprimorar os resultados apresentados. Seria interessante explorar estratégias de otimização de parâmetros e combinação de modelos para potencializar ainda mais o desempenho dos sistemas de geração de áudio. Além disso, investigar técnicas de pré-processamento de dados e aquisição de conjuntos de dados mais diversificados poderia enriquecer a análise e proporcionar resultados mais abrangentes. Outra vertente de pesquisa promissora seria a exploração de modelos de aprendizado profundo mais avançados, como Redes Generativas Adversariais (GANs), para investigar seu potencial na geração de áudio de alta qualidade. Essas direções podem contribuir significativamente para o avanço da área e para o desenvolvimento de sistemas de geração de áudio mais robustos e eficazes em diferentes cenários de aplicação.

Além das abordagens mencionadas, há ainda espaço para explorar a interpretabilidade dos modelos de geração de áudio, buscando compreender melhor os padrões aprendidos e como eles influenciam na qualidade das previsões. A análise das representações latentes dos modelos pode revelar insights sobre as características acústicas mais relevantes para a geração de áudio de alta qualidade, possibilitando a identificação de áreas específicas para refinamento e aprimoramento dos algoritmos. Além disso, a integração de técnicas de explicabilidade pode contribuir para tornar os modelos mais transparentes e confiáveis, facilitando sua adoção em aplicações práticas onde a interpretabilidade é essencial, como em sistemas de assistência médica ou de análise de áudio forense, já mencionados anteriormente. Essa abordagem holística, que combina otimização de desempenho, diversificação de dados, exploração de modelos avançados e interpretabilidade, promete impulsionar significativamente o progresso na geração de áudio e sua aplicação em uma variedade de domínios.

Considerando os trabalhos futuros, além das estratégias mencionadas, seria interessante explorar ainda mais a adaptação dos modelos de geração de áudio para diferentes tipos de dados, como músicas com instrumentos específicos ou até

mesmo falar em diferentes idiomas e sotaques. A diversificação dos conjuntos de dados pode ajudar a garantir a robustez e generalização dos modelos em uma variedade maior de contextos, tornando-os mais úteis em cenários do mundo real. Ademais, a exploração de técnicas de transferência de aprendizado e aprendizado incremental pode ser vantajosa para aproveitar o conhecimento adquirido em um domínio específico e aplicá-lo em outro, reduzindo a necessidade de grandes conjuntos de dados para treinamento. Isso poderia facilitar a adaptação dos modelos a novas tarefas ou domínios com menor esforço computacional e tempo de treinamento.

Paralelo a tudo isso, algumas pesquisas também têm seguido um rumo interessante no âmbito de IA generativa voltada para a área musical, como o projeto MusicLM do Google, apresentado no paper *“MusicLM: Generating Music from Text”* de Andrea Agostinelli e outros (2023) onde ele cria músicas de alta qualidade, com fidelidade sonora de 24 kHz, e que se mantém coerente por vários minutos. O *MusicLM* é capaz de interpretar descrições textuais e transformá-las em peças musicais que condizem com a ambientação descrita. Para avaliar sua performance, foi utilizado o *MusicCaps*, um extenso banco de dados com mais de 5.500 pares de texto e música, cuidadosamente selecionados por músicos profissionais. Os resultados demonstram que o *MusicLM* supera outros modelos similares, tanto na qualidade do áudio quanto em sua fidelidade ao texto de entrada.

Sabendo disso, para trabalhos futuros podem se concentrar na geração de letras, juntamente com a melhoria do condicionamento do texto e da qualidade vocal. Outro aspecto é a modelagem da estrutura de músicas em níveis elevados, como introdução, estrofe e refrão. A modelagem da música a uma taxa de amostragem mais alta é um objetivo adicional.

Um outro trabalho que é bem interessante destacar atualmente, é o DDSP-Piano que foi apresentado no artigo *“DDSP-Piano: A Neural Sound Synthesizer Informed by Instrument Knowledge”* por Renault e outros (2023), que é um novo método de sintetizar sons de piano utilizando deep learning. Embora alcance melhor qualidade de som do que alguns outros métodos baseados em redes neurais,

reconhece-se que as técnicas de modelagem física ainda podem produzir resultados superiores.

Trabalhos futuros é que o app poderia se beneficiar da interpretabilidade e diferenciação do DDSP-Piano para lidar com outras tarefas relacionadas à música polifônica, como separação de fontes sonoras e transcrição multi-pitch auto-supervisionada.

Em última análise, a combinação de abordagens inovadoras com uma compreensão mais profunda dos modelos e de seus efeitos na geração de áudio pode levar a avanços significativos na área. A contínua colaboração entre pesquisadores e a comunidade científica, juntamente com o acesso a recursos computacionais cada vez mais poderosos, oferece um cenário promissor para o desenvolvimento de sistemas de geração de áudio mais sofisticados e eficazes. O trabalho pode ser baixado gratuitamente no repositório: https://github.com/lucaspiano/ml-frevo.

REFERÊNCIAS

ABDULWAHAB, S. et al. (2017). **Deep Learning Models for Paraphrases Identification.** 10.13140/RG.2.2.15743.46240.

AGOSTINELLI, A. et al. **Musiclm:** Generating music from text. arXiv preprint arXiv:2301.11325, 2023.

ARAÚJO, M. V. A. de. **Métodos de Clustering em Aprendizado de Máquinas Não Supervisionado.** 2021. 89 f. Trabalho de Conclusão de Curso (Graduação de Estatística) - Instituto de Matemática e Estatística, Universidade Federal Fluminense, Niterói, 2021. Disponível em: https://app.uff.br/riuff/handle/1/26201 Acesso em: 15 fev. 2024.

BISHOP, C.M. (2006) **Pattern Recognition and Machine Learning.** Springer Verlag, Singapore

BRIOT, J.-P.; PACHET, F. Deep learning for music generation: challenges and directions**. Neural Computing and Application**, v. 32, n. 4, p. 981-993, 2020. Disponível em: https://doi.org/10.1007/s00521-018-3813-6. Acesso em: 06 fev. 2024.

BROWNLEE, J. **Long Short-Term Memory Networks With Python:** Develop sequence prediction models with deep learning. [S.l.]: Jason Brownlee, 2017.

CHUNG, J. et al. **Empirical evaluation of gated recurrent neural networks on sequence modeling.** arXiv preprint arXiv:1412.3555 (2014).

CUNHA, J. M.; MERINO, G. S. A. D.; MERINO, E. A. D. Design para saúde e qualidade de vida: desenvolvimento e avaliação de requisitos de projeto para fone de ouvido inclusivo. **Revista online de la Red Internacional de Investigación en Diseo**, v. 2, p. 198, 2016. 10.4995/IFDP.2016.3153.

DAAT. GITHUB. Introdução. **GitHub**, online, 2024. Disponível em: https://dataat.github.io/introducao-ao-machine-learning/introdu%C3%A7%C3%A3o.html. Acesso em: 05 fev. 2024.

DE OLIVEIRA, A. C. S. et al. Aplicação de redes neurais artificiais na previsão da produção de álcool. **Ciência e Agrotecnologia**, v. 34, n. 2, p. 279–284, mar. 2010. Disponível em: https://www.scielo.br/j/cagro/a/HWFRGpHsMrQkShH8WDXggbM/?lang=pt# Acesso em: 15 fev. 2024.

DO CARMO, Filipe Braida. **Transformando o problema de filtragem colaborativa em aprendizado de máquina supervisionado.** 2013. 108 f. Dissertação (Mestrado em Engenharia). Universidade Federal do Rio de Janeiro, 2013. Disponível em: https://www.cos.ufrj.br/uploadfile/1362165575.pdf Acesso em: 15 fev. 2024.

DO CARMO, Filipe Braida. **Transformando o problema de filtragem colaborativa em aprendizado de máquina supervisionado.** 2013. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

DONG, H.-W.; HSIAO, W.-Y.; YANG, L.-C.; YANG, Y.-H. **MuseGAN:** Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. 2019. arXiv preprint arXiv:1909.06298.

DONG, Hao-Wen et al. **MusPy:** A toolkit for symbolic music generation. arXiv preprint arXiv:2008.01951, 2020.

FAUL, A. C. **A Concise Introduction to Machine Learning.** Boca Raton: CRC Press, 2019.

FERREIRA, M. et al. Avaliação do Uso Redes Neurais Convolucionais para Identificação de Lesões Cariosas Dentárias. In: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE (SBCAS), 23, 2023, São Paulo/SP. **Anais [...].** Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 473-478. ISSN 2763-8952. DOI: https://doi.org/10.5753/sbcas.2023.229626.

FISH, Randall; HU, Qian; BOYKIN, Stanley**. Using audio quality to predict word error rate in an automatic speech recognition system.** Unpublished, 2006.

GHAHRAMANI, Z. Unsupervised Learning. In: BOUSQUET, O., VON LUXBURG, U., RÄTSCH, G. (eds.). **Advanced Lectures on Machine Learning**. ML 2003. Lecture Notes in Computer Science, v. 3176. Springer, Berlin, Heidelberg, 2004. https://doi.org/10.1007/978-3-540-28650-9_5

ICHIGAYA, Atsuro et al. A method of estimating coding PSNR using quantized DCT coefficients. **IEEE Transactions on circuits and systems for video technology**, v. 16, n. 2, p. 251-259, 2006.

LATAH, M.; TOKER, L. Artificial intelligence enabled software-defined networking: a comprehensive overview. **IET Networks**, v. 8, n. 2, p. 79–99, 2018.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LIU, X. et al. **CNN based music emotion classification**. arXiv preprint arXiv:1704.05665, 2017.

MCFEE, B. **Artificial intelligence in music**: A survey. 2019. arXiv preprint arXiv:1902.02744.

MENDES, Pedro Henrique Rodrigues; MALVEZZI, William R. **Geração de músicas polifônicas utilizando redes neurais artificiais.** Programa de Iniciação Científica-PIC/UniCEUB-Relatórios de Pesquisa, 2019.

MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C. (eds.). **Machine Learning, Neural and Statistical Classification.** New York: Overseas Press, 1994.

NGUYEN, R. H. Machine Learning Mandolin, 2019. **Honors Projects**. 760. p. 2. Disponível em: https://scholarworks.gvsu.edu/honorsprojects/760. Acesso em: 6 fev. 2024.

OORD, Aaron van den et al. **Wavenet:** A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

RANA, Rajib. **Gated recurrent unit (GRU) for emotion classification from noisy speech.** arXiv preprint arXiv:1612.07778, 2016.

RENAULT, Lenny; MIGNOT, Rémi; ROEBEL, Axel. DDSP-Piano: a Neural Sound Synthesizer Informed by Instrument Knowledge. **Journal of the Audio Engineering Society**, v. 71, n. 9, p. 552-565, 2023.

SAMI, M.; MOBIN, I. **A Comparative Study on Variational Autoencoders and Generative Adversarial Networks.** In: 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, p. 1-5, doi: 10.1109/ICAIIT.2019.8834544.

SANTIAGO, Diana (Org.). **Prática musical, memória e linguagem.** Salvador, BA: EDUFBA, 2018. 286 p. (Paralaxe ; 4). ISBN 9788523217143.

SERRA, Arthur et al. (2021). **Quality Enhancement of Highly Degraded Music Using Deep Learning-Based Prediction Models for Lost Frequencies**. 205-211. 10.1145/3470482.3479635.

SHEN, Guizhu et al. Deep learning with gated recurrent unit networks for financial sequence predictions. **Procedia computer science**, v. 131, p. 895-903, 2018.

SHULEI JI, J. L.; XINYU, Y. **A Comprehensive Survey on Deep Music Generation:** Multi-level Representations, Algorithms, Evaluations, and Future Directions. 2020. arXiv preprint arXiv:2011.06801.

STEURER, M.; HILL, R. J.; PFEIFER, N. Metrics for evaluating the performance of machine learning based automated valuation models. **Journal of Property Research**, Graz, v. 38, n. 2, p. 99-129, nov. 2020.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning:** An introduction. Cambridge, London: MIT Press, 2018.

THE KERAS BLOG. **Building Autoencoders in Keras.** Online, 2024. Disponível em: https://blog.keras.io/building-autoencoders-in-keras.html?source=post_page-----405ab73afa05---------- Acesso em: 05 fev. 2024.

UNESCO. **Frevo: artes cênicas do Carnaval do Recife.** Disponível em: https://ich.unesco.org/en/RL/frevo-performing-arts-of-the-carnival-of-recife-00603 . Acesso em: 05 fev. 2024.

WIKIPEDIA CONTRIBUTORS. **Recurrent neural network** — Wikipedia, The Free Encyclopedia. 2020.

WYSE, Lonce. **Audio spectrogram representations for processing with convolutional neural networks.** arXiv preprint arXiv:1706.09559, 2017.

XU, Siwei. **A Structural Overview of Reinforcement Learning Algorithms.** Junho 2020. Towards Data Science.

YANG, L.-C.; CHOU, S.-Y.; YANG, Y.-H. (2017**). MidiNet: A convolutional generative adversarial network for symbolic-domain music generation.** In: International Society for Music Information Retrieval Conference (ISMIR), 2017, Suzhou, China. Proceedings of the 2017 International Society for Music Information Retrieval Conference, p. 477-484.

ZHENG, Kaitong et al. **EmotionBox:** a music-element-driven emotional music generation system using Recurrent Neural Network. arXiv preprint arXiv:2112.08561, 2021.