UNIVERSIDADE DE TAUBATÉ

RAMOM DIAS DOS SANTOS LANDIM DE SOUZA
SARAH REJANE RIBEIRO LORENÇON

TREINAMENTO DE MÁQUINA PARA CLASSIFICAR OS LATIDOS ATRAVÉS DE ONDAS SONORAS

Taubaté - SP
2022

RAMOM DIAS DOS SANTOS LANDIM DE SOUZA
SARAH REJANE RIBEIRO LORENÇON

TREINAMENTO DE MÁQUINA PARA CLASSIFICAR OS LATIDOS ATRAVÉS DE ONDAS SONORAS

Trabalho de Conclusão de Curso apresentado com requisito parcial para a conclusão do curso de Engenharia de Computação da Universidade de Taubaté.
Área: Aprendizagem de Máquina
Orientador: Prof. Dr. Luis Fernando de Almeida

Taubaté - SP

2022

Grupo Especial de Tratamento da Informação - GETI
Sistema Integrado de Bibliotecas – SIBi
Universidade de Taubaté - Unitau

S729t Souza, Ramom Dias dos Santos Landim de
Treinamento de máquina para classificar os latidos através de ondas Sonoras / Ramom Dias dos Santos Landim de Souza , Sarah Rejane Ribeiro Lorençon. -- 2022.
51 f. : il.

Monografia (graduação) – Universidade de Taubaté, Departamento de Informática, 2022.
Orientação: Prof. Dr. Luis Fernando de Almeida, Departamento de Informática.

1. Inteligência Artificial. 2. Sons. 3. Latidos. 4. MFCC. 5. Redes Neurais. I. Lorençon, Sarah Rejane Ribeiro. II. Universidade de Taubaté. Departamento de Informática. Graduação em Engenharia de Computação. III. Titulo.

CDD – 006.3

Ficha catalográfica elaborada **por Shirlei Righeti – CRB-8/6995**

RAMOM DIAS DOS SANTOS LANDIM DE SOUZA
SARAH REJANE RIBEIRO LORENÇON

TREINAMENTO DE MÁQUINA PARA CLASSIFICAR OS LATIDOS ATRAVÉS DE ONDAS SONORAS

Trabalho de Conclusão de Curso apresentado com requisito parcial para a conclusão do curso de Engenharia de Computação da Universidade de Taubaté.
Área: Aprendizagem de Máquina
Orientador: Prof. Dr. Luis Fernando de Almeida

Data: 15/12/2022

Resultado: Aprovado

BANCA EXAMINADORA

Prof. Dr. Luis Fernando de Almeida	**Universidade de Taubaté**
Prof. Fabio Rosindo Daher de Barros	**Universidade de Taubaté**
Prof. Luiz Ricardo Arantes Filho	**Universidade de Taubaté**

DEDICATÓRIA

Dedicamos este trabalho a todos os professores que fizeram parte do nosso crescimento acadêmico, profissional e acima de tudo o crescimento pessoal, a todos os demais que estiveram ao nosso lado, nos apoiando para chegarmos até aqui, aos que nos incentivaram a nunca desistir.

A todos vocês, os nossos mais sinceros carinho e admiração.

AGRADECIMENTOS

Primeiramente gostaríamos de agradecer por estarmos concluindo este ciclo tão importante em nossas vidas. Dedicando este agradecimento inicial a Deus, pois foi responsável por proporcionar oportunidade e condição para que pudéssemos cursar a Universidade, vendo que é uma jornada longa e nem todos conseguem chegar ao fim deste ciclo, seja por falta de acessibilidade de ingressar na universidade ou para aqueles que não conseguiram concluir.

Nosso segundo agradecimento se destina às nossas famílias, um pilar de suma importância para nós, pois sem eles esse caminho não seria possível, sendo nossa base para todos os momentos, em especial para aqueles que estavam cada dia mais próximos e nos apoiando, disponibilizando seu tempo, nos ajudando como podiam, nos dando força para continuar e que não nos deixaram nem por um momento desistirmos dos nossos sonhos.

Agradecemos também aos nossos amigos, que ao longo de todo o curso nos ajudaram, partilharam bons momentos conosco, passamos por poucas e boas juntos, mas com sucesso conseguimos chegar até aqui.

E por fim, mas não menos importante, gostaríamos de agradecer ao grande impulsionador de todo este projeto, sendo o nosso orientador Luis Fernando de Almeida que aceitou trilhar esta jornada conosco, nos apoiou, seguiu conosco em todos os momentos fáceis e difíceis do projeto e nos encaminhou para que no final fosse concluído com sucesso.

RESUMO

Este trabalho demonstra a pesquisa e implementação de técnicas de aprendizado de máquina em aplicações de reconhecimento de sons de cães em termos de expressões animais: fome, chamar a atenção, brincar etc. Para a realização deste, houve o estudo de diversas técnicas de processamento de áudio e análise de alguns ambientes de desenvolvimento. A base de dados utilizada está disponível no site Freesounds, onde os áudios dos latidos utilizados foram armazenados. Para este trabalho foi utilizado, sete classes de latidos, criados a partir dessa base e foram separados em folders locais a fim de, variar a quantidade de classes disponíveis e o tempo de duração dos sinais de áudio nos testes, selecionando somente vocalizações do áudio original. Além disso, foi utilizado a linguagem de programação Python no ambiente do Visual Studio Code. Foi utilizado também métodos de decisão para o reconhecimento do padrão acústico, com a aplicação do coeficiente cepstral de frequência mel (MFCC), o qual é uma maneira de representar o som e foi criado em seguida o modelo e estrutura da RNA, utilizando especificamente Redes Neurais Convolucionais (CNN) através da biblioteca de código aberto TensorFlow. Após aplicação do modelo selecionado, o presente projeto conseguiu realizar o seu objetivo, no qual é o desenvolvimento de uma ferramenta capaz de classificar os padrões de latidos de um cachorro com a base já pré-classificada.

Palavras-chave: Inteligência Artificial, Sons, Latidos, MFCC, Redes Neurais Convolucionais.

ABSTRACT

This work demonstrates the research and implementation of machine learning techniques in applications of sound recognition of dogs in terms of animal expressions: hunger, calling attention, playing etc. To accomplish this, there was the study of several audio processing techniques and the analysis of some development environments. The database used is available on the Freesounds site, where the audios of the barks used were stored. For this work we used seven classes of barks, created from this database, and separated in local folders in order to vary the amount of classes available and the duration time of the audio signals in the tests, selecting only vocalizations of the original audio. In addition, the Python programming language was used in the Visual Studio Code environment. Decision methods were also used for acoustic pattern recognition, with the application of the honey frequency cepstral coefficient (MFCC), which is a way to represent sound, and then the ANN model and structure was created, specifically using Convolutional Neural Networks (CNN) through the open source TensorFlow library. After applying the selected model, the present project achieved its goal, which is the development of a tool capable of classifying the barking patterns of a dog with an already pre-classified base.

Keywords: Artificial Intelligence, Sounds, Barks, MFCC, Convolutional Neural Networks.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O Reconhecimento de Padrões pode ser definido como a classificação de objetos (padrões) em um número de categorias ou classes (THEODORIDIS; KOUTROUMBAS, 1999). Trata-se de uma área ampla, podendo englobar o reconhecimento de imagens, sons, anomalias, padrões de compra, dentre outros.

No que diz respeito ao foco deste trabalho, reconhecer sons pode ser tratado como um problema de se analisar pequenos trechos sonoros e encontrar, em sua sonoridade e não em seu conteúdo, padrões de repetição.

O estudo do som, e mais precisamente o estudo das ondas sonoras, apresenta-se como um argumento importante e interessante que pode se basear e conectar vários conceitos matemáticos intermediários, é relativamente fácil de entender, e pode abrir um leque de possibilidades para que possamos iniciar nosso estudo de caso, proporcionando a interdisciplinaridade da física ao tratar do conteúdo de ondas e acústica.

O Reconhecimento de sons pode ser aplicado em sistemas de segurança, avisos para pessoas com problema auditivo entre outros. Por exemplo, na criação do iOS14 a Apple (Apple,2022) adicionou essa novidade ao seu smartphone, a fim de ajudar os usuários, principalmente para aqueles com algum problema auditivo, a reconhecer sons ambiente, como campainha, choro de bebê, buzina de veículos, cães e entre outro.

A evolução da Inteligência Artificial (IA) agregada ao avanço dos computadores, vem disponibilizando, cada vez mais, técnicas eficazes para a solução de problemas de reconhecimento de padrão, principalmente quando se fala em Aprendizado de Máquina.

Para aprimorarmos os estudos de reconhecimento de sons e suas problemáticas utilizamos como base alguns trabalhos correlatos pertinentes que fizeram parte do processo para a criação deste trabalho, como da Muniz (2009) utiliza o reconhecimento de Áudio para que se possa identificar sons repetitivos, através de um trecho sonoro armazenado. Os autores, Manaro, Larco e Lopes (2021) utiliza uma rede neural a fim de classificar sons em ambientes urbanos para que possa avisar o

usuário com deficiência auditiva de possíveis perigos ao seu redor. Por fim, Zottesso (2017) utilizou o reconhecimento de sons com o objetivo de apresentar uma proposta para a identificação de espécies de pássaros utilizando espectrogramas e a abordagem de dissimilaridade, em uma base de dados com alta quantidade de espécies.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho apresenta um estudo e implementação de técnicas de Aprendizado de Máquina na aplicação de reconhecimento de sons de cachorros de acordo com aquilo que o animal manifestando: fome, chamar a atenção, brincar, dentre outros.

1.1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Criar uma base de dados de sons de um cachorro previamente mapeada com o seu significado.
- Efetuar uma pesquisa dos melhores métodos que poderíamos utilizar para o recebimento dos áudios e suas futuras conversões, para aplicação a técnicas de Aprendizado de Máquina.
- Aplicação de métodos computacionais para se obter o espectro resultante do áudio e realizar uma comparação de padrões dentro de uma base já alimentada pelo próprio tutor.
- Criar um modelo classificador baseado em técnica e Aprendizado de Máquina.

1.1.3 Justificativa

Após o encontro de um cachorro com dificuldades em sua locomoção, pode-se perceber um certo padrão em seus latidos para determinadas ações, com isso foi pensado e elaborado uma das principais justificativas deste trabalho, o qual foi imaginar como seria se nossos animais falassem e compreender qual a necessidade ele quer mostrar, chamando a atenção do tutor, pelo latido. A respeito disso foi analisado também a referência a esse livro: A cabeça do cachorro, o que seu amigo mais leal vê, fareja, pensa e sente, Alexandra Horowitz, onde a psicóloga mostra, cientificamente, que nossos animais têm sentimentos e tem habilidade de pensar e tomar decisões.

1.2 Estrutura do Trabalho

Este trabalho está dividido em seis capítulos:

- O presente capítulo é composto por contextualizar o problema e os objetivos a serem alcançados com este projeto.
- O Capítulo 2, aborda de maneira detalhada sobre a pesquisa feita para se criar toda a estrutura computacional do programa, visando a arquitetura completa do som, a percepção do som, como os cálculos ou amostragens que seriam usuais no decorrer do projeto e finalizando-o sobre a área da Inteligência artificial junto com o uso de Deep Learning em redes neurais.
- O Capítulo 3, apresenta a aplicação desenvolvida, os materiais que foram utilizados, bem como os métodos necessários para o desenvolvimento
- O Capítulo 4 contém o que foi abordado para a criação do programa e as subdivisões necessárias para validarmos os áudios gravados.
- O Capítulo 5 trata-se de mostrar os testes e resultados que tivemos para concluir a eficiência que o programa obteve.
- No Capítulo 6, são apresentadas as considerações finais, informando o resultado e propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

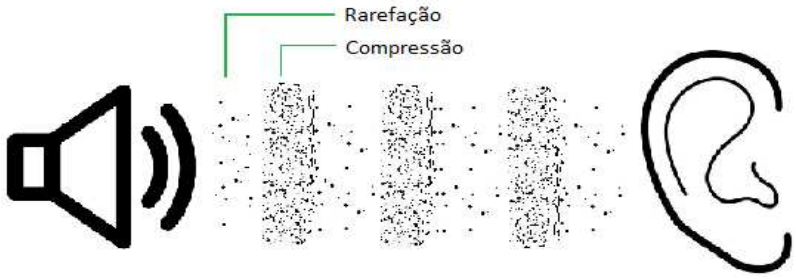
Este capítulo apresenta conceitos relevantes para auxiliar no entendimento deste trabalho: som e suas características; onda sonora; conversão analógico-digital; padrões de codificação e compressão de áudio; transformada de Fourier; Inteligência Artificial; Redes Neurais Artificiais.

2.1 O Som e Suas Características

No universo existe diversas formas de energia e o som, como tal, obedece às leis fundamentais da física. O som é uma onda mecânica, longitudinal e tridimensional, quando um objeto vibra, as moléculas de ar oscilam para frente e para trás a partir de sua posição de repouso e transmitem sua energia para as moléculas vizinhas. Isso resulta na transmissão de energia de uma molécula para outra, que por sua vez produz uma onda sonora.

As partes onde o ar é empurrado para mais perto são chamadas de compressões, e as partes onde ele é empurrado para longe são chamadas de rarefações (Figura 2.1). Essas ondas que atravessam o espaço usando compressões e rarefações são chamados de ondas longitudinais.

Figura 2.1: Representação de onda de compressão e rarefação.



Fonte: Os autores.

A mais simples das ondas sonoras é conhecida como onda senoidal (Figura 2.2), gerada por um “Movimento Harmônico Simples”. O som gerado por uma onda

senoidal é um som com uma forma de onda repetitiva mais básica, com uma amplitude oscilando em ambos os lados de um valor central e seguindo uma curvatura senoidal.

No entanto, essas ondas sonoras não são naturais porque não pode haver uma fonte sonora que vibre com intensidade estritamente constante em frequências audíveis, e por existir entre o emissor sonoro e o receptor um meio material não uniforme, devido ao movimento das partículas e variação da densidade, que, por menor que seja, resulta na variação do período e da amplitude máxima em distintos intervalos de tempo. Mesmo que não seja natural, movimentos harmônicos simples podem ser sintetizados eletronicamente, por exemplo, em um gerador de voz.

2.2 Percepção do Som

A percepção auditiva é a interpretação e a compreensão do ambiente sonoro que nos rodeia realizado pelo nosso cérebro, as características fundamentais da sensação auditiva são: a sonoridade, a tonalidade e o timbre.

- A sonoridade é a sensação de intensidade, permite-nos afirmar se o som é mais forte ou mais fraco.
- A tonalidade é a sensação ligada à frequência, permite-nos saber se o som é mais agudo ou mais grave.
- O timbre é a característica que nos permite diferenciar da mesma intensidade e tonalidade.

Junto a percepção auditiva, pode-se classificar os sons das seguintes maneiras:

- Classificação de dados acústicos - O local que foi coletado ou gravado o áudio.
- Classificação de som ambiente - Classificação de diferentes sons no ambiente, por exemplo: Veículos, animais, músicas etc.
- Classificação musical - Se classifica o estilo da música: Rock, Samba etc.

- Classificação de linguagem natural - Forma como as máquinas entendem e lidam com as linguagens humanas, por exemplo: se tradução simultânea, escrita automática etc.

2.3 Componentes de uma Onda Sonora

Ondas são frutos de perturbações causadas em um sistema qualquer, que se propagam no espaço, logo não são capazes de transportar matéria, apenas energia. O som, como dito já anteriormente, é uma forma de se transportar a energia e obedece às leis da física. Independente de qual for a natureza, forma de propagação ou perturbação, todas as ondas possuem as mesmas propriedades.

De acordo com a ABNT (1959), a definição dá para o som é: "toda e qualquer vibração ou onda mecânica em um meio elástico dentro da faixa de audiofrequência". Em resumo, as ondas apresentam as seguintes características básicas (Figura 2.3):

- Comprimento de onda: Uma onda sonora, provocada por um movimento de perturbação no meio material, se propaga longitudinalmente através da variação de pressão, ou seja, a onda sonora se propaga na mesma direção do movimento perturbador do meio material. O comprimento de onda, em movimento harmônico simples, é representado pelo símbolo λ e equivale à distância que as ondas percorrem entre dois vales consecutivos ou dois picos consecutivos, ou seja, até que realizem uma oscilação completa.
- Período: O período τ de uma onda é o tempo necessário para a produção de uma onda. Pode-se relacionar a velocidade do som v_s, com o comprimento λ e o período τ por:

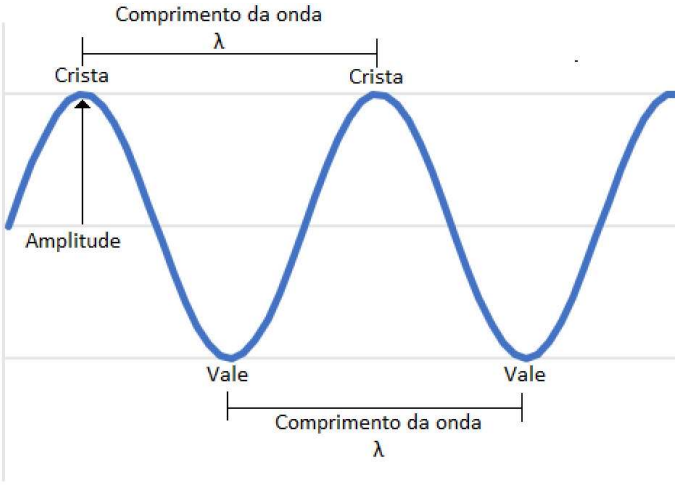
$$v_s = \frac{\lambda}{\tau}$$

- Frequência: A frequência da onda é dada pela quantidade de ondas produzidas dentro de um intervalo de tempo, ou seja, o número de oscilações que ela realiza a cada segundo. Quando a unidade de tempo é o segundo, representamos a frequência em Hertz (Hz). Desta forma, o período τ é o inverso da frequência f, isto é,

$$\tau = \frac{1}{f}$$

- Pico e Vale: A pico(s) são os pontos mais altos das ondas, enquanto o vale(s) é os pontos mais baixo das ondas.
- Amplitude: A amplitude da onda está relacionada à sua intensidade, é medida como a distância da posição central da onda até a altura de um pico ou de um vale. Mas, quando som quando se propaga, provoca um deslocamento de ar que, por consequência, provoca diferenças de pressão no ambiente. Essa diferença de pressão em acústica é geralmente lida em termos de nível de pressão sonora, ou NPS, e dada em decibel (dB), assim que medimos a amplitude de uma onda sonora.

Figura 2.2: Exemplos de alguns elementos compostos nas ondas.



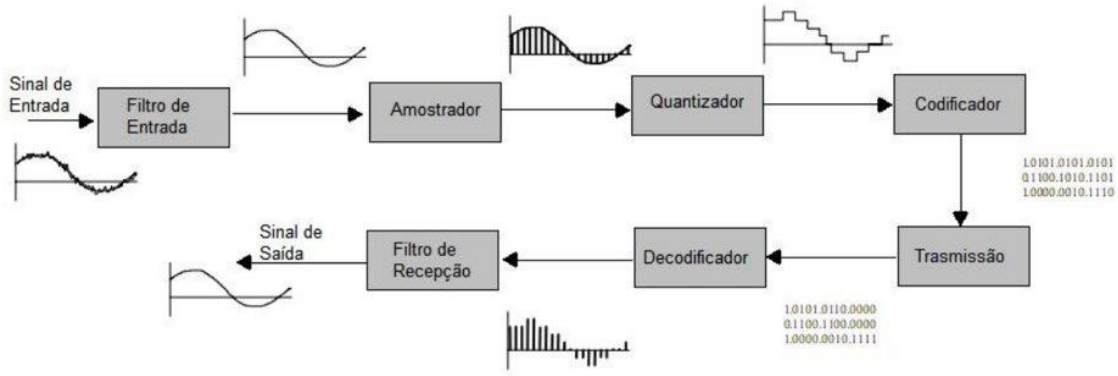
Fonte: Os autores.

2.4 Conversão Analógico-Digital

O sinal obtido é considerado analógico por apresentar variações contínuas no tempo e em amplitude. A digitalização consiste na restrição do sinal analógico, este é um sinal contínuo com infinitas possibilidades de amplitude. Esta restrição gera um sinal discreto com finitas possibilidades de amplitude. Este procedimento é realizado ao utilizar um amostrador e com um quantizador.

O primeiro é responsável pela discretização no tempo, e o segundo realiza a discretização das amplitudes de saída. A parte da codificação tem como finalidade realizar a transformação da onda digital em uma sequência binária. Dessa forma é possível transformar o sinal analógico em digital como ilustra a Figura 2.4.

Figura 2.3: Exemplo do funcionamento de um CAD.



Fonte: Muniz (2009, p.17).

2.5 Padrões de Codificação e Compressão de Áudio

A codificação refere-se à representação de um sinal digital em notação binária, com a finalidade de transmiti-lo ou armazená-lo, com a qualidade exigida pela aplicação. A compressão é definida pela redução do número de bits usados para representar cada amostra.

A codificação, juntamente com a compressão, caracteriza o que define um formato de áudio (para armazenamento e transmissão). O codificador afeta o tamanho

do arquivo e o nível de processamento necessário para decodificar o som. Isso ocorre pois o codificador executa todos os processos de compactação necessários, além de converter as amostras em palavras binárias. Assim há diversas características a serem analisadas nos codificadores de áudio (RIBEIRO, 2008), e tais são:

- Qualidade de áudio medida em PSNR (Peak Signal-to-Noise Ratio) ou MOS (Mean Opinion Score).
- Clareza.
- Possibilidade de identificação do locutor.
- Quantificação.
- Latência total do sistema.
- A complexidade do processo de codificação e decodificação.
- Memória necessária para codificação e decodificação.
- Suscetibilidade a erros de transmissão.
- Reduzir a quantidade de informações que representam o áudio (compressão binária).

Além dessas características, o conteúdo sonoro precisa ser analisado para extrair de seu sinal apenas as informações necessárias para a aplicação e assim selecionar a codificação mais adequada. A seguir, são listados alguns formatos de arquivo que contêm informações sonora.

2.5.1 Modulação por Código de Pulso (Pulse Code Modulation – PCM)

É o formato padrão para áudio digital em computadores e mídias como DVDs e CDs. É a mais antiga tecnologia de digitalização de som e muitos formatos digitais sem perdas são baseados nela. É um método de codificação que demonstra a forma de uma onda sonora e pode usar quantização linear ou não linear (lei μ ou A). (MOECKE, 2006).

2.5.2 Wafeform áudio format (WAV)

É um formato de áudio sem perdas e de alta qualidade, baseado em PCM, usando um método de armazenamento de áudio não compactado. WAV permite a gravação de áudio com diferentes taxas de amostragem e bits, inclusive com a mesma qualidade de um CD de áudio. É adequado para edição de trabalhos profissionais, podendo ser facilmente editado e manipulado por este tipo de software. Tendo apenas um empecilho, o arquivo se limitando a apenas 4 GB. Criado em conjunto pela Microsoft e pela IBM, o padrão foi pioneiro no armazenamento de som em computadores (IBM e MICROSOFT, 1991).

2.5.3 MPEG-1 Audio Layer 3 (MP3)

Este é o mais popular de todos os formatos porque combina qualidade de áudio com uma boa taxa de compressão, o que o torna muito pequeno em tamanho. Ele foi desenvolvido pelo Moving Pictures Experts Group e é, comumente, usado para armazenar arquivos de música em PCs, telefones celulares e até sites de compartilhamento na Internet. Embora a qualidade seja boa, há uma pequena perda em relação ao áudio original, mas em um nível quase imperceptível para a maioria dos ouvidos. Conta com o apoio dos principais players do mercado.

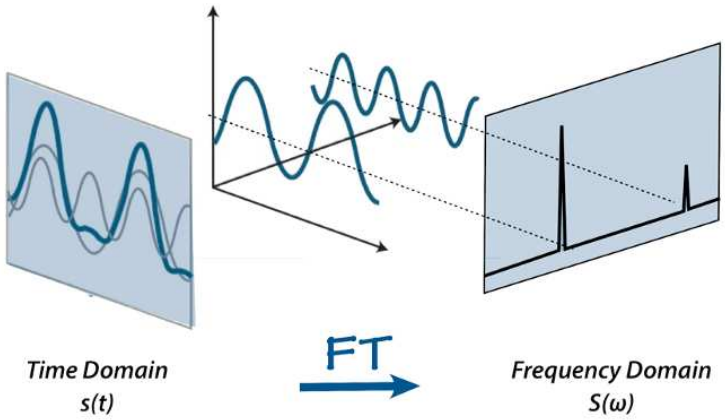
2.5.4 Vorbis (OGG)

Possui uma taxa de compressão semelhante ao MP3, mas com melhor qualidade de som. É um formato de código aberto, criado pelo Xiph.org, mas devido à maior difusão do MP3, este formato ainda tem muitas dificuldades de difusão. Um de seus pontos fortes é a inclusão de metadados, como informações do artista e do álbum. (XIPH.ORG FOUNDATION, 2020).

2.6 Amostragem do Sinal de Áudio (FOURIER)

A análise de Fourier é uma família de técnicas matemáticas, todas elas baseadas na decomposição de sinais em senoides (FOURIER, 1807). Um sinal de áudio é um sinal complexo composto de múltiplas “ondas sonoras de frequência única” que viajam juntas como uma perturbação (mudança de pressão) no meio. Quando o som é gravado, capturamos apenas as amplitudes resultantes dessas múltiplas ondas. A Transformada de Fourier não fornece apenas as frequências presentes no sinal, mas também a magnitude de cada frequência presente no sinal. A Figura 2.5 ilustra um exemplo da Transformada de Fourier.

Figura 2.4: Exemplo da Transformada de Fourier.



Fonte: Gajdošová, 2020, p.20 (adaptado).

Porém, se tem uma desvantagem com a aplicação da FT, na transformação para o domínio da frequência, a informação do tempo é perdida. Ao analisar para uma transformada de Fourier de um sinal, é de grande dificuldade informar a ocorrência de um determinado evento.

Pode-se afirmar que o sinal de voz é uma sequência de intervalos estacionários dentro dos quais a distribuição espectral de potência é mais ou menos constante, o que faz a análise espectral de Fourier uma possibilidade para a parametrização do sinal de áudio (MUNIZ, 2009).

Para esse tipo de parametrização, vale salientar que a resolução em frequência é dada por $\Delta f = \frac{1}{n.t}$ sendo T o período de amostragem. Assim a resolução

em frequência aumenta de forma diretamente proporcional ao número de amostras por quadro. Na parametrização temporal, contudo, a resolução aumenta à medida que o tamanho dos quadros diminui, e o número de amostras por quadro se tornam mais concentradas.

Para diminuir as consequências da divisão do áudio em pequenos quadros é feita a filtragem das extremidades destes de modo a evitar descontinuidades entre os quadros, tal técnica é conhecida como janelamento e tem demonstrado resultados eficientes. Tal análise também pode ser realizada através do algoritmo de Transformada Rápida de Fourier (FFT - Fast Fourier Transform), tendo o mesmo tipo de resultado e ganho na diminuição da complexidade computacional.

2.7 Coeficientes Cepstral de Frequência Mel (MFCC)

A faixa de frequência de áudio da escala Mel é dimensionada não linearmente de um valor mais alto para um mais baixo. Isso resulta em uma faixa de frequência de áudio que soa imperceptivelmente diferente para o ouvinte. O método de escala baseia-se no princípio de que sons de distância igual na escala são percebidos como sendo de distância igual para os humanos.

O alcance da audição humana é apenas entre 20Hz e 20KHz, a percepção humana de ouvir a diferença entre a onda sonora de '100 Hz' e '200 Hz' é muito maior em comparação com a diferença de identificação entre '10100 Hz' e '10200Hz', embora a diferença no valor de Hz seja a mesma, ou seja, é muito mais difícil para os humanos serem capazes de diferenciar entre frequências mais altas e mais fácil para frequências mais baixas.

É isso que torna a Escala de Mel fundamental em aplicações de Aprendizado de Máquina para áudio, pois imita a própria percepção de som. A fórmula para transformação da escala Hertz para escala Mel é a seguinte:

$$m = 1127 \times \log(1 + \frac{f}{700})$$

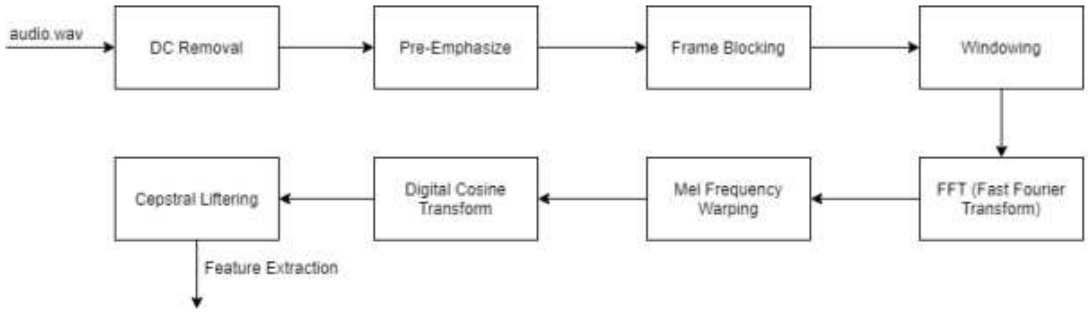
O Coeficiente Cepstral de Frequência Mel (MFCCs) tornou-se uma maneira popular de representar o som. Cepstral é a informação da taxa de mudança nas

bandas espectrais (transformada de Fourier no sinal de tempo). A ideia básica de como obtém-se o cepstral é a seguinte:

- Aplica-se uma transformada no próprio espectro de frequência;
- O espectro resultante não está no domínio de frequência, nem no domínio do tempo, por isso, Bogert (1963) decidiu chamá-lo de domínio “quefrency”.
- Este espectro do logaritmo do espectro do sinal de tempo foi denominado cepstrum.

As etapas apresentadas na Figura 2.6 são um resumo para o cálculo do MFCC.

Figura 2.5: Processo do MFCC.



Fonte: WIBAWA, DARMAWAN (2014, p.3).

Onde:

- DC Removal: Ao processar os dados do sinal de voz, foi necessário um processo de normalização de dados. No processo do DC Removal, a normalização dos dados foi realizada calculando a média dos dados da amostra de voz e subtraindo o valor de cada amostra de voz deste valor médio.
- Pre-Emphasize: O objetivo deste processo era reduzir o ruído no som de entrada, aumentando assim a precisão durante a extração do som.
- Frame Blocking: É o processo de cortar um sinal sonoro em vários quadros
- Windowing: Para evitar os efeitos descontínuos sobre o sinal, devido ao Frame blocking do quadro, foi necessário fazer um processo de janelas para reduzir este efeito de perda.
- Fast Fourier Tranform: Neste processo da FFT, cada quadro com n amostras foi convertido do domínio do tempo para o domínio da frequência. FFT foi um

algoritmo rápido para implementar a Transformada Discreta de Fourier (DFT) que operava sobre um sinal discreto de tempo composto por N amostras.

- Mel Frequency Warping: Nessa etapa que se aplica a conversão faixa de frequência do áudio de entrada de Hz para escala Mel
- Digital Cosine Transform: é usada para decorar o espectro do mel para produzir uma boa representação das propriedades espectrais do som. O resultado foi chamado de Coeficiente de Ceptro de frequência Mel (MFCC)
- Cepstral Liftering: Os coeficientes cepstral de baixa ordem tinham suas características que eram muito sensíveis à inclinação espectral, enquanto as peças de alta ordem eram muito sensíveis ao ruído. Portanto, a elevação cepstral foi uma das técnicas padrão aplicadas para minimizar esta sensibilidade.

2.8 Inteligência Artificial

A Inteligência Artificial, ou de modo abreviado IA, é um campo da computação multidisciplinar; ela apresenta diversos desafios em multimidia, não somente em aprimorar a experiencia na interação computador-humano, como possibilita a construção de sentidos com base em uma proposta comunicacional.

Os autores Russell e Norvig (2013, p. 7) discriminam a Inteligência Artificial como "o estudo de agentes que recebem percepções do ambiente e executam ações", os autores classificam estes agentes de acordo com sua dificuldade computacional. Para ambos, o agente racional é aquele que age para alcançar o melhor resultado ou, quando há dúvida.

Desta forma, para cada sequência de percepções, um agente deve selecionar uma ação que se espera e que venha potencializar sua medida de desempenho, dado o sinal oferecido pela sequência de informações e por qualquer conhecimento interno do agente (RUSSELL; NORVIG, 2013, p. 66).

As ações que o agente, anteriormente mencionado, realiza pode ter uma abordagem multimídia, já que para interagir com o ser humano, é importante corresponder a linguagem para ser mais bem compreendido. Dessa forma, uso de

imagens, áudio, entre outros, permite que os agentes inteligentes passem melhor uma mensagem.

Escolher um ponto de vista comunicativa logo na geração de um agente inteligente é importante pois permite localizar as possíveis conexões entre significados que indicam significações.

De acordo com Russell e Norving (2013), aparece uma abordagem com base na análise estrutural chamado PEAS, cujo significado em inglês aparece como Performance, Environment, Actuators e Sensors, e em português, Desempenho, Ambiente, Atuadores e Sensores.

De acordo com a análise inicial atribuída pelos autores torna-se mais pertinentes as escolhas sobre os elementos que o agente necessitará para entender o mundo e agir acerca dele, inclusive os algoritmos. Após a definição do agente é que será possível roteirizar suas ações como efeitos secundários.

2.9 Redes Neurais Artificiais

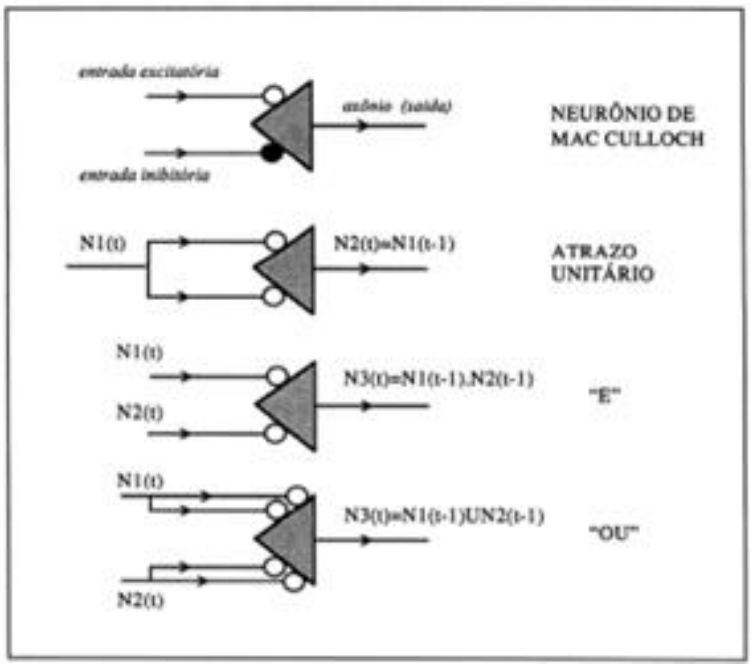
Segundo o autor Haykin (2001) as Redes Neurais Artificiais (RNA) são um sistema que permite simular a maneira que o cérebro executa uma tarefa, adquirindo conhecimento através de um aprendizado. É semelhante ao cérebro humano, em relação ao processo de aprendizagem e as conexões entre neurônios para armazenar o conhecimento adquirido, a parte de como é formada por unidades de processamento de informações é chamada de neurônios.

As RNAs, basicamente, possuem três tipos de estruturas, são elas: redes recorrentes, redes com camadas únicas e múltiplas camadas. Para Haykin (2001) a camada de redes neurais tem diversos neurônios, onde uma rede de camada haverá apenas uma camada, assim como a de múltiplas camadas irá possuir uma ou mais.

O surgimento das Redes Neurais Artificiais iniciou-se com o matemático Warren McCulloch e Walter Pitts em 1943 (McCULLOCH e PITTS, 1943) e seu modelo denominado neurônio MCP (McCulloch-Pitts), seu princípio era definir um valor n de

entradas, multiplicando-as por um determinado peso e na sequência os resultados eram somados e comparados a um limiar.

Figura 2.6: Neurônio booleano de McCulloch e algumas funções booleanas.



Fonte: KOVÁCS (2002, p.28).

As RNAs são comumente utilizadas hoje para resolver problemas complexos, tendo como uma de suas principais características o aprendizado por meio de exemplos e terem modelos generalizados das informações que foram aprendidas.

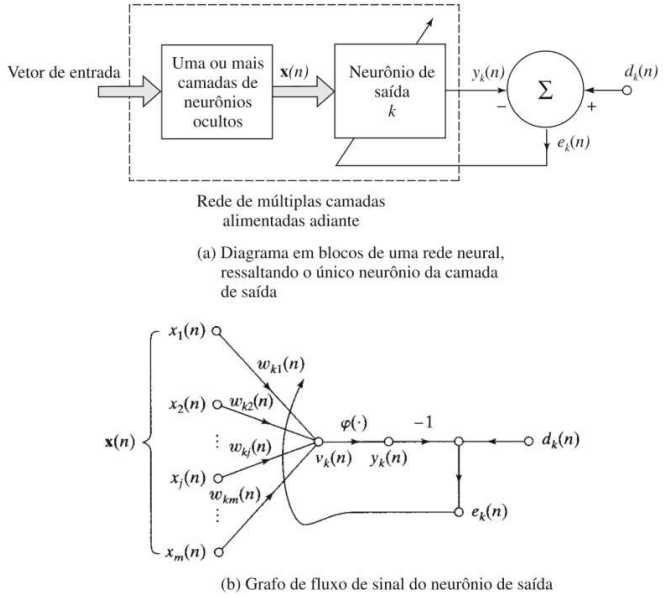
Em questões topológicas a parte mais complexa de se utilizar a RNA vem da escolha da melhor arquitetura que será trabalhada por ser um processo experimental e o tempo de execução varia de projeto para projeto. Sua prática começa com o teste de vários métodos de aprendizagem e as diferentes configurações que uma rede pode ter para a resolução do problema em questão. As subseções seguintes apresentam um pouco sobre as diferentes formas de aprendizagem.

2.9.1 Aprendizagem por Correção de Erro

É baseada em uma problemática na qual, informa-se um valor em um vetor de entrada sendo ele produzido por uma ou mais camadas de neurônios ocultos que,

por sua vez, acionam um vetor de entrada sendo ele o estímulo da rede neural que passa para o sinal de saída da rede que é representado por ser a única saída da rede neural que, é comprado com uma resposta esperada. Mas como não é o desejável na saída é produzido um sinal de erro. A Figura 2.8 ilustra o seu funcionamento.

Figura 2.7: Representação da aprendizagem por correção de erro.

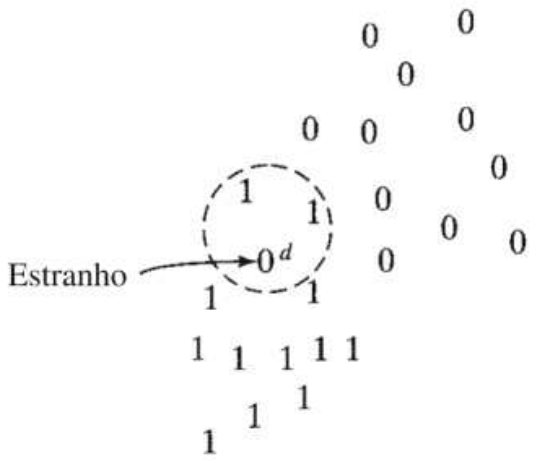


Fonte: HAYKIN (2001, p.77).

2.9.2 Aprendizagem Baseada em Memória

Neste modelo toda a informação (ou grande parte dela) passada é armazenada em uma grande memória que possuem exemplos de entrada-saída que estão classificadas corretamente, onde, o vetor de entrada é representado pela resposta desejada, esta resposta é tratada para que não perca sua informação em sua maioria. Isto se dá para que apenas esta classe/hipótese seja considerada e por isso ela assume o valor de 0 para que quando o vetor de teste seja chamado ele possa efetuar a busca e analisar dados de treinamento em uma base semelhante de teste, isto é, buscando material que já foi testado e que se assemelha ao que é pedido. A Figura 2.9 ilustra o seu funcionamento.

Figura 2.8: Representação da aprendizagem baseada em memória.



Fonte: HAYKIN, 2001, p.80.

2.9.3 Aprendizagem Hebbiana

Sendo a aprendizagem Hebb, a mais antiga e famosa de todas as regras de aprendizagem, a modificação tem como base a aprendizagem associativa, resultando em modificações permanentes do padrão permitindo agrupar células nervosas espacialmente distribuídas.

A forma mais simples de definir a aprendizagem hebbiana (Hipótese de Hebb) é pela seguinte equação:

$$\Delta w_{kj}(n) = \eta(y_k(n)x_j(n))$$

Na equação apresentada, anteriormente, $\Delta w_{kj}(n)$ significa a função dos sinais pré e pós-sinápticos, os sinais de $y_k(n)\ e\ x_j(n)$ são sinais tratados como adimensionais. O valor η tem como significado a constante positiva determinando a taxa de aprendizagem. Sendo referenciada como a regra do produto das atividades.

2.9.4 Aprendizagem Competitiva

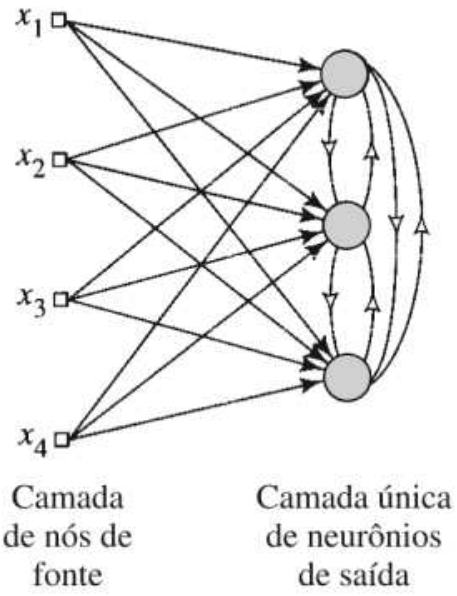
Como o nome da aprendizagem apresenta ela é definida pelos neurônios de saída de uma rede neural que competem entre si para se tornarem ativos. Enquanto na RNA hebbiana, apresentada anteriormente, inúmeros neurônios de saída podem ser vistos estarem ativos simultaneamente, na aprendizagem competitiva apenas um estará ativo em um determinado momento.

Este modelo de aprendizagem é adequado para quando se deseja descobrir características que se destaquem podendo ser utilizadas para classificar um conjunto de padrão de entrada.

No modelo de aprendizagem competitiva (Figura 2.10) é válido ressaltar três elementos básicos descritos pelo psicólogo Rumelhart e pelo professor Zisper (1985):

- Um conjunto de neurônios que são todos iguais entre si, exceto por alguns pesos sinápticos distribuídos aleatoriamente, e que por isso respondem diferentemente a um dado conjunto de padrões de entrada.
- Um limite imposto sobre a “força” de cada neurônio.
- Um mecanismo que permite que o neurônio compita pelo direito de responder a um dado subconjunto de entradas, de forma que somente um neurônio de saída, ou somente um neurônio por grupo, esteja ativo (i.e., “ligado”) em um determinado instante. O neurônio que vence a competição é denominado um neurônio vencedor leva tudo.

Figura 2.9: Grafo de representação de aprendizagem competitiva simples.



Fonte: HAYKIN (2001, p.84).

2.9.5 Aprendizagem de Boltzmann

Em homenagem a Ludwig Boltzmann a regra de aprendizagem Boltzmann levou seu nome, sendo um algoritmo de aprendizagem que deriva de ideias já enraizadas na área de mecânica estatística.

Na máquina de Boltzmann os neurônios em sua estrutura recorrente operam de maneira binaria, uma vez que estão ligados representando o valor de +1, ou no estado -1 que é como fica quando se está desligado.

Em suma, a máquina de Boltzmann possui uma função de energia com o valor, E, e esse valor é determinado pelos estados ocupados pelos neurônios individuais da máquina, podendo ser mostrado pela seguinte fórmula:

$$E = -\frac{1}{2}\sum_{j}\sum_{\substack{k \\ j \neq k}} w_{kj}\, x_k x_j$$

2.9.6 Aprendizagem de Atribuição de Créditos

Simplificando esta aprendizagem, é o problema de se atribuir a culpa (crédito) por resultados globais de acordo com as decisões internas que a máquina de aprendizagem tomou e que no final tenha contribuído para a obtenção do resultado.

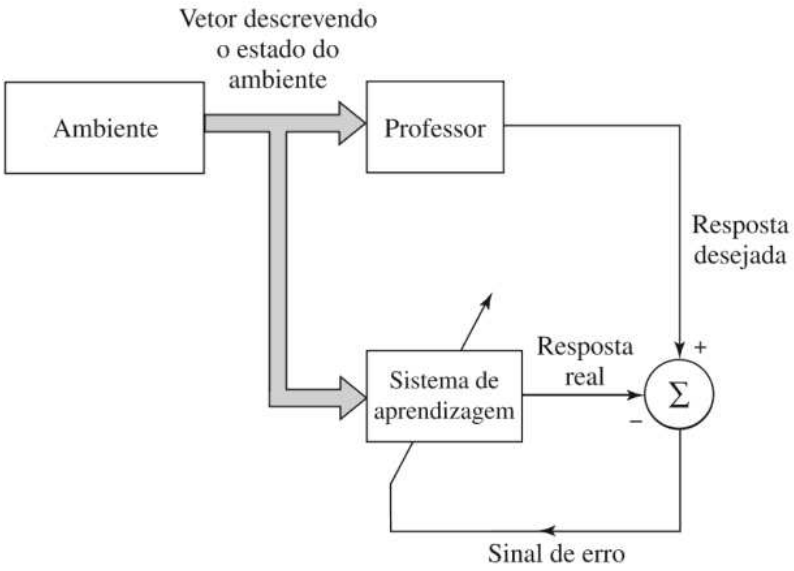
Em muitos casos os resultados dependem das decisões internas e eles são medidos por uma série de ações tomadas pela máquina, isto é, as decisões internas da máquina afetam diretamente os resultados gerais. Com base nessa explicação, o problema de atribuição de crédito pode ser dividido em dois subproblemas de acordo com Sutton (1984):

- Quando se tem a atribuição de crédito por resultados e ações, podemos denominar como problema de atribuição de crédito temporal, em resumo, se trata dos instantes de tempo quando as ações que realmente merecem o crédito são tomadas.
- Segundo subproblema é a atribuição de crédito por ações a decisões internas, mais conhecida como problema de atribuição de crédito estrutural, isto é, trata do fato de colocar crédito em estruturas internas das ações que o sistema gera.

2.9.7 Aprendizagem Supervisionada

Como o próprio nome diz é quando se temos um tutor/professor tem conhecimento, sendo representado por um conjunto de exemplos de entrada-saída sobre o ambiente, mas este ambiente é desconhecido pela rede neural trabalhada.

Figura 2.10: Diagrama de representação de aprendizagem com um professor.



Fonte: HAYKIN, 2001, p.88.

Sabendo que o professor possui o conhecimento, isto facilita para que ele forneça a resposta desejada para o vetor de treinamento, representando uma ação de sucesso para a rede neural. Pois esses parâmetros que a rede recebe são ajustados sob influência do vetor de treinamento e o sinal de erro.

2.9.8 Aprendizagem Não Supervisionada

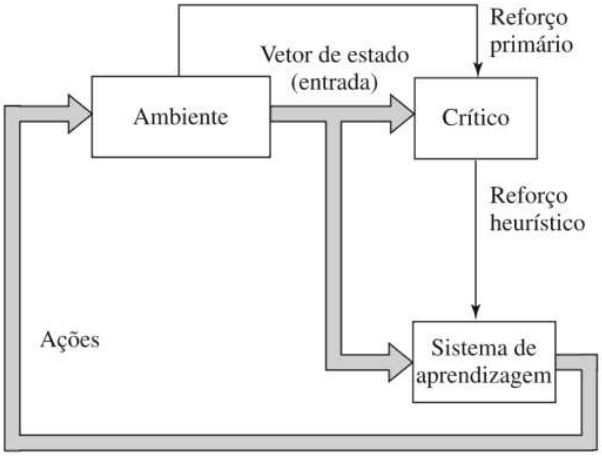
Como visto no tópico anterior, na aprendizagem supervisionada o processo de aprendizagem gira em torno de uma tutela. Já na aprendizagem não supervisionada como o nome já informa não há necessidade de um professor participar deste processo, isto é, não há vestígios de funções aprendidas pela rede que foram influenciadas.

Contudo, para contornar esta maneira de treinar a máquina pode-se incluir duas subdivisões que são trabalhadas: programação neurodinâmica e aprendizagem auto-organizada.

A primeira, comumente conhecida como aprendizagem por reforço, trata-se de um aprendizado mapeando a entrada-saída e é realizado pela interação contínua

com o ambiente, minimizando atingir o índice máximo de desempenho. Tendo como objetivo principal diminuir uma função de custo, definida como expectativa de custo cumulativo das ações tomadas ao longo de passos sequenciais, ao invés de um custo realizado de modo imediato. A Figura 2.12 ilustra o seu funcionamento.

Figura 2.11: Diagrama de representação de aprendizagem por reforço.

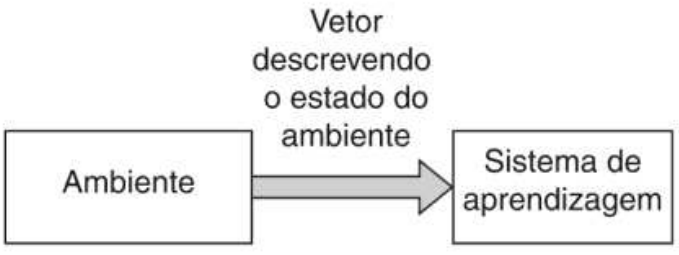


Fonte: HAYKIN, 2001, p.90.

Na aprendizagem auto-organizada não se tem ninguém para supervisionar o processo de aprendizado. Ao invés disso, são informadas condições para que sejam realizadas as tarefas de maneira independente de como uma rede deve aprender, os parâmetros recebidos são livres, porém otimizados em relação a esta medida.

Com a rede já ajustada aos dados de entrada, ela é capaz de desenvolver habilidades suficientes de formar representações para codificar as características de entrada fornecidas e criar classes automaticamente. A Figura 2.13 ilustra o seu funcionamento.

Figura 2.12: Diagrama de representação de aprendizagem não-supervisionada.

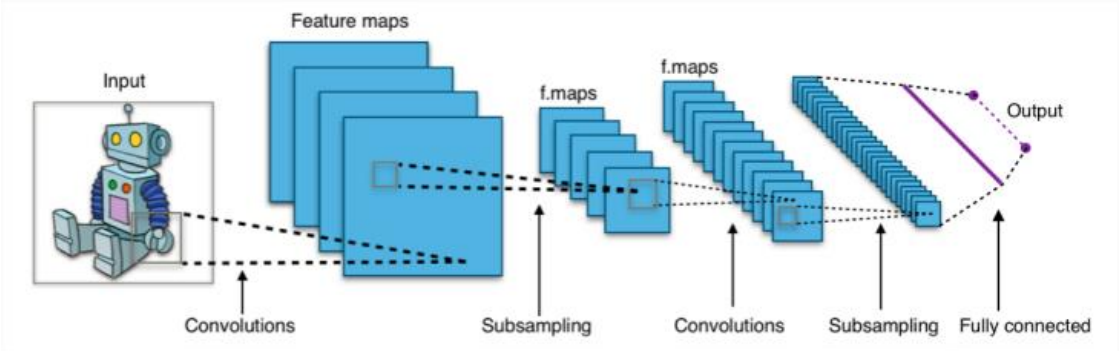


Fonte: HAYKIN, 2001, p.91.

2.10 Rede Neural Convolucional (CNN)

Uma CNN, pode ser classificada como a variação das redes de múltiplas camadas, por se tratar de uma inspiração do processo biológico de processamentos dos dados visuais. A rede neural convolucional hoje é uma das técnicas mais populares quando trabalhada em reconhecimento de imagem, mas atuam também na abordagem de Aprendizado Profundo (Deep Learning). O Deep Learning é comumente utilizado para que a máquina aprenda um tipo de representação para classifica e analisar os dados recebidos. A Figura 2.14 ilustra a arquitetura de uma CNN.

Figura 2.13: Arquitetura de uma CNN.



Fonte: Lama (2020, p.34).

Podem ser destacados dois pontos fundamentais para trabalhar com Aprendizado Profundo, sendo eles:

- Esse método funciona com múltiplos níveis de representação.
- São compostos por módulos simples e não lineares.

Importante destacar que as CNNs são um tipo específico de rede neural baseadas em algumas das variações das RNAs, são módulos feitos através de funções matemáticas simples, sendo trabalhados através de camadas de processamento. O aprendizado se dá em função da otimização dos números, enviando informações de um conjunto de treinamentos para que a tarefa informada seja realizada.

Seu desenvolvimento faz uso de pouco pré-processamento, tendo em vista outros algoritmos de classificação de imagem existentes. Resultando em um

aprendizado mais automatizado (com filtros), diferentemente dos algoritmos tradicionais que precisam ser alterados ou manipulados de maneira manual. Como descrito por Tianyi Liu (2015) possuir esse conhecimento prévio mais imparcial do esforço humano nos recursos ocasiona grandes vantagens.

Os algoritmos passam por mudanças, ao longo do tempo, e com a CNN não foi diferente, sua arquitetura evoluiu e com isso as redes passaram a terem análises de campos receptivos locais (Local Receptive Fields), isto é, o agrupamento em mapas é feito com neurônios da mesma camada.

Outra variável importante para a rede convolucional é o passo (stride), é o valor que representa quantos pixels que serão pulados em cada janela, informando o tamanho da camada seguinte desta mesma unidade.

3 MATERIAIS E MÉTODOS

3.1 Materiais

Como o projeto é voltado para a aprendizagem de máquina, todos os recursos utilizados são baseados em ferramentas de software, tais como:

3.1.1 Python

Python é uma linguagem de programação de código aberto (open source), com uma estrutura de dados com alto nível de eficiência e com abordagens simples para programação orientada a objetos (POO). Possuindo uma extensiva biblioteca padrão.

A aplicabilidade de Python é muito utilizada, para desenvolvimento web, gráficos numéricos, educação, desenvolvimento de software, aplicações comerciais e afins. Concentrando uma gama de documentações para uso dos programadores, o que facilita a compreensão da linguagem (PYTHON, 2022).

3.1.2 TensorFlow

Para uma abordagem mais rápida do machine learning, foi feito o uso de uma biblioteca de código aberto como o TensorFlow, pois nele foi trabalhado o treinamento de redes neurais para detecção e identificação de padrões, correlações de frequências sonoras.

Além disso, esta biblioteca oferece uma infinidade de soluções para adaptar o fluxo de trabalho, seja em preparação de dados, criação de modelos de ML, implementar modelos, seja em um dispositivo, navegador ou até mesmo na nuvem,

visto todos esses recursos, no final o uso foi imprescindível para a aplicação (TENSORFLOW).

3.1.3 Librosa

Librosa foi outra ferramenta que se fez necessária para a execução do programa, sendo ele um pacote Python para que sejam analisados áudios e músicas. Foi necessário seu uso pois, é possível trabalhar com dados de áudio e reconhecimento automático de fala.

Além disso, é fornecido os blocos de construção necessários para os que desejam criar sistemas de recuperação das informações dos sons enviados. Ajudando o programado a visualizar os sinais de áudios, efetuar extrações de recursos com diferentes técnicas de processamento de sinal (LIBROSA, 2022).

3.1.4 Clideo

Clideo é uma plataforma de edição online e de fácil uso muito utilizada para editar arquivos de vídeo, imagens, Gifs. Fazendo upload dos arquivos que queira editar, colocar filtros e outros recursos oferecidos pela ferramenta e no final das edições estará disponível a opção de baixar o arquivo para sua máquina no formato original do envio ou para outro padrão que assim desejar (CLIDEO, 2014).

3.1.5 Freesound

Freesound é um repositório que visa criar um banco de dados colaborativos, coletando e guardando amostras, trechos de áudio, gravações e outros. Além de ser gratuito, ele ainda oferece a vantagem de o usuário navegar pelas amostras incluídas por outras pessoas com palavras-chave, ouvir e até mesmo baixar os sons desejados e interagir com outros usuários se necessário (FREESOUND, 2013).

3.1.6 Visual Studio Code

O Visual Studio Code, mais conhecido como VSCode é um editor de código aberto que foi desenvolvido pela empresa Microsoft. Nesta ferramenta é possível criar códigos de software, efetuar testes e codificação, podendo customizar seja na aparência ou em suas funcionalidades, é uma ferramenta simples de se utilizar, possui várias funcionalidades e atalhos, sua arquitetura é algo bem planejado e possibilita também o programador desenvolver e publicar suas extensões (VISUAL STUDIO CODE, 2015).

3.2 Métodos

O desenvolvimento deste trabalho consistiu nas seguintes etapas:

- Elaboração de um banco de dados de latidos de cachorro o qual o tutor mapeou previamente seus significados.
- Investigar os melhores métodos que podemos usar para receber áudio e sua futura transformação para aplicação em técnicas de aprendizado de máquina.
- Aplicar métodos computacionais para obtenção do espectro de áudio final e comparar padrões dentro da base já fornecida pelo tutor.
- Criar modelos de classificadores baseados em técnicas e aprendizado de máquina.

4 SOLUÇÃO PROPOSTA

Para a realização desse trabalho foi necessário, inicialmente, entender e compreender o que o animal no qual foi estudado estava querendo transmitir ao seu dono. Analisando que o latido do cachorro possuía alguns padrões para determinadas ações, foi realizado a coleta dos áudios utilizando o gravador de áudios do WhatsApp que, por sua vez, era gravado no formato .ogg.

Foi necessário realizar uma conversão de formato, já que as ferramentas que foram estudadas e o formato inicialmente dos áudios não eram compatíveis. Utilizamos um site de conversão e edição de arquivos de áudio chamado Clideo, e foi realizar a conversão do arquivo original para o formato .wav.

Como os arquivos originais são arquivos grandes, para facilitar o processamento e a aplicação da técnica de janelamento, realizou-se, novamente, a edição do áudio recortando o arquivo original no formato .wav em arquivos menores. Para não se perder a origem do trecho recordado do áudio, realizamos o upload desses arquivos originais no site freesound (um repositório gratuito de áudio). Com isso o “id” do áudio no site foi utilizado como chave para se lembrar qual o arquivo original.

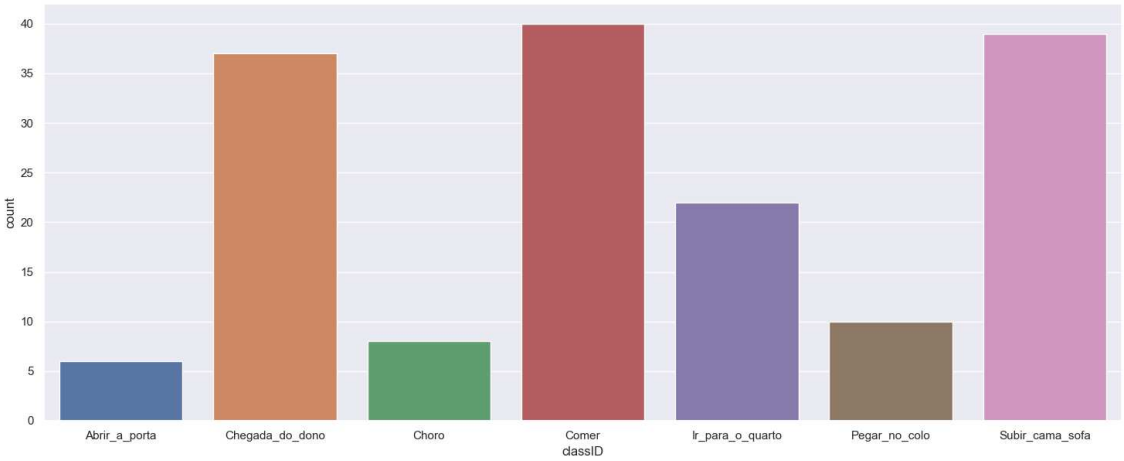
4.1 Classificação dos Áudios

Já com a base de áudios no formato correto, realizamos a classificação de cada áudio de acordo com o entendimento do tutor e foi separado em 7 classes:

- 0 = Choro.
- 1 = Abrir_a_porta.
- 2 = Chegada_do_dono
- 3 = Comer.
- 4 = Ir_para_o_quarto.
- 5 = Subir_cama_sofa
- 6 = Pegar_no_colo.

No gráfico apresentado da Figura 4.1, é possível visualizar a distribuição dos áudios em cada classe.

Figura 4.1: Distribuição dos áudios em suas classes.



Fonte: Os autores.

4.2 Separação e Nomeação dos Arquivos

A fim de classificar e separar os metadados de forma perfeita, realizou-se mais duas separações para nomear cada arquivo:

- ID do Freesound: um identificador numérico para informar o ID de registro do áudio original no site do Freesound;
- ID da Class: um identificador numérico para informar qual é a classe pertencente do áudio;
- ID de ocorrência: um identificador numérico para distinguir diferentes ocorrências do som dentro da gravação original;
- ID do fatiamento: um identificador numérico para distinguir diferentes fatias tiradas da mesma ocorrência;

Com isso, os áudios ficaram segmentados e separados da seguinte maneira:

[fsID]-[classID]-[occurrenceID]-[sliceID].wav

E, por fim, na parte da base de dados, foi realizada a criação de um arquivo metadados, contendo as informações dos áudios separados pelas seguintes colunas:

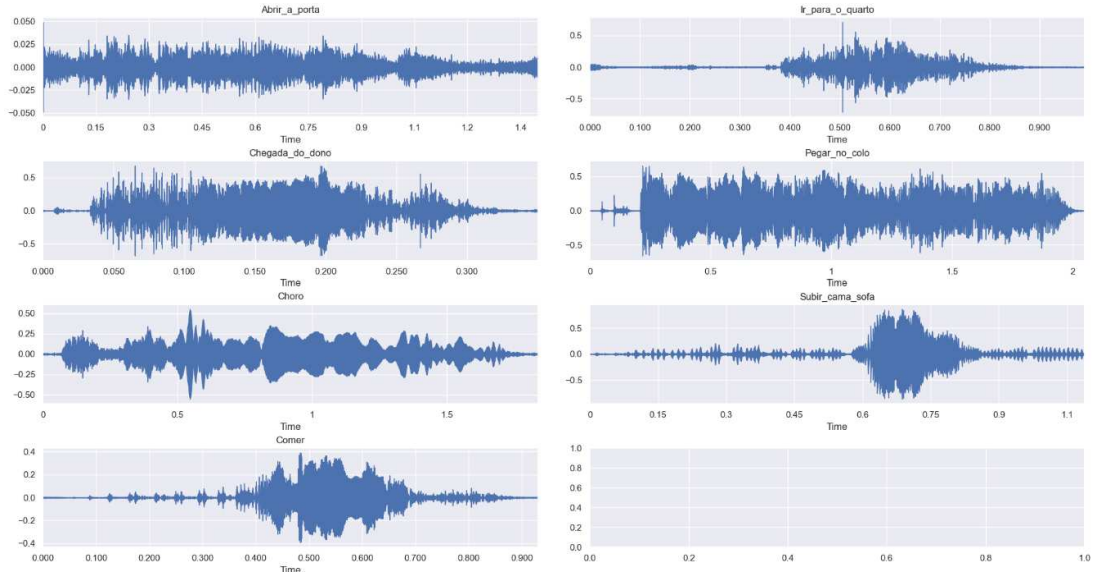
- slice_file_name: Nome final do arquivo.
- fsID: id do site freesound.
- start: tempo do início do áudio editado em relação ao arquivo original.
- end: tempo final do áudio editado em relação ao arquivo original.
- salience: uma classificação de saliência do som. 1 = primeiro plano, 2 = plano de fundo.
- Folder: o número da pasta (1-7) ao qual este arquivo foi alocado.
- classID: O id da classe
- class: descrição da classe

Com os metadados já criado e separados de forma que o algoritmo consiga realizar a diferenciação, iniciamos a programação.

4.3 Identificação dos Áudios

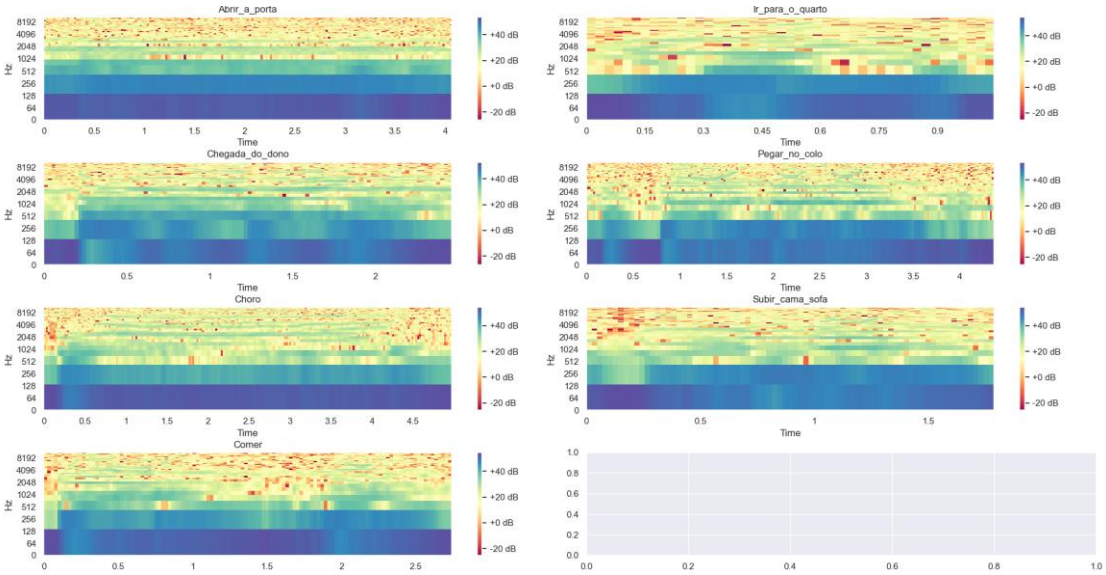
Foi realizada a leitura do arquivo metadado e salvo de forma que foi possível identificar cada arquivo e sua localização. Foi aplicado então a técnica do MFCC em todos os arquivos de áudios, extraindo recursos/características MFCCs de cada arquivo de áudio do dataset features. As Figuras 4.2 e 4.3 ilustram, respectivamente, exemplos de espectros de alguns áudios e de espectrogramas de MFCCs.

Figura 4.2: Espectro de alguns áudios aleatórios e suas classes.



Fonte: Os autores.

Figura 4.3: Espectrogramas de MFCCs.



Fonte: Os autores.

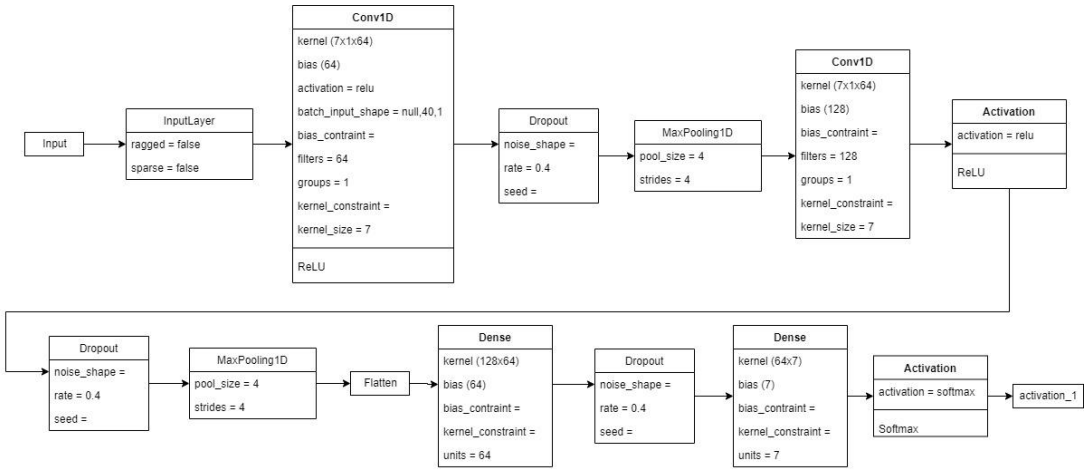
Após a realização da transformação, foi realizado a separação da base completa entre atributos classe(class) e atributos previsores(features). Para se obter uma base para o treinamento do modelo, os dados foram separados para treinar nossa máquina e dados para testar. A princípio, para iniciar o treinamento, foi divido então a base em 80% para treino e 20% para teste e validação.

4.4 Estrutura da RNA

Foi criado em seguida o modelo e estrutura da RNA, utilizando especificamente Redes Neurais Convolucionais (CNN), conforme ilustrado na Figura 4.4. A arquitetura desta rede neural foi definida com base em alguns testes realizados para obter o resultado esperado. A estrutura pode ser ajustada livremente e comparada aos resultados desta estrutura:

- Sequential: é a classe para criar a rede neural, pois uma rede neural nada mais é que uma sequência de camadas (camada e entrada, camadas ocultas, camada de saída).
- kernel_size: o tamanho do kernel (matriz) de convolução.
- activation: função de ativação.
- input_shape: na primeira camada este é o tamanho dos dados de entrada.
- Camada MaxPooling1D: que vai fazer a extração das características principais.
- Camada Conv1d: uma rede neural convolucional que realiza a convolução ao longo de apenas uma dimensão.
- Camada Flatten: para transformar de matriz em vetor.
- Camada Dense: quando um neurônio de uma camada está ligado a todas os outros neurônios das outras camadas.
- Dropout: técnica de regularização para diminuir o overfitting
- padding='same': indica que adicionamos uma nova coluna composta por somente 0 (zeros).

Figura 4.4 - Estrutura da rede neural elaborada.



Fonte: Os autores.

5 TESTES E RESULTADOS

Após a criação do modelo da rede neural, iniciou-se o treinamento da RNA. Conforme mencionado, anteriormente, foi dividida a base em 80% para treino e 20% para teste e validação.

Foram definidas, então, algumas variáveis:

- num_epochs = 80, número de épocas de treinamento.
- num_batch_size = 32, isto indica que vamos enviar de 32 em 32 recursos de áudio.
- ModelCheckpoint: para salvar o modelo enquanto faz o treinamento.
- Filepath: caminho onde será salvo o modelo. Para isto temos uma pasta nos arquivos chamado saved_models.
- verbose: mostrar mensagens enquanto a RNA é treinada
- save_best_only = True: para salvar o modelo somente quando houver uma melhora no resultado

Variáveis para efetuar a contagem do tempo de treinamento:

- start: pegando o horário atual de início do treinamento;
- duration: ao final do treinamento, subtrair a hora atual com hora de início do treinamento.
- model_history: para armazenar o histórico de treinamento.
- model.fit: para fazer o ajuste dos pesos ao longo do treinamento
- X_train, Y_train: dados de treinamento
- batch_size = num_batch_size: que definimos acima
- epochs = num_epochs: que também definimos acima
- validation_data= (X_test, Y_test): dados de teste para monitorarmos como está o percentual de acerto da rede neural a cada época
- callbacks=[checkpointer]: checkpointer definido anteriormente
- verbose = 1: para mostrar as mensagens

No início dos treinamentos foram obtidos os resultados apresentados na Tabela 5.1.

Tabela 5.1: Resultados preliminares da RNA.

Classe	Precision	Recall	f1-score	support
Abrir_a_porta	1.00	1.00	1.00	2
Chegada_do_dono	1.00	1.00	1.00	2
Choro	1.00	1.00	1.00	1
Comer	0.50	1.00	0.67	1
Ir_para_o_quarto	0.67	0.67	0.67	3
Pegar_no_colo	1.00	0.33	0.50	3
Subir_cama_sofa	0.60	0.75	0.67	4

Em uma segunda etapa, foi alterada a porcentagem de distribuição entre treinamento e validação, os novos parâmetros foram 40% para treinamento e 60% para teste e validação. Os resultados são apresentados na Tabela 5.2.

Tabela 5.2: Resultados com as distribuições alteradas.

Classe	Precision	Recall	f1-score	support
Abrir_a_porta	1.00	1.00	1.00	4
Chegada_do_dono	1.00	1.00	1.00	11
Choro	1.00	1.00	1.00	1
Comer	0.93	0.93	0.93	14
Ir_para_o_quarto	1.00	0.57	0.73	7
Pegar_no_colo	1.00	0.17	0.29	6
Subir_cama_sofa	0.73	1.00	0.84	6

Desta maneira consegue-se obter o melhor modelo de classificação, já que, com poucas épocas de treinamento se obteve, em média, os seguintes resultados:

- Treinamento

Acuracidade – 1.0000 = 100%

Perda – 1.6957e-04 = 0,016%

- Validação

Acuracidade – 0,9388 = 93,88%

Perda – 0,6964 = 6,1263%

- Teste

Acuracidade – 0,8163 = 81,63%

Perda – 0,1839 = 18,39%

6 CONSIDERAÇÕES FINAIS

Após toda a pesquisa realizada, pode-se concluir que o presente projeto conseguiu realizar o seu objetivo, no qual é o desenvolvimento de uma ferramenta capaz de classificar os padrões de latidos de um cachorro com a base já pré-classificada.

Durante toda a pesquisa, foi avaliado qual seria a melhor maneira de se obter tais resultados. Após toda a elaboração da base, como iriamos separar os áudios, de forma que separássemos em treinamento, teste e validação de forma aleatória, a forma na qual iriamos treinar a rede neural, conseguimos encontramos qual seria o melhor modelo.

Durante os testes obteve-se como resultado cerca de 90% de acuracidade em acertos em relação a novos áudios. Devido a quantidade de áudios para o treinamento da RNA ser relativamente baixa, cerca de 160 áudios, não foi possível obter resultados melhores. Porém. Acredita-se que, com essa porcentagem de acuracidade, pode-se concluir que o modelo elaborado tem capacidade de classificação muito boa e satisfez a resolução do nosso problema.

Portanto, os resultados sugerem que o modelo e o algoritmo elaborados têm capacidade de classificar o latido do animal que foi estudado, podendo ter uma base completa ou não, teremos sempre um nível alto de sucesso para aprendizado e para teste.

Para trabalhos futuros propõe-se:

- Analisar outros casos e tipos de latidos de cães, já que o modelo pensado tem a capacidade de se treinar de forma rápida e precisa, desde que se tenha uma base já pré-classificada desse outro possível caso.
- Possibilidade de criação de algum aplicativo ou até mesmo uma coleira já com a RNA introduzida nela, realizando, desta forma, a classificação do latido para a linguagem humana em tempo real.

7 REFERÊNCIAS

CLIDEO. **Ferramentas de Vídeo Online**. 2014. Disponível em: <https://clideo.com/pt>. Acesso em: 13 ago. 2022.

COVO, C. **Modelagem Matemática e Computacional de Efeitos em Ondas Sonoras**. Orientador: Paula Couto. 2016. Dissertação (Mestrado Profissional em Matemática) - Universidade Federal do Paraná, Curitiba, 2016. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/69650/R%20-%20D%20-%20CARLOS%20CESAR%20DE%20CARVALHO%20COVO.pdf?sequence=1&isAllowed=y>. Acesso em: 24 out. 2022.

FECHINE, J. M. **A Transformada de Fourier e Suas Aplicações**. UFCG, 21 maio 2010. Disponível em: <http://www.dsc.ufcg.edu.br/~pet/ciclo_seminarios/tecnicos/2010/TransformadaDeFourier.pdf>. Acesso em: 22 ago. 2022.

FREESOUND. **Freesound**. 2012. Disponível em: <https://freesound.org/>. Acesso em: 13 ago. 2022.

JESUS, M. **Análise de Áudio de Voz Para Identificação do Emissor Utilizando Técnicas de Processamento de Sinais e Redes Neurais Artificiais**. Orientador: Max de Oliveira. 2021. Monografia de conclusão de curso (Graduação em Ciência da Computação) - Pontifícia Universidade Católica de Goiás, Goiânia, 2021. Disponível em: <https://repositorio.pucgoias.edu.br/jspui/bitstream/123456789/3541/1/TCC-Maria_Regina-2021.pdf>. Acesso em: 15 set. 2022.

LIBROSA. **Librosa 0.8.1 documentation**. 2013 - 2022 Disponível em: <http://librosa.org/doc/0.8.1/index.html>. Acesso em: 13 dez. 2022.

LOPES, R. R. **Processamento Digital de Sinais**. 1. [S. l.], 15 fev. 2007. Disponível em: <https://www.decom.fee.unicamp.br/~rlopes/EA614/masterea614.pdf>. Acesso em: 24 out. 2022.

MANARO, A.; LARCO, L.; LOPES, T. **I.A. embarcada classificadora de sons para deficientes auditivos**. 2021. Trabalho de Conclusão de Curso (Bacharel em Engenharia de Controle e Automação) - Centro Universitário do Instituto Mauá de

Tecnologia, São Caetano do Sul, 2021. Disponível em: <https://repositorio.maua.br/bitstream/handle/MAUA/269/AUGUSTO%20MANARO.pdf?sequence=1&isAllowed=y>. Acesso em: 24 out. 2022.

MICROSOFT. **Multimedia Programming Interface and Data Specifications 1.0**. 1991. Disponível em: <https://www.aelius.com/njh/wavemetatools/doc/riffmci.pdf>. Acesso em: 22 ago. 2022.

MICROSOFT. **Visual Studio Code**. 2016. Disponível em: <https://code.visualstudio.com/>. Acesso em: 15 set. 2022.

MOECKE, Marcos. **Princípios de Sistemas de Telecomunicações: Unidades de medidas logarítmicas em telecomunicações**. São José - SC: Instituto Federal de Santa Catarina, 2006. Disponível em: <https://wiki.sj.ifsc.edu.br/wiki/images/1/1d/Apostila_de_PRT_2014-1_(Material_Professores_Saul-Moecke).pdf>. Acesso em: 25 set. 2022.

MUNIZ, D. **Estudo Sobre Reconhecimento de Áudio Repetitivo: Desenvolvimento de um Protótipo**. 2009. Monografia de conclusão de curso (Tecnólogo em Sistemas de Telecomunicações) - Instituto Federal de Santa Catarina, 2009. Disponível em: <https://wiki.sj.ifsc.edu.br/images/2/24/ProjetoFinal_DaianaNascimentMuniz.pdf>. Acesso em: 15 set. 2022.

NAIR, P. **The dummy's guide to MFCC**. 24 jul. 2018. Disponível em: <https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd>. Acesso em: 25 set. 2022.

POSITIVO. **MP3, FLAC, WAV: quais são os principais formatos digitais de música e qual opção eu devo escolher**. Disponível em: <https://www.meupositivo.com.br/doseujeito/dicas/principais-formatos-musica-digital-mp3-flac-wav/>. Acesso em: 30 jul. 2022.

PYTHON. **Applications for Python**. 2022. Disponível em: <https://docs.python.org/pt-br/3/tutorial/index.html>. Acesso em: 30 jul. 2022.

TENSORFLOW. **Criar modelos de machine learning no nível de produção com o TensorFlow**. Disponível em: <https://www.tensorflow.org/?hl=pt-br>. Acesso em: 5 ago. 2022.

XIPH. **Vorbis I specification**. 4 jul. 2020. Disponível em: <https://xiph.org/vorbis/doc/Vorbis_I_spec.pdf>. Acesso em: 30 jul. 2022.

ZOTTESSO, R. **Identificação de espécies de pássaros utilizando espectrogramas e dissimilaridade**. 2017. Dissertação (Mestre em Ciência da Computação) - Universidade Estadual de Maringá, Maringá, 2017. Disponível em: <http://repositorio.uem.br:8080/jspui/bitstream/1/2528/1/000227789.pdf>. Acesso em: 10 out. 2022.

8 BIBLIOGRAFIA CONSULTADA

ABNT. **Quantidades, unidades e símbolos das grandezas acústicas fundamentais**. Rio de Janeiro, v.7, n.45, p. 50-72, jul./ago. 1959.

LACHAMBRE, H.; André-Obrecht., R; Pinquier, J. **Singing voice characterization for audio indexing**, 2007 15th European Signal Processing Conference, 2007, pp. 1536-1540.

HALLIDAY, D.; RESNICK, R.; WALKER, J. **Fundamentos de Física: Gravitação, Ondas e Termodinâmica**. 9ª edição. ed. [S. l.]: LTC, 2012. 328 p. v. Volume 2. ISBN 9788521619048.

HAYKIN, S. **Redes Neurais Artificiais: Princípios e Práticas**. 2ª. ed. Porto Alegre: Bookman, 2001. 900 p. ISBN 0132733501.

HOROWITZ, A. **A cabeça do cachorro: o que seu amigo mais leal vê, fareja, pensa e sente**. 1. ed. [S. l.]: BEST SELLER, 2010.

KOUTROUMBAS, K.; THEODORIDIS, S. **Pattern Recognition**. 4. Ed., Nova Iorque: Academic Press, 2008.

KOVÁCS, Z. **Redes Neurais Artificiais: Fundamentos e Aplicações**. 4ª. ed. São Paulo: Ed. Livraria da Física, 2002. 174 p. ISBN 8588325144.

NUSSENZVEIG, H. M. **Curso de física básica, 1: Mecânica**. São Paulo: E. Blucher, 2013.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2013.

TEIXEIRA, J. (2019). O que é inteligência artificial (3a ed.). e-galáxia.

WIBAWA, I. D. G. Y. A.; DARMAWAN, I. D. M. B. A. **Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini**. Journal of Physics: Conference Series, v. 1722, n. 1, p. 012014, 1 jan. 2021.