INSTITUTO FEDERAL DO ESPÍRITO SANTO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

RAPHAEL MACEDO BERNARDINO

IDENTIFICAÇÃO DE TONALIDADE MUSICAL UTILIZANDO REDES NEURAIS ARTIFICIAIS

Cachoeiro de Itapemirim
2023

RAPHAEL MACEDO BERNARDINO

IDENTIFICAÇÃO DE TONALIDADE MUSICAL UTILIZANDO REDES NEURAIS ARTIFICIAIS

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Sistemas de Informação do Instituto Federal do Espírito Santo, Campus Cachoeiro de Itapemirim, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Ricardo Maroquio Bernardo

Cachoeiro de Itapemirim

2023

(Biblioteca do Campus Cachoeiro de Itapemirim)

B523i Bernardino , Raphael Macedo.

Identificação de tonalidade musical utilizando redes neurais artificiais / Raphael Macedo Bernardino . - 2023.
56 f. : il. ; 30 cm..

Orientador: Ricardo Maróquio Bernardo

TCC (Graduação) Instituto Federal do Espírito Santo, Campus Cachoeiro de Itapemirim, Sistemas de Informação, 2023.

1. Redes neurais (Computação). 2. Inteligência artificial. 3. Tonalidade (Música). I. Bernardo, Ricardo Maróquio. II.Título III. Instituto Federal do Espírito Santo.

CDD: 006.32

Bibliotecário/a: Renata Lorencini Rizzi CRB6-ES nº 085

MINISTÉRIO DA EDUCAÇÃO
INSTITUTO FEDERAL DO ESPÍRITO SANTO
CAI - COORDENADORIA DO CURSO DE BACHARELADO EM SISTEMAS DE INFORMACAO

FOLHA DE APROVAÇÃO-TCC Nº 20 / 2023 - CAI-CCSI (11.02.18.01.08.02.13)

Nº do Protocolo: 23151.004356/2023-99

Cachoeiro De Itapemirim-ES, 15 de dezembro de 2023.

RAPHAEL MACEDO BERNARDINO

IDENTIFICAÇÃO DE TONALIDADE MUSICAL UTILIZANDO REDES NEURAIS ARTIFICIAIS

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Sistemas de Informação do Instituto Federal do Espírito Santo, Campus Cachoeiro de Itapemirim, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Ricardo Maroquio Bernardo

Aprovado em 12 de dezembro de 2023

COMISSÃO EXAMINADORA

Prof. Dr. Ricardo Maroquio Bernardo
Instituto Federal Do Espírito Santo
Orientador

Prof. Dr. Rafael Silva Guimarães
Instituto Federal Do Espírito Santo

Prof. Dr. Lucas Poubel Timm do Carmo
Instituto Federal Do Espírito Santo

(Assinado digitalmente em 15/12/2023 16:33)
LUCAS POUBEL TIMM DO CARMO
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO
CAI-CCSI (11.02.18.01.08.02.13)
Matrícula: 2417426

(Assinado digitalmente em 15/12/2023 13:02)
RAFAEL SILVA GUIMARAES
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO
CAI-CCSI (11.02.18.01.08.02.13)
Matrícula: 1919203

(Assinado digitalmente em 15/12/2023 12:33)
RICARDO MAROQUIO BERNARDO
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO
CAI-CCSI (11.02.18.01.08.02.13)
Matrícula: 2152606

DECLARAÇÃO DO AUTOR

Declaro, para fins de pesquisa acadêmica, didática e técnico-científica, que este Trabalho de Conclusão de Curso pode ser parcialmente utilizado, desde que se faça referência à fonte e ao autor.

Cachoeiro de Itapemirim, 20 de Novembro de 2022.

Raphael Macedo Bernardino

Dedico esse trabalho primeiramente a Deus e a todos que de alguma forma contribuíram para que o mesmo fosse realizado.

AGRADECIMENTOS

Primeiramente agradeço a Deus que sem ele eu não conseguiria literalmente nem viver pois eu nasci sem respirar por 5 minutos. Agradeço minha família que me deu todo apoio necessário para que eu pudesse me formar.

"Música é arte de expressar os diversos afetos da nossa alma através do som"
Jô Bernardino

RESUMO

A análise e identificação de elementos musicais desempenham um papel fundamental na compreensão e classificação de músicas, sendo que a tonalidade musical é um aspecto importante da estrutura melódica e harmônica de uma música, influenciando sua sonoridade e expressividade. Neste cenário, o objetivo deste trabalho é explorar o uso de redes neurais para identificar a tonalidade musical de músicas que são formadas por uma determinada escala maior. Por meio da aplicação de técnicas de aprendizado de máquina, mais especificamente redes neurais, busco correlacionar dados musicais e seus respectivos tons para posteriormente treinar um modelo de rede neural para executar a tarefa de prever a tonalidade de músicas previamente nunca antes vista pelo modelo. Esta abordagem apresenta a vantagem de lidar com a complexidade e diversidade das características musicais, proporcionando uma análise mais precisa e automatizada. Embora existam diversos estudos e aplicações de redes neurais na área de processamento de áudio e música, a identificação da tonalidade musical por meio dessas técnicas ainda carece de uma quantidade significativa de referências e estudos específicos. Portanto, este trabalho se propõe a preencher essa lacuna explorando a eficácia das redes neurais nesse contexto. O modelo obtido conseguiu uma acurácia de 97,5% para um áudio não visto anteriormente na etapa de treinamento da rede neural. Além do modelo treinado, foi elaborado para o treinamento da rede neural o *dataset* que foi de grande importância e relevância para a concepção deste trabalho.

Palavras-chave: aprendizado de máquina, modulação, tonalidade musical

ABSTRACT

The analysis and identification of musical elements play a crucial role in understanding and categorizing music, with musical key being an important aspect of a song's melodic and harmonic structure, influencing its sound and expressiveness. In this context, the objective of this work is to explore the use of neural networks to identify the musical key of songs that are composed in a specific major scale. Through the application of machine learning techniques, specifically neural networks, I aim to correlate musical data and their respective keys to subsequently train a neural network model to predict the key of songs that have never been seen by the model before. This approach has the advantage of dealing with the complexity and diversity of musical features, providing a more precise and automated analysis. Although there are numerous studies and applications of neural networks in the field of audio processing and music, the identification of musical key through these techniques still lacks a significant amount of specific references and studies. Therefore, this work aims to fill this gap by exploring the effectiveness of neural networks in this context. The obtained model achieved an accuracy of 97.5% for audio that had not been seen previously during the neural network training phase.

Keywords: machine learning, modulation, musical key.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A percepção musical e a capacidade de discernir variações tonais podem variar significativamente de pessoa para pessoa. De acordo com Wise e Sloboda (2008), aproximadamente 17% dos adultos ocidentais se autodefinem como "surdos para tons", ou seja, têm dificuldade em discernir com precisão variações tonais musicais.

Em um estudo realizado com 40 estudantes dos cursos de música da UFSM (Universidade Federal de Santa Maria), uma parte considerável dos discentes relatou ter muitos impedimentos em relação à percepção musical. Esses impedimentos podem se manifestar de várias formas, como dificuldade em cantar no tom correto, identificar notas musicais ou até mesmo diferenciar diferentes tons (NASCIMENTO, 2020). Além disso, um iniciante em música possui dificuldades em assimilar os intervalos musicais de uma determinada música, consequêntemente não tendo um bom discernimento para identificação da tonalidade musical. É importante ressaltar que a habilidade musical é influenciada por uma combinação de fatores genéticos e ambientais, conforme afirma Andrade (2004) em sua abordagem: "A música, com certeza, é a forma de arte mais subjetiva e a que mais se presta à abstração de nossos sentimentos, à relação do homem com o sobrenatural e à nossa religiosidade". Algumas pessoas podem ter uma predisposição natural para a percepção musical, enquanto outras podem precisar de mais prática e treinamento para desenvolver suas habilidades musicais.

No entanto, mesmo que algumas pessoas tenham dificuldades iniciais, é possível melhorar a percepção musical por meio de treinamento e prática adequados. O cérebro é altamente adaptável e pode se especializar na codificação da tonalidade musical ao longo das escalas musicais (PERETZ, 2002). Peretz (2002) afirma que essa codificação da tonalidade musical é considerada um componente fundamental para a especialização do cérebro na música. A capacidade de discernir variações tonais e compreender a tonalidade de uma música é essencial para a percepção e a produção musical.

1.1 OBJETIVOS

Desenvolver um modelo de rede neural capaz de realizar a classificação da tonalidade musical para músicas que possuem a escala musical maior e não apresentam modulação. Para alcançar o objetivo geral, os seguintes objetivos específicos devem ser alcançados:

a) Criar um *dataset* rotulado composto por músicas e suas tonalidades musicais;

b) Criar a arquitetura de rede neural que será usada para classificar as amostras de áudio;

c) Realizar o pré-processamento dos arquivos de áudio, de forma a extrair *features* representativas e compatíveis com a entrada da rede neural criada;

d) Treinar a rede neural com o dataset devidamente preparado e ajustar os parâmetros da rede buscando máxima convergência;

1.2 JUSTIFICATIVA

A compreensão da tonalidade musical é uma habilidade crítica na formação de músicos e no desenvolvimento da percepção musical. Esta habilidade não apenas facilita a identificação e a criação de linhas melódicas improvisadas, mas também é fundamental para o entendimento mais profundo das estruturas musicais. Como destacado por Peretz (2002), a codificação do tom ao longo das escalas musicais é um componente essencial para a apreciação e produção musical efetiva. Além disso, considerando que uma parcela significativa de indivíduos, aproximadamente 17% dos adultos ocidentais, conforme citado por Wise e Sloboda (2008), se identifica como tendo dificuldades em discernir variações tonais – um desafio que também se reflete entre os estudantes de música, conforme indicado pela pesquisa de Nascimento (2020) –, ressalta-se a importância de se desenvolver métodos e técnicas eficazes para melhorar essa percepção. Este trabalho portanto, visa construir uma ferramenta que irá ajudar todas as pessoas que estão iniciando no mundo da música a identificar melhor a tonalidade

musical de forma que com a correta identificação, o estudante possa obter a escala musical de acordo com à tonalidade musical identificada.

2 REFERENCIAL TEÓRICO

Este capítulo oferece uma visão abrangente sobre a fusão da teoria musical e inteligência artificial, explorando como conceitos musicais como notas, escalas e tonalidades interagem com aprendizado de máquina e redes neurais. O foco está em como a teoria musical pode enriquecer a tecnologia de IA para análise e processamento musical. Essencial para estudantes e profissionais, o capítulo destaca a importância da interdisciplinaridade entre música e tecnologia, proporcionando uma perspectiva valiosa para a aplicação de IA na música.

2.1 TEORIA MUSICAL

A teoria musical é um conjunto de conhecimentos que visam compreender e explicar os elementos e princípios que regem a música, incluindo a notação musical, a harmonia, a melodia, o ritmo, a forma, entre outros aspectos.

Segundo Castro e Ribeiro (2018), a teoria musical é fundamental para o desenvolvimento da percepção auditiva e da compreensão musical. Os autores ressaltam que a teoria musical permite ao músico entender o que está tocando, compreender as relações entre os diferentes elementos da música e, assim, interpretar de forma mais consciente e precisa.

Sousa e Rodrigues (2019) abordam a importância da teoria musical no contexto acadêmico. Os autores destacam que a teoria musical deve ser ensinada de forma integrada com a prática musical, a fim de que o aluno possa aplicar o conhecimento teórico na execução musical e, assim, desenvolver sua técnica e expressividade.

Outro aspecto importante da teoria musical é a análise musical, que consiste em estudar a estrutura e os elementos de uma obra musical. A análise musical é essencial para a compreensão da linguagem musical e para o desenvolvimento da sensibilidade interpretativa do músico (SANTOS; SANT'ANNA, 2019).

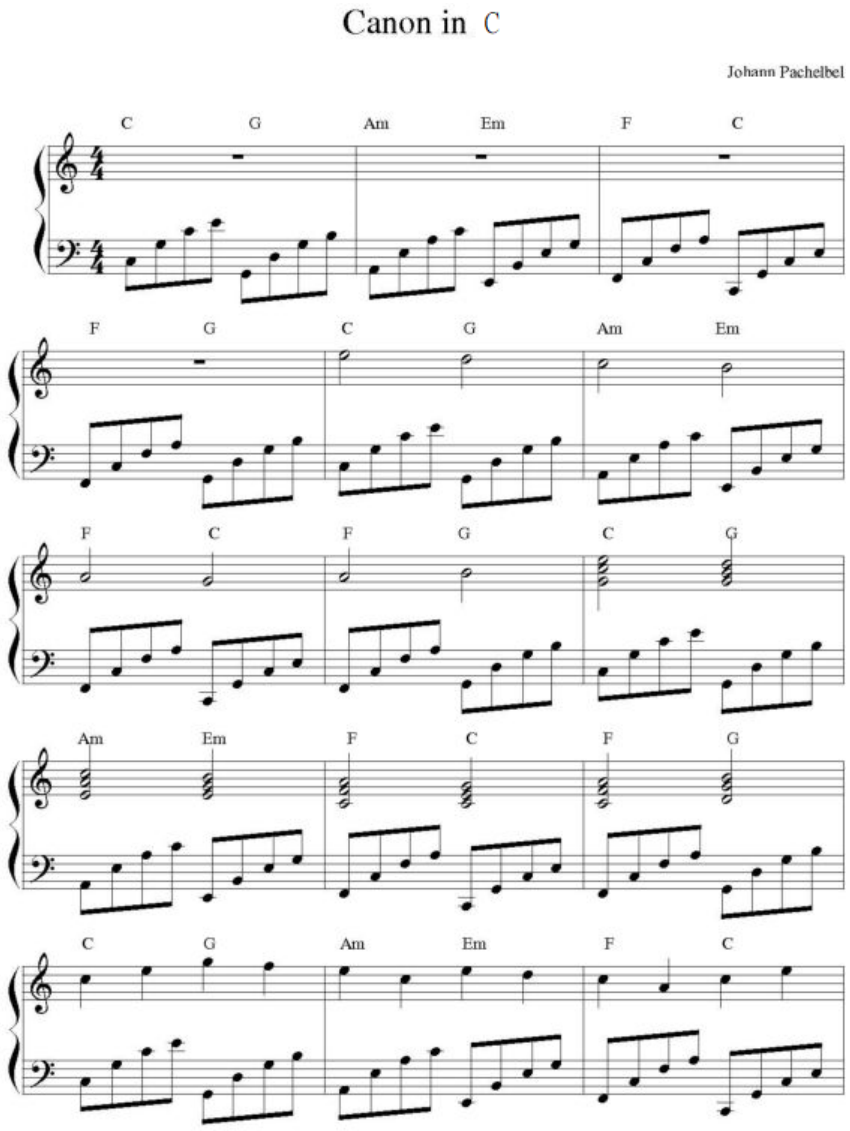
2.1.1 Nota musical

Em música, o termo nota, geralmente é usado para se referir a um símbolo musical quando se fala em representação de partitura ou em alguma sonoridade quando se fala em representação de áudio.Se tratando de notas musicais como símbolos musicais usados na notação musical ocidental, cada nota pode possuir vários atributos que determinam algumas características. Dentre as características de uma nota estão: duração relativa e a altura de uma nota a ser executado por um músico. No caso de um pianista por exemplo, a altura de uma nota se refere a qual tecla o pianista deve pressionar no piano, e a duração da nota indica a quantidade de tempo que essa tecla deverá ser pressionada (MÜLLER, 2015).

2.1.2 Representação musical

As músicas podem ser representadas de muitas maneiras e formatos diferentes. Como por exemplo, um compositor de músicas pode escrever uma composição na forma de uma partitura musical. Em uma partitura, há símbolos musicais que são usados para visualmente codificar as notas musicais e também símbolos que identificam a forma como essas notas devem ser executadas por um músico com um determinado instrumento musical (MÜLLER, 2015).

Figura 1 – Representação em partitúra musical de um trecho da música Canon em C de Johann Pachelbel



Fonte: Autor

Outro tipo de representação musical são as cifras musicais. As cifras musicais são uma notação simplificada usada para representar acordes em músicas. Elas são amplamente utilizadas em estilos populares, como o jazz, o rock, o pop e muitos outros. Embora a notação de cifra seja mais simplificada do que a notação musical tradicional (partituras), ela é eficaz para músicos que desejam acompanhar músicas e tocar acordes em instrumentos com a base harmônica como por exemplo teclado,contra baixo, violão, guitarra. O nome do acorde corresponde ao elemento principal de uma cifra e indica qual acorde deve ser executado por um músico em um determinado momento na música. Os acordes são representados por letras maiúsculas. Além dos

nomes, símbolos adicionais podem ser incluídos para indicar extensões ou alterações do acorde. Por exemplo, "Cmaj7"indica um acorde de Dó maior com a sétima maior (B) adicionada, enquanto "G7"indica um acorde de Sol com uma sétima menor (F) adicionada.

Figura 2 – Representação de cifras musicais naturais com o nome referente à cifra

C	- Dó maior	Cm	- Dó menor
C#	- Dó sustenido maior	C#m	- Dó sustenido menor
D	- Ré maior	Dm	- Ré menor
D#	- Ré sustenido maior	D#m	- Ré sustenido menor
E	- Mi maior	Em	- Mi menor
F	- Fa maior	Fm	- Fa menor
F#	- Fa sustenido maior	F#m	- Fa sustenido menor
G	- Sol maior	Gm	- Sol menor
G#	- Sol sustenido maior	G#m	- Sol sustenido menor
A	- Lá maior	Am	- Lá menor
A#	- Lá sustenito maior	A#m	- Lá sustenido menor
B	- Si maior	Bm	- Si menor

Fonte: Autor

2.1.3 Tonalidade musical

A tonalidade é um sistema central na música ocidental e é fundamental na compreensão e análise da maioria das obras musicais produzidas nos últimos quatro séculos. A tonalidade pode ser considerada como um sistema de tensão e resolução que cria expectativas e emoções específicas na audiência. Também permite que os compositores explorem uma ampla gama de possibilidades expressivas, desde a criação de harmonias ricas e complexas até a criação de efeitos emocionais sutis através do uso de modulações e variações rítmicas (SCHENKER, 1994).

Geralmente definida pela escala principal utilizada em uma composição, que é baseada em uma nota fundamental ou tônica. Essa nota é geralmente considerada como a mais estável da escala, e as outras notas da escala são organizadas hierarquicamente de acordo com sua relação de tensão e resolução em relação à tônica. Por exemplo, na escala de Dó Maior, a nota Dó é a tônica, e as outras notas da escala (Ré, Mi, Fá, Sol, Lá, Si) são organizadas de acordo com sua relação de tensão e resolução em relação à nota Dó. Alguns dos principais modos tonais incluem a escala maior e a escala menor harmônica e melódica. Além disso, existem tonalidades menores, como a tonalidade frígia e a tonalidade lídia, que têm uma sonoridade diferente da tonalidade maior e menor tradicionais (SCHENKER, 1994).

A tonalidade também é fundamental para a teoria musical, e muitos sistemas de análise musical são baseados em suas relações tonais. A teoria de Schenker, por exemplo, destaca a importância da estruturação hierárquica dos acordes e das linhas melódicas na música tonal. Outros sistemas de análise, como a análise harmônica funcional, enfatizam a relação entre acordes e as funções harmônicas que eles desempenham na tonalidade(SCHENKER, 1994).

2.1.4 Escala musical

Uma escala musical pode ser especificada pela divisão do espaço da oitava em um certo número de passos da escala. Existem diversos tipos de escala musical em que cada uma é determinada para um tipo de gênero musical. As escalas principais são respectivamente escala maior natural e escala menor natural (MÜLLER, 2015).

Figura 3 – Escala maior natural representada por cifras musicais

ESCALA MAIOR NATURAL							
GRAUS	I	II	III	IV	V	VI	VII
TOM	C	D	E	F	G	A	B
TOM	C#	D#	E#	F#	G#	A#	B#
TOM	D	E	F#	G	A	B	C#
TOM	Eb	F	G	Ab	Bb	C	D
TOM	E	F#	G#	A	B	C#	D#
TOM	F	G	A	Bb	C	D	E
TOM	F#	G#	A#	B	C#	D#	F
TOM	G	A	B	C	D	E	F#
TOM	Ab	Bb	C	Db	Eb	F	G
TOM	A	B	C#	D	E	F#	G#
TOM	Bb	C	D	Eb	F	G	A
TOM	B	C#	D#	E	F#	G#	A#

Fonte: Autor

Figura 4 – Escala menor natural representada por cifras musicais

ESCALA MENOR NATURAL							
GRAUS	I	II	III	IV	V	VI	VII
TOM	A	B	C	D	E	F	G
TOM	A#	B#	C#	D#	E#	F#	G#
TOM	B	C#	D	E	F#	G	A
TOM	C	D	Eb	F	G	Ab	Bb
TOM	C#	D#	E	F#	G#	A	B
TOM	D	E	F	G	A	Bb	C
TOM	D#	E#	F#	G#	A#	B	C#
TOM	E	F#	G	A	B	C	D
TOM	F	G	Ab	Bb	C	Db	Eb
TOM	F#	G#	A	B	C#	D	E
TOM	G	A	Bb	C	D	Eb	F
TOM	G#	A#	B	C#	D#	E	F#

Fonte: Autor

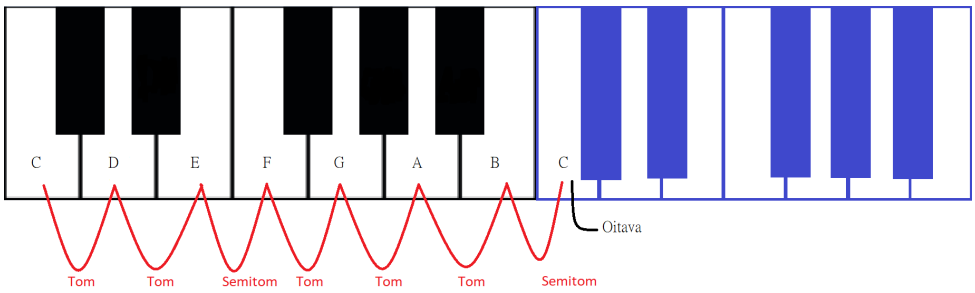
2.1.5 Intervalos musicais

As notas que formam uma determinada escala musical são chamadas de intervalos musicais. Os intervalos musicais são divididos em intervalos harmônicos e intervalos melódicos. Os intervalos harmônicos ou verticais dizem respeito à distância entre duas notas tocadas ao mesmo tempo. Por sua vez, os intervalos melódicos representam a distância entre duas notas musicais tocadas uma após a outra. Intervalos pequenos entre uma nota e outra são chamados de semitons, enquanto intervalos maiores são chamados de tons A junção entre determinados intervalos separados por tom e semitom, forma uma determinada escala. Escalas diatônicas possuem a característica de terem cinco intervalos de tom e dois de semitom distribuídos por uma oitava (ALDWELL; SCHACHTER; CADWALLADER, 2018).

2.1.6 Escala maior natural

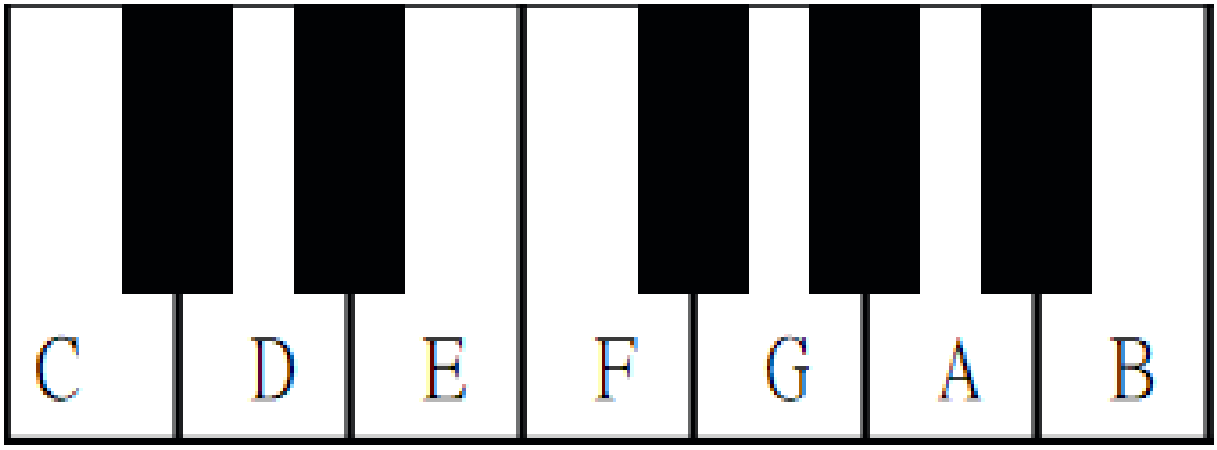
A escala maior representa uma variante das escalas diatônicas, que são caracterizadas por terem cinco intervalos de tom completo e dois semitons distribuídos ao longo de uma oitava. A localização desses semitons varia entre as diferentes formas de escalas diatônicas. A tradição da música clássica ocidental, desde o período dos gregos antigos até o século XIX, predominantemente fundamentou-se nestas escalas diatônicas (ALDWELL; SCHACHTER; CADWALLADER, 2018).

Figura 5 – Escala maior natural de dó representada por cifras musicais e a estrutura de intervalos da escala (tom e semitom)



Fonte: Autor

Figura 6 – Escala maior natural da tonalidade dó representada nas teclas de um teclado

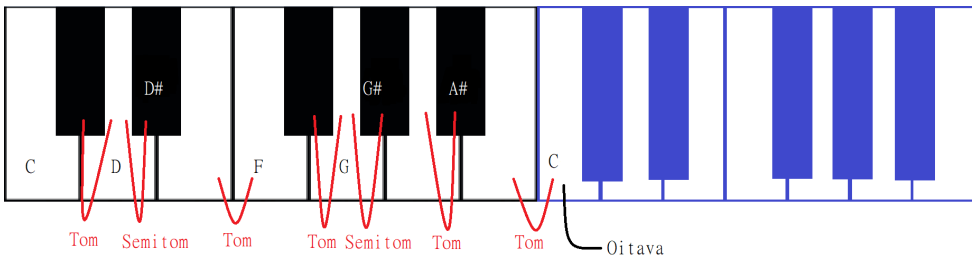


Fonte: Autor

2.1.7 Escala menor natural

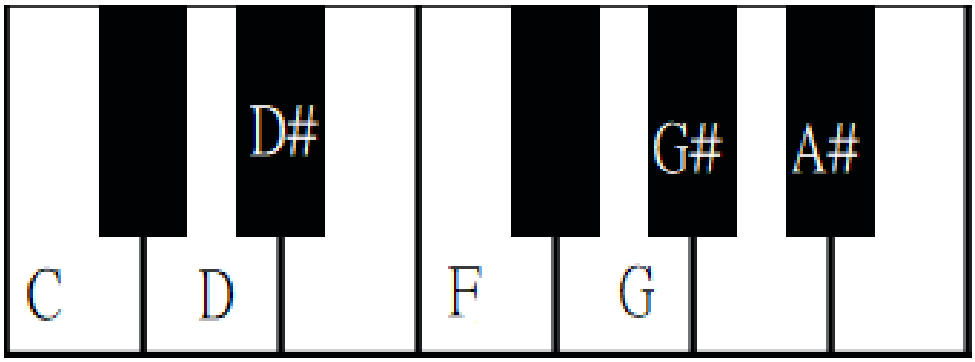
A escala menor por sua vez também é uma variante das escalas diatônicas, possuindo cinco intervalos de tom completo e dois semitons distribuídos ao longo de uma oitava. A maior diferença é a forma de distribuição dos intervalos da escala relativa à escala maior. Aldwell, Schachter e Cadwallader (2018)

Figura 7 – Escala menor natural de dó representada por cifras musicais e a estrutura de intervalos da escala (tom e semitom)



Fonte: Autor

Figura 8 – Escala menor natural da tonalidade dó representada nas teclas de um teclado

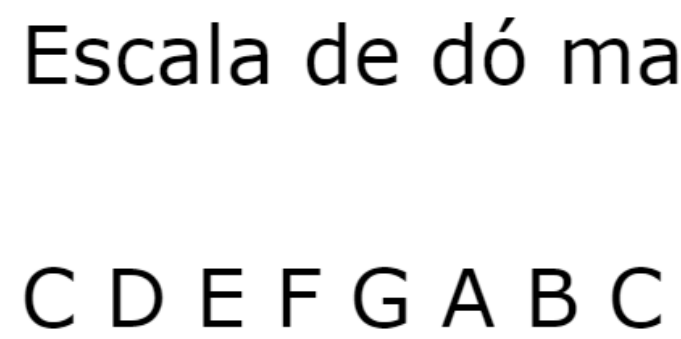


Fonte: Autor

2.1.8 Escala relativa

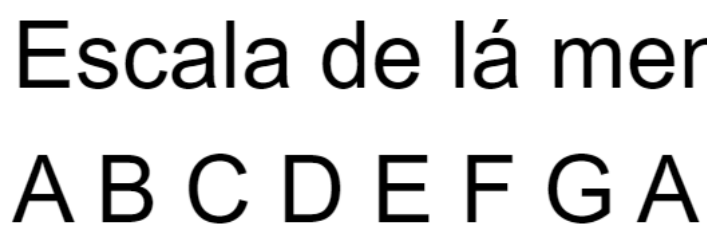
O termo relativo em música diz respeito à uma tonalidade maior com a mesma assinatura de uma determinada tonalidade menor e vice-versa. Exemplificando, dó maior seria a tonalidade relativa de lá menor, assim como lá menor é a tonalidade relativa de dó maior (ALDWELL; SCHACHTER; CADWALLADER, 2018).

Figura 9 – Representação da escala relativa de dó maior até a oitava



Fonte: Autor

Figura 10 – Representação da escala relativa de lá menor até a oitava



Fonte: Autor

2.1.9 Teoria da modulação musical

Modulação ou mudança de tonalidade, é o termo usado para identificar a mudança da tonalidade musical referente a um trecho musical ou alguma parte da música em questão, fazendo com que a música possua mais de uma tonalidade em um determinado período de tempo da própria música (MAZZOLA et al., 2016).

2.2 INTELIGÊNCIA ARTIFICIAL

Inteligência artificial é um campo de estudo que visando pesquisas e novos projetos de dispositivos computacionais capazes de simular o intelecto humano, no que diz respeito

ao modo de pensar, perceber, tomar decisões e na resolução de problemas (SILVA, 2013).

2.2.1 Aprendizagem de máquina

A aprendizagem de máquina é uma subárea do campo de estudo da inteligência artificial que tem como principal objetivo desenvolver diversas técnicas computacionais relacionadas ao aprendizado, consequentemente concebendo sistemas capazes de adquirir conhecimento sem interferência humana, ou seja, automáticamente. Em suma, um sistema de aprendizado é um programa de computador que é capaz de tomar decisões baseado em experiência acumuladas através da solução bem sucedida de outros problemas anteriores (MONARD; BARANAUSKAS, 2003).

2.2.2 Redes neurais artificiais

De acordo com Haykin (2001) uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples (neurônios artificiais) que possuem a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. O conhecimento da rede neural é feito a partir da arquitetura da rede através de um algoritmo que será usado para treinar a rede neural. Os pesos sinápticos são utilizados para armazenas o conhecimento adquirido através desse processo de aprendizagem.

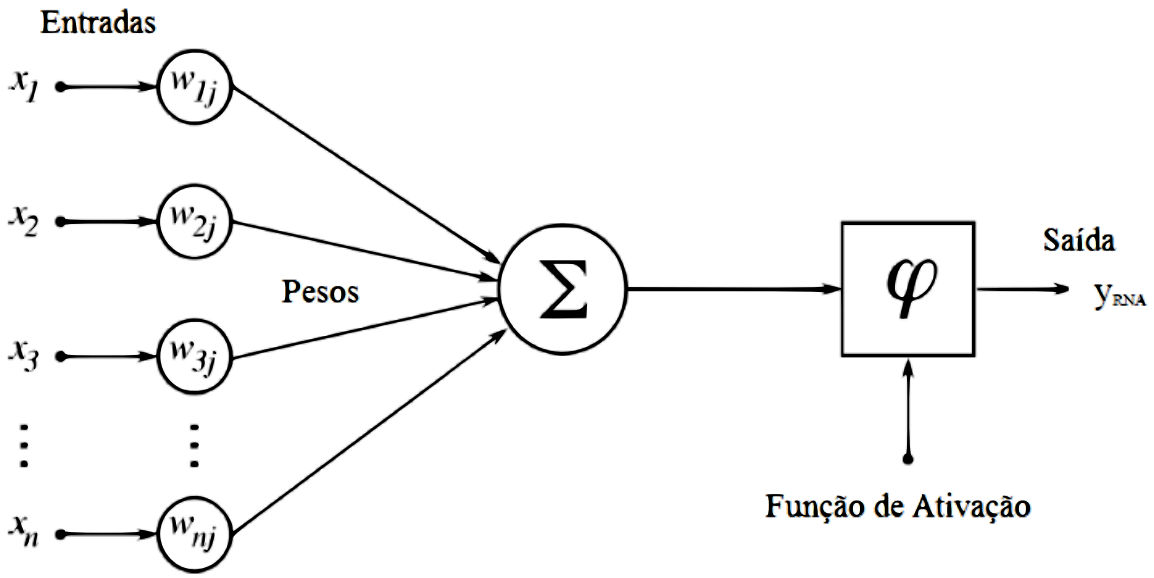
O neurônio artificial ou matemático, é uma simulação simplificada da estrutura e o funcionamento de um neurônio biológico. De forma concisa, um neurônio em uma rede neural artificial é um elemento que realiza a soma ponderada de múltiplas entradas posteriormente passa por uma função matemática e por fim encaminha o resultado para o próximo neurônio.

2.2.3 Funções de ativação

Existem muitas funções de ativação para decidir se um determinado neurônio irá ativar ou não. Funções de ativação são funções matemáticas. Uma das funções bastante utilizadas atualmente é a função ReLU.

A função ReLU ou unidade linear rectificada *(rectified linear unit)* é uma função não

Figura 11 – Representação do neurônio artificial



Fonte: Autor

linear, possibilitando muitas camadas de neurônios pois ela não ativa todos os neurônios ao mesmo tempo. Além disso, é muito importante promover a não linearidade para à rede neural, pois é crucial para a rede aprender padrões complexos nos dados. Se a entrada for um número negativo, ele será convertido em 0 ou seja, o neurônio que recebeu a entrada negativa não irá ser ativado. Se o valor for maior do que zero, a função trará como resultado da ativação o próprio valor recebido.

Fórmula da Função ReLU

$$f(x) = \max(0, x) \tag{1}$$

Outra função que é bastante útil quando principalmente há problemas que possuem determinadas classes, a função softmax tem como entrada um vetor de tamanho N e para cada valor em N, é retornado um valor entre 0 e 1. Se tratando da camada de saída de uma rede neural, quanto mais próximo de 0 for o valor computado pela função softmax, pode-se assumir que a resposta não será aquela determinada classe, porém, se a função softmax retornar um número próximo de 1, provavelmente o resultado é aquele determinado neurônio em que a função retornou o número próximo de 1 (GOODFELLOW; BENGIO; COURVILLE, 2023).

Fórmula da função softmax

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{2}$$

2.2.4 *Overfitting*

Overfitting ou sobreajuste, acontece em um modelo de rede neural quando ele consegue fazer previsões acertivas para os dados existentes que foram utilizados no treinamento, porém, o modelo não conseguirá generalizar para novos dados e novas situações que serão impostas. Para testar um modelo de rede neural é necessário testar a sua capacidade de fazer novas previsões para dados e situações que não foram expostas antes, ou seja, dados que não foram passados para o modelo treinar e efetivamente aprender os pesos. Uma das técnicas para evitar o sobreajuste dos pesos da rede neural, existem diversas técnicas. Uma delas é a técnica do *dropout* (GOODFELLOW; BENGIO; COURVILLE, 2023).

2.2.5 Dropout

Conforme Goodfellow, Bengio e Courville (2023) dropout é uma técnica de regularização utilizada em redes neurais. A técnica consiste em desativar aleatoriamente (e temporariamente) alguns neurônios ocultos da rede durante o treinamento. Isso é feito sem afetar os neurônios de entrada e saída. A lógica por trás do Dropout é que ao desativar aleatoriamente neurônios, a rede neural é forçada a aprender padrões mais robustos e menos dependentes de neurônios específicos, reduzindo assim a co-adaptação entre neurônios e o *overfitting*.

2.2.6 Redes neurais profundas

Aprendizagem profunda utiliza estruturas de neurônios matemáticos para processar dados, compreender a linguagem humana e identificar objetos visualmente. A informação é transmitida através de cada estrutura, onde a saída da estrutura anterior serve como entrada para a próxima. A primeira estrutura em uma rede é denominada camada de entrada, enquanto a última é identificada como camada de saída. Todas as estruturas intermediárias são conhecidas como camadas ocultas. Cada estrutura geralmente consiste em um algoritmo simples e uniforme que incorpora um tipo de função de

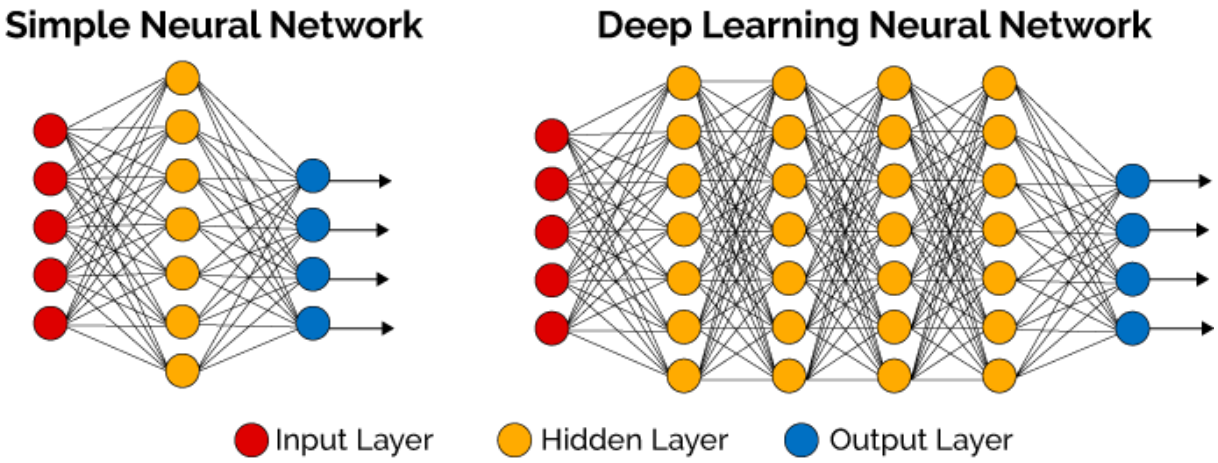


Figura 12 – Representação de uma rede neural simples e uma rede neural profunda

Fonte: Adaptado de Goodfellow, Bengio e Courville (2023)

ativação que determinada se aquele determinado neurônio irá ativar ou não. A maior diferença entre uma rede neural simples e uma rede neural profunda é a quantidade de camadas ocultas entre a camada de entrada e a camada de saída.

2.3 PROCESSAMENTO MUSICAL

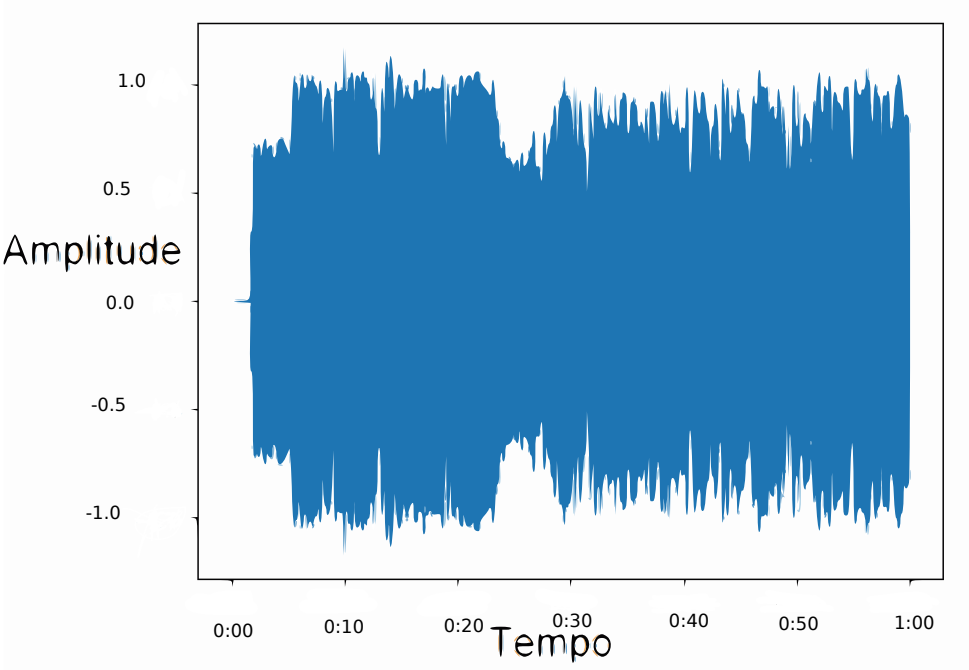
Processamento musical é uma área de pesquisa que tem por objetivo principal fazer contribuições com diversos conceitos, modelos, algoritmos, implementações e técnicas de avalição para lidar com problemas de análise e recuperação de elementos musicais em áudios (MÜLLER, 2015).

2.3.1 Som

O som é dado pela vibração nas moléculas de ar sendo deslocadas e osciladas causadas por um objeto vibrante como por exemplo as cordas vocais de um cantor, o canto de um pássaro, uma nota tocada por um pianista ou o barulho de uma britadeira sendo utilizada. Essas vibrações viajam pelo ar como uma onda sonora, desde a origem em que o som foi emitido até um ouvinte ou um microfone. No caso de um ouvinte, o ouvido capta a onda sonora e repassa para o tímpano, que por sua vez passa a vibrar de acordo com as oscilações de pressão. Após processamento adicional no ouvido médio e interno, a onda sonora é transformada em impulsos nervosos, que são finalmente enviados e interpretados pelo cérebro. É possível visualizar a mudança na pressão do ar em um determinado local por meio de um gráfico que representa a

forma da onda sonora (MÜLLER, 2015).

Figura 13 – Representação do gráfico da forma de onda da música Sorriso resplandescente



Fonte: Autor

2.3.2 Sinal analógico

Basicamente os sinais analógicos de áudio são os que ocorrem ao redor do mundo físico. São representados com valores reais e por conta disso, é possível modelar mudanças infinitamente pequenas tanto no tempo quanto na amplitude. Por se tratar de um conjunto dos números reais, os sinais analógicos são sinais de tempo contínuo Müller (2015).

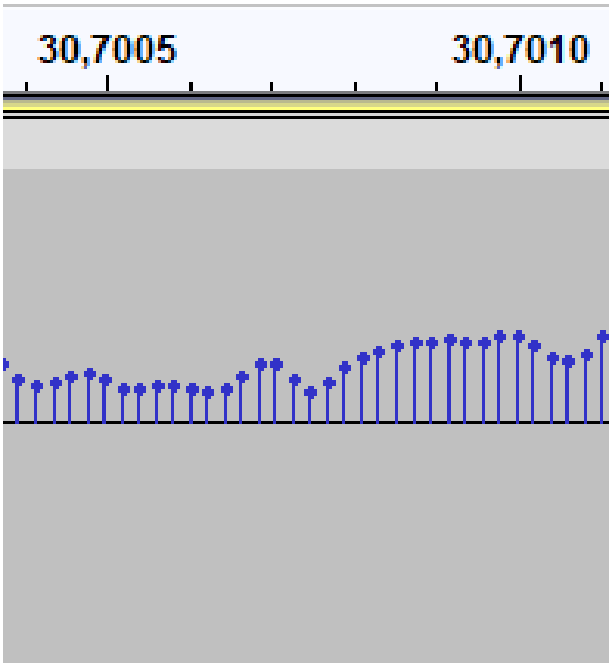
2.3.3 Sinal digital

As características dos sinais analógicos, por fazerem parte do conjunto numérico dos números reais, utilizam uma faixa contínua de valores em ambos os eixos do sinal de áudio (amplitude e tempo). Por conta disso, esse tipo de sinal leva a um número infinito de valores. Em contra partida, um computador não possui uma capacidade infinita de armazenamento, sendo impossível de guardar um sinal contínuo. Portanto, é necessário converter a forma de onda em alguma representação de números finitos, ou seja, uma representação discreta. Esse processo de conversão é chamado de digitalização. O processo para digitalizar um sinal analógico possui duas partes em seu cerne, nomeadas de amostragem e quantização Müller (2015).

2.3.4 Amostragem e quantização

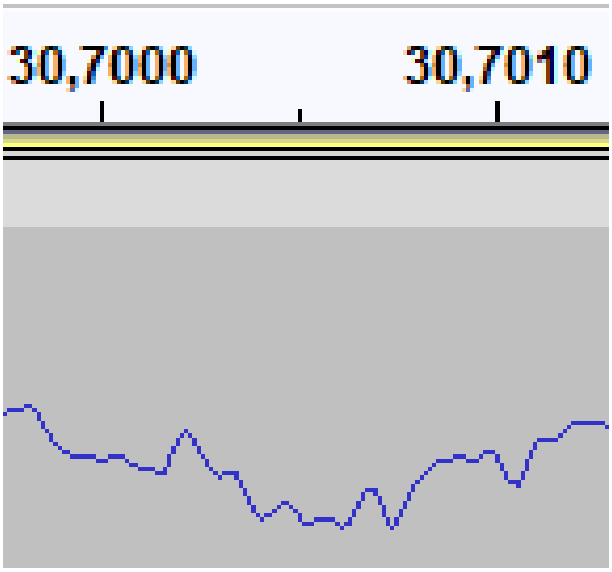
Amostragem é o processo para reduzir um sinal de tempo contínuo a um sinal de tempo discreto, com valores finitos para serem representados e computados em um computador. Quantização envolve a atribuição de um valor discreto a cada amostra do sinal, geralmente representado em um número finito. Em resumo, amostragem transforma um sinal contínuo em um sinal discreto no domínio do tempo, enquanto quantização limita os valores dessas amostras a um conjunto finito de valores discretos Müller (2015).

Figura 14 – Representação de amostragem e quantização de noventa milissegundos de um sinal de áudio (aproximado)



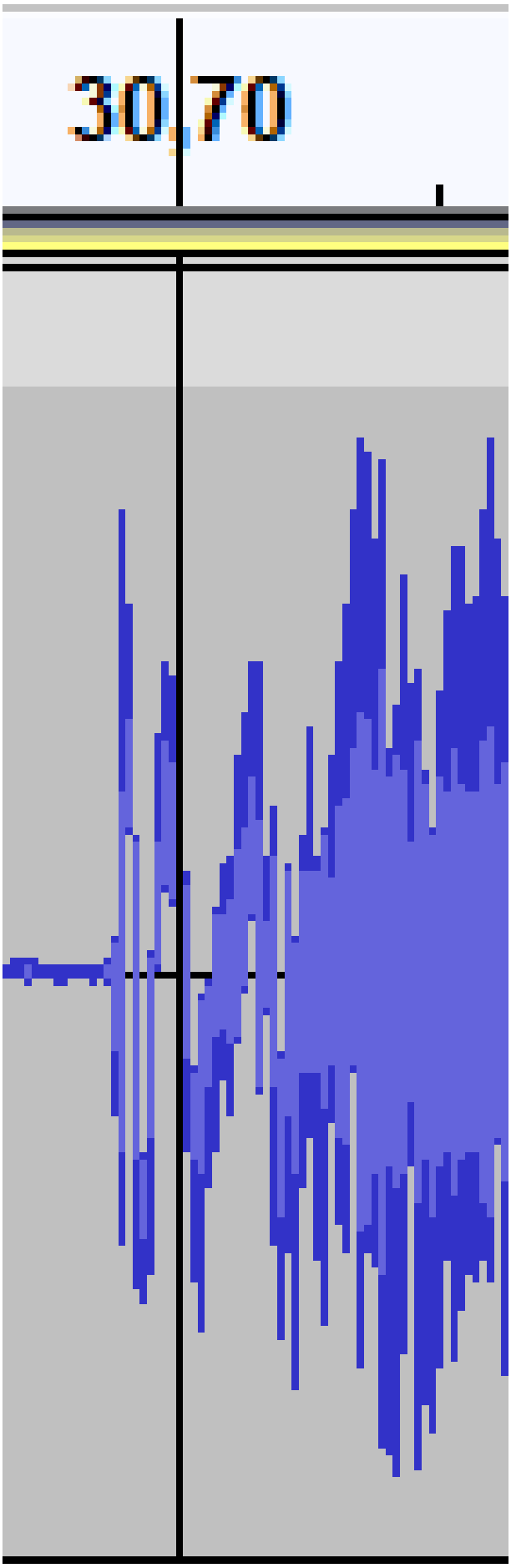
Fonte: Autor

Figura 15 – Representação de amostragem e quantização de noventa milissegundos de um sinal de áudio (média aproximação)



Fonte: Autor

Figura 16 – Representação de amostragem e quantização de noventa milissegundos de um sinal de áudio

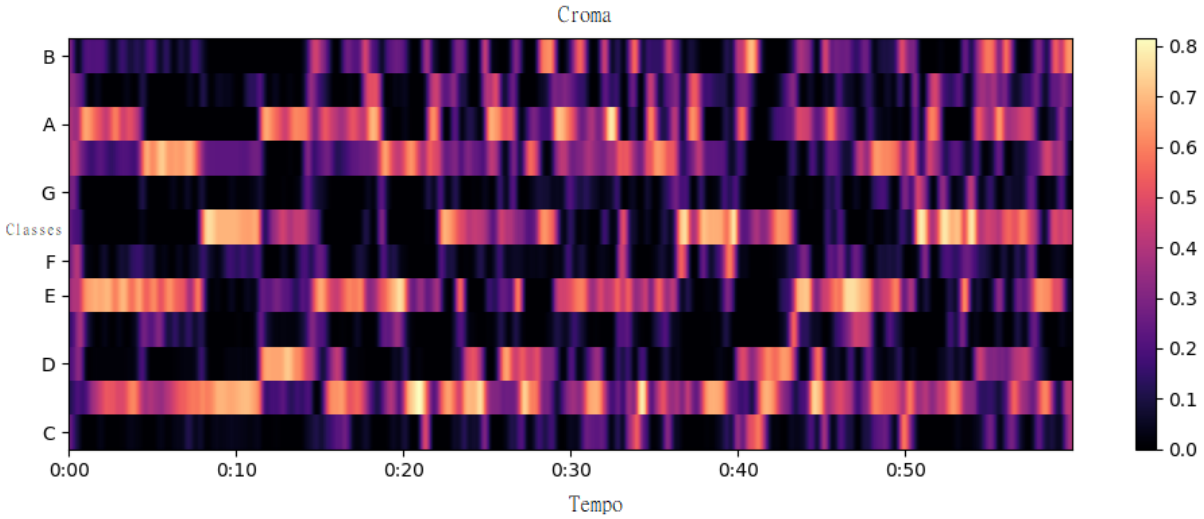


Fonte: Autor

2.3.5 Cromagrama

Os recursos de áudio baseados em croma estão intimamente ligados com o aspecto de harmonia musical. Além disso, é uma das ferramentas mais estável e estabelecida no processamento e análise de dados musicais. Há diversas formas de calcular e aprimorar recursos de croma, que por consequência, resulta em um grande número de variantes de croma com propriedades diferentes. Não existe uma variante única de croma que funcione melhor em todas as aplicações de processamento e análise de dados musicais. A percepção humana de *pitch* é periódica, no sentido de que dois *pitchs* são percebidos similar a distinção de cor se diferindo em uma oitava. De acordo com Shepard (1964), um *pitch* se difere em dois componentes, que são referidos como altura de tom e croma. Os cromas da escala musical ocidental fundamental maior corresponde ao conjunto {C, C# D, D# E, F, F# G, G# A, A# B} (respectivamente Dó, Dó sustenido, Ré, Ré sustenido, Mi, Fá, Fá sustenido, Sol, Sol sustenido, Lá, Lá sustenido, Si). Sendo assim, pode-se correlacionar o vetor de cromas com as determinadas classes de croma. Por exemplo, para os cromas da classe C, se assume que o vetor tem índice zero, para C, um, ... até o índice onze. Um cromagrama é a representação gráfica de um determinado áudio. No eixo x é dado o tempo, no eixo y a qual classe pertence cada amostra de áudio. Quanto mais branco a faixa de amostra, mais intensidade possui aquela amostra em relação a uma determinada classe de croma (EWERT, 2011).

Figura 17 – Representação de um cromagrama de uma música em Lá maior



Fonte: Autor

3 TRABALHOS RELACIONADOS

Neste capítulo será abordado dois trabalhos relacionados com alguns conceitos abordados no capítulo anterior, mais especificamente o conceito de *pitch* ou tonalidade e o conceito de *deep learning*.

3.1 *TRADITIONAL MACHINE LEARNING FOR PITCH DETECTION*

A detecção de frequência fundamental é uma tarefa importante no processamento digital de sinais e análise de música, sendo amplamente realizada por algoritmos de aprendizado de máquina. Esses algoritmos utilizam técnicas de extração de características do sinal sonoro e algoritmos de classificação para determinar o tom de um determinado sinal sonoro. Embora eficazes em diferentes contextos, ainda podem apresentar limitações em situações de ruído ou variação nas características do sinal sonoro.

Drugman et al. (2018) propõe um algoritmo de detecção de tom baseado em recursos de engenharia e Machine Learning tradicional, com o objetivo de aumentar a precisão o máximo possível em gravações limpas, pois o caso de uso final é a síntese de fala de alta qualidade.

3.2 *ERROR CORRECTION IN PITCH DETECTION USING A DEEP LEARNING BASED CLASSIFICATION*

Khadem-Hosseini et al. (2020) aborda a aplicação de técnicas de aprendizado profundo (deep learning) para corrigir erros na detecção de altura (*pitch*) em sinais de áudio. A detecção precisa da altura é crucial em muitas aplicações musicais e de processamento de áudio. O estudo propõe um modelo de classificação baseado em redes neurais profundas para identificar e corrigir erros de detecção de altura. O modelo é treinado em um conjunto de dados que contém exemplos de sinais de áudio com as alturas corretas e as alturas erroneamente detectadas.

3.3 *AUDIO-BASED MACHINE LEARNING MODEL FOR TRAFFIC CONGESTION DETECTION*

Rong (2016) aborda a avaliação inteligente do tráfego e a detecção de congestionamentos usando sensores de som e aprendizado de máquina. Ele foca em dois problemas principais: a avaliação das condições de tráfego a partir de dados de áudio e a análise de áudio em ambientes não controlados. Utilizando a modelagem dos parâmetros de tráfego e a geração de som pelos veículos em movimento, o estudo explora o uso do áudio produzido como fonte de dados para aprender padrões de áudio de tráfego. Uma solução é proposta para lidar com o tempo, o custo e as limitações inerentes ao monitoramento do tráfego. Fontes de ruído externas foram introduzidas para criar cenas acústicas mais realistas e testar a robustez dos métodos apresentados. O monitoramento baseado em áudio é proposto como uma opção simples e de baixo custo, comparável a outros métodos baseados em laços de detecção ou GPS, e tão eficaz quanto soluções baseadas em câmeras, evitando problemas comuns do monitoramento baseado em imagem, como oclusões e condições de iluminação. A abordagem é avaliada com dados de análise de áudio do tráfego registrados em locais ao redor da cidade de São José dos Campos, Brasil, e arquivos de áudio de locais ao redor do mundo, baixados do YouTube. A validação demonstra a viabilidade do monitoramento automático de áudio do tráfego e o uso de algoritmos de aprendizado de máquina para reconhecer padrões de áudio em ambientes ruidosos.

3.4 *EXPLORING MACHINE LEARNING FOR AUDIO-BASED RESPIRATORY CONDITION SCREENING: A CONCISE REVIEW OF DATABASES, METHODS, AND OPEN ISSUES*

Xia, Han e Mascolo (2022) faz uma revisão narrativa focada no papel crucial da ausculta na clínica médica e como a comunidade científica vem explorando o aprendizado de máquina (Machine Learning, ML) para possibilitar a ausculta remota e automática para triagem de condições respiratórias através de sons. O objetivo principal é oferecer uma visão geral dos desenvolvimentos recentes nesse campo. Nesta revisão, são descritas bases de dados de áudio publicamente disponíveis que podem ser usadas em experimentos, ilustram-se os métodos de ML desenvolvidos até o momento, e destacam-se algumas questões subconsideradas que ainda necessitam de atenção.

Em comparação com pesquisas existentes sobre o tema, esta revisão abrange a literatura mais recente, especialmente os estudos de detecção de COVID-19 baseados em áudio, que receberam atenção considerável nos últimos dois anos. O trabalho tem o intuito de facilitar a aplicação da inteligência artificial no campo da ausculta respiratória, fornecendo uma compreensão abrangente dos avanços e desafios atuais na integração de tecnologias de ML na prática da ausculta.

3.5 *TOWARD AUDIO BEEHIVE MONITORING: DEEP LEARNING VS. STANDARD MACHINE LEARNING IN CLASSIFYING BEEHIVE AUDIO SAMPLES*

Kulyukin, Mukherjee e Amlathe (2018) se concentra em monitoramento eletrônico de colmeias usando técnicas de aprendizado de máquina e aprendizado profundo. O objetivo é extrair informações críticas sobre o comportamento e a fenologia das colônias de abelhas sem a necessidade de inspeções invasivas e os custos de transporte associados.

Os autores projetaram várias redes neurais convolucionais e compararam o desempenho delas com quatro métodos padrões de aprendizado de máquina (regressão logística, vizinhos mais próximos, máquinas de vetores de suporte e florestas aleatórias) na classificação de amostras de áudio obtidas de microfones posicionados acima das plataformas de pouso de colmeias Langstroth. Foram utilizadas duas bases de dados: a primeira com 10.260 amostras de áudio e a segunda, mais desafiadora, com 12.914 amostras, diferenciadas por colmeia, localização, tempo e raça de abelhas.

Os resultados mostraram que, na primeira base de dados, uma rede neural convolucional mais simples para áudio bruto com uma camada personalizada teve um desempenho superior a três redes mais complexas sem camadas personalizadas e foi comparável aos quatro métodos de aprendizado de máquina. Na base de dados mais desafiadora, todas as redes neurais convolucionais para áudio bruto superaram os métodos de aprendizado de máquina padrão e uma rede treinada para classificar imagens de espectrograma de amostras de áudio.

3.6 *A SURVEY ON MACHINE LEARNING TECHNIQUES FOR AUTO LABELING OF VIDEO, AUDIO, AND TEXT DATA*

Zhang, Jafari e Nagarkar (2021) aborda como o aprendizado de máquina vem sendo utilizado em diversos domínios, como classificação, detecção de objetos, segmentação de imagens e análise de linguagem natural. Um ponto central no aprendizado de máquina é a rotulagem de dados, essencial para treinar modelos eficazes. No entanto, rotular grandes volumes de dados pode ser custoso do ponto de vista financeiro.

Devido a isso, pesquisadores têm se concentrado em reduzir os custos de anotação e rotulagem de dados. Uma abordagem eficiente para mitigar o impacto de dados limitados e reduzir os custos de preparação de dados é o uso de transferência de aprendizado. Esta técnica permite transferir conhecimento prévio de um domínio fonte para um domínio alvo, reduzindo a quantidade de dados necessários neste último. Contudo, grandes volumes de dados anotados ainda são necessários para construir modelos robustos e melhorar a precisão das previsões.

Por isso, tem crescido o interesse por métodos automáticos de anotação e rotulagem de dados. Neste artigo de revisão, os autores fornecem uma análise das técnicas anteriores focadas na otimização da anotação e rotulagem de dados para vídeo, áudio e texto. O objetivo é oferecer uma visão abrangente das abordagens atuais e potenciais melhorias nessa área essencial do aprendizado de máquina.

4 METODOLOGIA

Neste capítulo irá ser abordado a metodologia para de fato alcançar os objetivos deste trabalho.

4.1 TECNOLOGIAS

4.1.1 Linguagem de programação Python

Python é uma linguagem de programação criada em 1991 por Guido van Rossum. A linguagem possui aplicação em diversas áreas como por exemplo desenvolvimento *web*, desenvolvimento de sistemas de interface para usuários, matemática, sistema que como base *scripts*, desenvolvimento para redes de computadores e muitas outras diversas aplicações. Python funciona em diferentes sistemas operacionais existentes atualmente como por exemplo *Windows*, *Mac*, *Linux*, *Raspberry Pi*, entre outros (ROSSUM; DRAKE, 2009).

4.1.2 Librosa

Librosa é um módulo python para realizar análise e processamento musical em áudio. Ele fornece uma gama de ferramentas computacionais para construir sistemas que possam recuperar informações musicais. (MCFEE et al., 2023)

4.1.3 Numpy

Numpy é um dos módulos fundamentais para computação científica em Python. A biblioteca fornece um conjunto de vetores multidimensional com variações e uma incrível gama de ferramentas computacionais para operações rápidas em vetores e matrizes, incluindo matematica, lógica, manipulação de *shape*, classificação, seleção, operações para computar entradas e saídas, transformadas discretas de Fourier, operações básicas sobre algebra linear, e muito mais. (HARRIS et al., 2020)

4.1.4 Tensorflow e Keras API

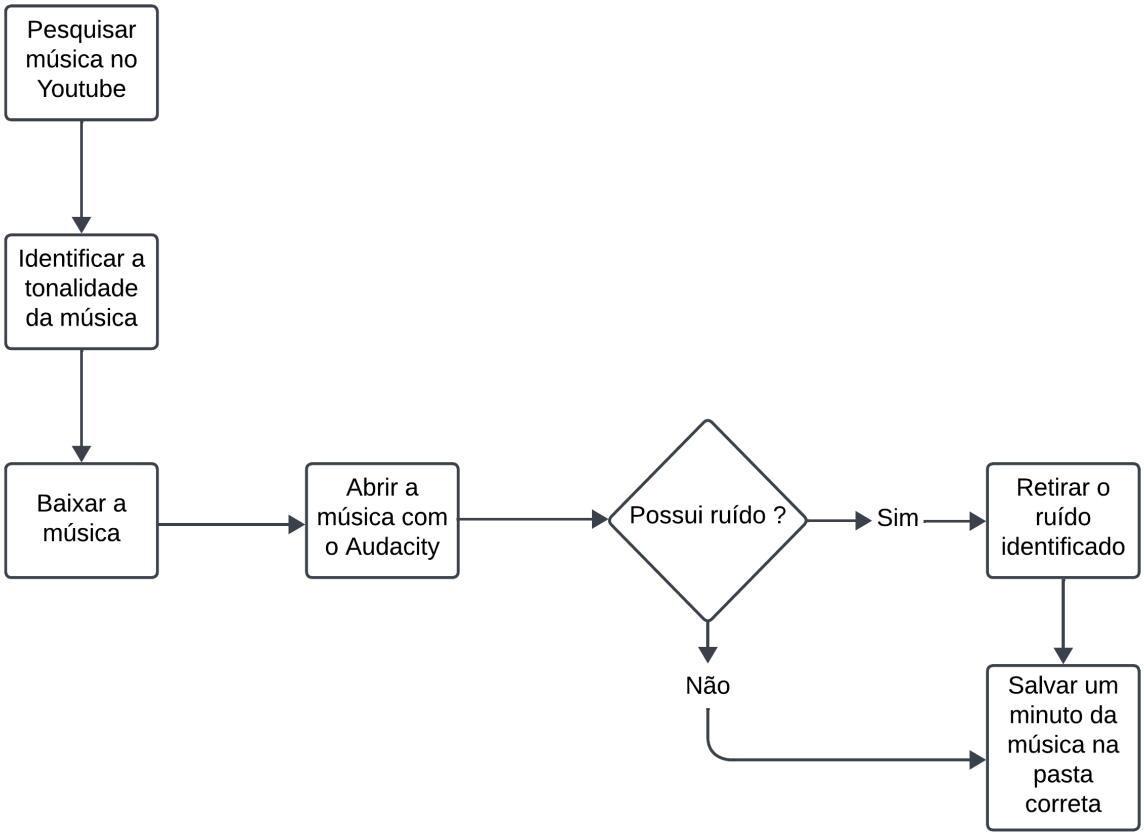
O módulo TensorFlow é uma biblioteca de software abrangente para computação numérica, usando gráficos de fluxo de dados. É amplamente utilizado para tarefas de

aprendizado de máquina e redes neurais profundas porque permite a construção de modelos complexos de forma eficiente e escalável. Integrado a biblioteca TenforFlow , Keras simplifica a criação de redes neurais profundas. Ele oferece uma API mais amigável e de mais alto nível, permitindo que os usuários construam e treinem modelos com menos código e complexidade. (ABADI et al., 2015) (CHOLLET et al., 2015)

4.1.5 *Dataset*

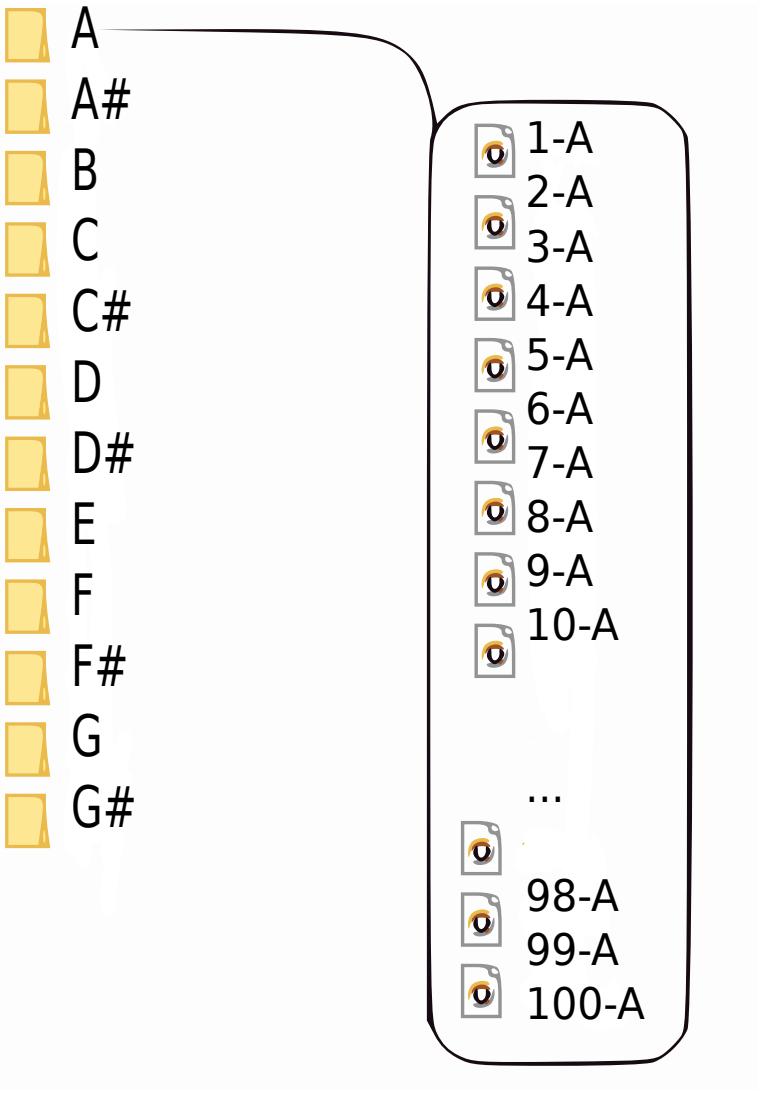
Para encontrar o conjunto de músicas utilizadas para construção do *dataset* foi utilizado um canal da tecnologia *YouTube* aonde foi encontrada diversas *playlists* em que cada *playlist*, haviam músicas rotuladas no título da playlist. (Jonathan Cruz Garcia, 2023) Foi necessário baixar todas as playlists em que no título estavam as tonalidades fundamentais ocidentais e além disso, foi necessário converter os vídeos para um formato de áudio. O conjunto de dados criado para efetivamente treinar, validar e testar a rede neural foi baseado no artigo de Tzanetakis e Cook (2002). O *dataset* criado contém 1200 arquivos de áudio separados por classes de croma, ou seja, 100 arquivos para a classe C, 100 para classe C# e assim por diante, até contabilizar as 12 classes de croma. Além disso, para cada música, foi extraído as classes de croma por meio da biblioteca *librosa* que efetivamente são os dados serviram de entrada para o modelo de rede neural. Primeiramente foi feito uma etapa de reconhecimento das músicas para rotular as músicas de acordo com cada classe fundamental de croma. Após o mapeamento, foi feito o *download* de cada arquivo de áudio. Após essa etapa, foi feito o pré-processamento manualmente dos dados, para evitar padrões de aprendizagem indesejáveis no futuro modelo treinado. O *dataset* foi nomeado e rotulado em cada arquivo com as respectivas tonalidade. Com base no *dataset* de Tzanetakis e Cook (2002), foi separado um minuto de cada arquivo de áudio para obter um tamanho padronizado. Foi utilizado o *software Audacity* que é um *software* livre para edição digital de áudio, para realização do pré-processamento dos dados da base de dados.

Figura 18 – Fluxograma do processo de construção do *dataset*



Fonte: Autor

Figura 19 – Resumo de uma das pastas do *dataset*



Fonte: Autor

Para cada amostra de áudio, foi extraído as classes de cromagrama totalizando trinta e um mil e oito *features* que são efetivamente as características em probabilidade de cada classe de croma. Todas as amostras de áudio em conjunto, foram salvas em um arquivo *numpy*, juntamente com os rótulos de cada arquivo.

4.1.6 Pré-processamento do *dataset*

Primeiramente é carregado o caminho de todos os arquivos existentes no *dataset* e salvos em uma lista. Após o carregamento dos dados, é feito o processo de randomização para que a rede não aprenda padrões de ordem nos arquivos de áudio. O rótulo de cada arquivo de áudio está associado ao diretório em que o arquivo se

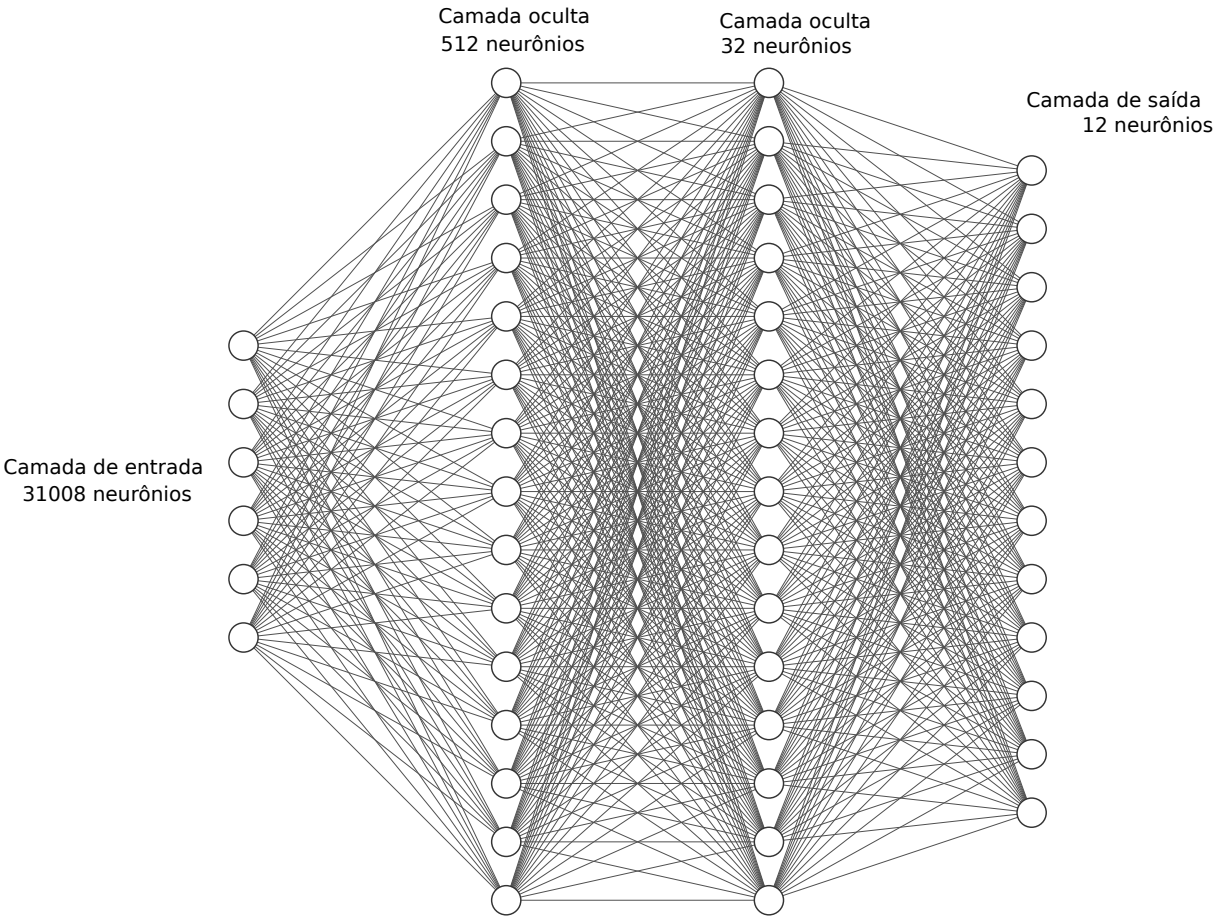
encontra. Por conta disso, a extração dos rótulos é feito a partir do diretório em cada localização do arquivo e posteriormente, é salvo os arquivos em uma lista chamada tonalidade_de_cada_arquivo. Após os rótulos serem extraídos para cada arquivo , foi criado um *dataframe* para facilitar a manipulação dos dados para passar para função que irá extrair as classes de croma para cada arquivo de áudio. Em cada arquivo de áudio do *dataset* criado, foi necessário separar a melhor faixa de áudio em que exista elementos relevantes de cromas. Após feito o pré-processamento manual, foi extraído os recursos de croma necessários para criar os dados de treinamento e seus respectivos rótulos. Para cada arquivo de áudio foi extraído as classes de croma Primeiramente é carregado o sinal de áudio com a função *librosa.load* da biblioteca librosa. Essa função retorna os dados de áudio e a taxa de amostras por segundo. Ambos foram salvos respectivamente em duas variáveis: data e *sample_rate*. Após é extraído as classes de croma do arquivo passado através da função*librosa.feature.chroma_cens* da biblioteca librosa que é uma variante de croma normalizado para melhorar a comparação entre as classes de croma. Por fim é retornado as classes de croma computadas pela função. Posteriormente o formato dos dados são padronizados de forma que todos tenham a mesma forma para serem adequados ao modelo de rede neural criado. Primeiro é encontrado o vetor com o maior formato (linhas,colunas). Após, todos os vetores com a forma inferior ao maior formato é preenchido com zeros de forma que todos os vetores após essa etapa, tenham o mesmo formato (12,2584). Após o preenchimento, é criado as variáveis X e y que são respectivamente os dados de entrada e os rótulos para cada dado de entrada. Após a normalização, as saídas rotuladas foram convertidas para adequar na configuração da rede neural.

4.1.7 O modelo

Foi utilizado a arquitetura MLP (*Multi Layer Perceptron*) para compor a estrutura da rede neural. Cada arquivo de música foi necessário aplicar um padding para que todo dataset ficasse com o mesmo tamanho sendo possível analisar os dados com a mesma forma (*shape*) que mais especificamente, cada áudio ficou com o formato de (12,2584), sendo que o número 12 representa cada classe de croma e o número 2584 representa as características de croma em cada classe de croma, extraídos pela função *chroma_cens* da biblioteca librosa. Para os valores da construção da rede neural, foram feitos alguns testes para chegar em uma melhor configuração para se obter alta taxa de

convergência. O modelo possui cinco camadas distribuídas respectivamente em uma camada de entrada com 31.008 neurônios, uma camada densa com 512 neurônios, uma camada de *dropout* com 20%, uma camada densa com 32 neurônios e uma camada de saída com 12 neurônios. Foi utilizada a arquitetura *Feed Forward*. De acordo com Goodfellow, Bengio e Courville (2023), as redes neurais *feedforward*, as informações movem-se apenas em uma direção, para frente, da camada de entrada, através das camadas ocultas, até a camada de saída. Não há ciclos ou *loops* na rede, e cada camada se conecta apenas à próxima camada. Este tipo de arquitetura é comum em redes neurais densas onde cada neurônio em uma camada se conecta a todos os neurônios na próxima camada. Para o treinamento da rede neural foi utilizado ambiente de execução do google colab com acesso à placa gráfica da empresa NVIDIA Tesla V100. Só foi possível o treinamento diverso para encontrar os melhores parâmetros das camadas densas da rede neural devido ao alto desempenho e eficiência da placa gráfica Tesla V100 pois o tempo de treinamento médio de cada época foi de 1 segundo.

Figura 20 – Representação do modelo de rede neural



Fonte: Autor

A definição do parâmetro de entrada se deu pelo cálculo da extração do recurso cromagrama referente a função *librosa.feature.chroma_cens*. O resultado da função são características de cromas normalizados de forma que eles se tornem mais comparáveis (EWERT, 2011). Cada áudio é passado como entrada para a rede neural como uma matriz de 12 linhas indicando cada cromagrama (0 a 11) e 2584 colunas que são efetivamente as características extraídas em cada faixa de croma. Em relação ao código para construção do modelo, o código inicia com a definição de um modelo sequencial, indicando que as camadas são empilhadas uma após a outra. Após é definido a camada de entrada que como dito anteriormente, espera receber dados na forma de um *array* com 12 linhas e 2584 colunas. Depois é aplicado a operação de achamento (*Flatten*) para transformar o vetor em uma dimensão, para que satizfaça a configuração da rede neural pois se tratando de camadas densas, é necessário que os dados estejam em uma dimensão. É adicionado uma camada densa totalmente conectada com 512 neurônios e sua função de ativação ReLU. Após, é aplicado a regularização por *dropout* com uma taxa de 20%, que ajuda a prevenir o *overfitting*. É adicionada outra camada densa com 32 neurônios juntamente com a função de ativação ReLU e finalizando a arquitetura com uma camada densa de 12 neurônios que são relativamente cada classe de croma dentro do conjunto {C, C#, D, D#, E, F, F#, G, G#, A, A#, B}, com a função de ativação *softmax* para obter uma probabilidade para cada classe, pois a função *softmax* como dito anteriormente, é adequada para problemas de classificação multiclasse.

4.1.8 Compilação do modelo

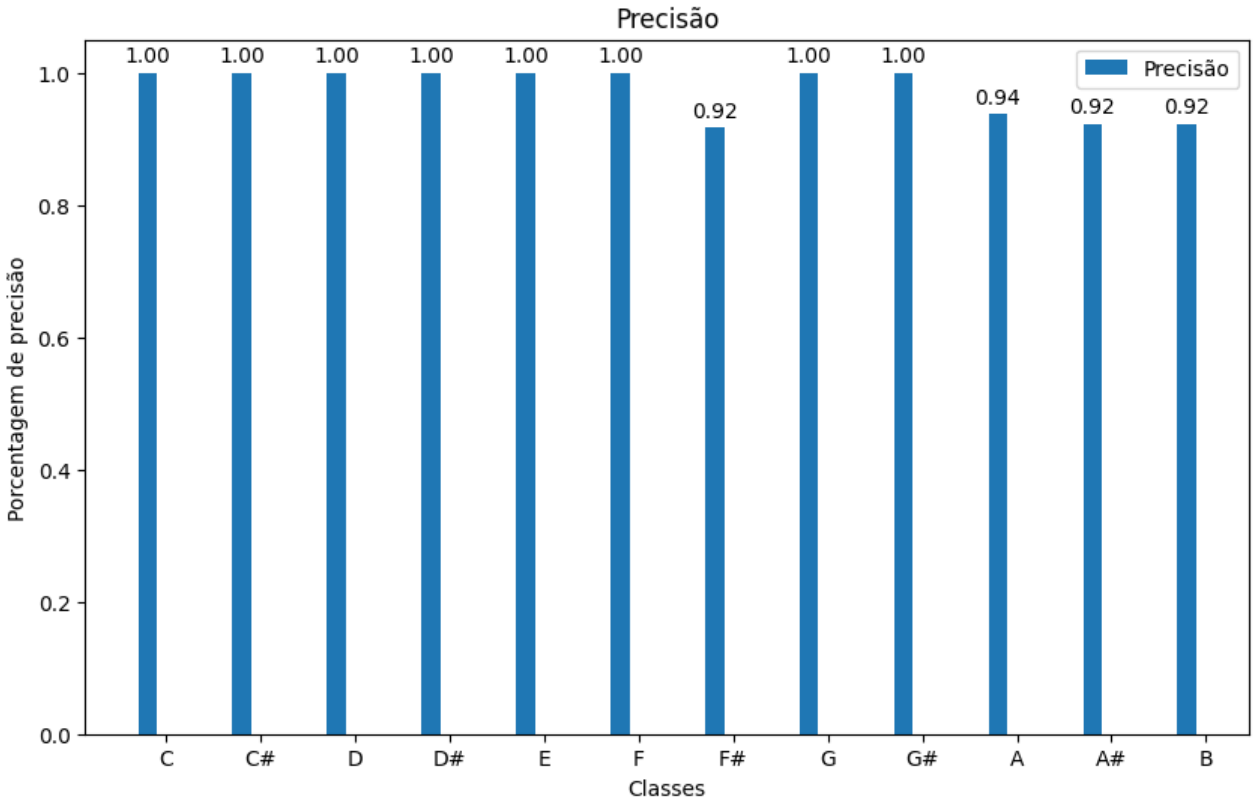
É configurado o modelo para treinamento usando o otimizador otimizador usado para atualizar os pesos da rede durante o treinamento *Adam*, a função de perda *categorical_crossentropy* é utilizada em problemas de classificação onde as classes são exclusivas, ou seja, cada entrada só pode pertencer a uma classe. As métricas são usadas para monitorar o treinamento e o teste do modelo. A precisão (accuracy) mede a porcentagem de previsões corretas do modelo. O modelo foi treinado por 100 épocas com um tamanho de número de exemplo de dados de treinamento de 32 para cada época.

5 RESULTADOS

Neste capítulo irá ser abordado os resultados que o modelo de rede neural conseguiu atingir nas métricas: acurácia, precisão, revocação, *f1-score* e a matriz de confusão.

A acurácia é uma métrica que mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões (GOODFELLOW; BENGIO; COURVILLE, 2023). A precisão é a proporção de previsões positivas corretas (verdadeiros positivos) em relação ao total de previsões positivas feitas (verdadeiros positivos + falsos positivos), indicando quão confiáveis são as previsões positivas do modelo. Uma precisão alta significa que o modelo é bom em não classificar negativos como positivos (GOODFELLOW; BENGIO; COURVILLE, 2023).

Figura 21 – Precisão do modelo



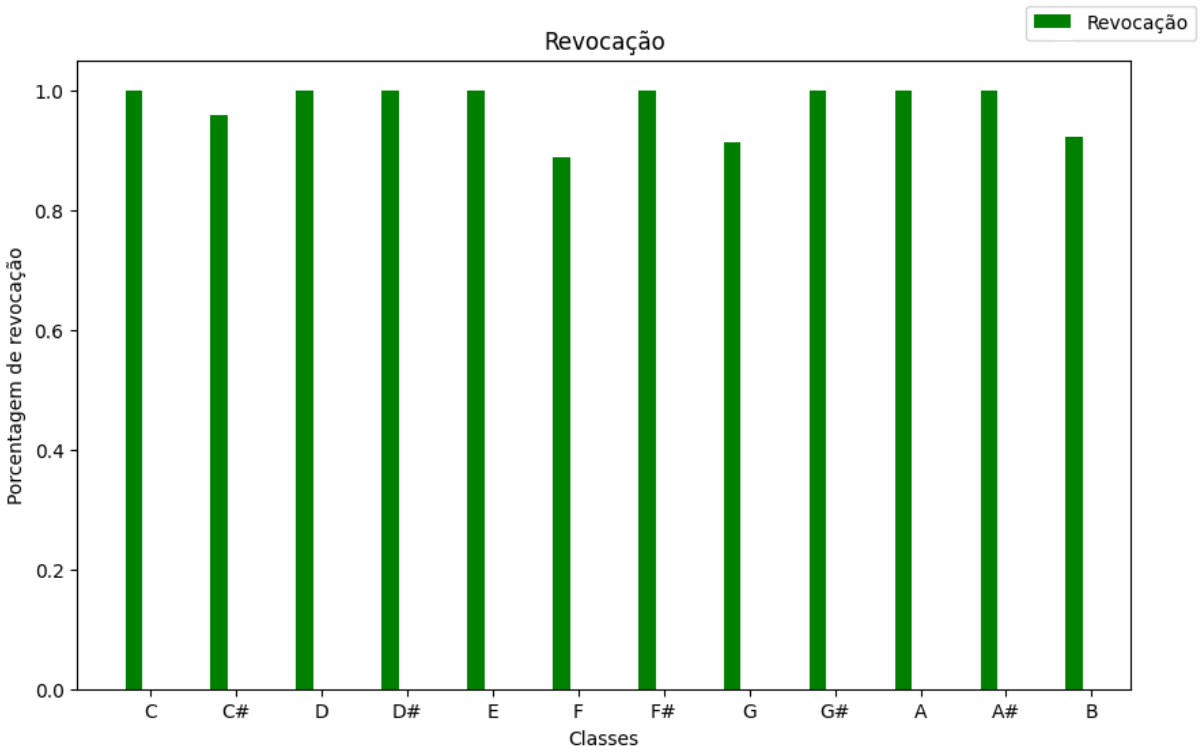
Fonte: Autor

Pode-se na Figura 21 observar que o modelo treinado obteve uma precisão muito boa para cada classe em que ele foi aplicado. Nos dados de validação, o modelo conseguiu

atingir 100% de precisão em oito classes. Além disso, em todas as classes a precisão foi acima de 90%¨.

A revocação é a proporção de positivos verdadeiros identificados corretamente pelo modelo em relação a todos os casos que são realmente positivos no conjunto de dados(verdadeiros positivos + falsos negativos). A métrica revocação (*recall*) mede a capacidade do modelo de encontrar todas as instâncias positivas, ou seja, uma revocação alta é importante quando os custos dos falsos negativos são altos (GOODFELLOW; BENGIO; COURVILLE, 2023).

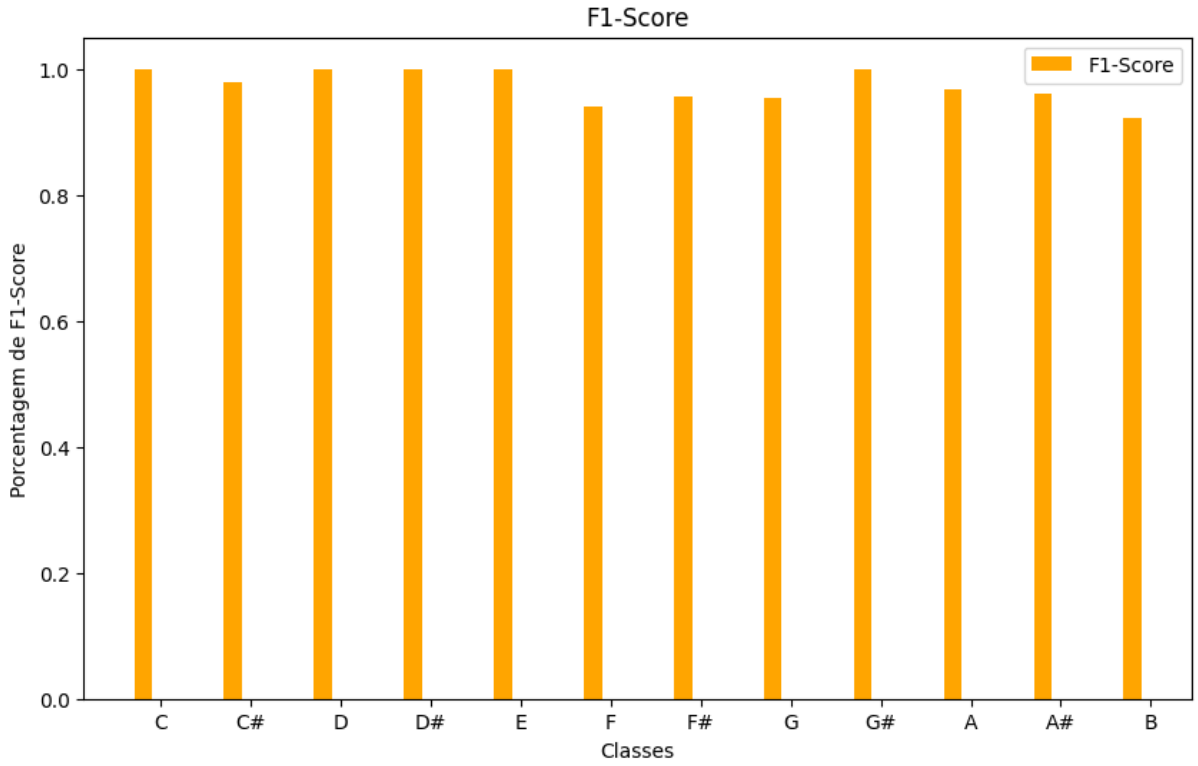
Figura 22 – Fórmula da revocação



Fonte: Autor

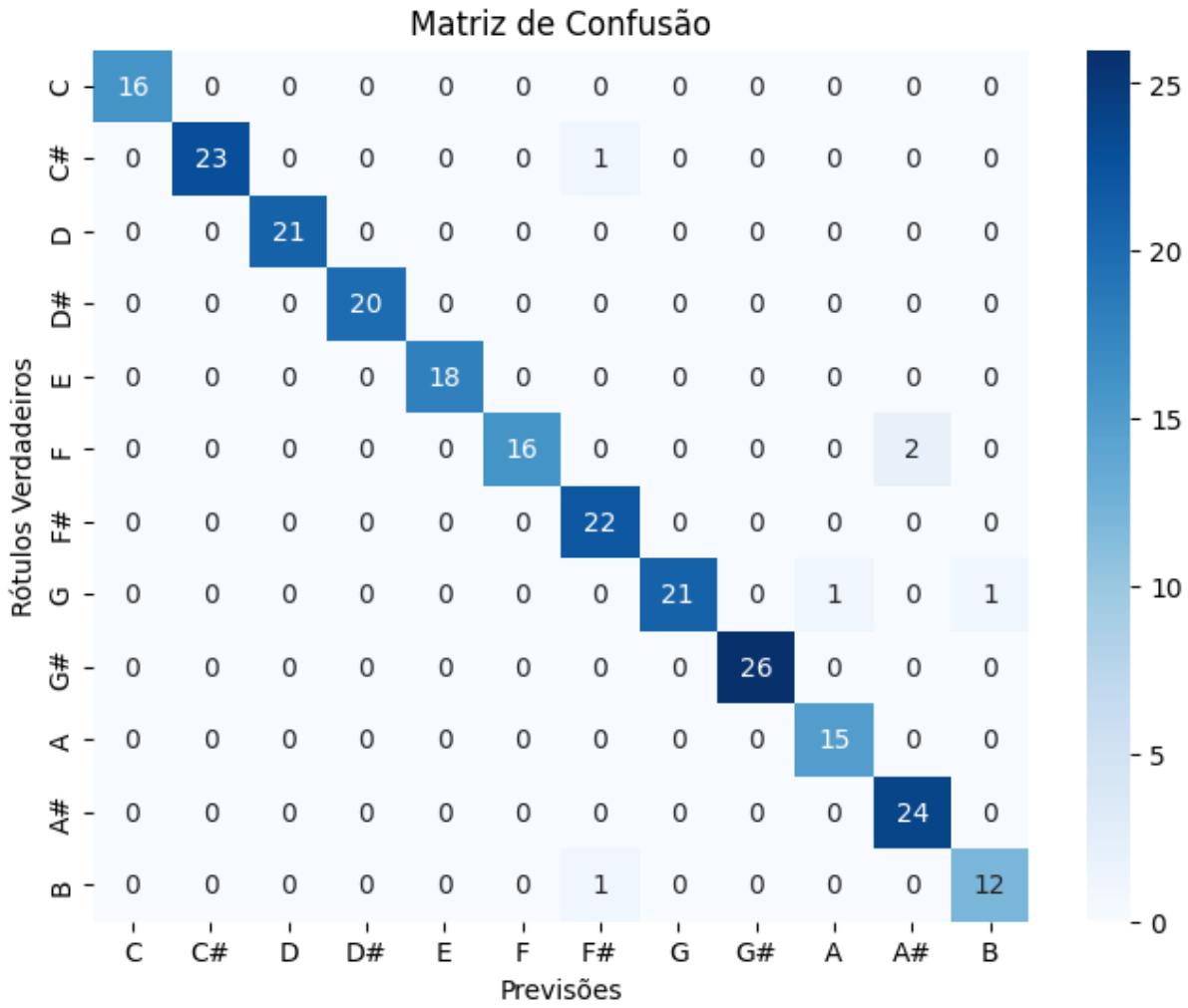
A partir da Figura 22, é notório que dos dados reais previstos, o modelo atingiu bons resultados. Além disso, em quase todas as classes o modelo conseguiu atingir acima de 90% de revocação. O F1-*score* é uma métrica que combina precisão e revocação em um único valor, fornecendo uma medida geral do desempenho do modelo. É a média harmônica das duas métricas, respectivamente precisão e revocação (GOODFELLOW; BENGIO; COURVILLE, 2023).

Figura 23 – F1-*score* do modelo



Assim como nas outras duas métricas de avaliação o modelo conseguiu ir bem na métrica F1-*score*, de acordo com a Figura 23 sendo possível observar uma porcentagem muito boa em relação a cada classe em que o próprio modelo previu, atingindo mais de 90% na métrica em questão. Fornece uma visão mais completa do desempenho do modelo, permitindo a análise de onde o modelo está acertando e onde está cometendo erros, especificamente relacionados a falsos positivos e falsos negativos (GOODFELLOW; BENGIO; COURVILLE, 2023).

Figura 24 – Matriz de confusão do modelo



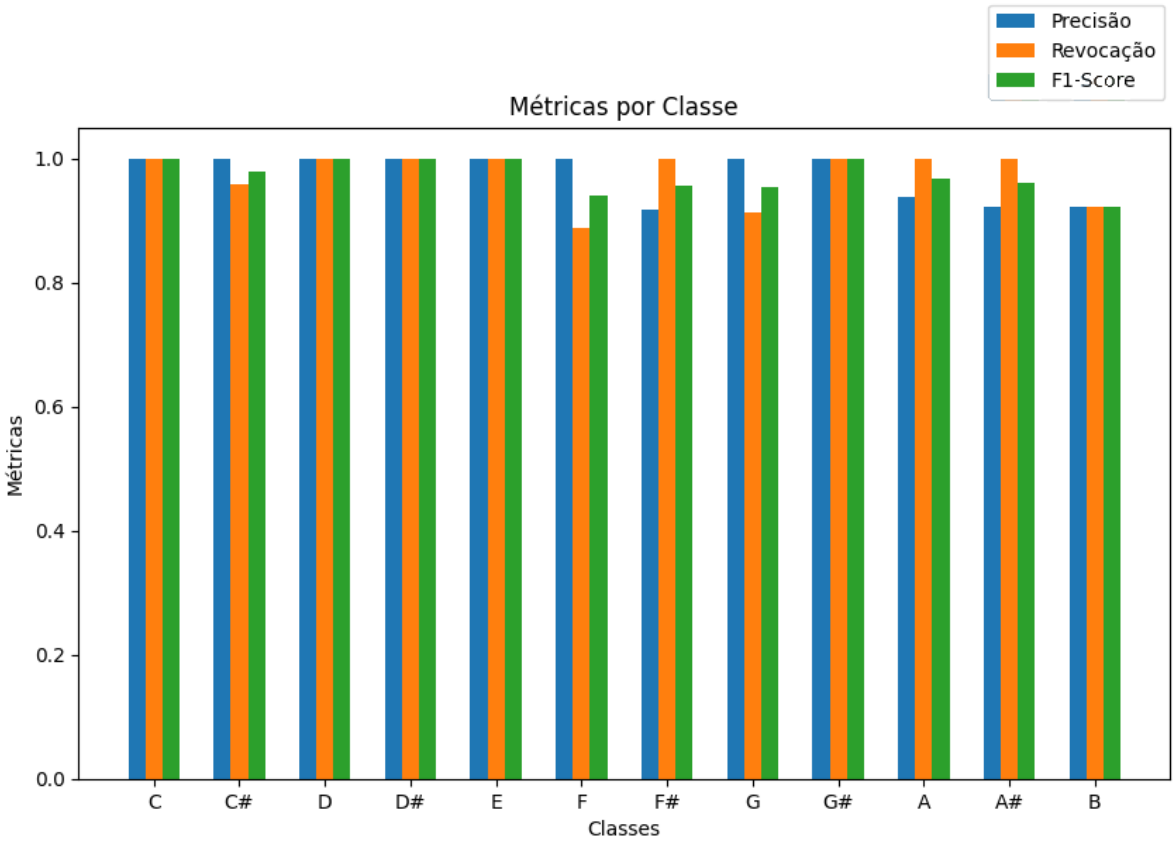
Fonte: Autor

Pode-se observar de acordo com a Figura 24 que os dados estão mais concentrados na diagonal principal da matriz de confusão, indicando que o modelo treinado teve uma boa taxa de acerto para cada classe rotulada. A tabela abaixo mostra o resumo de todas as métricas presente para avaliar o modelo treinado, juntamente com o gráfico de barras de todas as métricas, lado a lado.

Tabela 1 – Resumo das métricas por classe

Classe	Precisão	Revocação	F1-Score	Suporte
C	1.00	1.00	1.00	16
C#	1.00	0.96	0.98	24
D	1.00	1.00	1.00	21
D#	1.00	1.00	1.00	20
E	1.00	1.00	1.00	18
F	1.00	0.89	0.94	18
F#	0.92	1.00	0.96	22
G	1.00	0.91	0.95	23
G#	1.00	1.00	1.00	26
A	0.94	1.00	0.97	15
A#	0.92	1.00	0.96	24
B	0.92	0.92	0.92	13
Acurácia			0.97	240

Figura 25 – Precisão, Revocação, F1-*Score*



Fonte: Autor

Tanto a Tabela 1 quanto a Figura 25, evidenciam as métricas citadas anteriormente em conjunto, reforçando a qualidade do modelo de rede neural treinado neste trabalho,

pois em sua grande maioria das classes, o modelo conseguiu altos valores para as métricas avaliadas.

6 CONCLUSÃO

No desenvolvimento do projeto de conclusão de curso, foi criado um *dataset* abrangente, contendo doze classes com cem amostras cada. Essas classes representam as doze tonalidades musicais fundamentais do mundo ocidental, associando cada música à sua tonalidade correspondente. O *dataset* foi crucial para o treinamento de um modelo de rede neural dedicado à identificação da tonalidade musical em músicas não moduladas.

O resultado impressionante do modelo treinado, alcançando uma acurácia de 97,5%, destaca a viabilidade do ajuste de pesos em uma rede neural, desde que os dados sejam selecionados e rotulados corretamente.

Considera-se, como próxima etapa, a expansão do *dataset*. Este aumento de dados visa capacitar o modelo a prever tonalidades menores, superando uma das limitações atuais de identificar apenas tonalidades maiores. Isso será possível ao empregar técnicas adequadas de extração de recursos, enriquecendo o aprendizado da rede neural.

Para aumentar os dados do *dataset* pode-se utilizar a técnica chamada de *Data augmentation* que por meio dessa técnica é possível aumentar os dados do *dataset* de maneira artificial. Para exemplificar melhor, é possível fazer a alteração da tonalidade musical de forma de uma música em específico, de forma de essa determinada música tenha as doze classes de croma, sendo possível extrair mais onze amostras de áudio para cada amostra de áudio.

REFERÊNCIAS

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

ALDWELL, E.; SCHACHTER, C.; CADWALLADER, A. *Harmony and voice leading*. [S.l.]: Cengage Learning, 2018.

ANDRADE, P. E. Uma abordagem evolucionária e neurocientífica da música. *Neurociencias*, v. 1, n. 1, p. 21–33, 2004.

CASTRO, C.; RIBEIRO, F. A importância da teoria musical na formação de músicos. *Revista Brasileira de Educação Musical*, v. 25, n. 2, p. 25–37, 2018.

CHOLLET, F. et al. *Keras*. GitHub, 2015. Disponível em: <https://github.com/fchollet/keras>.

DRUGMAN, T. et al. Traditional machine learning for pitch detection. *IEEE Signal Processing Letters*, IEEE, v. 25, n. 11, p. 1745–1749, 2018.

EWERT, S. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In: *Proc. ISMIR*. [S.l.: s.n.], 2011.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning Book - Versão em Português Brasileiro*. 2023. <https://www.deeplearningbook.com.br>. Acessado em: 9 de novembro de 2023.

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

HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2001.

Jonathan Cruz Garcia. *Playlists - Canal do YouTube*. 2023. <https://www.youtube.com/@jonathancruzgarcia5304/playlists>. [Online; acessado em 1 de Abril de 2023].

KHADEM-HOSSEINI, M. et al. Error correction in pitch detection using a deep learning based classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, v. 28, p. 990–999, 2020.

KULYUKIN, V.; MUKHERJEE, S.; AMLATHE, P. Toward audio beehive monitoring: Deep learning vs. standard machine learning in classifying beehive audio samples. *Applied Sciences*, MDPI, v. 8, n. 9, p. 1573, 2018.

MAZZOLA, G. et al. Modulation theory. *Cool Math for Hot Music: A First Introduction to Mathematics for Music Theorists*, Springer, p. 191–202, 2016.

MCFEE, B. et al. *librosa/librosa: 0.10.1*. Zenodo, 2023. Disponível em: <https://doi.org/10.5281/zenodo.8252662>.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

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

NASCIMENTO, J. M. Percepção musical: o desenvolvimento do ouvido relativo e suas particularidades. *Música em Foco*, v. 2, n. 1, 2020.

PERETZ, I. Brain specialization for music. *Neuroscientist*, Baltimore, MD: Williams & Wilkins, c1995-, v. 8, n. 4, p. 372, 2002.

RONG, F. Audio classification method based on machine learning. In: IEEE. *2016 International conference on intelligent transportation, big data & smart city (ICITBS)*. [S.l.], 2016. p. 81–84.

ROSSUM, G. V.; DRAKE, F. *Python*. 2009.

SANTOS, R.; SANT'ANNA, B. A análise musical e a formação do músico. *Opus*, v. 25, n. 1, p. 1–16, 2019.

SCHENKER, H. *Free Composition*. [S.l.]: Praeger, 1994.

SHEPARD, R. N. Circularity in judgments of relative pitch. *The journal of the acoustical society of America*, Acoustical Society of America, v. 36, n. 12, p. 2346–2353, 1964.

SILVA, R. Inteligência artificial. Amigos da Enciclopédia, 2013.

SOUSA, J. d. D.; RODRIGUES, R. d. C. O ensino da teoria musical nos cursos de graduação em música. *Per Musi*, v. 40, p. 44–57, 2019.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, IEEE, v. 10, n. 5, p. 293–302, 2002.

WISE, K. J.; SLOBODA, J. A. Establishing an empirical profile of self-defined "tone deafness": Perception, singing performance and self-assessment. *Musicae Scientiae*, SAGE Publications Sage UK: London, England, v. 12, n. 1, p. 3–26, 2008.

XIA, T.; HAN, J.; MASCOLO, C. Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine*, SAGE Publications Sage UK: London, England, v. 247, n. 22, p. 2053–2061, 2022.

ZHANG, S.; JAFARI, O.; NAGARKAR, P. A survey on machine learning techniques for auto labeling of video, audio, and text data. *arXiv preprint arXiv:2109.03784*, 2021.