Classificação de Eventos Sonoros em Ambientes Assistidos: um Estudo Experimental para Detecção de Tosse

Caio Souza de Oliveira

JUIZ DE FORA

JULHO, 2023

Classificação de Eventos Sonoros em Ambientes Assistidos: um Estudo Experimental para Detecção de Tosse

Caio Souza de Oliveira

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA

JULHO, 2023

Classificação de Eventos Sonoros em Ambientes Assistidos: um Estudo Experimental para Detecção de Tosse

Caio Souza de Oliveira

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação pela UFRJ

Heder Soares Bernardino
Doutor em Modelagem Computacional pelo LNCC

Luciana Brugiolo Gonçalves
Doutora em Ciência da Computação pela UFF

JUIZ DE FORA
13 DE JULHO, 2023

Aos pais, pelo apoio e sustento.

A minha irmã e meus amigos.

Resumo

Ambientes assistidos são capazes de monitorar e facilitar a vida de seu residente, por meio de um ecossistema composto por uma rede de sensores e diversos dispositivos IoT. Com foco em sensores de captura de áudio, este trabalho almeja avaliar a capacidade que modelos de aprendizagem de máquina possuem para lidar com a classificação de eventos sonoros registrados por tais sensores. Para tal, foi realizada uma comparação de dois modelos de redes neurais artificiais, neste contexto: Convolucional e Long Short-Term Memory. Utilizando Coeficientes Cepstrais de Frequência-Mel como atributos de dados acústicos, que são produtos de uma técnica estabelecida como a melhor para extração de atributos de áudios. Os modelos foram testados com o conjunto de dados focado para tosses, COUGHVID, e alguns dados de instrumentos, para serem o outro rótulo de classificação, com resultados que apresentaram acurácias acima de 80% e bons F1-Scores para ambos modelos, dado que estão próximos do valor máximo na escala.

Palavras-chave: Ambientes assistidos, classificação de áudio, aprendizado de máquina.

Abstract

Assisted environments are capable of monitoring and facilitating the life of their resident, through an ecosystem composed of a network of sensors and various IoT devices. Focusing on audio capture sensors, this work aims to evaluate the capacity that machine learning models have to deal with the classification of sound events recorded by such sensors. For this, a comparison of two models of artificial neural networks was carried out, in this context: Convolutional and Long Short-Term Memory. Using Mel Frequency Cepstral Coefficients as acoustic data features, which are products of a technique established as the state-of-art for extracting audio features. The models were tested with the dataset focused on coughs, COUGHVID, and some instrument data, to be the other classification label, with results that showed accuracies above 80% and good F1-Scores for both models, since they are close to the maximum value at the scale.

Keywords: Assisted ambient living, audio classification, machine learning.

Agradecimentos

Aos meus pais, Ana Cláudia e Marcelo, pelo investimento na minha educação, assim como o apoio e incentivo no meu sonho de graduação.

A minha irmã pelo encorajamento e amizade.

Ao professor Victor pela orientação, amizade e pela paciência, durante a construção deste trabalho.

Aos amigos que tornaram momentos difíceis em momentos descontraídos durante a graduação.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que contribuíram de algum modo para o enriquecimento pessoal e profissional.

"O ontem é história, o amanhã é um mistério, mas o hoje é uma dádiva. É por isso que se chama presente".

Mestre Oogway (Kung Fu Panda)

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AAL	Ambient Assisted Living
AmI	Ambient Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DCC	Departamento de Ciência da Computação
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
GPGPU	General Purpose Graphics Processing Unit
IoT	Internet of Things
kNN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
MFCC	Mel Frequency Cepstral Coefficients
MMED	Max-margin Early Event Detectors
OMED	Online Framework with Max-margin Early Event Detection
RNN	Recurrent Neural Network
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

Devido a um grande avanço na área de saúde, a população idosa aumentou muito nos últimos anos e continuará aumentando com o passar dos anos, de acordo com o Prospecto da População Mundial das Nações Unidas[1].

As pessoas desta faixa etária, acima de 65 anos, necessitam de muita atenção e cuidados, por terem organismos mais frágeis que jovens e adultos, sendo expostos a diversos riscos. Algumas doenças são bem frequentes na população idosa, como Artrite, Alzheimer, Diabetes e Hipertensão[2].

Sendo assim, idosos precisam cada vez mais de acompanhamento em suas atividades diárias ou em tratamentos. Casas de repouso são uma excelente alternativa para tal problema, cuidando e monitorando os residentes, tanto em suas necessidades de saúde quanto em atividades básicas como alimentação e lazer.

Entretanto, existem alguns idosos que preferem preservar suas privacidades e suas liberdades, morando em sua própria residência de forma independente. Em Navarro et al. (2018), os autores acreditam também que há um motivo econômico, para que governos incentivem que tais pessoas continuem morando por conta própria. Nesse caso, tem-se a opção de se ter um acompanhante presente em sua casa, em determinados períodos do dia, para auxiliar em suas tarefas e poder cuidar de tal pessoa.

Mesmo com a presença de um acompanhante em sua moradia, não é garantido que o idoso estará sempre seguro, pois o mesmo pode sofrer de algum evento em um momento em que o acompanhante esteja ausente. Com isso, é apresentado o conceito de Moradias Assistidas (em inglês Ambient Assisted Living, AAL), definido por Patel e Shah (2019) como: "a junção de dois sistemas *smart*, a *smart home* e *smart health*, de forma que idosos possam viver de forma independente em suas casas".

Trata-se então, de um ecossistema composto por diversos aparelhos e sensores, em uma residência, que visam monitorar a saúde de quem o habita, para ajudar e fazer

[1] ⟨https://population.un.org/wpp/⟩

[2] ⟨https://www.ncoa.org/article/the-top-10-most-common-chronic-conditions-in-older-adults⟩

o acompanhamento do idoso em sua própria casa. Câmeras de vídeo, microfones para captação de áudio, sensores de movimento e dispositivos vestíveis, como *smartwatches* e *smartbands*, são exemplos de aparelhos e sensores que compõem este sistema AAL.

Dessa forma, o tema do trabalho é detectar, através de sensores de áudio em um AAL, eventos fora do padrão comportamental do idoso,e prever o que pode significar esse comportamento anormal no futuro.

Um AAL está sempre gerando dados, por se tratar de uma rede de diversos dispositivos IoT. Neste sentido, surgem problemas que se relacionam diretamente a preocupações quanto à transmissão, armazenamento, processamento e segurança dos dados (JANJUA et al., 2019). A preocupação com relação aos dados é um dos vários pontos que devem ser considerados quando se busca trabalhar com reconhecimento de atividades em AAL.

Os autores em Patel e Shah (2019) fazem um levantamento dos desafios mais frequentes encontrados na literatura, para ajudar na melhor avaliação de usabilidade e eficiência de cada abordagem. Um dos principais desafios, senão o principal, é a privacidade do usuário, dada a necessidade de utilização de dispositivos externos para monitorar as atividades do usuário, como câmeras, sensores de áudio, sensores vestíveis, etc. Com ênfase para sensores de áudio, como gravadores e microfones, que são os principais captadores que geram os dados sonoros, os quais este trabalho busca tratar. Estes são considerados invasivos e a escolha dos mesmos é dependente da aceitação por parte do usuário, morador do AAL (PANICO et al., 2020).

A detecção de eventos fora do padrão também foi identificada como um dos desafios mais comuns. Esta se torna muito complexa de se detectar, uma vez que comportamento humano não é consistente e periódico, sendo assim imprevisível (PATEL; SHAH, 2020) necessitando de uma implementação mais flexível a essa característica de atividades e comportamentos variantes.

Com esses, foram identificados outros 15 desafios, como o número de residentes, atividades paralelas, conjunto de dados de treinamento para métodos que necessitem de treinamento, entre outros (PATEL; SHAH, 2019).

A detecção de anomalias é importante para auxiliar no reconhecimento de si-

tuações de perigo em potencial (JANJUA et al., 2019).

Tendo em vista que os ambientes AAL são voltados para pessoas idosas que estão expostas a diversas situações que colocam sua saúde em risco, este trabalho busca atuar na detecção de eventos sonoros gerados pelo usuário, a fim de identificar eventos sonoros que possam representar risco a saúde ou bem-estar do mesmo.

Dessa forma, com os sons registrados pelo sistema, através dos dispositivos captadores, será possível fazer a rotulação dos áudios, de acordo com o que o mesmo representa. Partindo da identificação destes eventos sonoros, será possível verificar os acontecimentos que se deram em um ambiente de moradia assistida através de relatórios. Esses relatórios visam auxiliar os cuidadores, equipe médica de apoio e familiares no monitoramento dos residentes, para determinar em qual momento os mesmos começaram apresentar mal-estar. Um exemplo é a possibilidade de demarcar quando um usuário começou a dar sinais como tosses, olhando para os relatórios de saída, de modo a assistir no entendimento do que pode ter causado tal condição.

O presente trabalho tem como objetivo geral averiguar a aplicação de modelos de aprendizado de máquina em um contexto de rotulação de dados sonoros.

O processo de classificação de dados acústicos depende de registros de áudio, capturados pelos sensores, produzidos pelo usuário ou o próprio ambiente do sistema. Tais dados acústicos são classificados de acordo com o modelo de aprendizado de máquina selecionado, possibilitando definir o que ocorreu dentro do ambiente assistido. A classificação feita por modelos de aprendizado de máquina necessita de atributos a respeito dos dados e a classificação de dados sonoros não é diferente, dessa maneira é preciso um método de extração de atributos em cima desses dados.

Assim, para este trabalho, temos os seguintes objetivos específicos: (i) extrair os atributos dos dados sonoros a fim de serem usados nos modelos de aprendizado de máquina, (ii) investigar a capacidade de diferenciação dos sons dos modelos e (iii) analisar a performance dos modelos neste contexto acústico.

Em um primeiro momento é importante verificar na literatura como se encontra o cenário de classificação de áudios em ambientes assistidos, para melhor entender quais são as técnicas de aprendizado de máquina mais utilizadas, e as principais abordagens e

desafios quando se busca trabalhar neste tema.

A partir do mapeamento sistemático da literatura em (PATEL; SHAH, 2019), é possível entender quais são os principais desafios e pontos de atenção quando se busca trabalhar com análise de dados, detecção de eventos, no cenário de ambientes assistidos. O mapeamento apresenta 17 pontos que vão da infraestrutura, *Hardware*, ao processamento dos dados. Considerando, os principais desafios referentes à análise de dados listados pelos autores, pretende-se implementar uma abordagem para a rotulação de eventos sonoros, em um ambiente de moradia assistido.

Assim, a metodologia se inicia por meio de uma coleta de dados, gerados por sensores e outros dispositivos IoT. Após a coleta, os dados brutos devem ser pré-processados, ou seja, devem passar por uma etapa de limpeza, transformação, para estarem de acordo com os requisitos dos modelos de machine learning. Tais dados pré-processados devem passar por uma etapa de extração de atributos acústicos, para estarem prontos para serem usados nos modelos. Com os dados prontos, uma parte destes é usada para realizar o treinamento dos modelos de aprendizado de máquina e outra parte é destinada à validação, experimentação dos modelos.

Importante ressaltar que este trabalho busca uma experimentação para estabelecer um modelo de aprendizado de máquina capaz de detectar eventos de tosse para auxiliar pessoas em ambientes domiciliares assistidos. Assim, seguindo esses passos, pretende-se obter resultados para determinar qual modelo é mais adequado para a continuidade desta pesquisa em trabalhos futuros, assim como quais pontos de melhoria devem ser aplicados.

O restante deste trabalho está organizado da seguinte forma: O Capítulo 2 apresenta as definições de Ambientes Assistidos, Internet das Coisas e Detecções de eventos. Em seguida, o Capítulo 3 realiza uma revisão na literatura sobre o contexto de Detecção de Eventos em Ambientes Assistidos. O Capítulo 4 detalha o processo de Classificação de Dados Sonoros, o conjunto de dados, os atributos, definição dos modelos e seus resultados. Por fim, o Capítulo 5 apresenta as considerações finais a respeito de trabalho assim como possibilidades de trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta uma exposição de conceitos necessários para melhor entender o cenário de um sistema de ambiente assistido. São apresentados conceitos sobre os próprios ambientes de moradia assistida e uma visão de internet das coisas, para entender os conceitos de sensores e dispositivos que estão presentes nestes ambientes. Também são apresentados ideias e trabalhos que processam e extraem conhecimento por meio dos dados que estes sistemas de ambientes assistidos produzem.

2.1 Ambientes Assistidos

Os Ambientes Assistidos (AAL) são definidos como uma composição de sistemas técnicos que dão suporte às pessoas idosas e com necessidades especiais em suas rotinas diárias (DOHR et al., 2010). Tais sistemas têm o objetivo de fomentar a autonomia destas pessoas em suas próprias residências, assim atingindo diversos benefícios como segurança, bem-estar e uma melhor qualidade de vida. Estes sistemas são aplicados para auxiliar o dia-a-dia das pessoas, com as mudanças físicas e psicológicas decorridas no processo de envelhecimento (PANICO et al., 2020).

Dohr et al. (2010) apontam que a necessidade da utilização de tecnologias como essa surge em virtude da mudança demográfica em países industrializados, pois estes apresentam uma taxa de natalidade em declínio, enquanto a expectativa de vida aumenta. Em seus estudos, Navarro et al. (2018) acreditam que existe um incentivo por parte de governos, com propósito econômico, para que idosos vivam de maneira independente, reduzindo a uma necessidade mínima de serviços de saúde.

Diversas tecnologias relacionadas à saúde podem ser usadas no ecossistema, com ponto focal no residente, variando de monitoramento físico de saúde (e.g., batimentos cardíacos, tosse, temperatura corporal), monitoramento de interação com o espaço (e.g., quedas), monitoramento de interação social (e.g., conversa ao telefone, visitas), e também de assistência cognitiva (e.g., como lembretes de medicação). O monitoramento também

pode incluir o próprio ambiente, verificando princípios de incêndio ou enchente (NAVARRO et al., 2018).

Um conceito muito importante para AAL, é o de Ambiente Inteligente (AmI). Trata-se da paradigmas que buscam naturalizar as interações entre humanos e o seu ambiente, de forma ubíqua (CEDILLO et al., 2018; DOHR et al., 2010; COSTA et al., 2009). Assim, a casa se torna inteligente com a ajuda de dispositivos *smart*, na visão de Ambiente Inteligente.

Dohr et al. (2010) citam as principais necessidades do público idoso e algumas soluções, quando se utiliza um AAL, através destes aparelhos IoT:

- Saúde: monitoramento de doenças crônicas;
- Segurança: sistema de alarme;
- Independência: serviços de agenda e lembretes; e
- Contato Social: aplicativos de comunicação para parentes e amigos.

Todos estes pontos são atendidos com as características de conectividade, adaptabilidade e antecipação que um ambiente assistido proporciona. Para isso, AAL pode ser estruturado da seguinte forma (DOHR et al., 2010):

- *Hardware*: consistindo de diversos dispositivos IoT, sensores e as redes sem fio;
- *Middleware*: onde são feitos os devidos tratamentos dos dados, desde a captura à transmissão para o próximo nível; e
- Serviços: responsáveis pelo processamento de dados e extração de conhecimento a partir dos dados capturados.

Esta é uma definição básica de uma infraestrutura que serve de referência para as diversas propostas de AAL presente na literatura. Nas camadas de *Hardware* e *Middleware*, os dados, referentes ao ambiente e usuário, são gerados, tratados e transmitidos para o nível de análise. Na camada de Serviços é onde as informações consolidadas são divulgadas para médicos, assistentes ou familiares do residente, para que decisões acerca do bem-estar deste possam ser tomadas.

2.2 Internet das Coisas

Como foi descrito na seção anterior, os dispositivos IoT são responsáveis pela captura dos dados na residência do usuário. Dohr et al. (2010) denominam estes dispositivos como "objetos inteligentes", pois são capacitados para serem acessados de diversas formas e diferentes lugares e são capazes de registrar quaisquer mudanças no ambiente. Esta rede de objetos interligados, que se comunicam entre si, caracteriza a Internet das Coisas.

Dohr et al. (2010) ainda classificam estes objetos inteligentes como ativos, que permitem decisões específicas de forma local, e passivos, que são os sensores que estão monitorando sempre, gerando e transmitindo dados sem preocupar com o processamento desses dados. Existe uma gama de dispositivos inteligentes que podem ser utilizados em um ambiente assistido.

2.2.1 Sensores em ambientes assistidos

Erden et al. (2016) fazem um levantamento destes componentes essenciais, quando se trata de um sistema AAL, que são extremamente variados para o monitoramento dos usuários. Ele separa estes objetos em dois grupos: sensores estáticos (sensores infravermelhos, sensores de vibração, sensores de pressão, câmeras e microfones) e sensores móveis ou vestíveis (sensores térmicos, acelerômetro, oxímetro, *smart watches*).

Com essa gama variada de sensores disponíveis, os mesmos podem ser combinados a fim de satisfazer ou detectar diversos eventos, como detecção de queda, monitoramento em vídeo, automação, monitoramento de rotina e atividades diárias.

Como se trata de uma grande variedade de dispositivos, Cedillo et al. (2018) fazem uma revisão sistemática na literatura, para entender quais os sensores mais utilizados nos sistemas AAL e também quais são suas medidas, ou seja, quais informações que são extraídas por eles. Observa-se que boa parte dos estudos utilizam de dispositivos vestíveis e objetos de casa, como camisas, relógios, acessórios corporais, calçados e mais sensores distribuídos pela residência a fim de transformá-la em uma *smart home*. A maioria dos dados que são obtidos por tais sensores são os sinais vitais, tais como batimentos cardíacos, pressão arterial, temperatura, entre outras. Para isso, cita diferentes dispositivos, tais como fita de curativo que detecta atividade elétrica do coração através da pele, e um mini

computador de tamanho de um botão, que deve ser usado no peito coletando sinais vitais, padrões de respiração ou qualquer outro sinal na região torácica.

Erden et al. (2016) concluem que os sistemas de moradia assistida permitem que pessoas idosas vivam de forma independente e segura, diminuindo a carga de trabalho de assistentes e familiares. Entretanto, os dispositivos que compõem o sistema devem ser de fácil utilização, robustos e confiáveis. Panico et al. (2020) também concluem que os dispositivos devem interferir minimamente na vida de seus usuários, olhando também para o lado ético de quais tipos de dados estão sendo gravados e monitorados. Argumentando então que a implementação de sistemas AAL devem ser centrados nas pessoas que o utilizam e não na tecnologia em si. Dessa forma, propondo que a escolha de dispositivos seja customizada e de acordo com a necessidade de seu paciente.

2.3 Detecção de eventos

Um sistema que compõe a moradia assistida pode compreender uma coleção diversa e customizada de sensores e de dispositivos IoT para monitoramento, ocasionando na produção de um enorme volume de dados (JANJUA et al., 2019). Isto impacta diretamente no contexto de AALs, pois são compostos diretamente destes dispositivos e objetos inteligentes em seu nível *Hardware*. Esse fato gera a necessidade de melhor cuidado com transmissão, armazenamento, segurança e o processamento dos dados. Este último, sendo o ponto focal quando se trata da análise de dados e extração de conhecimento, que é o objetivo final dos ambientes assistidos.

Janjua et al. (2019) argumentam que a tarefa mais desafiadora, em análise de dados, é a de descoberta de padrões que indicam mudanças bruscas nos dados, podendo significar uma situação atípica ou evento. Com isso, ele define estes eventos raros como: “uma observação que ocorre de maneira indeterminada e inconsistente, em relação à outras observações, de modo que pode se suspeitar de uma anomalia ou irregularidade no conjunto de observações” . A detecção de eventos raros é muito importante no contexto de um sistema AAL, pois a ocorrência destas anomalias, na transmissão de dados dos sensores, podem indicar situações de perigo para o residente, sendo estas definidas pelo ambiente, como um incêndio, ou físicas, como identificação de baixa pressão arterial.

Em sua revisão sistemática, Patel e Shah (2019) enumeram 17 desafios quando se trata de reconhecimento de atividades e rotinas. Algumas diretamente relacionadas a padrões, como atividades paralelas feitas pelo residente; atividades intercaladas; e ambiguidade na interpretação de atividades que são presentes em várias tarefas. Dado que muitos estudos e pesquisas utilizam de algoritmos de aprendizado de máquina para identificação destes eventos fora do padrão, estes desafios são importantes para encontrar o melhor modelo, onde a definição de modelos supervisionados ou não também é um desafio, uma vez que, para um modelo supervisionado, é necessário um conjunto de dados de treinamento rotulado.

Janjua et al. (2019) escolhem em seu estudo de detecção de eventos um método não supervisionado, argumentando em cima da necessidade de dados de treinamento devidamente rotulados, o que demandaria tempo, além de ser uma tarefa difícil, pois depende de uma visão de um especialista para observar cada uma das instâncias e rotulá-las. Em contrapartida, Navarro et al. (2018) utilizam de um modelo de Rede Neural Artificial treinado para uma classificação primária dos dados transmitidos.

2.4 Considerações Finais do Capítulo

Os conceitos que foram apresentados estão interligados quando se busca discutir ambientes assistidos. Esses ambientes são, efetivamente, um grande suporte à vida cotidiana de seu usuário, para que este continue vivendo de forma independente em sua própria casa.

Devido à gama de sensores e dispositivos IoT que existem atualmente, tais sistemas podem ser extremamente customizados e personalizados, permitindo que esta tecnologia de assistência atinja diversas pessoas, para que possam se beneficiar do auxílio sem perceber a intervenção tecnológica (AmI). Além disso, os cuidadores e responsáveis do residente podem aproveitar de várias informações relacionadas diretamente à saúde e bem estar do mesmo, a partir dos modelos de extração de conhecimento que enviam alertas automáticos para que medidas preventivas ou corretivas sejam tomadas.

Visto que, de fato, os ambientes assistidos trazem múltiplos benefícios, se torna cada vez mais interessante buscar aprimoramentos na análise do grande volume de dados que estes geram. Da mesma maneira que existe uma variedade em dispositivos IoT,

também existem diversos modelos de processamento de dados usados no contexto de AAL (PATEL; SHAH, 2019). Portanto, é importante definir a forma de processamento e extração de conhecimento dos dados, pensando nos sensores que estão sendo utilizados, já sabendo também quais os formatos e padrões de dados esperados, advindo do monitoramento do usuário do sistema.

Para o funcionamento do ecossistema de um ambiente assistido é necessário que ele esteja bem formado, utilizando os dispositivos e sensores que melhor se adéquam às necessidades de seu usuário, ou seja, quais as métricas mais relevantes para que se possa usufruir e obter informações significativas sobre o residente.

A detecção de eventos é de suma importância para que os conhecimentos e informações sobre os comportamentos e ações dos residentes sejam cada vez mais precisos. Como foi visto, técnicas de aprendizado de máquina, que necessitam de bases de dados para treinamento, como agrupamento e classificação, são ótimas estratégias para que este objetivo seja alcançado. A aplicação de tais técnicas permite que eventos fora do padrão sejam encontrados na transmissão dos dados obtidos pelos sensores. Modelos de agrupamento não necessitam de treino e sim de uma especificação bem descrita do que pode caracterizar um evento considerado normal de um usuário para um evento que está fora do padrão (anomalia comportamental), permitindo que seja visto em conjunto de eventos atípicos. Já um modelo com base em classificação, necessita de um treinamento com dados consolidados e rotulados, para poder distinguir os eventos padrões e as anomalias. A escolha deve ser feita de acordo com as capacidades e necessidades do ecossistema assistido por completo.

3 Trabalhos Relacionados

Neste capítulo é apresentada uma visão de trabalhos que se relacionam, de alguma maneira, com este trabalho de conclusão de curso. Como o foco é a detecção de eventos sonoros em ambientes assistidos, são descritos estudos que atuam na detecção de eventos, seja com base em dados acústicos ou não, de tal modo que se possa entender tanto quais os tipos de aprendizagem de máquina estão sendo usados, quanto quais tipos de dados estão trabalhando.

3.1 Detecção de Eventos em Ambientes Assistidos

Navarro et al. (2018) realizam um trabalho de coleta e análise de dados em uma arquitetura distribuída *Fog* para identificar eventos acústicos que possam representar comportamento anormal ou condições de perigo, em cenários de Ambientes de Moradia Assistida de grande escala, como casas de repouso. O sistema proposto é composto por 3 camadas (NAVARRO et al., 2018): camada de sensores, camada de detecção prévia de eventos em tempo real, e camada de análise de evento em alto nível. Como este trabalho foca na detecção de eventos e análises, é interessante compreender de maneira mais detalhada as duas últimas camadas da estrutura proposta pelos autores, sabendo que a camada de sensores provê vetores de atributos acústicos para a detecção e análise dos dados.

Sua camada de detecção prévia de eventos em tempo real apresenta uma GPGPU, que permite diversas análises de dados em paralelo, com uma Rede Neural Artificial treinada. A saída da rede é um vetor ponderado com destino à camada superior de análise detalhada, onde cada elemento do vetor representa a probabilidade de cada evento, uma visão preliminar do evento que pode ter acontecido. Sendo assim, o resultado dessa primeira detecção não deve ser tomado como certo, a não ser que seja considerado como um evento extremo. Por sua vez, a camada de análise de evento em alto nível, trabalha com o contexto dos eventos acústicos e a localização deles, comparando resultados concatenados da camada inferior e uma visão histórica de casos anteriores, através de Raciocínio Baseado

em Casos.

Janjua et al. (2019) desenvolvem um trabalho de detecção de eventos raros, com base nos dados de dispositivos na camada *Edge*, existindo uma transmissão contínua de dados, possibilitando redução de custos de transmissão e análise de dados que existem no contexto de Nuvem. Os autores utilizam aprendizado de máquina não supervisionado, pela dificuldade em ter um conjunto de dados de treino bem estabelecido, por meio de rotulação manual. Assim, foi utilizada uma estratégia de detecção de eventos em duas etapas, que é composta por uma combinação de duas técnicas de agrupamento, para poder lidar com a alta velocidade que uma transmissão contínua de dados possui. A primeira etapa é o micro-agrupamento, que é feito de forma *online*, ou seja, trabalha na transmissão contínua dos dados. Já a segunda etapa é o macro-agrupamento, que extrai os eventos raros dos micro-grupos. O resultado final é apresentado em 2 grupos: grupo 1 mais denso e contendo dados de comportamento normais e o grupo 2 que contém eventos de comportamento anormais, de acordo com o intervalo de tempo dos dados transmitidos.

Ressaltando os modelos de agrupamento, na etapa micro, os autores utilizam uma técnica específica para agrupamento na transmissão, de forma que não seja necessário o armazenamento dos dados, sendo ele um algoritmo de agrupamento baseado em árvore. Para a etapa macro é utilizado um método aglomerativo *Ward* para agrupar minimizando a distância entre os micro-grupos.

Patel e Shah (2020) consideram em sua abordagem um cenário onde o comportamento das pessoas monitoradas é mais imprevisível. Além disso, os autores consideram uma visão mais generalizada para que a solução proposta possa ser adaptada para ambientes assistidos distintos. Para tal, o modelo utiliza um sistema de *feedback* e aprendizagem para adicionar diferentes comportamentos e ações do residente do ambiente assistido. Patel e Shah (2020) separam seu *framework* em 3 módulos: reconhecimento de atividade simples, reconhecimento de atividade complexa e modelagem comportamental com detecção de anomalia. Por se tratar de um *framework* generalizado, os dados gerados pelos sensores precisam ser agregados e pré-processados em um banco central, para então serem analisados nos módulos mencionados.

O primeiro módulo, reconhecimento de atividade simples, se baseia apenas em

experiência história para rotular de forma direta uma atividade. Dormir, sentar, ficar em pé, andar e correr, são exemplos de atividades predefinidas, que são classificadas nesse módulo. O segundo módulo, reconhecimento de atividade complexa, objetiva identificar a ação relacionada à primeira atividade, como comer, exercitar, escovar os dentes e outras tarefas de casa. Também nesse módulo é utilizada a mesma técnica de inferência de ações conhecidas, ou a criação de novo rótulo, caso a ação não esteja presente na lista predefinida.

Por fim, seu terceiro e último módulo apresenta a descoberta de comportamentos anormais do residente do ambiente assistido. Para isso, é necessária uma fase de treinamento para observar a rotina do usuário do sistema. Ao ser capaz de reconhecer a rotina, as atividades diárias são atribuídas à grupos diferentes com base no turno do dia: manhã, tarde, noite e madrugada. Sendo assim, qualquer atividade reconhecida que não se encaixe em seu grupo, é identificada como um comportamento fora do padrão, mas se a mesma ocorre de forma diária, o sistema deixa de reconhecer como anomalia e agrega ela em um dos 4 grupos, identificando uma nova atividade diária.

Thakur e Han (2021) apresentam uma visão bastante interessante e focada em detecção de quedas e ações relacionadas com quedas no contexto de ambientes assistidos, como deitar no chão, estar apoiado nos 4 quatro membros, levantar de uma queda. Seu modelo trabalha com sensores que captam ângulos e aceleração do corpo para analisar o evento ocorrido. O estudo se baseia na validação cruzada *K-Fold*, que divide K vezes o conjunto de dados em dois subconjuntos: um subconjunto de teste e um subconjunto de treinamento, de diversas maneiras. Isto é feito para que sejam obtidos resultados diferentes a cada iteração de K, a fim de saber a performance e comportamento do modelo a cada iteração, de modo que se consiga melhor determinar e avaliar o comportamento do modelo. Também utiliza o algoritmo de aprendizagem de máquina *AdaBoost*, que busca melhorar a performance de modelos de classificação. Assim, conseguem melhorar os resultados do algoritmo de aprendizagem máquina k-NN, utilizado na proposta do trabalho. Sua abordagem também é capaz de detectar longadas quedas, ou seja, o residente continuar deitado após ter sofrido uma queda, uma vez que idosos podem sofrer diversos impactos em sua saúde.

Hoai e Torre (2012) propõem um modelo de detecção de eventos (MMED), que tem como focos dados em vídeo e imagens, que por sua vezes são tipos de dados que também podem ser gerados em ambientes assistidos, através de dispositivos com câmera. Tal modelo é baseado em SVM, sendo este um modelo supervisionado. O interessante deste detector é a capacidade do mesmo identificar a duração de um evento, mesmo que o evento não tenha terminado, isso é possível devido a qualidade dos conjuntos de treinamento.

O modelo proposto por Hoai e Torre (2012), tem como característica a necessidade de uma base de dados já rotulada e consolidada para o seu treinamento e suas análises são feitas de maneira *offline*, necessitando de um armazenamento dos dados para então poder trabalhar sobre eles e realizar a extração de conhecimento. Xie et al. (2019) propõem uma melhoria para tal modelo para detecção de eventos em um ambiente de *streaming* de dados, para reduzir os gastos relacionados ao modelo anterior, anteriormente mencionados, que dificultam sua aplicação em cenário real. O modelo proposto pelos autores (OMED) lida com sequências de dados de treinamento momentâneas, não necessitando de dados históricos e sim das restrições específicas para os dados trabalhados. Os marcadores de início e fim de evento são determinados por um limite predefinido, com base nas razões de verdadeiros positivos, quantas amostras positivas foram classificados de forma positiva, e falsos positivos, quantas amostras negativas foram classificadas de forma positiva.

3.2 Comparações nos trabalhos

Como foi visto, existem abordagens diferentes para a solução de um mesmo problema e, por isso, se faz necessário entender como estão organizadas as soluções na literatura. Patel e Shah (2019) fazem uma pesquisa na literatura acerca de implementações de reconhecimento de atividades com base em sensores, nos Ambientes de Moradia Assistidos. Em seu trabalho, os mesmos classificam os estudos encontrados com base em seus modelos (Generativo, Discriminativo, Mineração de Dados e baseados em Lógica). Como é um estudo comparativo, eles identificam os desafios encontrados na questão de reconhecimento de atividade, que devem ser considerados para um trabalho nesta área de pesquisa. Tais desafios são usados para mostrar também o que cada abordagem contempla.

A fim de comparação entre os estudos mostrados anteriormente, foram considerados estes desafios com foco na análise de dados. É importante observar que os trabalhos referentes à detecção prévia de eventos não estão contextualizados ao cenário de moradias assistidas, mas apresentam propostas relacionadas à análise e processamento de dados, que buscam reconhecer eventos e suas causas, da mesma maneira que os outros estudos também objetivam a identificação de eventos. A seguir são apresentados os principais pontos comparativos entre os trabalhos:

- Tipo de dados (A): se refere a quais os dados utilizados pelos trabalhos, tais como dados sonoros, imagens, vídeos, espacial, aceleração, temperatura, batimentos, entre outros;
- Necessidade de treinamento (B): se refere ao modelo ser supervisionado ou não. Este item indica se é necessário um conjunto de dados rotulados consolidado para treinar o modelo;
- Técnica de aprendizagem de máquina (C): se refere a qual técnica base para identificação dos eventos foi utilizada, como agrupamento ou classificação;
- Processamento dos dados (D): se refere a forma que é feito o processamento dos dados, se é feito de maneira *offline* ou *online*, i.e., em tempo real.
- Tipo da base de dados (E):
    1. Utiliza um ou mais conjuntos de dados consolidados, rotulados, existentes na literatura; ou
    2. Utiliza um conjunto de dados obtido por implementação e captura própria em seu trabalho.

Na Tabela 3.1 são apresentadas as comparações entre os trabalhos relacionados. Mesmo que modelos supervisionados possuam custo para seu treinamento e um custo de armazenamento para os dados de treinamento dos mesmos, os estudos em sua maioria optam por ele. Alguns buscando contornar a característica de armazenamento por um processamento *online*, reduzindo gastos de treino e armazenamento dos conjuntos dedicados a isto, ajustando e treinando os modelos juntamente com a transmissão de dados

Tabela 3.1: Diferenças na detecção de eventos entre os trabalhos

Estudo	A	B	C	D	E
(NAVARRO et al., 2018)	Acústicos	Sim	Classificação	*Online*	2
(JANJUA et al., 2019)	Acústicos	Não	Agrupamento	*Online*	1
(PATEL; SHAH, 2020)	Diversos	Não	Agrupamento	*Offline*	2
(THAKUR; HAN, 2021)	Aceleração espacial	Sim	Classificação	*Offline*	2
(HOAI; TORRE, 2012)	Vídeos	Sim	Classificação	*Offline*	1
(XIE et al., 2019)	Vídeos	Sim	Classificação	*Online*	1

que já está sendo realizada. Por se tratarem de proposições de modelos, muitos não implementam sua própria infraestrutura de sensores para a extração de dados e utilizam de múltiplas bases de dados, com características bem definidas para utilizarem em suas abordagens.

A maioria dos trabalhos se especializa em um único tipo de dado. Isto é um ponto importante a se considerar, pois seus modelos são ajustados, especializados, para os sensores e dispositivos que compõem o ambiente assistido, que na maioria das vezes são propostas que apresentam suas próprias camadas de *Hardware*, não permitindo uma customização dos dispositivos. Caso seja uma abordagem que apresenta somente o detector de eventos e não possua uma infraestrutura própria, em sua camada de *Hardware* torna-se mais fácil trabalhar com vários tipos de dados e os dispositivos que fazem estas capturas, resultando na característica de ser mais generalizado, possibilitando a customização de sensores e que outros dispositivos IoT componham o ecossistema.

É interessante observar a preocupação dos modelos apresentados em trabalhar de maneira *online*, em tempo real, permitindo que estes sejam aplicados, de fato, em cenários reais, visto que a geração de dados é constante e que o armazenamento dos mesmos é complexo, e também permitindo uma tomada de decisão mais próxima do acontecimento.

4 Modelo de Classificação para Dados Sonoros

Este capítulo tem como objetivo descrever o método de extração de atributos dos dados, a criação dos modelos e dos parâmetros das redes neurais desenvolvidas. Para a extração de atributos foi utilizada a linguagem de programação Python e as bibliotecas *python_speech_features* e *librosa*. Para criação dos modelos foi utilizada a linguagem de programação Python, o framework Keras que utiliza como base a biblioteca TensorFlow.

4.1 Tratamento dos dados

Como referência aos pontos mencionados na seção 3.2, optou-se por seguir o padrão em tratar um único tipo de dado neste trabalho de conclusão de curso, que utiliza dados acústicos. Para alinhar com o contexto de moradia assistida, foi escolhido o som de tosses como os sons a serem reconhecidos pelos modelos. Assim, este trabalho busca entender se é possível que os modelos diferenciem se um som corresponde ou não a uma tosse. Dessa maneira, para realizar o treinamento dos modelos se faz necessário uma base de dados consolidada com dados acústicos rotulados referentes a tosses e dados que não sejam referentes a tosses.

Com o objetivo de classificar tosses, foi utilizado o conjunto de dados COUGHVID, apresentado em (ORLANDIC; TEIJEIRO; ATIENZA, 2021). O conjunto de dados COUGHVID possui mais de 30.000 gravações de tosses obtidas de maneira colaborativa, *crowdsourcing*, com envios de pessoas de várias idades, gêneros e localidades, contribuindo com uma variedade de gravações de tosses para treinamento de modelos de aprendizado de máquina. A fim de adicionar uma nova classe de áudios para diferenciação de tosses no treinamento dos modelos, foi escolhido trabalhar com dados de áudio de instrumentos. Estes dados foram obtidos com uma combinação de dois conjuntos de dados disponibilizados em repositórios do Kaggle, sendo o primeiro referente a uma competição de rotulação

de áudios com diversas classes[3] e também um conjunto de dados de instrumentos musicais[4]. Destas bases de dados foram utilizados áudios dos seguintes instrumentos para a formação da classe Instrumento: Violão, Bumbo, Violoncelo, Clarineta, Contrabaixo, Flauta, Chimbal, Saxofone, Caixa e Violino. Em um primeiro momento, devido a grande variedade de áudios disponíveis no conjunto, foi feita uma filtragem de tipos de extensão nos mesmos para separar somente os áudios do tipo *wavfile*, para padronizar a entrada de dados.

Para utilização dos áudios, foi feito um tratamento em cima dos dados para remoção de trechos de baixa magnitude, ou seja, foram removidos trechos que possuíam pouca ou nenhuma informação sonora. Essa remoção é feita partindo da extração do envelope do sinal de áudio, que contorna os extremos do áudio. Um exemplo pode ser observado na Figura 4.1, onde temos representada curva do envelope de um sinal, que delimita o que é mantido ou removido deste.

Após esta limpeza, o conjunto de dados contém somente os dados mais relevantes para serem usados pelos modelos de Deep Learning. Todos os dados de áudio usados neste trabalho passaram por este processo de limpeza inicial.

O conjunto de áudios de instrumentos referente a competição de rotulação apresentou uma maior diversidade de instrumentos quando comparado ao segundo, sendo este conjunto selecionado para compor o treinamento dos modelos. Desta maneira, teve-se a necessidade de realizar um corte nos dados da base de tosse com o intuito de obter um conjunto de treino balanceado. Após o corte, o conjunto de treino passou a ter 300 áudios de tosse escolhidos aleatoriamente e os 300 áudios de instrumentos presentes no conjunto de instrumentos mencionado anteriormente.

O segundo conjunto de dados de instrumentos conta com 1460 instrumentos, todos estes são utilizados para testes. A fim de utilizar todo o volume restante de áudios de tosses, foram criados 18 diretórios contendo arquivos desta classe, escolhidos aleatoriamente. Cada diretório representa um lote de experimentação, com os dados de instrumentos e tosses. Isso foi feito para que os experimentos tenham dados balanceados.

[3] ⟨https://www.kaggle.com/competitions/freesound-audio-tagging/overview⟩
[4] ⟨https://www.kaggle.com/datasets/soumendraprasad/musical-instruments-sound-dataset⟩

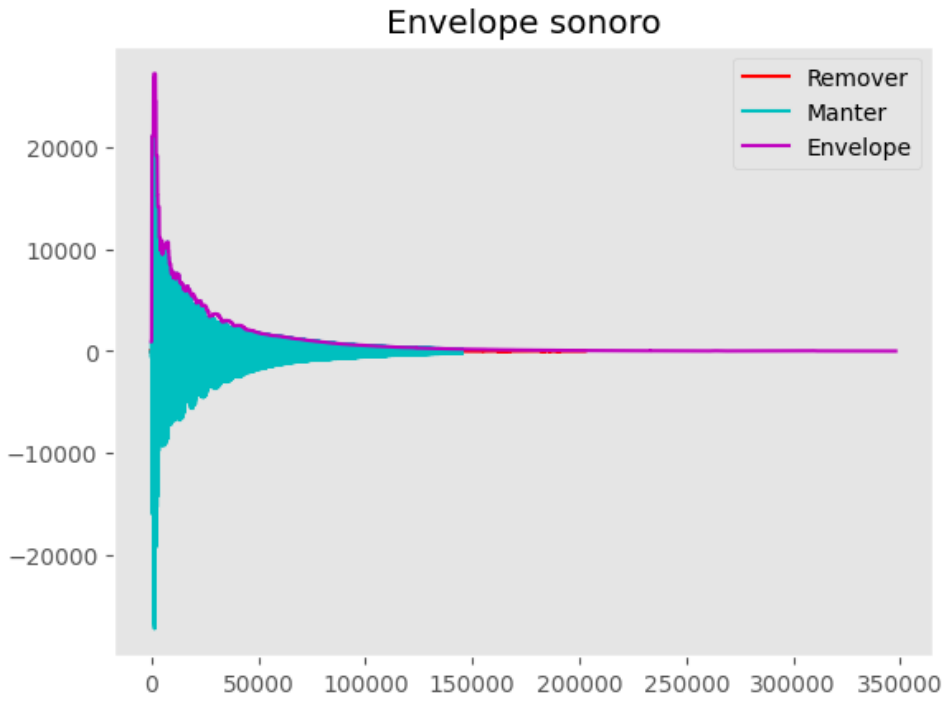

Figura 4.1: Exemplo de envelope sonoro em um sinal de violão

4.2 Extração de atributos

Nesta seção são apresentadas as etapas necessárias para a extração de atributos dos dados acústicos. Para este trabalho foram escolhidos os Coeficientes Ceptrais de Frequência-Mel, que são os atributos mais populares para extração de informação para modelos de aprendizado de máquina no contexto de dados sonoros (NAVARRO et al., 2018; JANJUA et al., 2019).

4.2.1 Transformada de Fourier

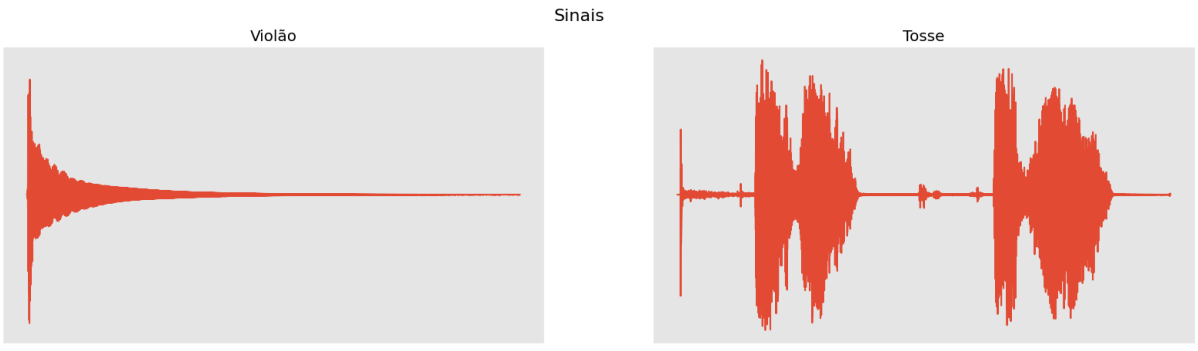

Figura 4.2: Exemplos de sinais presentes no conjunto de dados

Usualmente, áudios são visualizados a partir de um gráfico de sinais, ondas, no domínio de Tempo(s), como visto nas amostras presentes na Figura 4.2. Entretanto, é muito difícil discernir qual som cada sinal representa, olhando somente para esses gráficos, devido a característica da frequente alteração de amplitude em poucos intervalos de tempo.

Uma outra representação matemática do som é a Transformada Rápida de Fourier (FFT), uma função que recebe de entrada um sinal no domínio do Tempo e emite como saída sua decomposição em domínio de Frequência (Hz). Esta transformação pode ser observada na Figura 4.3, que foi realizada em cima de uma amostra de som de violão e uma amostra de som de tosse. Na figura é possível observar que existe uma grande quantidade de informação em frequências mais baixas, assim como obter de forma precisa essas informações, que são importantes para auxiliar no reconhecimento dos diferentes sons.

Entretanto, não é recomendada a transformação em cima do sinal por completo, pois é perdida a informação de contorno de frequência, devido a característica de mudança de frequências ao longo do tempo. Nesse sentido, é necessário utilizar a FFT em cada quadro do sinal de áudio para calcular o espectro de frequência. Esses quadros são obtidos por uma divisão do sinal em tempos curtos. Assumindo que as frequências em um sinal são estáveis durante um período de tempo curto. Assim, realizando uma transformada de Fourier sobre esses quadros de tempo curto e concatenando os mesmos é possível obter uma aproximação dos contornos de frequência do sinal. O tamanho padrão de um quadro é de 25 ms com sobreposição entre os adjacentes e um passo de 10 ms. Este processo se chama Transformada de Fourier de tempo curto (STFT), ou seja, é feita uma FFT de N pontos em cada quadro para calcular o espectro de frequência, onde N é sempre um número de base 2.

Partindo deste ponto, podemos usar a Transformada de Fourier de tempo curto, para criar um periodograma com um empilhamento dos espectrogramas obtidos (FAYEK, 2016). Tais espectrogramas são apresentados na Figura 4.4.

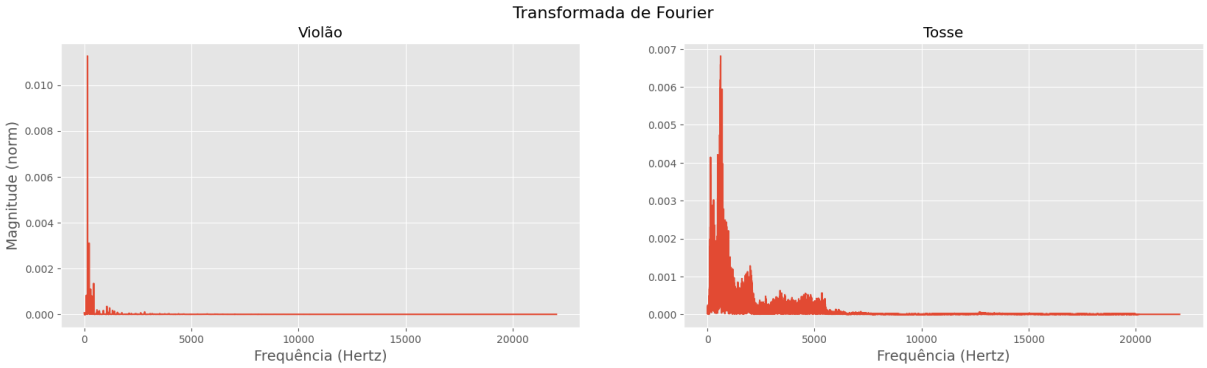

Figura 4.3: Exemplos da realização da Transformada de Fourier.

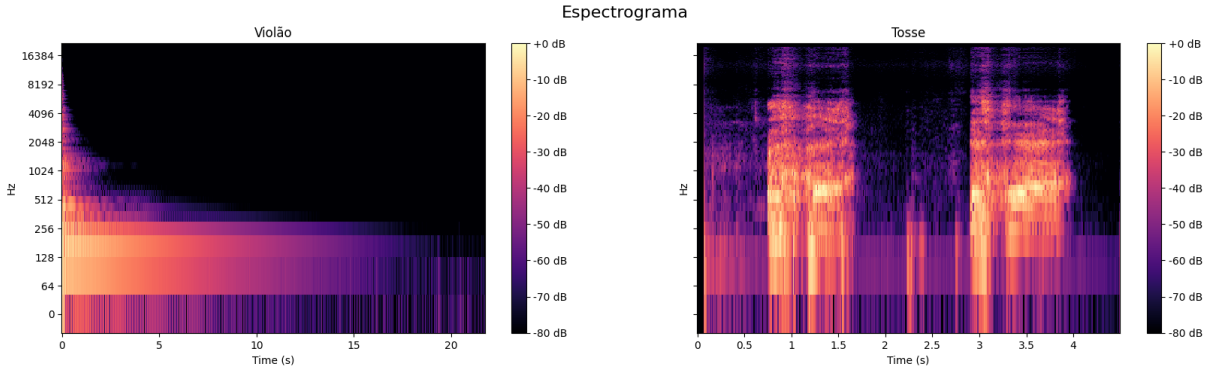

Figura 4.4: Exemplos de espectrogramas obtidos pela STFT.

4.2.2 A escala Mel

A escala Mel busca imitar a percepção não linear pelo ouvido humano sobre o som, sendo mais discriminativa em frequências mais baixas e menos discriminativa em frequências mais altas. Assim, a escala Mel dá mais importância na variação de frequências mais baixas e menos importância na variação de frequências mais altas, como pode ser observado no gráfico representado na Figura 4.5. A conversão de frequência (Hz) para Mel é feita pela Equação 4.1 e a conversão de Mel para frequência pela Equação 4.2[5].

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{4.1}$$

$$F(m) = 700\,(\exp(m/1125) - 1) \tag{4.2}$$

Esta importância permite a criação de bancos de filtro que são aplicados nos

[5] ⟨http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/⟩

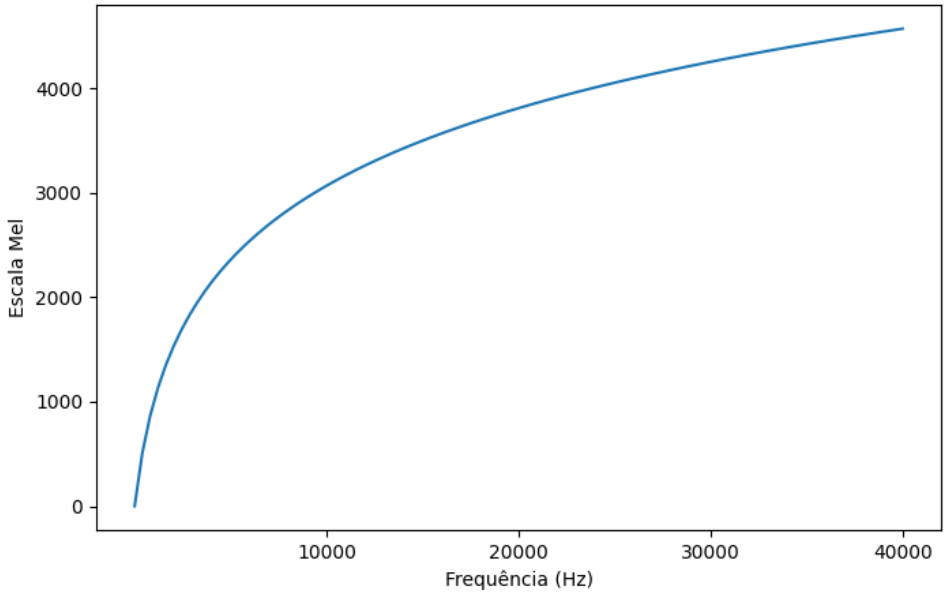

Figura 4.5: Gráfico referente a escala Mel.

periodogramas gerados a partir das FFTs, a Figura 4.6 representa um banco de filtros da escala Mel com 26 filtros.

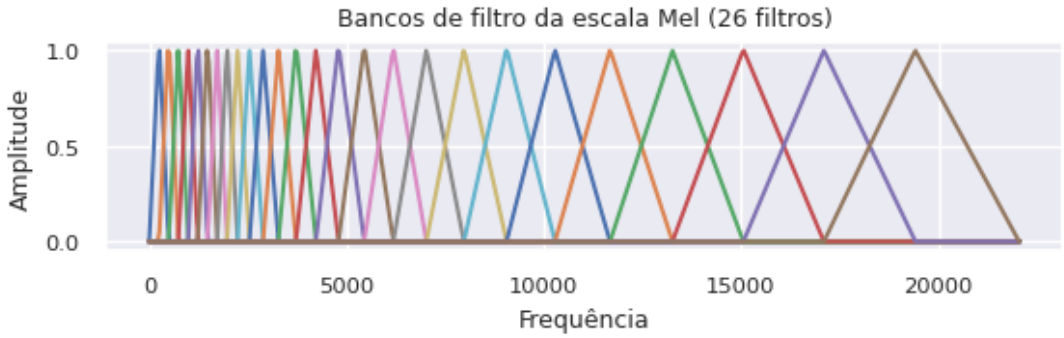

Figura 4.6: Bancos de filtro da escala Mel com 26 filtros.

A aplicação dos filtros da escala Mel resulta em um *downsampling* nos áudios para que os dados sejam mais relevantes ao serem passados para os modelos de aprendizado de máquina. Na Figura 4.7 é possível verificar que o espectrograma Mel apresenta dados mais condensados, comprimidos, quando comparado com os espectrogramas visto anteriormente na Figura 4.4.

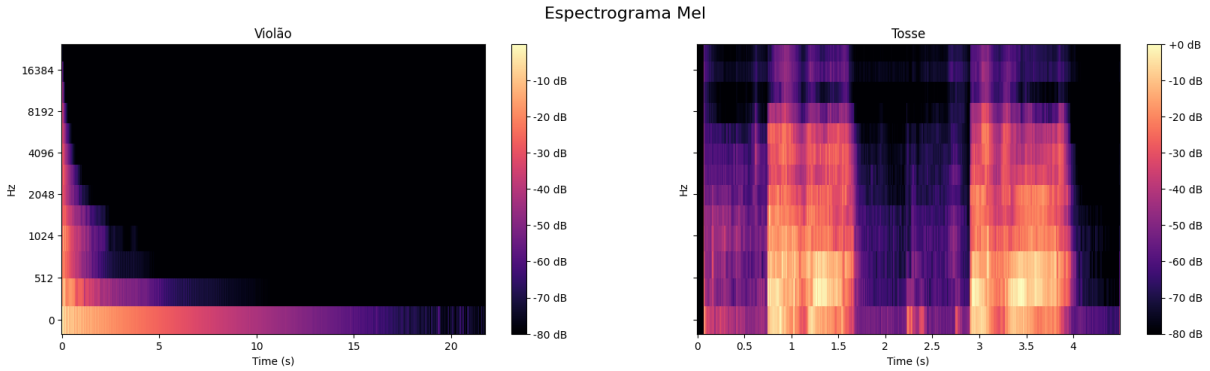

Figura 4.7: Exemplos de espectrogramas Mel.

4.2.3 Coeficientes Cepstrais de Frequência-Mel

Os coeficientes dos bancos de filtros podem ser usados como atributos para uso em aprendizado de máquina. Entretanto, eles são altamente correlacionados, o que poderia impactar alguns algoritmos. Nesse sentido, pode ser aplicada uma Transformada Discreta de Cosseno (DCT)[6] para tirar essa correlação e produzir uma representação comprimida dos bancos de filtro. Assim, são gerados os Coeficientes Cepstrais de Frequência-Mel(MFCCs), que podem ser observados na Figura 4.8, onde se tem uma representação em espectrograma com o eixo-y variando de acordo com o número de MFCCs desejado (FAYEK, 2016).

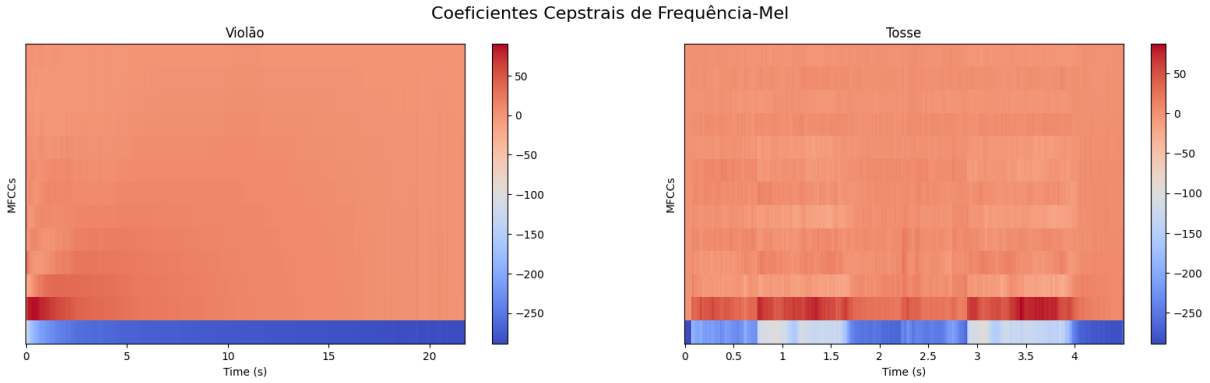

Figura 4.8: Exemplos de Coeficientes Cepstrais de Frequência-Mel.

Os Coeficientes Cepstrais de Frequência-Mel foram introduzidos em 1980 por Davis e Mermelstein (1980) e são considerados como os melhores atributos que podem ser extraídos de áudios até os dias de hoje, sendo amplamente utilizados na área de

[6] ⟨http://datagenetics.com/blog/november32012/index.html⟩

reconhecimento de fala (FAYEK, 2016)[7].

4.2.4 Parâmetros para extração de atributos

Como foi visto anteriormente, para que seja realizada a extração de atributos é necessária a definição de certos parâmetros. Os parâmetros, assim como seus valores, são especificados na Tabela 4.1. Neste trabalho, utilizamos 26 filtros, sendo suficiente para se trabalhar com a taxa de amostragem de 16 kHz. O tamanho e o passo dos quadros segue o padrão visto em Fayek (2016) de 25 ms para o tamanho, significando que em cada quadro se tem $0.025 * 16000 = 400$ pontos e 10 ms para o passo. Com isso, o número de FFTs deve ser de 512, por ser o número de base 2 mais próximo da quantidade de pontos. O número de atributos, MFCCs, a serem extraídos é 13, seguindo padrão visto em Fayek (2016)[8]. Os 13 MFCCs são frequentemente usados porque capturam as informações espectrais mais importantes (POORJAM, 2018).

Tabela 4.1: Parâmetros para extração de atributos

Parâmetro	Valor
Taxa de amostragem	16 kHz
Nº de filtros	26
Tamanho dos quadros	25 ms
Passos nos quadros	10 ms
Nº de FFTs	512
Nº de MFCCs	13

4.3 Construção dos modelos

Foram construídos modelos de redes neurais utilizando a linguagem Python com a biblioteca TensorFlow[9] e através dessa foi usado o framework Keras. Os modelos escolhidos para realização dos testes foram o Convolucional 2D e o Long short-term memory (LSTM).

[7] ⟨http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/⟩

[8] ⟨http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/⟩

[9] ⟨https://www.tensorflow.org⟩

4.3.1 Convolucional 2D

Devido a características matriciais dos atributos extraídos, pela semelhança com a possibilidade de tratamento dos dados como imagem, o modelo de Rede Neural Convolucional (CNN) 2D foi escolhido por já ser bem conhecido no contexto de classificação de imagens[10]. Sendo assim, só se faz necessária a formatação dos dados para serem capazes de alimentar o modelo. Como a CNN espera como entrada uma imagem, este trabalho apresenta como entrada os dados acústicos como uma imagem em escala de cinza, que possui somente um canal, resultando no seguinte formato de entrada: (9, 13, 1), onde 9 é o tamanho máximo do dado, 13 é o número de atributos e o 1 adicionado para representar a imagem de escala de cinza. O valor 9 pode ser observado que vai ao encontro com o tamanho do passo de cada quadro de 10 ms.

As camadas definidas para o modelo são encontradas na Tabela 4.2. As camadas Convolucionais (Conv2D) são responsáveis por aprender os padrões do dado. A camada MaxPooling2D reduz pela metade os atributos para evitar que sejam criados muitos nós e também impedindo que o modelo fique superajustado ao conjunto de treino. Também responsável pelo cuidado contra o superajuste ao conjunto de treino, a camada Dropout define de maneira aleatória os pesos de uma parte dos dados para zero. Já as camadas Dense estão ligadas ao ajuste do modelo aos dados. Por fim, a última camada usada foi a Flatten, que comprime todas as informações dos atributos em uma coluna única para que as camadas Dense sejam capazes de classificar os áudios.

Tabela 4.2: Camadas escondidas do modelo Convolucional 2D

Camada	Saída
Conv2D	(9, 13, 16)
Conv2D	(9, 13, 32)
Conv2D	(9, 13, 64)
Conv2D	(9, 13, 128)
MaxPooling2D	(4, 6, 128)
Dropout	(4, 6, 128)
Flatten	(3072)
Dense	(128)
Dense	(64)
Dense	(2)

[10] ⟨https://medium.com/theleanprogrammer/2-dimensional-convolution-189abb174d92⟩

O modelo foi treinado com o conjunto de treino definido anteriormente e com 50 épocas. A Figura 4.9 apresenta os gráficos a respeito do treinamento do modelo. Pelo gráfico de perda, podemos ver que o modelo apresenta uma estabilidade na curva de treino e a curva de validação também se encontra estável com uma pequena distância entre as curvas, significando que o mesmo se encontra bem treinado. É interessante observar que a acurácia, pelo gráfico da direita, também se estabiliza. Nesse sentido, não foi necessário utilizar mais épocas para o treinamento.

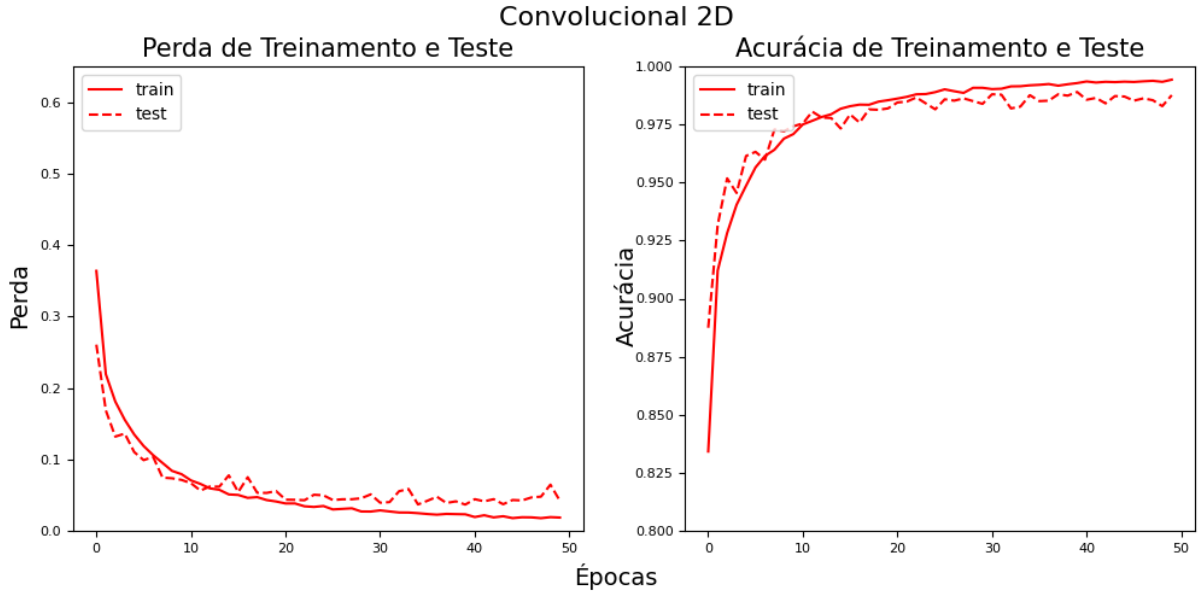

Figura 4.9: Gráficos de treino do modelo Convolucional 2D.

4.3.2 Long Short-Term Memory

As Redes Neurais Recorrentes (RNNs) tem como característica um algoritmo de aprendizado profundo capaz de recordar sequências, usando informações passadas para auxiliar na predição. Como áudios podem ser interpretados como séries temporais, se torna interessante o trabalho com tal modelo, devido a possibilidade dos sons tratados, instrumentos e tosses, apresentarem padrões repetitivos. Neste trabalho foi escolhido o modelo com Long Short-Term Memory (LSTM) que é capaz de conectar informações de longo prazo, lembrando informações de longos períodos de tempo[11].

As camadas definidas para o modelo são encontradas na Tabela 4.3. A camada

[11] ⟨https://medium.com/@web2ajax/redes-neurais-recorrentes-lstm-b90b720dc3f6⟩

Permute está presente para rearranjar os dados no padrão (MFCCs, Tempo, Canal) e a camada Reshape age na remoção da dimensão Canal que não é necessária para modelos RNN. A camada BidirectionalLSTM é responsável por olhar o dado na ordem padrão quanto na ordem reversa, permitindo extrair e guardar informações a respeito do dado atual, avançando em dois sentidos olhando pontos passados e futuros. Este modelo compartilha de algumas camadas presentes no modelo Convolucional, como MaxPooling, Dense, Dropout e Flatten, que desempenham os mesmos papéis apresentados anteriormente.

Tabela 4.3: Camadas escondidas do modelo LSTM

Camada	Saída
Permute	(13, 9, 1)
Reshape	(13, 9)
BidirectionalLSTM	(13, 64)
Dense	(13, 64)
MaxPooling1D	(6, 64)
Dense	(6, 32)
Flatten	(192)
Dropout	(192)
Dense	(2)

Pelo gráfico de perda, podemos ver que o modelo apresenta uma estabilidade na curva de treino e a curva de validação também se encontra estável com uma pequena distância entre as curvas

Diferentemente do modelo apresentado anteriormente, o LSTM, no gráfico de perda na Figura 4.10, não demonstra uma estabilidade nas curvas de treino e validação com 50 épocas, indicando a necessidade de mais épocas de treinamento. Por isso, este modelo foi treinado com 100 épocas, quando foi possível observar sua estabilidade. Assim como o modelo Convolucional, o LSTM apresentou boas curvas a respeito a perda e acurácia no treinamento, que diz ter sido um modelo bem treinado, com a distância entre as curvas de treino e validação bem menores do que o Convolucional.

4.4 Experimentação

Nesta seção é apresentado o fluxo de testes realizados nos modelos apresentados. As predições são realizadas em cima do conjunto completo de teste. Para cada áudio presente

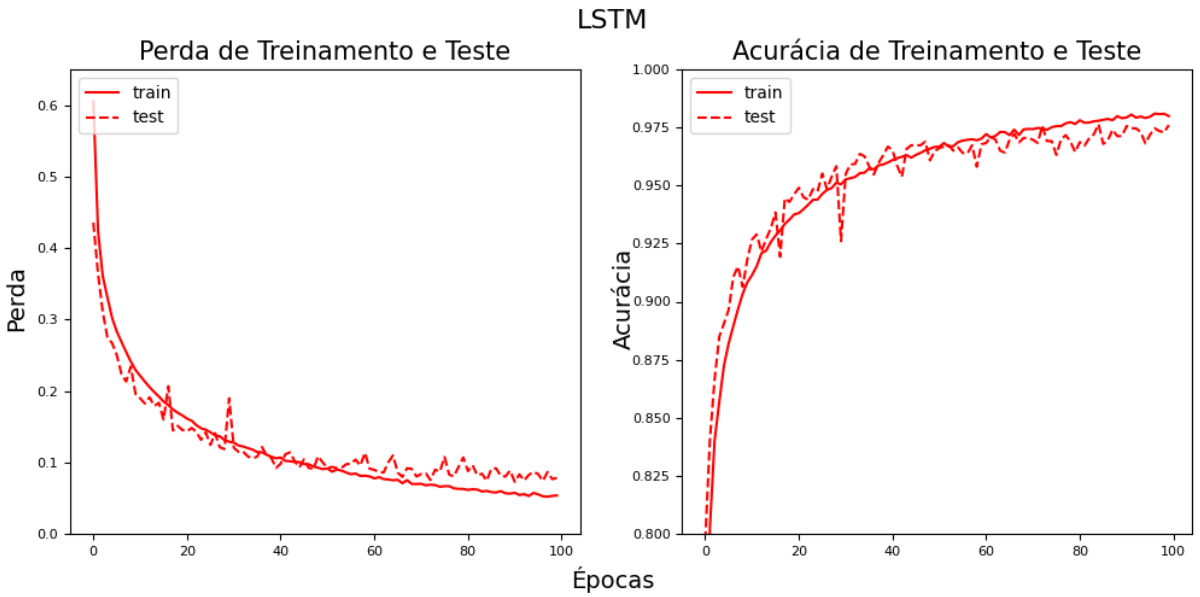

Figura 4.10: Gráficos de treino do modelo LSTM.

no conjunto, a cada segundo de sua duração, são extraídos os atributos por MFCCs e a predição é feita sobre eles. Após as predições sobre o conjunto de dados, é gerado um arquivo *csv* para cada arquivo *wavfile* com as seguintes informações: nome do arquivo, classe, probabilidade de ser Tosse, probabilidade de ser Instrumento, classe predita. Uma amostra de saída pode ser observado na Figura 4.11.

	fname	label	Cough	Instruments	y_pred
0	cf981fbe-eb7e-4a5f-b8eb-a1cdec1086b1.wav	Cough	0.882653	0.117347	Cough
2	a38b5cb8-7ee4-4a4a-90d6-1152f3dd3abb.wav	Cough	0.824630	0.175370	Cough
3	b11f7fd4-6919-4d2b-a2c0-600709606a59.wav	Cough	0.999714	0.000286	Cough
4	a62caf3c-edd5-4498-8c1a-90d090dc2df8.wav	Cough	0.999998	0.000002	Cough
5	79fa4dfa-37e2-4fcd-a0c2-b2d6e330f7de.wav	Cough	0.943827	0.056173	Cough

Figura 4.11: Exemplo de saída da experimentação

4.5 Resultados

Partindo da experimentação dos modelos definidos no conjunto de dados definidos na seção anterior, é necessária uma análise para identificar qual modelo se destacou na classificação das classes de áudio propostas neste trabalho.

No contexto de um conjunto de experimentação composto por 18 lotes, foram ex-

traídas as médias das métricas de acurácia, precisão, *recall* e *F1-score*, em cima das saídas apresentadas pelos modelos, as mesmas podem ser observadas na Tabela 4.4. Estão marcadas em negrito os melhores valores quando comparados entre os modelos. É importante ressaltar que ambos os modelos apresentaram bons resultados com acurácias acima de 80%. Entretanto, podemos afirmar que o modelo Convolucional apresenta melhores resultados contra o LSTM, devido a diferença nos F1-Scores.

Tabela 4.4: Métricas dos modelos

Métrica	Conv2D	LSTM
Acurácia	**0.841096 ±0.001658**	0.809437 ±0.001350
Precisão	**0.985281 ±0.004679**	0.984158 ±0.004160
Recall	**0.691781 ±0.000000**	0.628082 ±0.000000
F1-Score	**0.812844 ±0.001589**	0.766796 ±0.001265

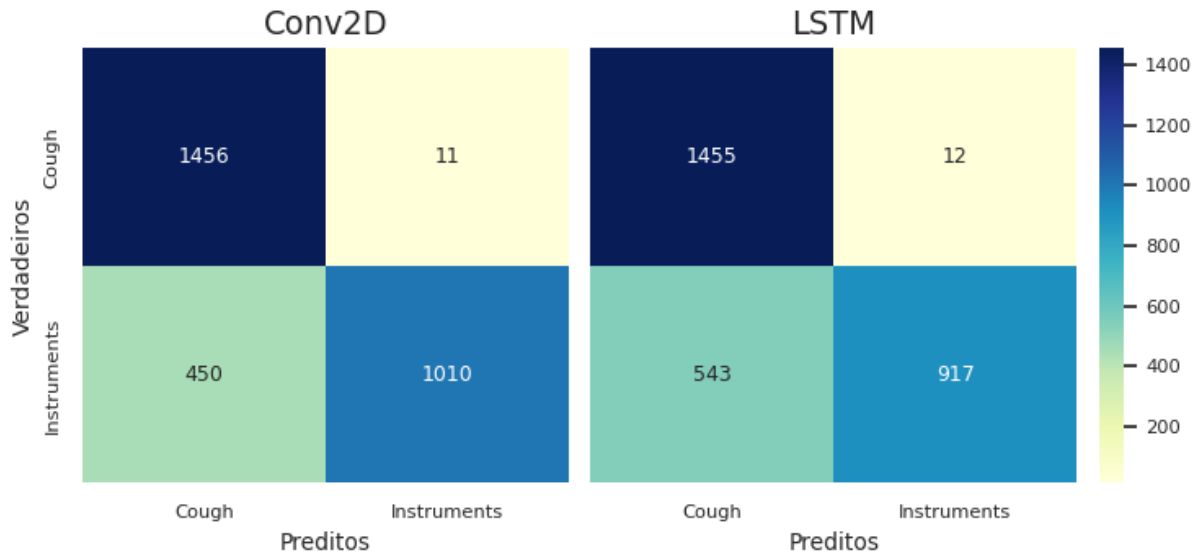

Figura 4.12: Matrizes de confusão Lote 18.

Olhando somente para um dos lotes, por exemplo o Lote 18, conseguimos gerar as matrizes de confusão, presentes na Figura 4.12. Nessa figura, pode ser observado que a classificação para Tosse está apresentando excelentes resultados para ambos os modelos. Em contrapartida, aproximadamente um terço de Instrumentos estão sendo preditos de maneira errada. Este comportamento pode se dar devido a uma similaridade dos sons de alguns instrumentos com tosse. Por exemplo, a característica sonora de instrumentos de percussão, como o Bumbo e o Chimbal, que produzem sons semelhantes a tosses.

Os resultados satisfazem a questão principal proposta por este trabalho acerca da aplicação de modelos de aprendizado de máquina na rotulação de dados sonoros, pois ambos os modelos apresentaram acurácias acima de 80% e bons F1-scores neste contexto. Tais modelos se mostraram bem capazes de diferenciar os diversos sons que integram o conjunto de dados proposto. Estes resultados são encorajadores para a aplicação em saúde, por meio dos ambientes assistidos para o acompanhamento de pessoas idosas, onde uma boa precisão é fundamental, i.e., os modelos não confundem tosse com instrumento, ainda que precisem melhorar para não confundir instrumento com tosse.

Por fim, acredita-se que os atributos extraídos, MFCCs, sejam os principais responsáveis para estes bons resultados, por ser considerada a melhor técnica de extração de atributos de dados sonoros desde 1980. A utilização desta técnica, consolidada e reconhecida na literatura, que retorna atributos significantes sobre os dados de áudio, pôde impactar positivamente no desempenho dos modelos. Alcançando, diretamente, um dos objetivos específicos deste trabalho, onde se tinha interesse em realizar a extração de atributos de dados acústicos da melhor forma. Tendo em vista que os modelos foram criados para um formato de entrada bem definido e é a partir dos atributos, dados, que estes modelos aprimoram a capacidade de discernimento das classes dos dados.

5 Considerações Finais e Trabalhos Futuros

Esse trabalho apresentou uma comparação entre modelos de redes neurais artificiais, Convolucional 2D e LSTM (*Long Short-Term Memory*) na classificação de áudios. Com foco em verificar a capacidade de reconhecimento destas redes neurais em um contexto de áudios de tosses, para que possam ser aplicadas em ambientes de moradias assistidas. Os modelos foram testados com o conjunto de dados COUGHVID e sons de instrumentos, para simular uma entrada com áudios relevantes e ruídos.

Os resultados obtidos são encorajadores para a aplicação destes modelos de redes neurais artificiais neste cenário de moradias assistidas. Contudo, é preciso reconhecer que existe muitas vias de extensão e melhorias a serem feitas nos modelos. É interessante ressaltar algumas limitações, como a diferença na qualidade dos áudios trabalhados, onde os dados de tosse possuem ruídos e sons externos por serem gravados por diversas pessoas no mundo de maneira livre, enquanto os áudios de instrumentos são mais controlados e gravados sem pertubação acústica.

Uma das possibilidades para prosseguimento deste estudo é a adição de mais classes sonoras a serem reconhecidas pelos modelos, como áudios referentes a atividades cotidianas de um lar e sons que dizem respeito da condição de saúde e estado do residente. Para entender como os modelos reagiriam no contexto de multi-classes em contraste ao contexto binário, que foi trabalhado.

Outro trabalho futuro que pode dar continuidade a este trabalho é a aplicação desses modelos em infraestruturas de ambiente assistido, para que sejam treinados de acordo com cada necessidade do usuário. Permitindo uma customização priorizando sons mais relevantes para cada usuário.

Em trabalhos futuros pretende-se ainda desenvolver uma extensão com um modelo que faça análises do histórico de saúde do usuário, onde será possível combinar com os modelos de reconhecimentos sonoros para que sejam feitas predições de condições de saúde futuras para alertar o usuário, os responsáveis ou equipe médica. Essa extensão vai permitir que tais condições possam ser corrigidas de maneiras antecipadas ou que haja

um melhor preparo para o tratamento das mesmas. Em suma, este trabalho serve como base para diversas perspectivas de continuidade e expansão.

Bibliografia

CEDILLO, P.; SANCHEZ, C.; CAMPOS, K.; BERMEO, A. A systematic literature review on devices and systems for ambient assisted living: Solutions and trends from different user perspectives. In: *2018 International Conference on eDemocracy eGovernment (ICEDEG)*. [S.l.: s.n.], 2018. p. 59–66.

COSTA, R.; CARNEIRO, D.; NOVAIS, P.; LIMA, L.; MACHADO, J.; MARQUES, A.; NEVES, J. Ambient assisted living. In: CORCHADO, J. M.; TAPIA, D. I.; BRAVO, J. (Ed.). *3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 86–94. ISBN 978-3-540-85867-6.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, p. 357–366, 1980.

DOHR, A.; MODRE-OPSRIAN, R.; DROBICS, M.; HAYN, D.; SCHREIER, G. The internet of things for ambient assisted living. In: *2010 Seventh International Conference on Information Technology: New Generations*. [S.l.: s.n.], 2010. p. 804–809.

ERDEN, F.; VELIPASALAR, S.; ALKAR, A. Z.; CETIN, A. E. Sensors in assisted living: A survey of signal and image processing methods. *IEEE Signal Processing Magazine*, v. 33, n. 2, p. 36–44, 2016.

FAYEK, H. M. *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between*. 2016. Disponível em: ⟨https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html⟩.

HOAI, M.; TORRE, F. De la. Max-margin early event detectors. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2012. p. 2863–2870.

JANJUA, Z. H.; VECCHIO, M.; ANTONINI, M.; ANTONELLI, F. Irese: An intelligent rare-event detection system using unsupervised learning on the iot edge. *Eng. Appl. Artif. Intell.*, Pergamon Press, Inc., USA, v. 84, n. C, p. 41–50, 9 2019. ISSN 0952-1976. Disponível em: ⟨https://doi.org/10.1016/j.engappai.2019.05.011⟩.

NAVARRO, J.; VIDAñA-VILA, E.; ALSINA-PAGèS, R. M.; HERVáS, M. Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios. *Sensors*, v. 18, n. 8, 2018. ISSN 1424-8220. Disponível em: ⟨https://www.mdpi.com/1424-8220/18/8/2492⟩.

ORLANDIC, L.; TEIJEIRO, T.; ATIENZA, D. *The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms*. Zenodo, 2021. For more information about the data collection, pre-processing, validation, and data structure, please refer to the following publication: https://www.nature.com/articles/s41597-021-00937-4 The cough pre-processing and feature extraction code is available from the following c4science repository: https://c4science.ch/diffusion/10770/. Disponível em: ⟨https://doi.org/10.5281/zenodo.7024894⟩.

PANICO, F.; CORDASCO, G.; VOGEL, C.; TROJANO, L.; ESPOSITO, A. Ethical issues in assistive ambient living technologies for ageing well. *Multimedia Tools and Applications*, v. 79, n. 47, p. 36077–36089, Dec 2020. ISSN 1573-7721. Disponível em: ⟨https://doi.org/10.1007/s11042-020-09313-7⟩.

PATEL, A.; SHAH, J. Sensor-based activity recognition in the context of ambient assisted living systems: A review. *J. Ambient Intell. Smart Environ.*, v. 11, p. 301–322, 2019.

PATEL, A.; SHAH, J. Real-time human behaviour monitoring using hybrid ambient assisted living framework. 06 2020.

POORJAM, A. H. *Why we take only 12-13 MFCC coefficients in feature extraction?* 2018.

THAKUR, N.; HAN, C. Y. A study of fall detection in assisted living: Identifying and improving the optimal machine learning method. *Journal of Sensor and Actuator Networks*, v. 10, n. 3, 2021. ISSN 2224-2708. Disponível em: ⟨https://www.mdpi.com/2224-2708/10/3/39⟩.

XIE, L.; ZHAO, J.; WEI, H.; FAN, Z.; PANG, G. Efficient early event detector for streaming sequence. *IEEE Access*, v. 7, p. 85875–85886, 2019.