UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

# PREDIÇÃO DE MÚSICAS UTILIZANDO REDES NEURAIS RECORRENTES

**VITOR HUGO FACHINETTO CUNHA**

Porto Alegre
2025

**VITOR HUGO FACHINETTO CUNHA**

# PREDIÇÃO DE MÚSICAS UTILIZANDO REDES NEURAIS RECORRENTES

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística pela Universidade Federal do Rio Grande do Sul (UFRGS).

Orientador metodológico
Prof. Dr. João Henrique Ferreira Flores

Porto Alegre
2025

Instituto de Matemática e Estatística
Departamento de Estatística

**PREDIÇÃO DE MÚSICAS UTILIZANDO REDES NEURAIS RECORRENTES**

Vitor Hugo Fachinetto Cunha

Banca examinadora:

Prof. Dra. Márcia Helena Barbian
Doutora pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Prof. Dr. Márcio Valk, UFRGS
Doutor pela Universidade Estadual de Campinas (Unicamp), Campinas, SP

## AGRADECIMENTOS

Agradeço a todos com quem convivi durante o percurso da minha formação, que à sua maneira contribuíram para o meu desenvolvimento pessoal, acadêmico e profissional.

À minha mãe, Mirian, agradeço pelo seu apoio sempre incondicional e imutável ao longo dessa jornada. Seu fascínio pela forma com que pensamos e seu incentivo para que eu explorasse diversas áreas da ciência deixaram um legado duradouro em minha vida, sem o qual eu definitivamente sequer teria pretensão de me tornar um estatístico.

Ao meu pai, Gilnei, agradeço pelas incansáveis incursões às artes, e o por ter me ensinado a persistir no selvagem passa-tempo do viver bem comportado. Me inspiro no seu modo de viver, e não conseguiria me imaginar enquanto ser-humano sem cultivar o amor à vida e ao próximo que ele me ensinou a cultivar no dia a dia.

À minha família, agradeço por ter sido sempre um ambiente de amor e carinho, pelas mesas fartas, pelos risos fáceis, pelas palavras de encorajamento e entusiasmo que me proporcionaram desde cedo. Para minha sorte, sempre que estive entre meus familiares estive acolhido, e espero que esse cuidado sempre reflita em minhas ações.

Aos meus amigos de longa data Nelson, Marco e Victor, agradeço pelos incontáveis momentos de apoio e risadas que me proporcionaram desde que nos conhecemos. Da praia ao campo, do futebol a qualquer modalidade esportiva que o Parque Germânia permita, que nossas aventuras permaneçam tão alegres quanto a nossa amizade permitir.

Aos meus amigos universitários, os quais são tantos e tão generosos, agradeço por me acompanharem nesse capítulo importante da minha vida. Graças a vocês, tenho o privilégio de dizer que vivi ao seu lado os melhores dias da minha vida (até agora). Obrigado por terem permitido que eu fizesse parte das suas histórias também.

Aos meus mestres e professores, agradeço por todos os ensinamentos, técnicos e práticos, que me foram confiados. As lições permanecerão, bem como aprendizados para além da sala de aula. Em especial, agradeço ao Prof. Dr. João Henrique Ferreira Flores por ter aceitado o desafio de me orientar durante a elaboração deste trabalho, e espero que ele possa se orgulhar de ter participado desse projeto e da minha graduação. Agradeço ainda aos estimados Prof. Dr. Márcio Valk e Profª. Drª. Márcia Helena Barbian, por terem aceito compor a banca de avaliação desse projeto e por suas contribuições generosas na apreciação deste texto.

Ainda, estendo esses agradecimentos a todos que de alguma forma estiveram presentes na minha vida. Ser formado pela UFRGS é o grande sonho da minha vida, e absolutamente todos, sem exceção, foram imprescindíveis para que ele se concretizasse. Muito obrigado por me lembrarem todos os dias que o mundo pode ser um lugar melhor.

“No caminho dessa descoberta, não tem compromisso
Com trilhar o que já foi trilhado
Deixo tudo ali do lado e parto cego e sempre em frente
Um tanto quanto alucinado
E nado para estar presente
Nada como estar assim ciente
Sem estar cansado
Sem estar à margem
Sem estar doente
Sem faltar coragem”

Nico Nicolaiewsky
[Excerto da música *Feito um picolé no sol, 1995*]

## RESUMO

A predição musical, embora pouco explorada na literatura, oferece uma miríade de possibilidades de aplicações de metodologias estatísticas, especialmente métodos de aprendizado de máquina, dada a natureza criativa do processo de composição musical. Por se tratar de um tema incipiente, o desenvolvimento deste trabalho tem por objetivo propor e analisar a arquitetura de uma rede neural recorrente que interprete arquivos de áudio sob a perspectiva de séries temporais de frequências fundamentais, baseada em uma metodologia de modelos de classe *Long Short-Term Memory*. A avaliação do desempenho se deu pela utilização de medidas de erros quadráticos médios e de erros absolutos médios percentuais, bem como análises pontuais por gênero musical foram consideradas. Constatou-se que o modelo se ajusta melhor a músicas em que o gênero musical principal seja homogêneo em suas características, como ritmos e cadências, fornecendo dados de performance do modelo para comparações de uma maneira geral e segmentada por gênero. Além disso, discutiu-se sobre a influência de elementos intrínsecos do objeto de estudo sobre o treinamento do modelo e como diferentes configurações de taxas de amostragem de áudio e intervalos de frequência na estimação da frequência fundamental podem interferir no processo de aprendizagem de máquina.

**Palavras-chave:** Música. Predição musical. Redes neurais recorrentes.

## ABSTRACT

Music prediction, although little explored in the literature, offers a myriad of possibilities for the application of statistical methodologies, especially machine learning methods, given the creative nature of the music composition process. Since this is an incipient topic, the development presented in this work aims to propose and analyze the architecture of a recurrent neural network that interprets audio files as time series of fundamental frequencies, based on a Long Short-Term Memory class model methodology. The performance evaluation was carried out using mean squared error measures and mean absolute percentage error, and specific analyses by musical genre are considered. It was found that the model fits better to songs in which the main musical genre is homogeneous in its characteristics, such as rhythms and cadences, providing model performance data for comparisons, in a general way and segmented by genre. Furthermore, the influence of intrinsic elements of the object of study on the training of the model was discussed, and how different configurations of audio sampling rates and frequency ranges in the estimation of the fundamental frequency can interfere in the machine learning process.

**Keywords:** Music. Music prediction. Recurrent neural networks.

# SUMÁRIO

**1 INTRODUÇÃO** **9**
**1.1 Conceitos** **10**
**1.2 Objetivos** **11**
**1.3 Delimitações** **12**
**2 REFERENCIAL TEÓRICO** **14**
**2.1 Redes neurais artificiais** **14**
2.1.1 Neurônio 15
2.1.2 Função de ativação 16
2.1.3 Arquitetura de redes neurais recorrentes 16
2.1.4 RNN padrão e Long Short-Term Memory (LSTM) 18
2.1.5 Função Perda e Descida Estocástica do Gradiente (SGD) 20
**2.2 Teoria musical básica** **21**
**2.3 Manipulação de arquivos de áudio** **23**
2.3.1 Taxa de amostragem de arquivos de áudio 25
2.3.3 Algoritmo YIN para detecção de frequências fundamentais 25
**3 METODOLOGIA** **27**
**3.1 Implementação computacional** **27**
**3.2 Base de dados** **28**
**3.3 Modelo** **28**
3.3.1 Arquitetura 29
3.3.2 Parâmetros e ajuste da rede 29
**4 RESULTADOS** **31**
**5 CONSIDERAÇÕES FINAIS** **39**
**REFERÊNCIAS** **41**

# 1 INTRODUÇÃO

Compor e publicar uma obra musical nunca foi tão acessível. Com o advento de *softwares* de produção musical aliados à criação de plataformas próprias para publicação e consumo de áudio, como *Spotify*, *Deezer*, *Youtube Music* entre inúmeras outras, o fluxo de publicação deixa de passar obrigatoriamente por grandes gravadoras e passa a ser publicado diretamente pelo artista em uma biblioteca musical de sua escolha. Ou seja, tanto a composição e produção musical quanto a publicação estão agora ao alcance de qualquer pessoa com algum conhecimento musical. Essa autonomia resulta em um fluxo de composição e publicação que atinge patamares anuais cada vez maiores. Apenas considerando a plataforma *Spotify*, por exemplo, mais de 20 milhões de obras foram disponibilizadas em seu repositório digital em 2021 (Spotify, 2025), e segundo Daniel Ek, fundador da plataforma, caso a crescente tendência de *uploads* se mantenha, a perspectiva é que em 2025 o total de *uploads* ultrapasse a marca de 130 milhões de arquivos (Ingham, 2021), sejam eles lançamentos, regravações, remasterizações ou novas versões de músicas já existentes, representando, em média, mais de quatro obras publicadas por segundo no ano.

Assim, tendo em vista a crescente disponibilização de músicas e a tendência de incremento ainda maior da capacidade de produção dessa indústria, faz-se necessário algum método de identificação automatizado que possibilite a predição de frequências de forma ágil para que músicos, sejam eles amadores ou profissionais, possam ter estimativas da tendência tonal de uma música em tempo real, facilitando assim a sua compreensão sobre a obra em questão e a sua capacidade de reproduzi-la. Deste modo, é possível mapear músicas sob demanda, poupando o tempo que previamente seria destinado ao estudo minucioso de conceitos específicos de teoria musical e treinamento ostensivo de técnicas analógicas de reconhecimento de sons. Trata-se então de um tema de pesquisa que envolve predição de uma série de elementos (frequências de notas musicais) em que métodos de modelagem via aprendizado de máquina se tornam ferramentas valiosas.

## 1.1 Conceitos

De acordo com Goodfellow, Bengio e Courville (2016), um método de aprendizado de máquina é uma técnica que permite que sistemas computacionais melhorem o seu desempenho, e é a única abordagem viável para construir sistemas de inteligência artificial com potencial para operar em situações do mundo real. Dentre os diferentes métodos existentes na área do conhecimento de aprendizado de máquina, a arquitetura proporcionada pelas metodologias de Redes Neurais Recorrentes (RNN) se destaca devido à sua performance em situações semelhantes às encontradas em composições musicais. Temos, por exemplo, Ashraf et al. (2023) que propuseram um modelo híbrido entre redes neurais convolutivas (CNN) e redes neurais recorrentes para classificar músicas em gêneros musicais. Utilizando variações de RNN como *Long Short-Term Memory* (LSTM), *Gated Recurrent Unit* (GRU) juntamente com Espectrograma de Mel como base de comparação de desempenho, esse modelo atingiu acurácia de 89,30% para a classificação de músicas em gêneros musicais. Sigtia et al. (2014) também propuseram um modelo híbrido, mas diferentemente de Ashraf e seus colegas, tinham como objetivo implementar um método de transcrição automático de música utilizando um estimador autorregressivo de distribuição na composição do modelo de RNN (RNN-NADE). A combinação de um modelo de RNN com um modelo de transcrição acústica se mostrou mais bem adaptado à transcrição de melodias polifônicas, ou seja, nas melodias em que mais de um som é executado simultaneamente.

Luo e Yu (2023) propuseram o uso de RNN para “desmixar” frequências misturadas de arquivos de áudio, o que significa separar os diferentes componentes que, unidos, resultam na única frequência registrada no arquivo de interesse. Com uma largura de banda de frequência definida pelo usuário para que o programa separe e estime as diferentes faixas de áudio implícitas, os pesquisadores foram capazes de criar um *software* mais eficiente que os disponíveis até então para esse propósito. Apesar de terem encontrado resultados expressivos, os próprios autores ressaltam no artigo que para um bom desempenho é necessário que a largura de banda seja delimitada por um especialista, não se tratando, então, de um método de estimação

completamente automatizado. Em abordagem ao tema de transcrição musical utilizando redes neurais, Dua et al. (2020) propuseram a utilização de RNN para transcrição automática de arquivos de áudio para cifras. Nesse tipo de uso, o usuário insere no modelo o arquivo de áudio que deseja transcrever e então o modelo estima o resultado, gerando o arquivo de saída, utilizando métodos de GRU para a partir daí decompor o arquivo final em quatro partes: contrabaixo, percussão, melodia e vocais. Já Hahn (2022), estudando o tema de composição algorítmica, utilizou uma RNN-LSTM para gerar músicas algoritmicamente, lançando mão de diferentes parametrizações do modelo para que fossem gerados resultados a partir dos bancos de dados *Irish* e *ABC Notation*. Com base na métrica de perplexidade, ele concluiu que valores menores nessa medida contribuem para a geração de peças musicais via algoritmo.

Cherla (2016) e, posteriormente, Yu (2022) abordaram diretamente a predição de sequências melódicas utilizando modelos distintos. Enquanto o primeiro utilizou uma variação do modelo de máquina de Boltzmann Restrito (*Recurrent Temporal Discriminative Restricted Boltzmann Machine* [RTDRBM]), propício para análise e categorização de dados recorrentes que apresentem forte correlação temporal, o segundo utilizou um Modelo de Markov de Ordem Variável (MMOV) em conjunto com um modelo "*Bar-Cycle*" para modelar as características melódicas e temporais das observações. Cherla (2016) construiu seu modelo RTRDBM de modo a estimar em um mesmo construto o tempo e a frequência da próxima nota, enquanto Yu (2022) uniu as predições do comprimento dos intervalos de repetições do modelo MMOV com as predições de repetições dentro de um intervalo fixo de tempo. Ambas são alguns dos exemplos das possíveis combinações de métodos já existentes, e apesar de não utilizarem diretamente modelos que sigam a estrutura de RNN, ainda assim inspiram a elaboração e teste de outras combinações de técnicas para a predição de sequências melódicas.

## 1.2 Objetivos

O presente trabalho tem como objetivo principal estimar qual a próxima frequência fundamental musical a ser executada a partir de um trecho curto de áudio,

propondo uma arquitetura e uma parametrização de rede neural que sejam eficientes tanto do ponto de vista qualitativo quanto do ponto de vista de tempo computacional. Partindo da perspectiva de publicação de grande quantidade de novas obras musicais que se projeta num futuro próximo, o intuito desse trabalho é o de apresentar uma metodologia capaz de predizer frequências fundamentais a partir de arquivos de áudio que possa ser útil em contextos em que não há tempo hábil para consultar outros recursos, ou ainda para que essa técnica possa ser aplicada em obras musicais ainda desconhecidas no momento de sua execução.

### 1.3 Delimitações

Este trabalho foi executado a partir de delimitações práticas para proporcionar uma execução mais simples e acessível do modelo proposto. A base de dados utilizada para treinamento e teste do modelo contém apenas músicas livres de direito autoral, respeitando, assim, a Lei de Direito Autoral. O tempo de treinamento da rede neural foi estipulado em 5 segundos com o intuito de reduzir as possibilidades de predição da sequência harmônica a partir de um número limitado de informações. O tempo de teste foi fixado em 5 segundos para que fosse possível obter predições de forma rápida, de modo a reduzir a margem para alucinações do modelo, cujas ocorrências são mais prováveis em intervalos mais longos de predição. Dadas as conversões de arquivo necessárias para a implementação do método, cada segundo de áudio inserido no modelo pode ser interpretado como uma sucessão de 43 pontos observados de uma mesma série temporal.

Ainda, considerando os objetivos deste trabalho, prezou-se pela elaboração de um modelo robusto e ágil, que pudesse predizer frequências fundamentais advindas de arquivos de qualquer gênero musical, e que, ao mesmo tempo, gerasse as predições em tempo adequado para terem utilidade, obtendo o resultado antes da execução da frequência em tempo real. A partir da praticidade de utilização almejada, também se optou por não utilizar no algoritmo nenhum dispositivo de áudio externo, seja ele físico ou digital, durante o processo de estimação e predição. Dessa forma, decidiu-se pela

construção do algoritmo apenas com recursos nativos do sistema operacional e da linguagem de programação utilizados

# 2 REFERENCIAL TEÓRICO

## 2.1 Redes neurais artificiais

Redes Neurais Artificiais (RNAs) são estruturas de aprendizado de máquina projetadas inicialmente para remeter ao processo de aprendizagem de seres vivos (Aggarwal et al., 2018), cuja arquitetura pode ser similar a modelos estatísticos bem estabelecidos na literatura, de modo que sua proposição e execução pode ser considerada análoga à aplicação de técnicas estatísticas de modelagem (Hahn, 2022). A partir disso, pode-se traçar um paralelo entre os procedimentos de especificação da arquitetura da rede e a especificação de um modelo estatístico, e entre os procedimentos de treinamento da rede com um determinado conjunto de dados de treinamento e a estimação dos parâmetros do modelo. Para tornar esse paralelo mais compreensível, Hahn (2022) apresentou uma lista de equivalências, o que pode contribuir para o entendimento da relação entre as duas perspectivas (Quadro 1).

Quadro 1 – Relação de equivalências entre a terminologia estatística e a da literatura sobre redes neurais

| Literatura estatística | Literatura de RNAs |
|---|---|
| Variáveis | Atributos |
| Variáveis independentes | Entrada (*Input*) |
| Valores preditos | Saída (*Output*) |
| Variáveis dependentes | Variáveis-alvo / Valores de treinamento (*Target*) |
| Resíduos | Erros |
| Estimação | Treinamento / aprendizagem |
| Critérios de estimação | Função perda / Função custo / Função erro |
| Observações | Pares de treinamento (*Training Pairs*) |
| Parâmetros estimados | Pesos (Sinápticos) |
| Transformações | Links funcionais (*Functional Links*) |
| Interpolação / extrapolação | Generalização |

Fonte: Elaborada com base em Hahn, 2022

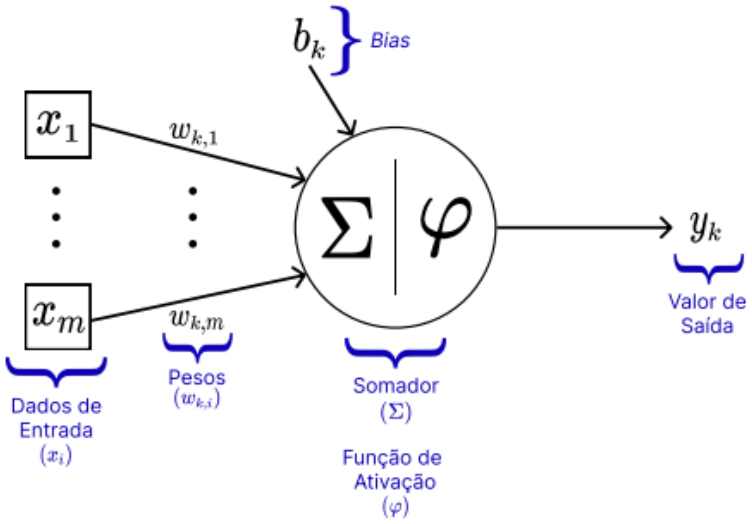
A seguir, serão apresentados alguns conceitos fundamentais para compreensão do modelo de RNN proposto neste trabalho aplicado à área musical.

### 2.1.1 Neurônio

Redes Neurais Artificiais são constituídas por neurônios, que são unidades de processamento de informação (Haykin, 2009). É possível definir quatro elementos fundamentais que compõem um neurônio (Figura 1):

1. Um conjunto de sinapses que são caracterizadas individualmente por um peso próprio e que ponderam os dados de entrada. Esse peso pode ser tanto positivo quanto negativo.
2. Uma função agregadora que soma os dados de entrada ponderados pelos pesos sinápticos, formando uma combinação linear.
3. Um intercepto – denominado Bias − aplicado ao resultado da função agregadora para modificar a entrada da função de ativação.
4. Uma função de ativação responsável por restringir a amplitude de saída do neurônio, garantindo que um valor finito seja gerado.

Figura 1 − Modelo do *k*-ésimo neurônio.



Fonte: adaptada de Hair *et al.* (2005) e Haykin (2009) *apud* Hahn (2022, p. 15).

A partir uma sequência $x_1, \cdots, x_m$ de dados de entrada (valores observados) e de uma função de ativação φ(·), o *k*-ésimo neurônio em uma rede neural pode ser definido por meio de uma função que leva o vetor de entrada ($x_1, \cdots, x_m$) na k-ésima saída yk. Um exemplo, conforme apresentado por Hahn (2022), é dado pelo sistema de equações

$$u_k := \sum_{i=1}^{m} w_{k,i} x_i;$$

$$y_k := \varphi(u_k + b_k), \quad (2.1)$$

em que $b_k$ representa o bias e $w_{k,i} \in \mathbb{R}$, correspondendo aos pesos associados ao neurônio k.

### 2.1.2 Função de ativação

A função de ativação de um modelo neuronal tem por objetivo transformar o valor de $v_k := u_k + b_k$ em um valor condizente com a distribuição de $y_k$. Para o presente trabalho optou-se por utilizar a função tangente hiperbólica (*tanh*), uma vez que a sua utilização é mais comum em estruturas de redes recorrentes. A função $f : R \longrightarrow (-1, 1)$ é dada por

$$f(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2.2)$$

### 2.1.3 Arquitetura de redes neurais recorrentes

A utilização de um único neurônio geralmente não é muito efetiva na resolução de problemas complexos, então a sua utilização em camadas interligadas muitas vezes é considerada.

Uma rede neural é formada por três tipos de camadas:

1. **Camada de entrada:** nesta camada os dados “brutos” são inseridos na rede sem função de ativação.
2. **Camadas intermediárias ou ocultas:** nelas ocorre o processamento dos dados de entrada, os quais podem ser oriundos tanto da camada de entrada quanto de uma camada oculta prévia. Sua saída pode ser tanto utilizada para alimentar uma camada oculta posterior quanto a própria camada de saída, gerando assim os valores preditos. É pela natureza dessas camadas que as redes neurais são costumeiramente classificadas como modelos “caixa-preta”, pois usualmente não há uma interpretação clara das operações nessas camadas que possam traçar comparações com os dados na vida real (Rojas, 1996).
3. **Camada de saída:** esta camada retorna um ou mais valores definitivos, conforme a arquitetura definida. Caso seja um modelo preditivo, o dado resultante será a própria predição; caso seja um modelo utilizado para problemas de classificação, o dado resultante será utilizado segundo alguma regra de classificação.

Segundo Hagan *et al.* (2014), as RNAs podem ser classificadas em duas arquiteturas distintas: Redes Alimentadas Adiante (FFNs) e Redes Neurais Recorrentes (RNNs).

A primeira classificação se refere a redes em que o fluxo de informação entre a camada de entrada e a camada de saída é unidirecional (Goodfellow; Bengio; Courvill, 2016), isto é, a informação que sai de um neurônio é sempre direcionada a um neurônio seguinte, seja ele de outra camada oculta ou da camada de saída. Já a segunda classificação é destinada a redes similares, porém com um diferencial importante: a existência de laços de realimentação. Nesse tipo de rede, a informação já processada por um neurônio pode ser reprocessada pelo mesmo neurônio durante o processo de aprendizagem. Esse tipo de rede neural foi desenvolvido para processar séries temporais e outros tipos de dados sequenciais (Fan; Ma; Zhong, 2021), razão pela qual foi a arquitetura escolhida para a proposição de implementação do modelo apresentado neste trabalho.

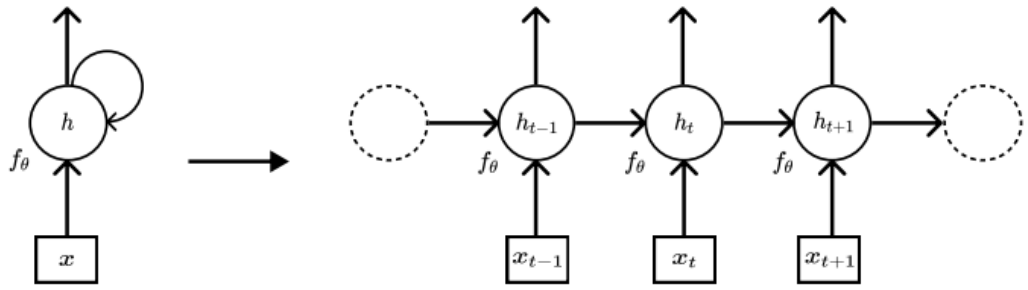
### 2.1.4 RNN padrão e *Long Short-Term Memory* (LSTM)

Sejam $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ os dados de entrada. Uma RNN padrão, também conhecida como *RNN Vanilla* (Stérin; Farrugia; Gripon, 2017) (Figura 2), modela o estado oculto $\boldsymbol{h}_t$ por meio da fórmula

$$\boldsymbol{h}_t = f_\theta(h_{t-1}, x_t) \tag{2.3}$$

em que $f_\theta$ geralmente é uma função não linear parametrizada por $\theta$.

Figura 2 − Diagrama de uma RNN Vanilla.

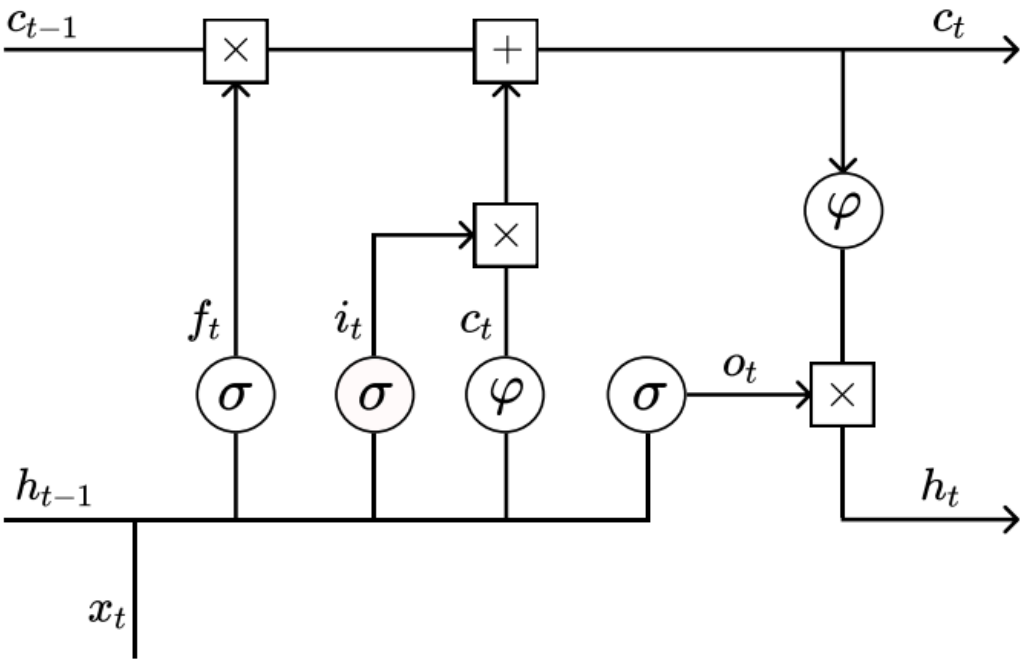



Fonte: adaptada de Goodfellow, Bengio e Courville (2016) e Kamath, Liu e Whitaker (2019) *apud* Hahn (2022, p. 19).

De maneira análoga à modelagem de séries temporais, nesse caso também há o compartilhamento dos parâmetros ao longo do tempo. Uma vantagem dessa abordagem é a possibilidade de fazer predições sem a necessidade de armazenamento de todas as informações na sequência de entrada até o tempo *t*, ou seja, são necessárias apenas informações suficientes para prever o restante da sentença (Hahn, 2022). Por outro lado, a arquitetura padrão encontra dificuldade em captar

dependências temporais longas nos dados; razão de ter sido proposta uma alternativa para contornar esse problema utilizando estruturas denominadas portões (*gates*) para manter ou descartar o estado oculto $\boldsymbol{h}_t$ a cada passo temporal ( Kamath; Liu; Whitaker, 2019) (Figura 3).

Figura 3 − Diagrama de uma LSTM, considerando *σ* como a função de ativação *logit.*




Fonte: adaptada de Kamath, Liu e Whitaker (2019) *apud* Hahn (2022, p. 20)

Considerando que o produto de Haddamard (⊙), corresponde à multiplicação elemento-a-elemento; que o estado de célula $c_t$ carrega informações da sequência; que o portão de esquecimento (*forget gate*) $f_t$ determina por quanto tempo os valores de $c_{t-1}$ são mantidos; que o portão de entrada (*input gate*) $i_t$ controla a quantidade de atualizações realizadas no estado de célula; e que o portão de saída (*output gate*) $o_t$ define quanta informação de $c_t$ é revelada para $h_t$, a arquitetura da rede pode ser descrita como

$$c_t = f_t \odot c_{t-1} + i_t \odot h_t, \tag{2.4}$$

$$h_t = o_t \odot \varphi(c_t), \tag{2.5}$$

em que $\varphi(\cdot)$ é uma função de ativação, geralmente a *tanh*.

Os elementos desses portões têm valores no intervalo (0, 1), garantido pelo uso de outra função de ativação, denominada *logit*, definida por $f : R \longrightarrow (0,\ 1)$

$$f\ (x) := \frac{1}{1+e^{-x}} \tag{2.6}$$

Com a adição do termo $f_t \odot c_{t-1}$ na fórmula de $c_t$, a rede captura longa dependência nos dados (Goodfellow; Bengio; Courville, 2016).

### 2.1.5 Função Perda e Descida Estocástica do Gradiente (SGD)

Para que ocorra o ajuste dos pesos do modelo é necessário que se use algum método que direcione a atualização dos pesos estimados de modo que o erro das predições do modelo seja minimizado. Para isso, uma função perda (*loss function*) é definida e deve ser minimizada para que os pesos sejam atualizados de modo a convergir para um valor ótimo, idealmente próximo de zero.

Seja $f : R^n \longrightarrow R^n$ uma função diferenciável, e seja representado por $\nabla_x f(x)$ o gradiente de $f$ por meio do método da descida estocástica do gradiente (*gradient descent*), um novo ponto ***x*** é dado por

$$x' = x - \epsilon \nabla_x f(x), \tag{2.7}$$

em que $\epsilon$ é a taxa de aprendizagem (*learning rate*), a qual é um valor escalar positivo que determina o tamanho do passo relativo ao valor do gradiente no ponto.

Considera-se que o método converge quando todos os elementos do gradiente ficam abaixo de uma tolerância predeterminada (Goodfellow; Bengio; Courville, 2016). Essa forma de otimização requer que todo o conjunto de treino seja utilizado na rede para que os pesos sejam atualizados. Assim, uma alternativa consiste em utilizar a SGD, a qual utiliza apenas uma parte dos dados – denominada lote – antes de aplicar o procedimento de minimização. Partindo de um valor inicial $\theta_0$, o SGD ajusta os parâmetros $\theta_t$ em direção ao gradiente negativo, minimizando a função perda com iterações de uma partição aleatória do conjunto de treino, que, pela Lei dos Grandes Números, é uma boa aproximação do gradiente verdadeiro. Depois que os dados dessa partição passam pela rede, uma época (*epoch*) é concluída, e depois de um determinado número de épocas preestabelecido, o treinamento é concluído (Goodfellow; Bengio; Courville, 2016). Uma prática comum é usar o erro de predição para calcular o gradiente, e através do SGD atualizar os pesos dos neurônios, em algoritmos denominados algoritmos de retropropagação (*backpropagation*). Para o caso das RNNs, o tempo de cada observação dentro do contexto temporal dos dados também é levado em consideração, e, dada essa particularidade, o algoritmo é denominado algoritmo de retropropagação ao longo do tempo (*Backpropagation Through Time* [BPTT]) (Goodfellow; Bengio; Courville, 2016).

## 2.2 Teoria musical básica

Para o entendimento das particularidades da aplicação do modelo de aprendizagem de máquina em música é fundamental que se conheça o objeto de pesquisa e suas características. Neste trabalho nos atemos a expor e definir alguns fundamentos relevantes para o entendimento da aplicação e discussão dos resultados a partir de estudos de Mascarenhas (1968), Gusmão (2012), de Alvarenga (2017) e Dantas e Cruz (2018).

- **Música** – É a arte de combinar os sons. Sua existência se dá pela união entre som, ritmo, melodia e harmonia

- **Som** – É o resultado de um movimento específico que se produz em um instrumento, na voz ou em um objeto qualquer, no ato de tocar, cantar ou bater no objeto. Esse movimento cria vibrações, que originam o som que ouvimos.
- **Ritmo** – É o movimento ordenado de sons no tempo.
- **Melodia** – É uma sequência de sons que forma um sentido.
- **Harmonia** – São sons simultâneos combinados.
- **Nota musical** – É a representação de um som.
- **Escala** – É uma série de notas sucessivas. Um exemplo de escala é a escala de Dó Maior: Dó – Ré – Mi – Fá – Sol – Lá – Si – Dó.
- **Tons e semitons** – Um tom é o intervalo entre duas notas, enquanto um semitom é a menor distância entre duas notas. Um tom sempre comporta uma nota intermediária, que o divide em dois semitons.
- **Alterações** – É um fenômeno que ocorre quando a entonação de uma nota é alterada. As alterações mais conhecidas são o sustenido e o bemol, os quais alteram a nota a que se referem um semitom acima e abaixo, respectivamente. Por exemplo, na escala de Dó Maior, um Dó sustenido e um Ré bemol têm o mesmo som.
- **Acorde** – É um conjunto harmônico de três ou mais notas. É o elemento que mais caracteriza ou confirma um tom.
- **Tema** – É a melodia principal de uma música.
- **Cadência** – É uma sequência de acordes que produz uma harmonia característica.
- **Oitava** – É o intervalo entre duas notas musicais que têm o mesmo nome, mas frequências diferentes. A frequência da nota de frequência superior é o dobro da frequência da nota de frequência inferior. Por exemplo, a primeira tecla branca de um piano ao ser pressionada produz a nota $Lá_0$, com frequência de 27,5 Hz, enquanto a oitava tecla branca produz a nota $Lá_1$, com frequência de 55 Hz quando pressionada.
- **Frequência fundamental** – É a nota mais grave que compõe um acorde, a partir da escala utilizada.

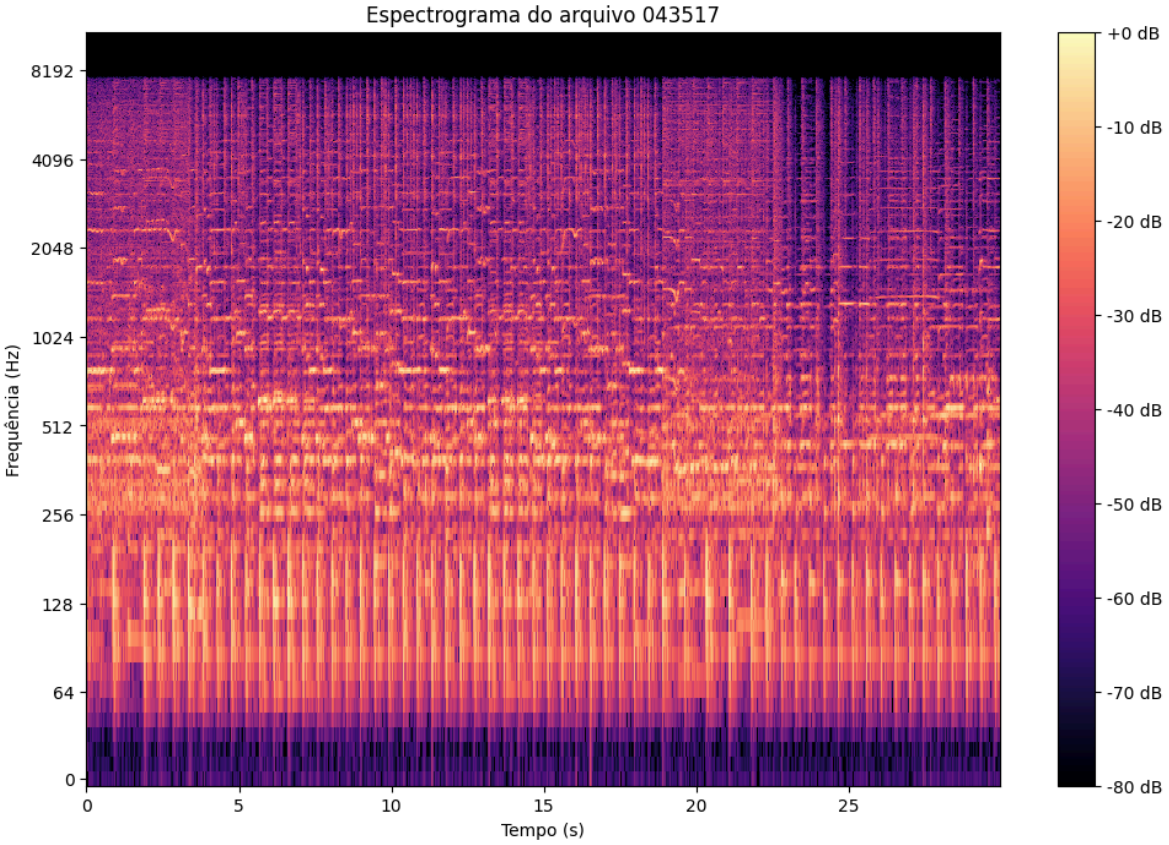
## 2.3 Manipulação de arquivos de áudio

Arquivos de áudio *.wav* caracterizam-se por serem arquivos sem compressão, em que não há perda de qualidade sonora, de forma que ocupa um maior espaço de armazenamento, com sua utilização sendo indicada em trabalhos profissionais, ao passo que o formato de áudio *.mp3* passa por um processo de compressão que o torna de 75% a 95% menor do que um arquivo sem compressão (Pereira, 2017). Esse processo de compressão consiste em três passos de codificação do arquivo que melhoram o desempenho de compressão ao custo de uma maior complexidade do codificador e perda de informação, priorizando a manutenção dos sons mais audíveis da gravação (Araujo; Ferreira; Peixoto, 2015).

Esses formatos de arquivo foram para este trabalho em razão de que o banco de dados utilizado apresenta os arquivos no formato *.mp3* e a biblioteca utilizada para a estimação da frequência fundamental apresentou melhores resultados com a entrada no formato *.wav*.

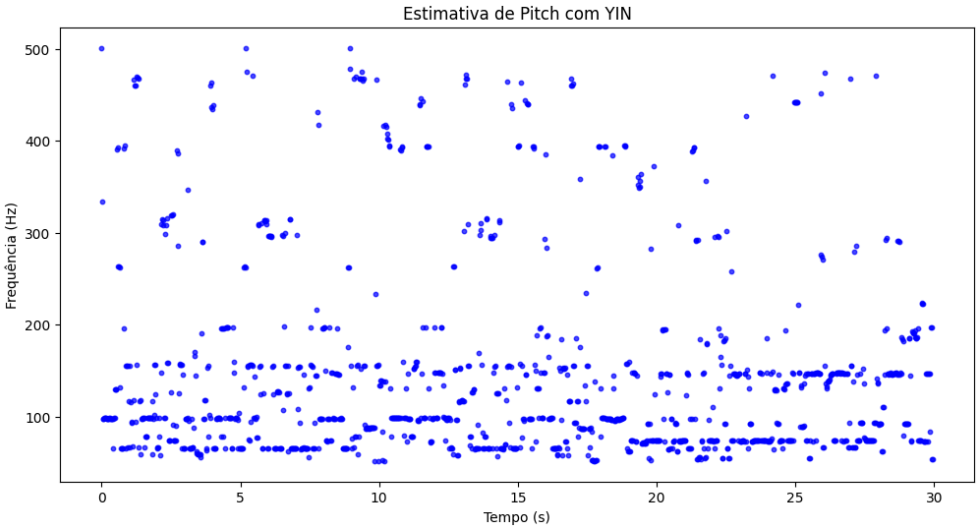
A manipulação de arquivos de áudio para a detecção da frequência fundamental foi realizada para possibilitar que os dados apresentados ao modelo fossem inseridos a partir da ótica de série temporal. A partir de arquivos *.mp3* polifônicos, com diversas medidas de intensidade de som por frequência e por tempo, foi possível estimar a frequência fundamental instante a instante de cada arquivo de áudio, resumindo a informação de diversas camadas de áudio em uma única medida de frequência fundamental para cada $t \in T$, em que $T$ é o número de observações extraídas de uma música. O banco de dados utilizado é padronizado, com todas as músicas tendo a mesma duração (30 segundos). Assim, todas as músicas têm $T = 1.290$ observações. Para exemplificar, utilizarei o arquivo 043517 presente na base de dados disponibilizada por Deferrard et. Al (2017)

Figura 4 − Espectrograma do arquivo 043517




Fonte: adaptada de Defferrard et. al (2017)

Figura 5 − Estimativa de pitch do arquivo 043517 utilizando o algoritmo YIN




Fonte: adaptada de Defferrard et. al (2017)

### 2.3.1 Taxa de amostragem de arquivos de áudio

De acordo com McFee *et al.* (2015), a taxa de amostragem (*sample rate*) constitui-se na quantidade de amostras discretas de um sinal de áudio capturadas por segundo durante o processo de digitalização. Representada por $f_s$ (em Hz), ela define a resolução temporal do sinal digitalizado e afeta diretamente a fidelidade da conversão analógico-digital.

Para o presente trabalho, se determinou uma taxa de amostragem de 16 kHz, uma vez que essa é a frequência média máxima que o ouvido humano é capaz de identificar (aproximadamente 20 kHz) (Wuensche, 2006) e se optou por uma quantidade parcimoniosa de dados, de modo a evitar uma sobrecarga do processo computacional. Como o intervalo utilizado para o reconhecimento da frequência fundamental foi fixado num intervalo entre 50 Hz e 500 Hz, não há prejuízo na qualidade obtida dos arquivos.

### 2.3.3 Algoritmo YIN para detecção de frequências fundamentais

O algoritmo proposto por de Cheveigné e Kawahara (2002) parte da função de autocorrelação (ACF) de um sinal discreto $x_t$, sobre a qual é aplicada uma série de

melhorias para reduzir as taxas de erro, sendo que sua estimação se dá através do estabelecimento de limites possíveis para a frequência, da escolha de uma estimativa local e da otimização dessa estimativa, nessa ordem, resultando no valor final do processo.

Esse algoritmo desempenha papel imprescindível no desenvolvimento deste trabalho, tendo sido utilizado para estimar as frequências fundamentais utilizadas como valores de treinamento e de teste do modelo. Foi a partir de sua aplicação nos arquivos de áudio do banco de dados que se tornou possível obter a “discretização” e estimação das frequências de cada música em cada tempo. Além disso, por ter sido aplicado a todos os arquivos de treino e de teste da mesma forma, eventuais erros não impactaram diretamente na avaliação do modelo proposto, tendo em vista a uniformidade do tratamento dado a ambos os conjuntos de arquivos.

# 3 METODOLOGIA

O presente estudo tem por objetivo a implementação de uma rede neural adaptada para a estimação de sequências harmônicas a partir de fragmentos de arquivos de áudio para simular uma interação em tempo real do algoritmo com o ambiente prático. O resultado gerado será uma sequência de frequências fundamentais em formato numérico.

## 3.1 Implementação computacional

O processo de coleta de dados, desenvolvimento do modelo empregado e análise dos resultados deste trabalho foram desenvolvidos utilizando a linguagem de programação **Python**, com o auxílio da IDE (Integrated Development Environment) **VS Code**. As bibliotecas utilizadas ao longo do trabalho incluem principalmente as indicadas a seguir:[1]

- **Python**: versão 3.12.8
- **VS Code**: versão 1.97.2
- **TensorFlow**: versão 2.18.0
- **NumPy**: versão 2.0.2
- **Matplotlib**: versão 3.9.3
- **pandas**: versão 2.2.3
- **librosa**: versão 0.10.2.post1
- **Keras**: versão 3.7.0
- **scikit-learn**: versão 1.6.0
- **wakepi**: versão 0.10.1
- **PyAV**: versão 14.0.1

[1] As demais bibliotecas utilizadas e os códigos desenvolvidos podem ser vistos em github.com/Vitor32459X/TCC

## 3.2 Base de dados

Este trabalho foi realizado utilizando uma base de dados de 8.000 arquivos no formato *.mp3* livres de direitos autorais, objeto de estudo de Defferrard *et al*. (2017).[2]

Para que cada arquivo pudesse ser analisado pelo modelo, fez-se necessária a aplicação de dois tratamentos prévios para uniformização da entrada: conversão e extração das frequências fundamentais, cujo processo está descrito a seguir:

- **Conversão:** para que fosse possível aplicar os métodos da biblioteca **librosa**, foi necessário converter os arquivos de formato *.mp3* para arquivos *.wav*, a uma taxa de amostragem de 16 kHz. Isso foi feito utilizando as bibliotecas **PyAV** e **os**, gerando cópias dos arquivos originais em formato próprio para leitura da biblioteca **librosa**.
- **Extração das frequências fundamentais:** tendo em vista a natureza dos arquivos de áudio, foi utilizada a biblioteca **librosa** para extrair as frequências fundamentais de cada música e registrá-las em um vetor de modo que fosse possível inserir no modelo. O processo de extração foi construído levando em consideração a taxa de amostragem definida no processo de conversão do arquivo (16kHz), com os limites de frequência de 50 Hz e 500 Hz utilizando a função **librosa.yin()**, concentrando os valores de treino e teste nessa faixa para proporcionar economia computacional, não gerando prejuízo para a estimação devido à natureza cíclica das notas musicais.

## 3.3 Modelo

Um modelo de RNN-LSTM foi construído a partir de propostas apresentadas por Goodfellow, Bengio e Courville (2016), e a base de dados foi dividida de modo a ser 80% utilizada para treino e 20% utilizada para teste, com os tratamentos sendo aplicados para cada arquivo após essa divisão.

---

[2] Essa base de dados, compilada por Defferrard *et al*. (2017), está disponível em um repositório no GitHub (github.com/mdeff/fma/tree/master) dos autores, tendo sido utilizada com autorização.

### 3.3.1 Arquitetura

O modelo proposto constitui-se de três camadas: a primeira contém as sequências de frequências fundamentais extraídas de um arquivo específico, caracterizadas por uma duração total de 5 segundos, apresentando uma lista unidimensional de 215 observações de frequências fundamentais estimadas igualmente espaçadas ao longo do tempo a cada trecho. A segunda camada contempla os neurônios encarregados pela identificação dos padrões da série temporal fornecida, que geram as próximas 215 observações a partir dos dados de entrada. É Importante ressaltar que o processo de estimação é feito todo em uma única etapa, não havendo influência de valores preditos para um tempo $t$ na predição para um tempo $t + i$, sendo $i$ um intervalo de tempo qualquer. Ao final do processo, a camada de saída apresenta os resultados preditos pela rede neural, também com duração total de 5 segundos, ou 215 observações, com estrutura igual à da lista de entrada.

### 3.3.2 Parâmetros e ajuste da rede

Para o processo de treinamento da rede neural foram sorteados 6.400 arquivos do banco de dados que foram inseridos integralmente no ciclo de atualização dos pesos sinápticos do modelo, em que considerou-se 20 épocas por arquivo. Foi definida uma camada oculta de 125 neurônios LSTM. A cada um desses neurônios foi atribuída a função de ativação ***tanh***, e uma taxa de aprendizado de $10^{-3}$, padrão da biblioteca **keras**. O desenvolvimento proposto tem por base o exemplo construído em sala de aula a partir do estudo de Carvajal (2023)[3].

Já para o processo de teste foi inicialmente definida uma semente de aleatorização variável conforme o índice atual do *loop* para que amostras fossem extraídas dos 1.600 arquivos do conjunto de teste. Essa definição foi estabelecida para que o modelo pudesse ser testado em momentos aleatórios da música testada, de maneira análoga à utilização proposta na vida real. De cada música foram extraídas 10

[3] Disponível em https://github.com/andresca94/MasterCard-Stock-Price-Prediction-Using-LSTM-and-GRU

amostras de 10 segundos cada a partir de pontos aleatórios definidos pela aleatorização, que então foram divididas em 5 segundos de entrada no modelo e 5 segundos de teste da predição a partir dos dados inicializados na estrutura. Então, ao final do *loop*, a sequência predita e a sequência verdadeira para os 5 segundos finais de cada amostra foram concatenadas para que se pudesse calcular o erro percentual absoluto médio (MAPE) e a raiz do erro quadrático médio (RMSE) de forma geral. O intervalo de 5 segundos de entrada foi escolhido para que o modelo pudesse se ambientar com o tom e a escala da música inserida, dessa forma reduzindo as possibilidades de continuidade da sequência harmônica a ser predita que poderiam ser consideradas se esses elementos fossem ignorados. Por sua vez, o intervalo de 5 segundos de predição foi definido tendo em vista a utilização prática do modelo. Intervalos menores de predição fariam com que a taxa de atualização do resultado exibido fosse maior para um mesmo intervalo de tempo, tornando a duração da exibição da frequência predita mais breve, de modo que a execução das notas correspondentes em um instrumento musical seria mais difícil de acompanhar. Por outro lado, intervalos de predição maiores poderiam não ser tão úteis, posto que as observações relativas aos primeiros 5 segundos podem não seguir o mesmo encadeamento lógico de composição de observações mais distantes. Entre os exemplos de construções harmônicas “disruptivas” nesse sentido podem ser listadas mudanças de seção (verso, refrão, ponte, etc.), mudanças de tonalidade, mudanças pontuais em acordes de uma mesma sequência, entre outros. Em suma, a duração de 10 segundos de cada amostra, sendo 5 segundos para inserção no modelo e 5 segundos para a predição, foi estabelecida considerando esses dois fatores, e a experimentação de diferentes durações de amostra são encorajadas em estudos posteriores.

Ainda, a partir dos metadados disponibilizados de cada arquivo por Deferrard *et al*. (2017), foi construída uma categorização a partir do gênero musical predominante de cada música − “Electronic”, “Experimental”, “Folk”, “Hip-Hop”, “Instrumental”, “International”, “Pop” e “Rock” − a partir da qual foi possível analisar o desempenho da rede neural tendo em vista o gênero musical do arquivo.

# 4 RESULTADOS

Após o treinamento do modelo e sua teste utilizando intervalos sorteados aleatoriamente dentro de cada música no conjunto de teste foram obtidas as medidas para o erro percentual absoluto médio (MAPE) e raiz do erro quadrático médio (RMSE), conforme apresentado na Tabela 1.

**Tabela 1** - Medidas descritivas avaliadas para o modelo proposto estratificado por gênero musical e sua totalização ordenadas de acordo com os resultados do MAPE

| Gênero musical | N_arquivos | MAPE (%) | RMSE (Hz) |
|---|---|---|---|
| Rock | 203 | 38,25 | 62,37 |
| Instrumental | 209 | 43,07 | 72,78 |
| Folk | 173 | 45,72 | 77,05 |
| Pop | 210 | 46,65 | 78,23 |
| Experimental | 199 | 50,24 | 93,99 |
| International | 210 | 51,59 | 95,09 |
| Electronic | 219 | 52,71 | 80,98 |
| Hip-Hop | 176 | 56,31 | 83,23 |
| **Geral** | **1599** | **48,00** | **81,11** |

Fonte: do autor.
Nota: MAPE, erro percentual absoluto médio; RMSE, raiz quadrada do erro quadrático médio

**Tabela 2** - Medidas descritivas avaliadas para o erro do modelo proposto estratificado por gênero musical e sua totalização ordenadas por ordem alfabética do gênero musical.

| Gênero Musical | Máximo | Mínimo | Média | Mediana | Desvio Padrão | 1º Quartil | 3º Quartil |
|---|---|---|---|---|---|---|---|
| Electronic | 217,60 | -451,14 | -0,32 | 19,04 | 80,98 | -10,47 | 37,85 |
| Experimental | 213,94 | -439,54 | -13,45 | 7,77 | 93,02 | -33,11 | 36,69 |
| Folk | 201,89 | -425,61 | -10,78 | 8,13 | 76,29 | -33,91 | 32,91 |
| Hip-Hop | 449,26 | -449,81 | -2,09 | 18,89 | 83,20 | -21,52 | 39,97 |
| Instrumental | 211,48 | -449,42 | -5,65 | 9,81 | 72,55 | -19,78 | 31,47 |
| International | 218,39 | -449,89 | -13,08 | 12,58 | 94,19 | -40,20 | 40,00 |
| Pop | 211,47 | -434,04 | -5,15 | 13,40 | 78,06 | -18,23 | 34,22 |
| Rock | 450,35 | -440,72 | -3,46 | 9,76 | 62,27 | -14,79 | 27,29 |
| **Geral** | **450,35** | **-451,14** | **-6,69** | **12,29** | **80,84** | **-22,31** | **35,28** |

Fonte: do autor.

De forma geral, o modelo apresentou um MAPE de aproximadamente 48% e um RMSE de cerca de 80 Hz, cujas métricas podem ser interpretadas como a estimação de valores preditos, em média, 48% acima dos valores-alvo, e são em média 80 Hz distantes dos valores-alvo. Em relação aos resultados de MAPE, os gêneros que apresentaram a menor e a maior medida foram Rock e Hip-Hop, com 38,25% e 56,31%, respectivamente, ao passo que em relação à RMSE, os gêneros musicais que tiveram a menor e a maior medida foram Rock e International, com 62,37 Hz e 95,09 Hz, respectivamente.

Essa diferença de resultados entre gêneros musicais pode ser explicada pelas diferentes características de cada gênero, tais como qualidade de registro, instrumentação, cadências, timbres e temas populares de cada gênero, entre outros aspectos que podem caracterizar cada arquivo. O Rock, por exemplo, um gênero musical originado a partir do Blues estadunidense, alcançou popularidade de proporções mundiais, herdando estruturas bem definidas de tempo, harmonia, composição textual e melódica, as quais são mantidas até os dias atuais (Covach;

Flory, 2018), enquanto o gênero Hip-Hop é marcado por técnicas de *sampling* (utilização de recortes temporais de músicas já existentes na composição de uma nova obra), sobressaindo-se os elementos de voz e percussão (Price III, 2006), o que pode dificultar a percepção fluida do modelo sobre as frequências fundamentais executadas na base da música. Ainda, o gênero listado como International é constituído por um conjunto de gêneros regionais típicos de diversas regiões do mundo na base fornecida (Defferrard *et al*., 2017), as quais são listadas no Quadro 2.

Quadro 2 - Lista de subgêneros constituintes do gênero musical “International”

| **Gênero Principal** | **Subgênero** |
|---|---|
| International | Latin America |
| International | French |
| International | Reggae - Dub |
| International | Indian |
| International | African |
| International | Middle East |
| International | Polka |
| International | Balkan |
| International | Europe |
| International | Brasilian |
| International | Asia-Far East |
| International | Pacific |
| International | Celtic |
| International | Flamenco |
| International | Latin |

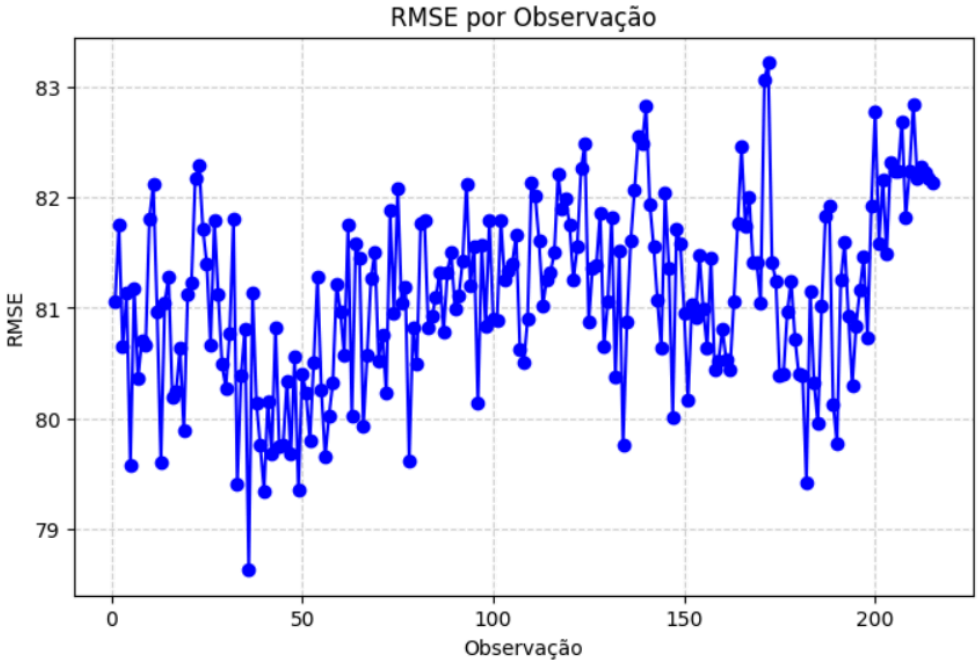
Fonte: elaborado pelo autor.

Essa variedade de subgêneros do gênero Internacional se traduz em uma variedade de elementos musicais característicos a cada um dos subgêneros listados no Quadro 2, adicionando heterogeneidade aos dados agrupados dentro do mesmo conjunto.

Avaliando todos os resultados, observa-se uma melhor resposta do modelo quando aplicado a gêneros notadamente mais homogêneos em relação às suas características, bem como uma pior resposta a gêneros com mais heterogeneidade de composições. Daí surge uma primeira conclusão: pode ser mais vantajoso e assertivo desenvolver uma treinamento que leve em consideração as particularidades de cada gênero, filtrando apenas os arquivos de um mesmo gênero musical para treino e teste.

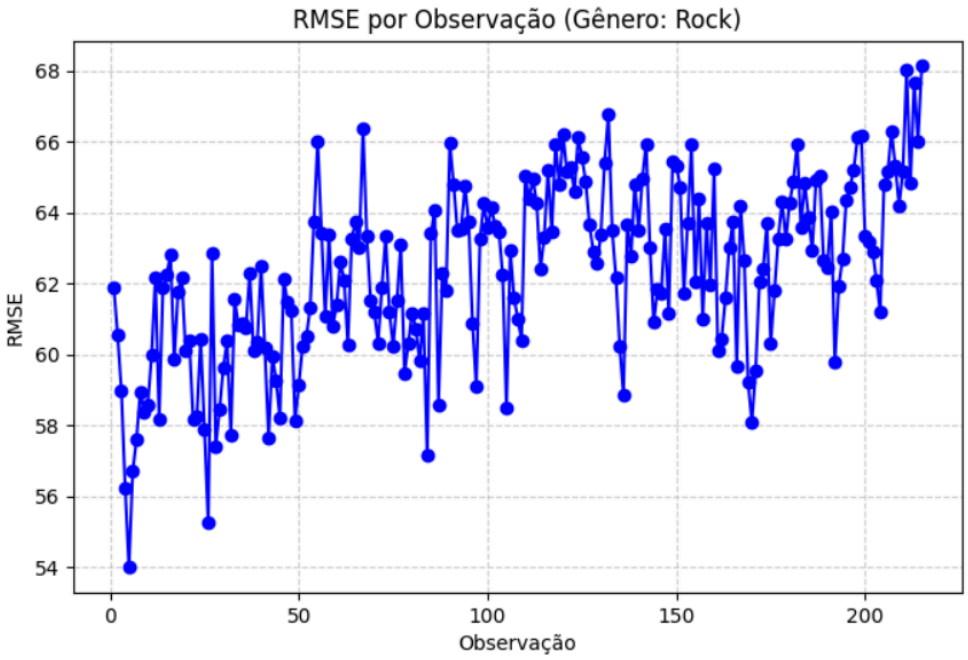
Além disso, a resolução (medida em taxa de bits (bitrate) e em número de bits por amostra) dos arquivos utilizados desempenha um papel significativo para o treino e teste do modelo. Considerando-se uma taxa de 43 observações por segundo, a predição de 215 unidades à frente do conhecido pelo modelo no momento de estimação (o equivalente a 5 segundos da duração do arquivo) e a estrutura recursiva de estimação, o erro proveniente de uma estimação no início do período predito é carregado e incrementado ao longo de todo período predito, de modo que uma revisão na taxa de observações por tempo é sugerida em desenvolvimentos futuros. Tal característica pode ser observada nas figuras 6, 7 e 8. Nas figuras 9 e 10, é possível observar os valores preditos e verdadeiros, bem como o erro do modelo em relação à frequência fundamental estimada pelo método YIN.

Figura 6 − RMSE conforme a passagem do tempo (em observações preditas)



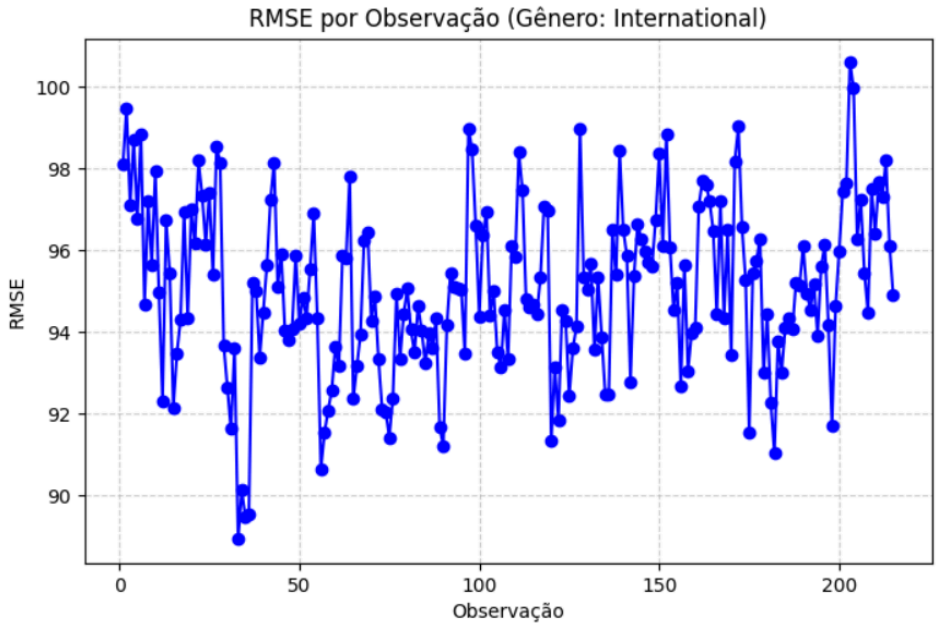

Fonte: elaborado pelo autor.

Figura 7 − RMSE conforme a passagem do tempo (em observações preditas) para o gênero “Rock”



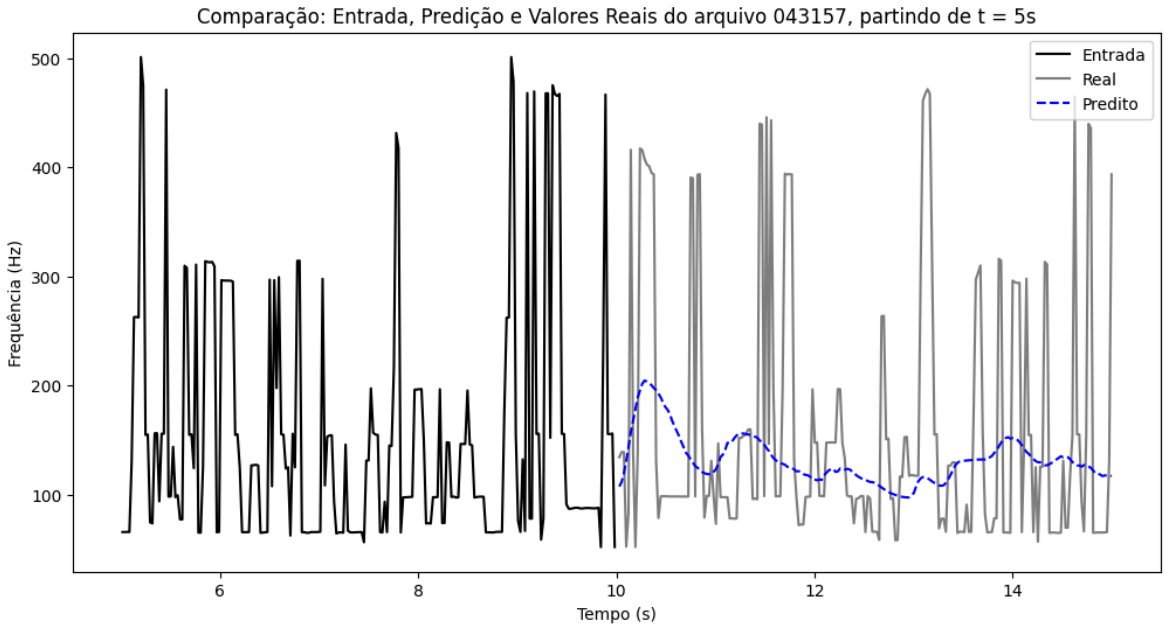

Fonte: elaborado pelo autor.

Figura 8 − RMSE conforme a passagem do tempo (em observações preditas) para o gênero “International”



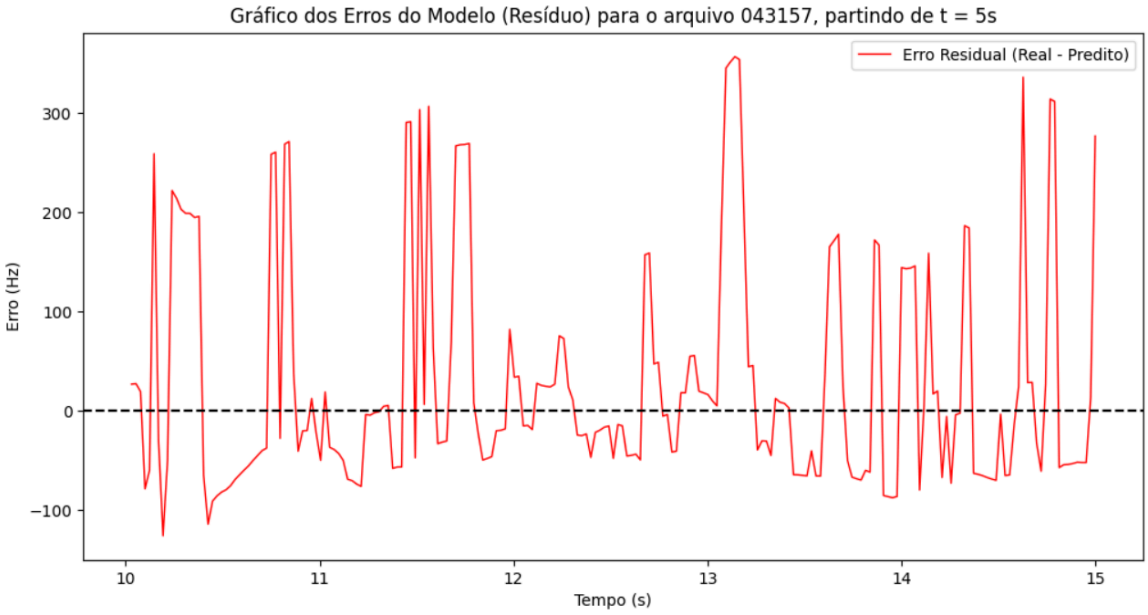

Fonte: elaborado pelo autor.

Figura 9 − Comparação: Entrada, Predição e Valores Reais do arquivo 043157, partindo de t = 5s




Fonte: elaborado pelo autor.

Figura 10 − Gráfico dos erros do modelo (Resíduo) para o arquivo 043157, partindo de t = 5s




Fonte: elaborado pelo autor.

Por outro lado, avaliando o trabalho a partir da ótica de eficiência computacional, foram obtidos resultados animadores em relação ao tempo empregado para a realização das predições. Para cada intervalo de 5 segundos estimados, o modelo levou em média 46 ms, utilizando um intervalo 100 vezes menor do que o proposto inicialmente, de 5 segundos. Considerando a metodologia de predição da biblioteca **keras** e a natureza dos dados inseridos para predição, é possível afirmar que a aplicação do modelo em larga escala produziria resultados utilizáveis pela lógica temporal, uma vez que as predições seriam realizadas e retornadas ao usuário antes que os valores reais se confirmassem na prática. No entanto, para futuras implementações em contextos cotidianos, sugere-se que o modelo seja executado em duas instâncias paralelas, com uma diferença de 5 segundos entre uma e outra, de modo que, enquanto uma instância exibe as predições para um intervalo de tempo de uma música sendo executada, essa mesma música, nesse mesmo intervalo de tempo esteja sendo utilizada para gerar as predições para os próximos 5 segundos. Dessa forma, as redes poderão trabalhar de maneira complementar ao longo do tempo.

Ainda, considerando a forma com que as músicas foram inseridas no modelo - enquanto observações de frequência ao invés de uma notação de acordes categórica - observa-se uma distinção entre uma mesma nota sendo executada em frequências diferentes entre oitavas. Por exemplo, o modelo pode prever a ocorrência de uma frequência fundamental de 440 Hz (Lá maior) em um determinado momento, enquanto a metodologia YIN (Cheveigné e Kawahara, 2002) determina a ocorrência de uma frequência fundamental de 220 Hz (Lá maior, uma oitava abaixo). Analisando esses resultados do ponto de vista puramente teórico, as notas previstas, de fato, são diferentes entre si. Entretanto, se partirmos de uma concepção mais prática, levando em consideração a natureza cíclica das oitavas, a execução de ambas as notas na sequência de um mesmo segmento musical resulta em segmentos equivalentes. Com isso em mente, uma outra possibilidade de desenvolvimento futuro a partir desse trabalho seria a utilização de marcadores qualitativos para a nota musical (como a notação em cifras) em conjunto com medidas de duração dessas ocorrências, de modo que a diferença entre oitavas seja desconsiderada no cálculo das métricas de avaliação.

Em termos das métricas avaliativas da performance do modelo, a configuração atual não se mostrou suficientemente adequada ao objetivo proposto se forem levadas em consideração as estatísticas descritivas calculadas a partir dos valores preditos. Contudo, dada a característica criativa e artística da música, em conjunto com os fatores elencados anteriormente nessa seção (gênero musical, resolução de arquivo, distinção entre oitavas e treinamento quantitativo da rede), é possível considerar que esse trabalho apresenta uma metodologia promissora, cujos resultados são compatíveis com a variação possível e esperada dentro de uma sequência musical harmônica.

Por fim, considerando a escassez de produção científica sobre o tema, seria oportuna a continuidade da pesquisa nessa área, utilizando o modelo aqui proposto e os resultados obtidos de modo a torná-lo mais assertivo ou mesmo para servir de base comparativa para os resultados de outros modelos que possam vir a ser propostos.

## 5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo propor uma metodologia para desenvolver modelos de predição musical envolvendo redes neurais recorrentes por meio de uma arquitetura simples, visando boa performance em termos de precisão das predições e de tempo computacional, utilizando arquivos de música a partir de uma concepção de série temporal de frequências fundamentais estimadas a partir de uma abordagem quantitativa do objeto de estudo.

Embora não sejam diretamente comparáveis a outros estudos nesta área do conhecimento ou correlatas por ainda não haver valores de referência estabelecidos para o contexto de predição musical via série de frequências fundamentais, os resultados, avaliados sob a ótica das medidas descritivas geradas, proporcionam uma estimação inicial da performance da metodologia empregada. Ainda, do ponto de vista de tempo de execução do algoritmo, esse trabalho permite um vislumbre positivo para pesquisas posteriores, uma vez que foi possível realizar predições em escalas de tempo 100 vezes menores do que o intervalo de predição proposto inicialmente.

Notou-se também que o gênero de música utilizada teve influência sobre as predições, assim como o formato e metadados dos arquivos e filtro utilizado nas manipulações de arquivos, além da natureza criativa e artística do processo de composição musical que por si só acrescenta uma variabilidade natural ao objeto de estudo.

Dessa forma, a implementação de modelos que considerem o gênero musical em suas predições, a manipulação de arquivos musicais seguindo diferentes metodologias de captação e tratamento dos dados de áudio e a proposição de novas arquiteturas e métodos de aprendizagem de máquina são sugeridos e encorajados para desenvolvimentos posteriores.

Finalmente, é importante destacar que o tema “predição musical” se encontra em estágio inicial de pesquisa, de forma que há possibilidade de muitas outras abordagens serem testadas e avaliadas. Sendo a música parte indissociável da experiência humana, enquanto houver humanidade haverá música, e havendo música, haverá a necessidade de ampliar os estudos que a definam e decodifiquem. Esse trabalho se

propõe a fornecer uma ferramenta ágil e eficiente para tal, e embora possa ser interpretado enquanto uma proposta inicial, fornece indícios de que a metodologia empregada apresenta grande potencial de desenvolvimento.

# REFERÊNCIAS


AGGARWAL, Charu. C. **Neural networks and deep learning**. Springer, 2018.

ALVARENGA, Luiz Gonzaga. **Tratado Geral Sobre o Som e a Música**. Clube de Autores, 2017.

ARAÚJO, Rodrigo Cenachi; FERREIRA, Flávia Magalhães; PEIXOTO, Zélia Myriam. Identificação da História de Compressão em sinais de Áudio nos Formatos WAV e MP3 utilizando o classificador Máquina de Vetor de Suporte. *In:* ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO 5.1. 2015. **Anais [...]**. p. 220-220.

ASHRAF, Mohsin *et al.* A hybrid CNN and RNN variant model for music classification. **Applied Sciences**, v. 13, n. 3, p. 1476, 2023.

CARVAJAL, Andrés. **MasterCard Stock Price Prediction Using LSTM and GRU**. 2023. Disponível em: https://github.com/andresca94/MasterCard-Stock-Price-Prediction-Using-LSTM-and-GRU Acesso em: 24 mar. 2025.

CHERLA, Srikanth. **Neural ProbabilisticModels for Melody Prediction, Sequence Labelling and Classification**. 2016. Tese de Doutorado. City, University of London.

COVACH, John; FLORY, Andrew. **What's That Sound?: An Introduction to Rock and its History**. 5th. W. W. Norton & Company, 2018.

DANTAS, Joseclécio Dutra; CRUZ, Sergio da Silva. Um olhar físico sobre a teoria musical. **Rev. Bras. Ensino Fís.** v. 41, n. 1, 2019.

DEFFERRARD, Michael; BENZI, Kirell; VANDERGHEYNST, Pierre; BRESSON, Xavier **FMA: A Dataset For Music Analysis**, arXiv preprint arXiv: 1612.01840, 2017.

CHEVEIGNÉ Alain; KAWAHARA Hideki. YIN, a fundamental frequency estimator for speech and music. **J Acoust Soc Am.** v. 111, n. 4, p. 1917-30. 2002.

DUA, Mohit et al. An improved RNN-LSTM based novel approach for sheet music generation. **Procedia Computer Science**, v. 171, p. 465-474, 2020.

FAN, Jianquing; MA, Cong; ZHONG, Yiqiao. A selective overview of deep learning. **Statistical science: a review journal of the Institute of Mathematical Statistics**. v. 36, n. 2, p. 264. 2021.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. MIT Press, 2016.

GUSMÃO, Pablo. Teoria elementar da música. Departamento de Música, Universidade Federal de Santa Maria Centro de Artes e Letras. Santa Maria, 2012. Apostila da disciplina de Teoria e Percepção Musical do primeiro semestre de 2012.

HAGAN, Martin T.; DEMUTH, Howard B; BEALE, Mark Hudson; JESÚS, Orlando. **Neural network design.**
Neural Networks in a Softcomputing Framework. 2014.

HAHN, Nicolas Mathias. **Composição Automática de Músicas utilizando Redes Neurais Recorrentes**. 2022. Trabalho de Conclusão de Curso (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2022.

HAIR, Joseph F.; BLACK, William C.; BABIN, Barry. J.; ANDERSON, Rolph E.; TATHAM, Ronald L. **Análise multivariada de dados**. Bookman Editora, 5th. 2005.

HAYKIN, Simon. **Neural networks and learning machine***s*. Pearson/Prentice Hall Toronto, 3th. 2009.

INGHAM, Tim. **Over 60,000 tracks are now uploaded to Spotify every day**. That's nearly one per second. Music Business Worldwide, 2021. Disponível em: www.musicbusinessworldwide.com/over-60000-tracks-are-now-uploaded-to-spotify-daily-thats-nearly-one-per-second/ Acesso em: 17 fev. 2025.

KAMATH, Uday; LIU, John; WHITAKER, James. **Deep learning for NLP and speech recognition**, v. 84. Springer, 2019.

LUO, Yi; YU, Jianwei. Music source separation with band-split RNN. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 31, p. 1893-1901, 2023.

MASCARENHAS, Mario. **Um vôo ao país da música**. Irmãos Vitale. 7 ed. 1968.

MCFEE, Brian, et al. librosa: Audio and music signal analysis in python." ***SciPy*** 2015 (2015) p. 18-24.

PEREIRA, Dimitri. **Quais são as diferenças entre os formatos de áudio WAV, MP3, AAC e FLAC?**. Disponível em: www.tecmundo.com.br/software/120043-diferencas-entre-formatos-audio-wav-mp3-aac-flac.htm Acesso em: 04 mar. 2025.

PRICE III, Emmett G. **Hip hop culture.** Bloomsbury Publishing USA, 2006.

ROJAS, Raul. **Neural networks: a systematic introduction**. Journal of The Springer Science & Business Media**,** 1996.

SIGTIA, Siddharth et al. An RNN-based music language model for improving automatic music transcription. **International Society for Music Information Retrieval,** 2014. p. 53-58.

SPOTIFY. About Spotify. **Resumo que destaca a essência e os principais objetivos do serviço**. c2025. Disponível em: https://newsroom.spotify.com/company-info/ Acesso em: 05 mar. 2025.

STÉRIN, Tristan; FARRUGIA, Nicolas; GRIPON, Vincent. **An intrinsic difference between vanilla RNNs and GRU models**. Cognitive**,** 2017. Ninth International Conference on Advanced Cognitive Technologies and Applications, Feb 2017, Athènes, Greece. pp.76 - 81.

WUENSCHE, Carlos Alexandre. **A Física da Música**. Itajubá. Trabalho de Conclusão de Curso (Curso de Física). 2006.

YU, Huiran. A Study of Statistical and Music-Theoretical Melody Prediction. 2022. Tese de Doutorado. Carnegie Mellon University Pittsburgh, PA.